

UTILITY PATENT APPLICATION TRANSMITTAL		Attorney Docket No. 32830/58872A	
		First Named Inventor Stephen Yip	
		Title ARTIFICIAL INTELLIGENCE SEGMENTATION OF TISSUE IMAGES	
(Only for new nonprovisional applications under 37 CFR 1.53(b))		Priority Mail Express® Label No.	

APPLICATION ELEMENTS See MPEP chapter 600 concerning utility patent application contents.	ADDRESS TO: Commissioner for Patents P.O. Box 1450 Alexandria, VA 22313-1450
1. ☐ Fee Transmittal Form (PTO/SB/17 or equivalent) 2. ☐ Applicant asserts small entity status. See 37 CFR 1.27 3. ☐ Applicant certifies micro entity status. See 37 CFR 1.29. Applicant must attach form PTO/SB/15A or B or equivalent. 4. ☒ Specification [Total Pages <u>45</u>] Both the claims and abstract must start on a new page. (See MPEP § 608.01(a) for information on the preferred arrangement) 5. ☒ Drawing(s) (35 U.S.C. 113) [Total Sheets <u>16</u>] 6. Inventor's Oath or Declaration [Total Pages _____] (including substitute statements under 37 CFR 1.64 and assignments serving as an oath or declaration under 37 CFR 1.63(e)) a. ☐ Newly executed (original or copy) b. ☐ A copy from a prior application (37 CFR 1.63(d)) 7. ☒ Application Data Sheet * See note below. See 37 CFR 1.76 (PTO/AIA/14 or equivalent) 8. CD-ROM or CD-R in duplicate, large table, or Computer Program (Appendix) ☐ Landscape Table on CD 9. Nucleotide and/or Amino Acid Sequence Submission (if applicable, items a. – c. are required) a. ☐ Computer Readable Form (CRF) b. ☐ Specification Sequence Listing on: i. ☐ CD-ROM or CD-R (2 copies); or ii. ☐ Paper c. ☐ Statements verifying identity of above copies	ACCOMPANYING APPLICATION PAPERS 10. ☐ Assignment Papers (cover sheet & document(s)) <div style="margin-top: 10px;"> Name of Assignee <div style="border: 1px solid black; height: 30px; width: 100%;"></div> </div> 11. ☐ 37 CFR 3.73(c) Statement (when there is an assignee) ☐ Power of Attorney 12. ☐ English Translation Document (if applicable) 13. ☐ Information Disclosure Statement (PTO/SB/08 or PTO-1449) ☐ Copies of citations attached 14. ☐ Preliminary Amendment 15. ☐ Return Receipt Postcard (MPEP § 503) (Should be specifically itemized) 16. ☐ Certified Copy of Priority Document(s) (if foreign priority is claimed) 17. ☐ Nonpublication Request Under 35 U.S.C. 122(b)(2)(B)(i). Applicant must attach form PTO/SB/35 or equivalent. 18. ☒ Other: <div style="border: 1px solid black; padding: 5px; display: inline-block;">Petition to Accept Color Drawings</div>

***Note:** (1) Benefit claims under 37 CFR 1.78 and foreign priority claims under 1.55 **must** be included in an Application Data Sheet (ADS).
 (2) For applications filed under 35 U.S.C. 111, the application must contain an ADS specifying the applicant if the applicant is an assignee, person to whom the inventor is under an obligation to assign, or person who otherwise shows sufficient proprietary interest in the matter. See 37 CFR 1.46(b).

19. CORRESPONDENCE ADDRESS			
☒ The address associated with Customer Number: <u>04743</u> OR ☐ Correspondence address below			
Name			
Address			
City	State	Zip Code	
Country	Telephone	Email	
Signature <u>/Paul B. Stephens/_</u>		Date	December 31, 2019
Name (Print/Type) Paul B. Stephens		Registration No. (Attorney/Agent)	47,970

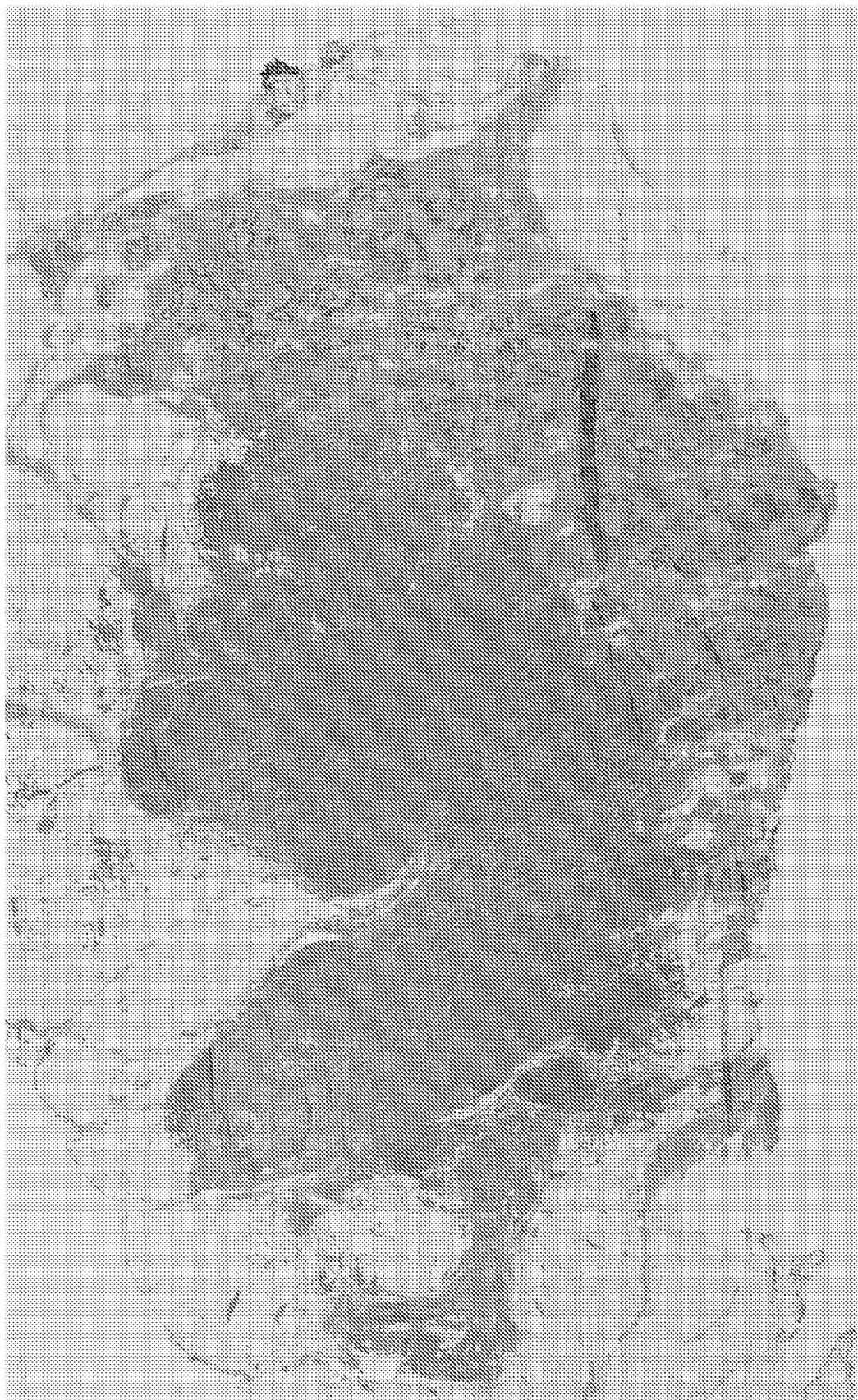


FIG 1

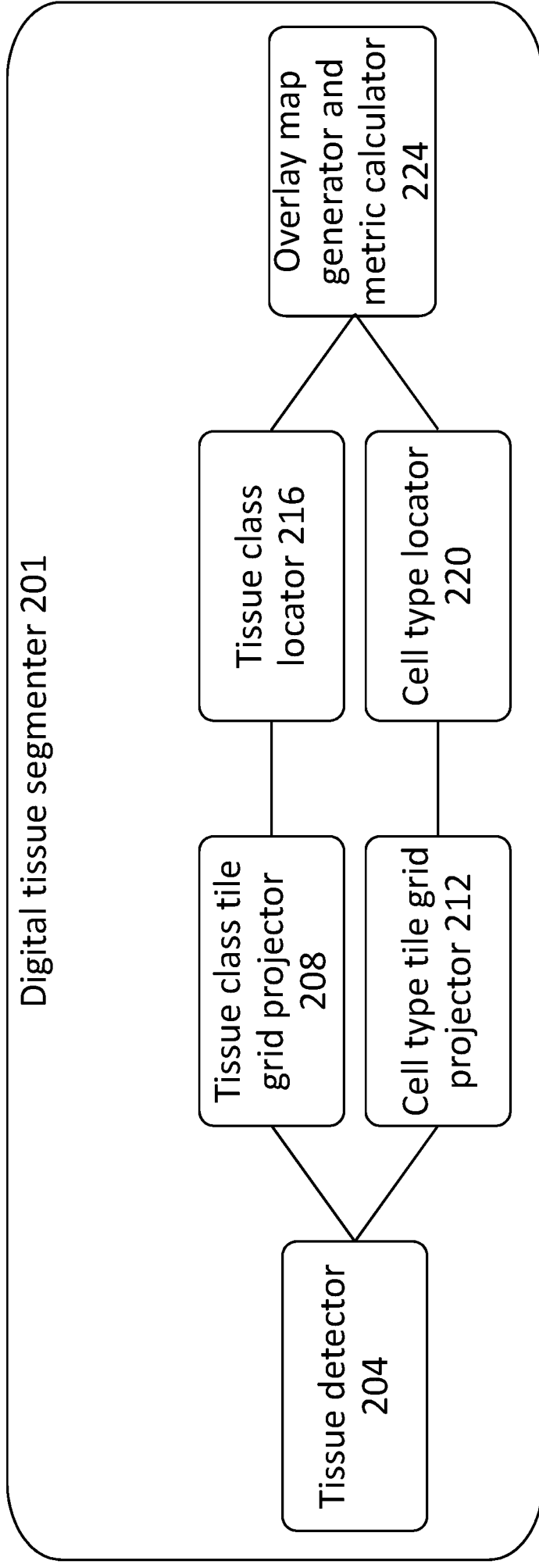


FIG 2

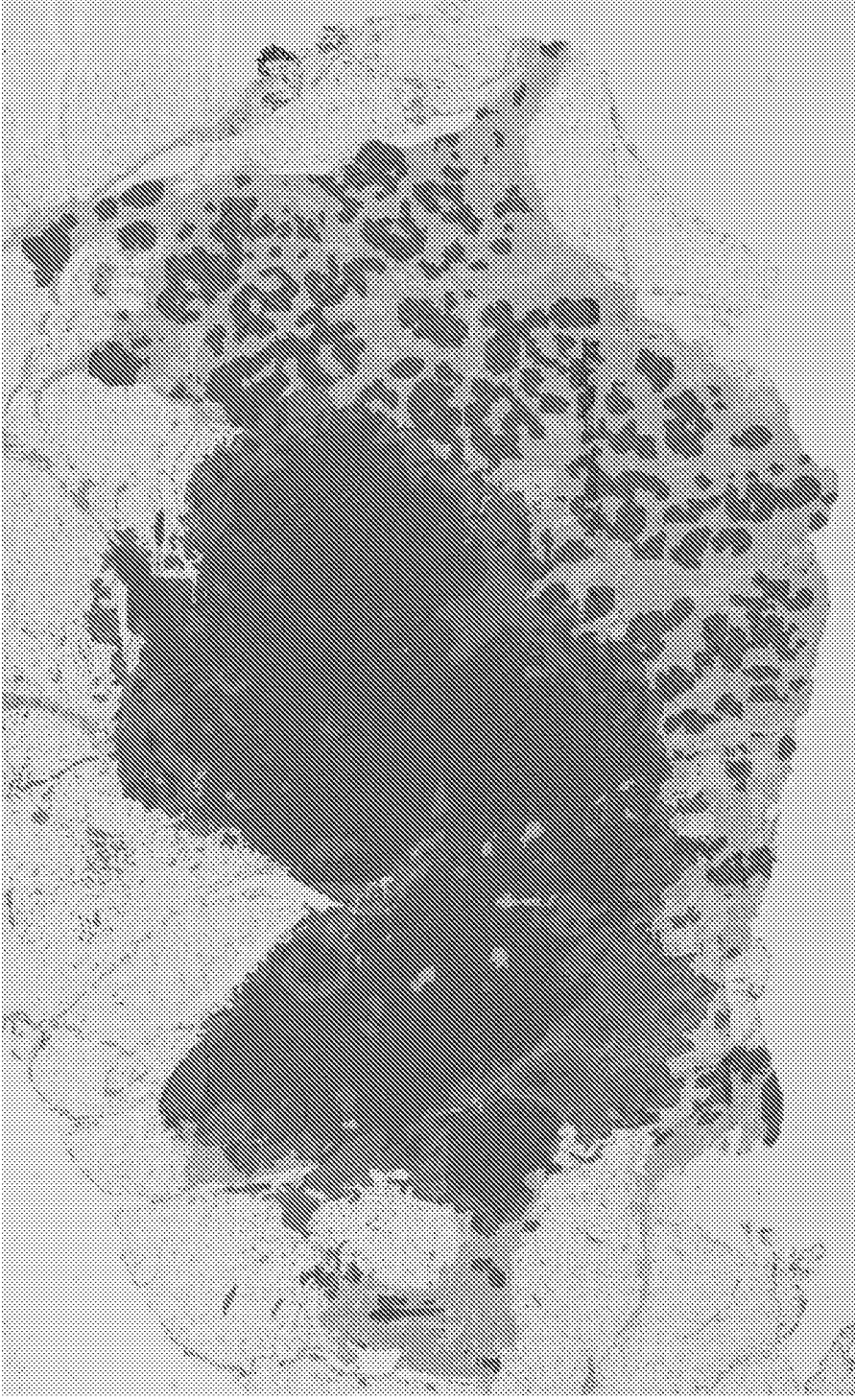
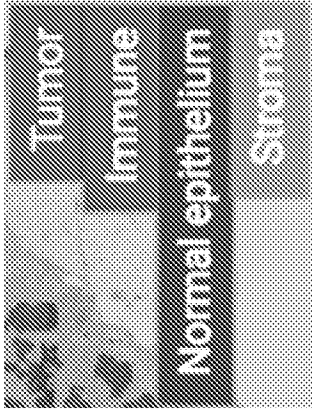


FIG 3A

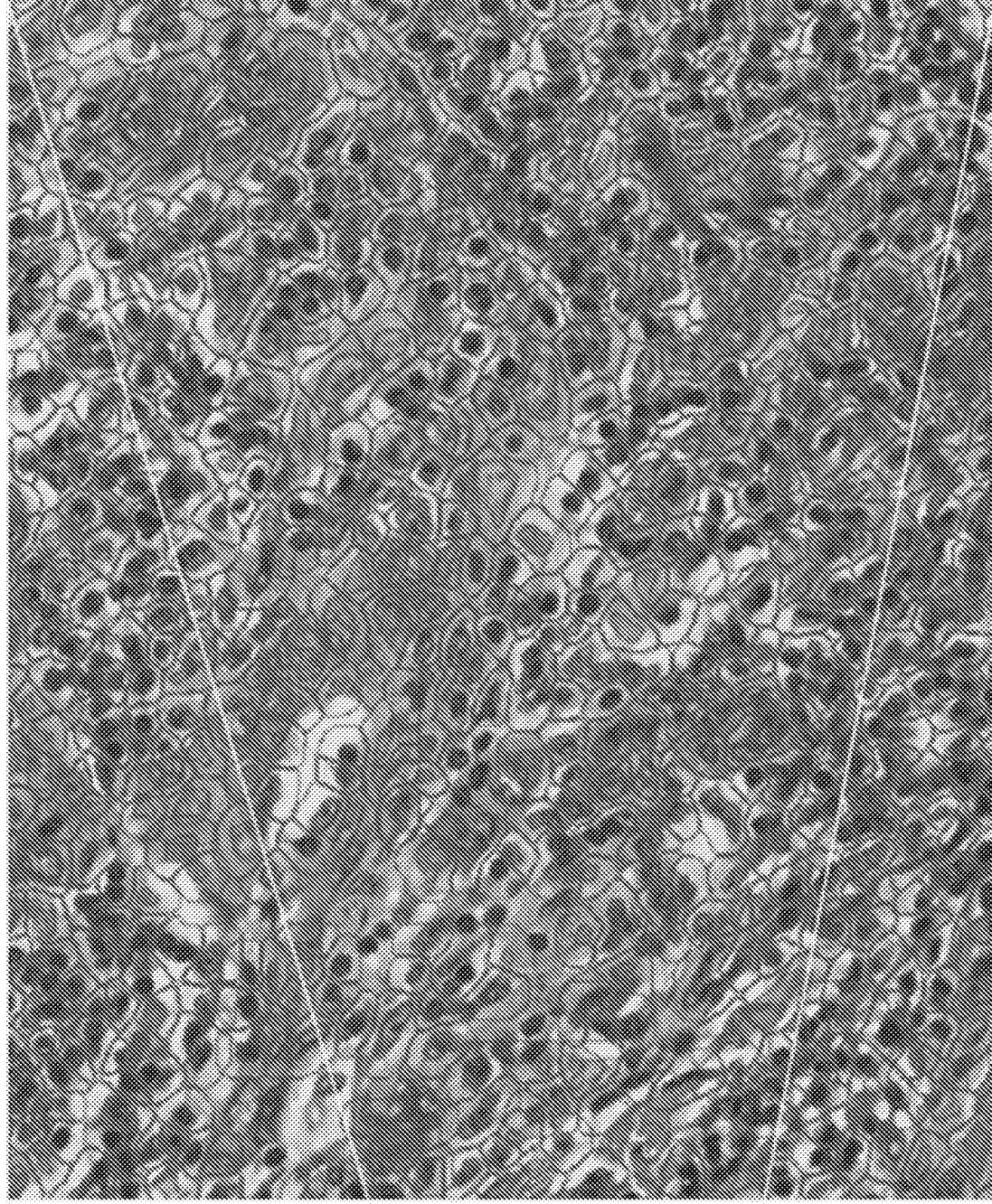


FIG 3B

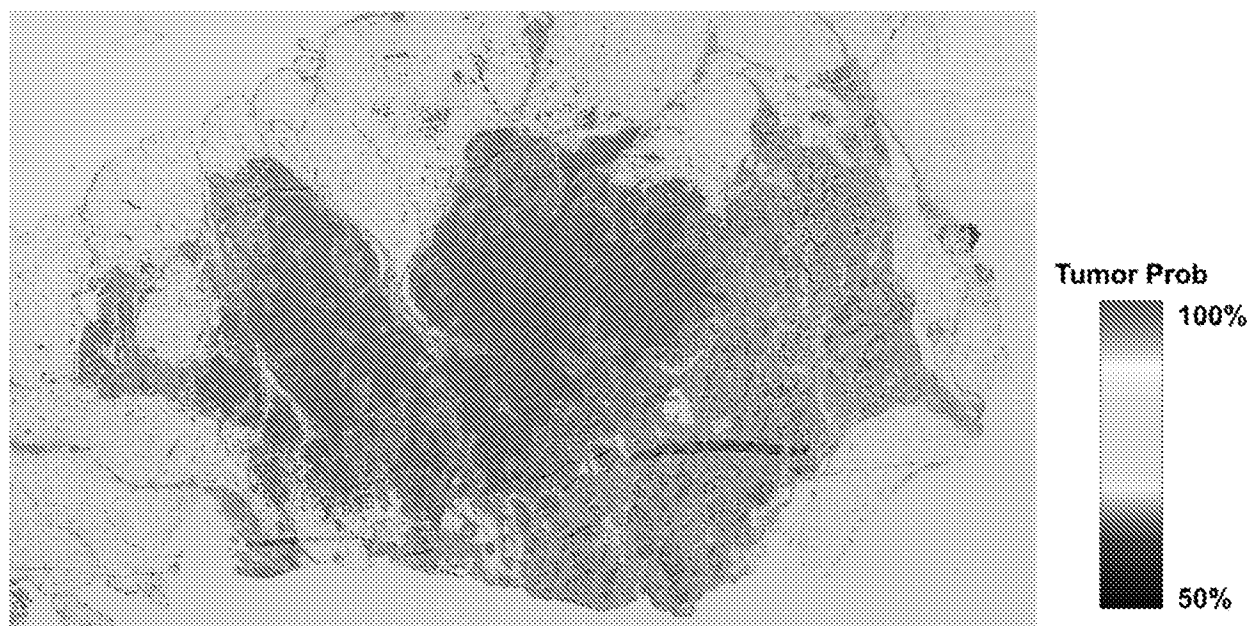


FIG. 3C

Receive digital image file →

Analyze metadata and retrieve image at known resolution →

(optional step) Downsample to create image with optimal resolution →

Identify and eliminate pixels that are too bright or too dark →

Identify and eliminate pixels with non-tissue objects

FIG 4

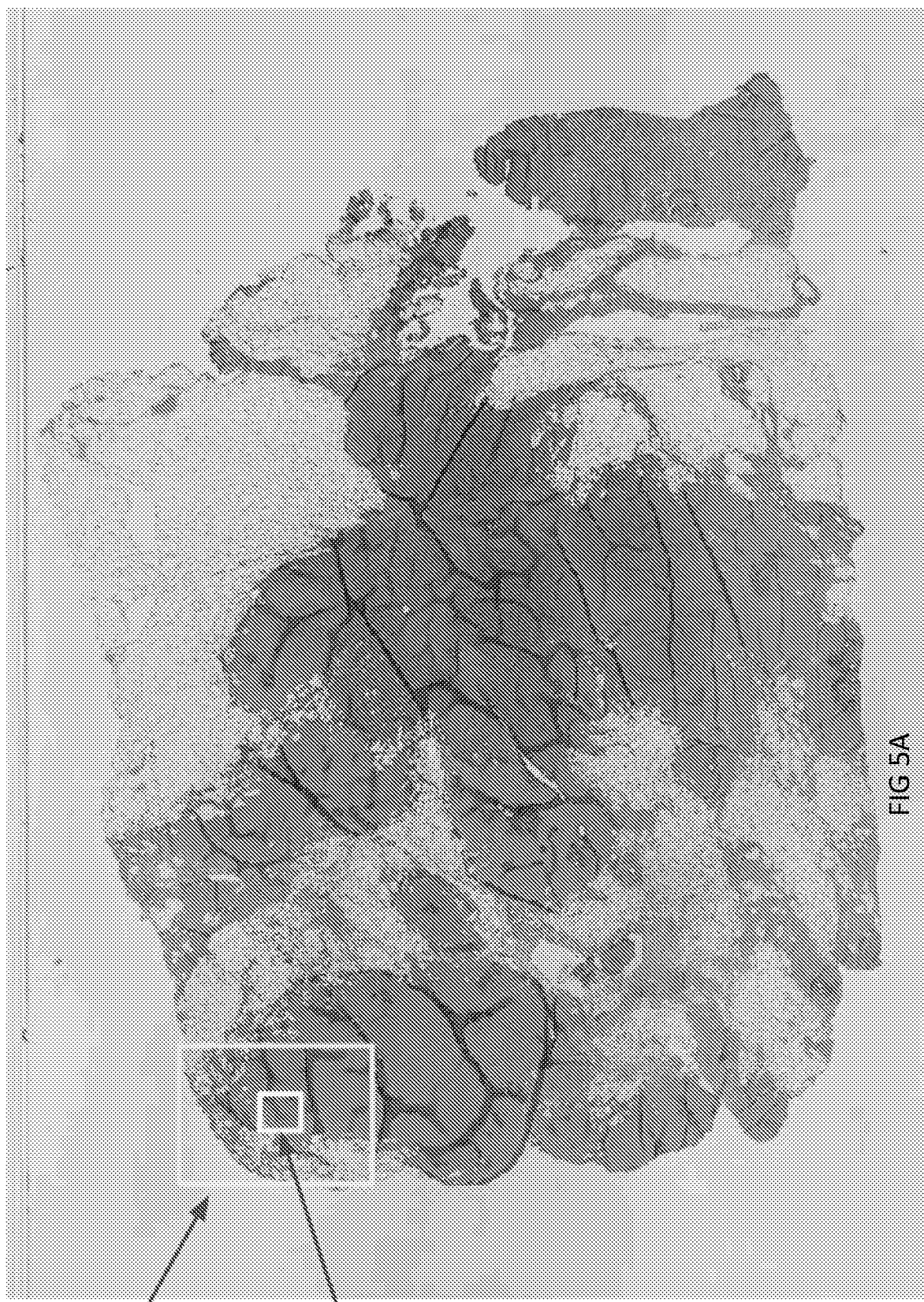


FIG 5A

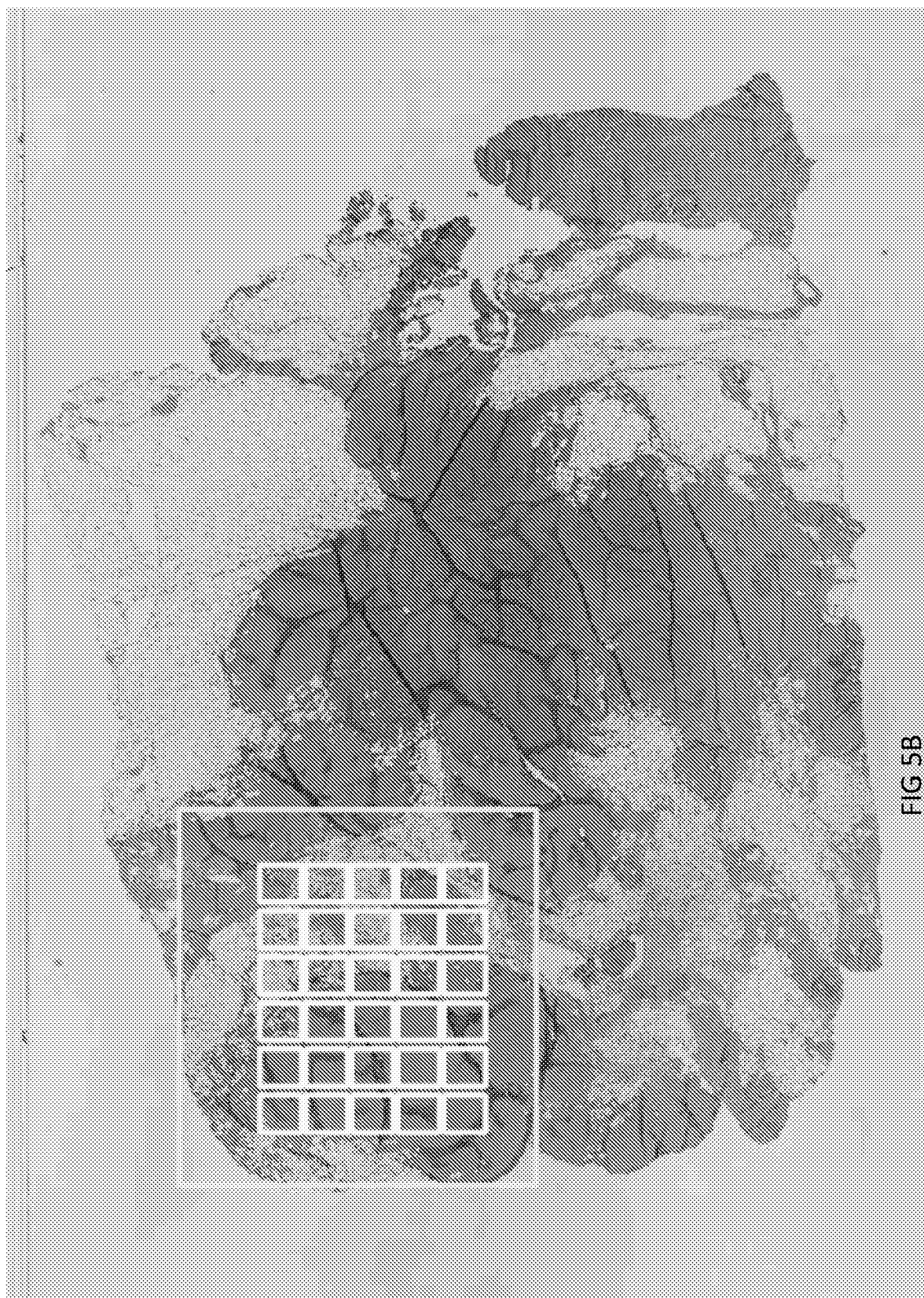
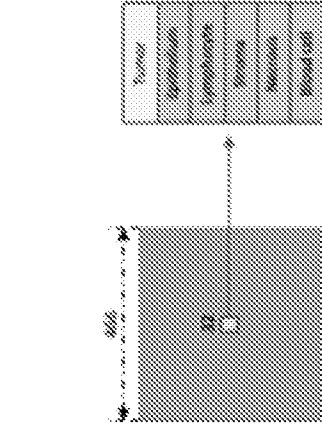
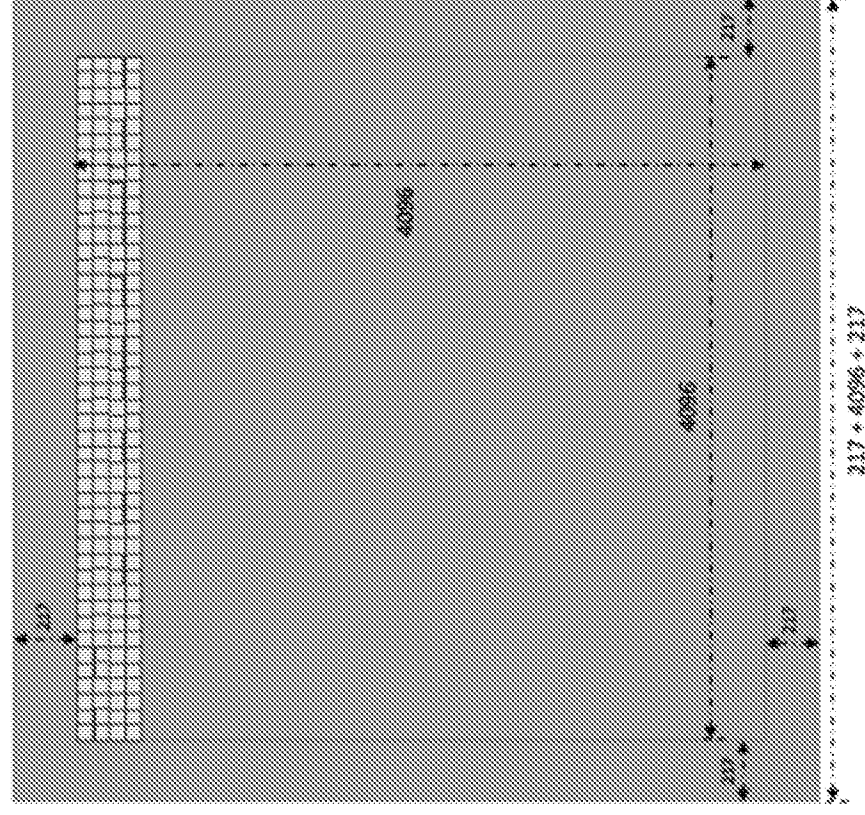


FIG 5B



Prediction of center tile label based on cell and its surrounding context.



Input: 4096 x 4096 large image tile with 217 padding on each size.

Output: 128 x 128 x N cube, with N equal to number of classes.

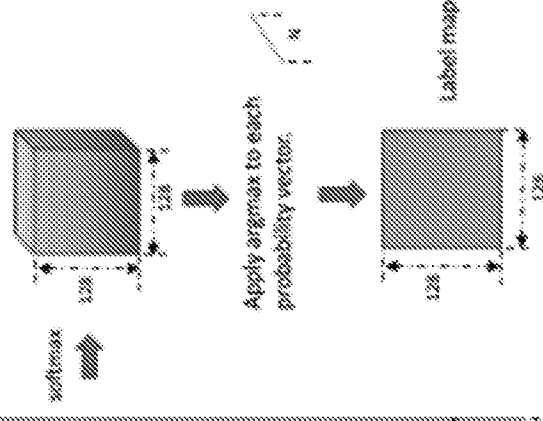


FIG 5C

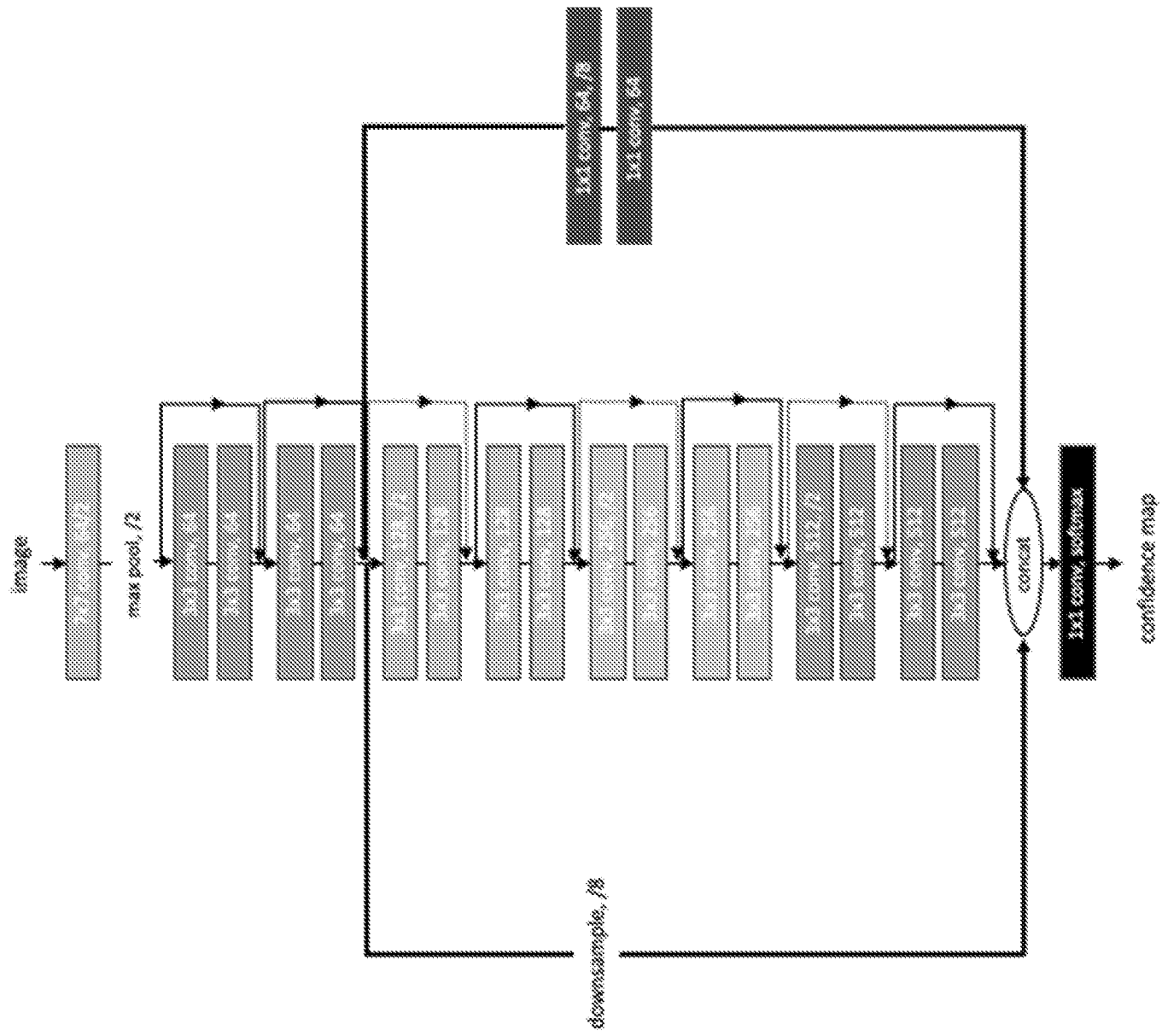


FIG 6A

ResNet-18

layer name	output size	sub layer
input layer	$224 \times 224 \times 3$	
conv1	$112 \times 112 \times 64$	
max pool	$56 \times 56 \times 64$	
conv2_x	$56 \times 56 \times 64$	Four 3x3 conv layers
conv3_x	$28 \times 28 \times 128$	Four 3x3 conv layers
conv4_x	$14 \times 14 \times 256$	Four 3x3 conv layers
conv5_x	$7 \times 7 \times 512$	Four 3x3 conv layers
avg. pool	$1 \times 1 \times 512$	
dense_layer	#classes	
softmax	#classes	

padding = 1 if filter_size = 3
padding = 3 if filter_size = 7

phiNet

layer name	output size	sub layer
input layer	$(434+32N) \times (434+32N) \times 3$	
conv1	$(214+16N) \times (214+16N) \times 64$	
max pool	$(106+8N) \times (106+8N) \times 64$	
conv2_x	$(98+8N) \times (98+8N) \times 64$	Four 3x3 conv layers
conv3_x	$(42+4N) \times (42+4N) \times 128$	Four 3x3 conv layers
conv4_x	$(14+2N) \times (14+2N) \times 256$	Four 3x3 conv layers
conv5_x	$N \times N \times 512$	Four 3x3 conv layers
concat	$N \times N \times (512+64+64)$	
1x1 conv	$N \times N \times \text{\#classes}$	
softmax	$N \times N$	

No padding

FIG 6B

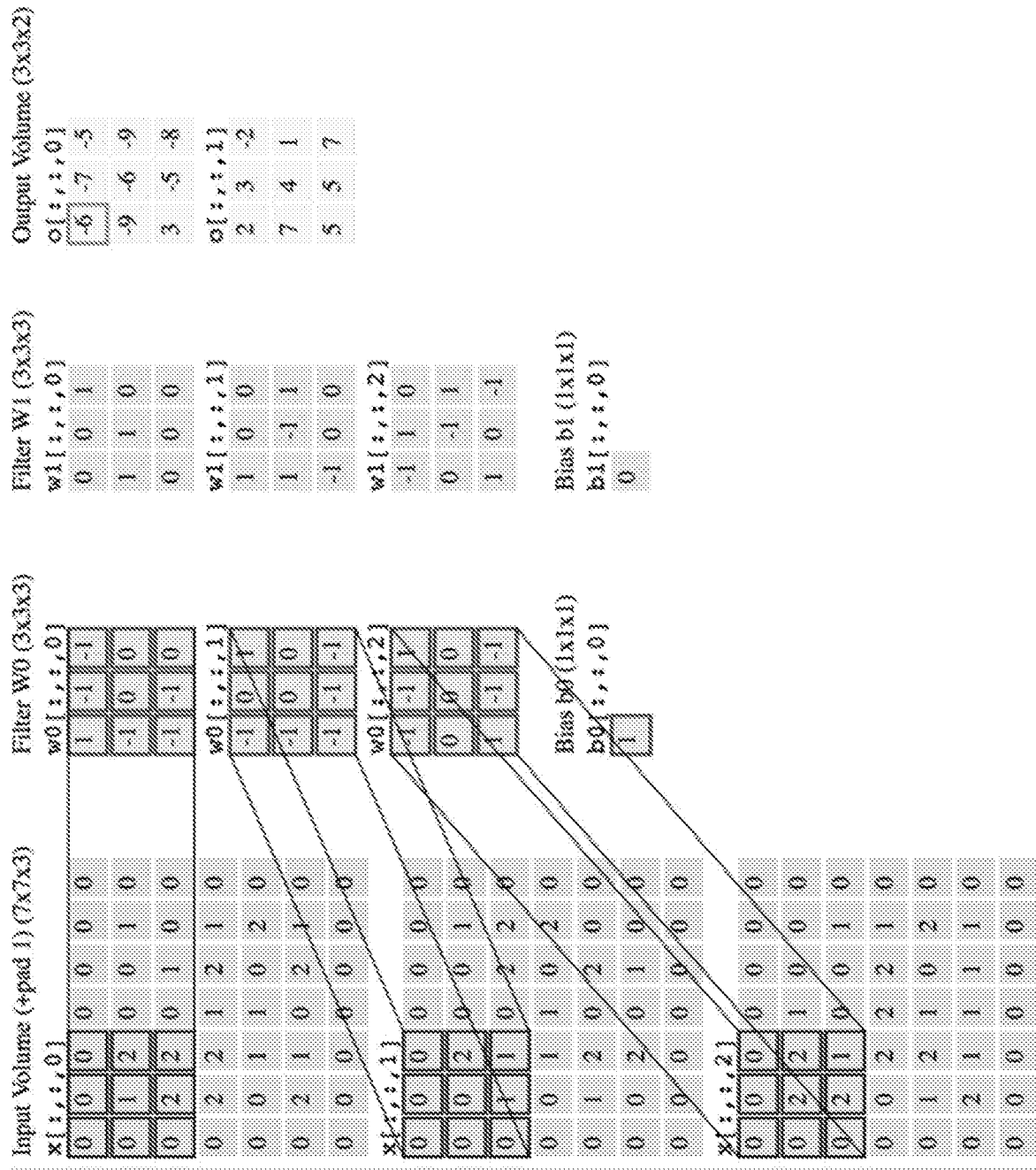
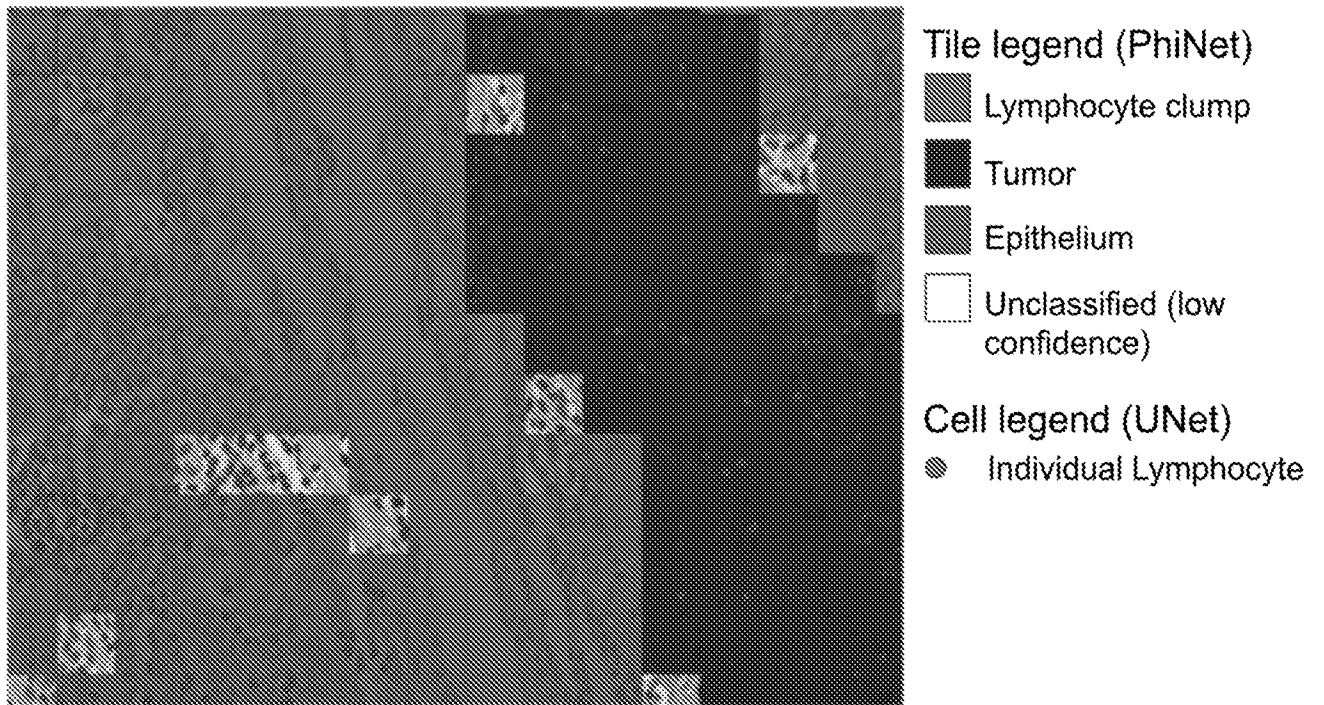


FIG 6C

Union of tile classification (PhiNet) and
cell segmentation (UNet)



Detected lymphocytes (appearing as red dots) can be seen throughout tissue. The lymphocytes within the dark blue tumor tiles are tumor-infiltrating lymphocytes.

FIG. 7

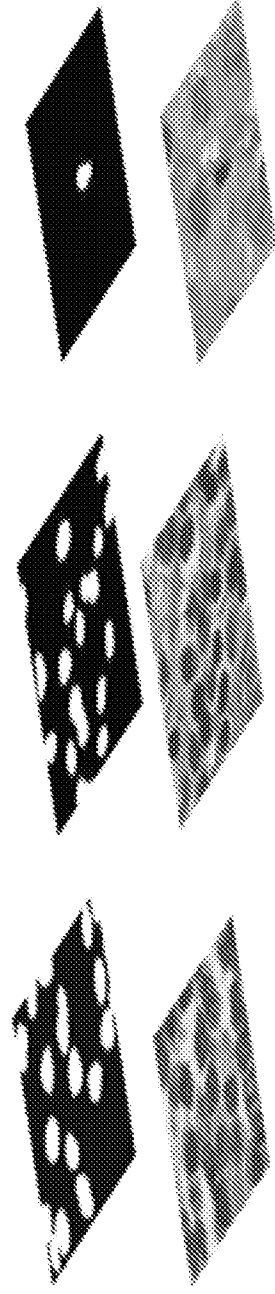
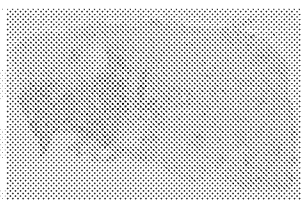
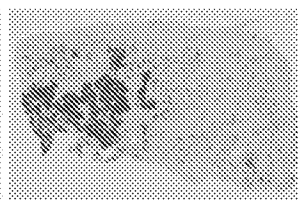


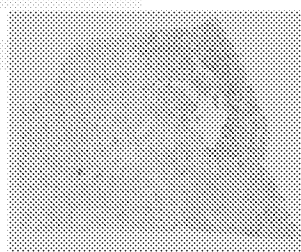
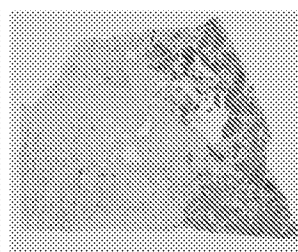
FIG 8

Predicted tissue composition maps

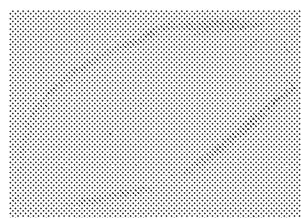
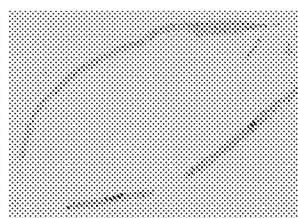
- Stroma
- Necrosis
- Epithelium
- Tumor -
- Tumor +
- Blood
- Lymph +
- Mac +
- Lymph -



Predicted TPS
96%



Predicted TPS
19%



Predicted TPS
0%

$$\text{Tumor positive score (TPS)} = \frac{\text{Tumor (+) area}}{\text{Total tumor area}}$$

FIG. 9

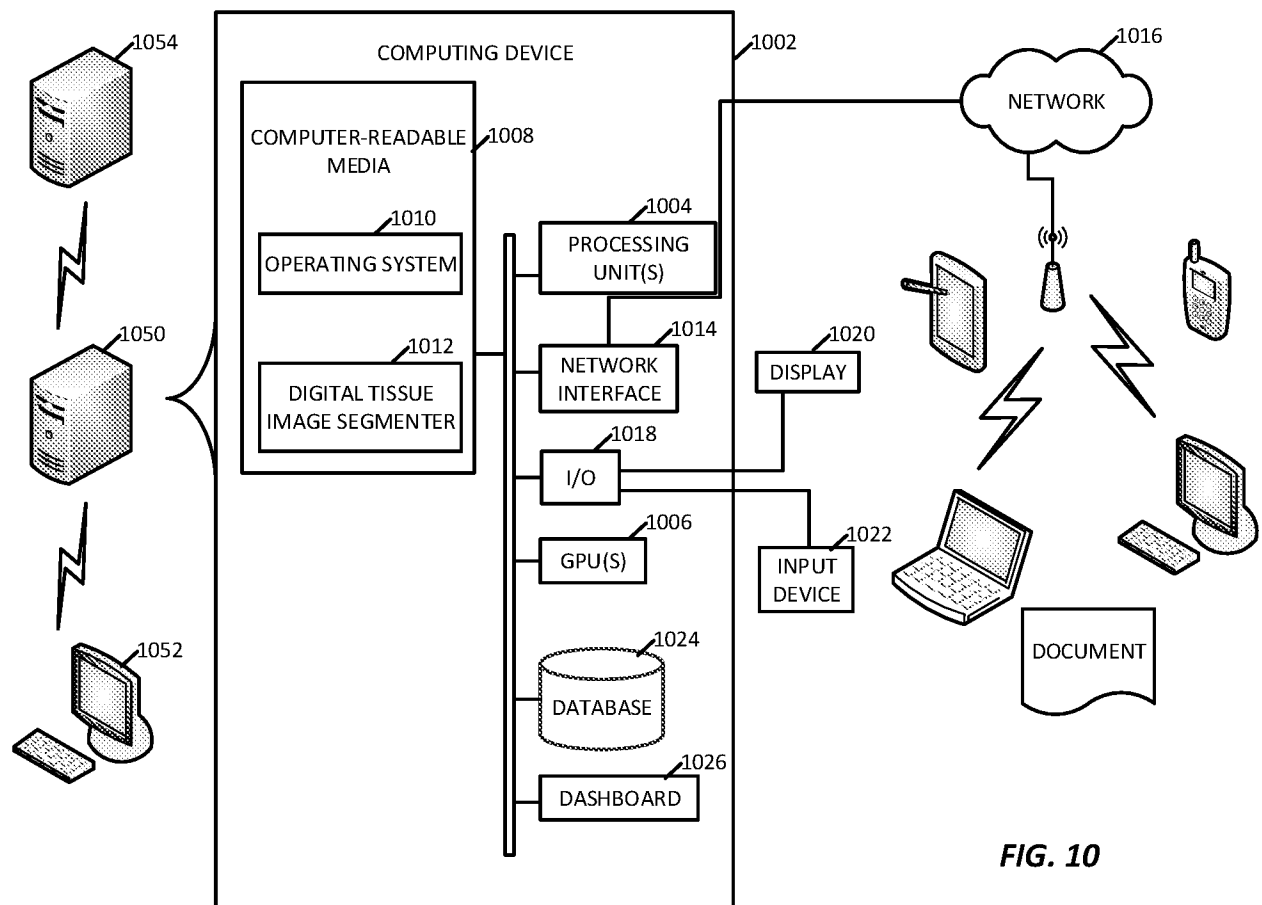


FIG. 10

ARTIFICIAL INTELLIGENCE SEGMENTATION OF TISSUE IMAGES

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims benefit of priority to and claims under 35 U.S.C. §119(e)(1) the benefit of the filing date of U.S. provisional application serial number 62/787,047 filed December 31, 2018, the entire disclosure of which is incorporated herein by reference.

FIELD OF THE INVENTION

[0002] The present disclosure relates to techniques for the analysis of medical images and, more particularly, to techniques for analysis of histological slides other images of cancerous tissue.

BACKGROUND

[0003] The background description provided herein is for the purpose of generally presenting the context of the disclosure. Work of the presently named inventors, to the extent it is described in this background section, as well as aspects of the description that may not otherwise qualify as prior art at the time of filing, are neither expressly nor impliedly admitted as prior art against the present disclosure.

[0004] To guide a medical professional in diagnosis, prognosis and treatment assessment of a patient's cancer, it is common to extract and inspect tumor samples from the patient. Visual inspection can reveal growth patterns of the cancer cells in the tumor in relation to the healthy cells near them and the presence of immune cells within the tumor. Pathologists, members of a pathology team, other trained medical professionals, or other human analysts visually analyze thin slices of tumor tissue mounted on glass microscope slides to classify each region of the tissue as one of many tissue classes that are present in a tumor sample. This information aids the pathologist in determining characteristics of the cancer tumor in the patient, which can inform treatment decisions. A pathologist will often assign one or more numerical scores to a slide, based on a visual approximation. Numerical scores assigned during microscope slide analysis include tumor purity, which is the percentage of the tissue that is formed by tumor cells.

[0005] Characteristics of the tumor may include tumor grade, tumor purity, degree of invasiveness of the tumor, degree of immune infiltration into the tumor, cancer stage, and anatomic origin site of the tumor, which is especially important for diagnosing and treating a metastatic tumor. These details about the cancer can help a physician monitor the progression

of cancer within a patient and predict which anti-cancer treatments are likely to be successful in eliminating cancer cells from the patient's body.

[0006] Manually analyzing microscope slides is time consuming and requires a trained medical professional. Furthermore, because numerical scores are assigned by approximation, these scores are subjective, not quantitative.

[0007] High-resolution, digital images of microscope slides make it possible to use artificial intelligence to analyze the slides and classify the tissue components by tissue class.

[0008] A Convolutional Neural Network ("CNN") is a deep learning algorithm that analyzes digital images by assigning one class label to each input image. Slides, however, include more than one type of tissue, including the borders between neighboring tissue classes. There is a need to classify different regions as different tissue classes, in part to study the borders between neighboring tissue classes and the presence of immune cells among tumor cells. For a traditional CNN to assign multiple tissue classes to one slide image, the CNN would need to separately process each section of the image that needs a tissue class label assignment. Neighboring sections of the image overlap, so processing each section separately creates a high number of redundant calculations and is time consuming.

[0009] A Fully Convolutional Network (FCN) can analyze an image and assign classification labels to each pixel within the image, so a FCN is more useful for analyzing images that depict objects with more than one classification. A FCN generates an overlay map to show the location of each classified object in the original image. However, FCN deep learning algorithms that analyze digital slides would require training data sets of images with each pixel labeled as a tissue class, which requires too much annotation time and processing time to be practical. In digital images of slides, each edge of the image may contain more than 10,000-100,000 pixels. The full image may have at least $10,000^2 - 100,000^2$ pixels, which forces long algorithm run times due to the intense computation required. The high number of pixels makes it infeasible to use traditional FCNs to segment digital images of slides.

SUMMARY OF THE INVENTION

[0010] In accordance with an example, a method for creating an overlay map on a digital image of a slide comprises: receiving the digital image; separating the digital image into a plurality of tiles; and identifying the majority class of tissue visible within each tile in the plurality of tiles, based on a multi-tile analysis.

[0011] In some examples, the method includes generating a digital overlay drawing of an outer edge of each cell in the image.

[0012] In some examples, the digital overlay drawing is prepared at the resolution level of an individual pixel.

[0013] In accordance with another example, a method for tissue classification of a digital image of a slide includes: receiving the digital image; generating a digital overlay drawing of a tissue region in the digital image; and displaying, in the digital overlay drawing, tiles over the tissue region and visually identifying predicted content of each tile, to generate a classification map classifying the digital image.

[0014] In some examples, the method further includes determining the predicted content for each tile using a classification model configured as a multi-resolution fully convolutional network, the multi-resolution fully convolution network configured to perform classification on digital images of different zoom levels.

[0015] In some examples, the method further includes using a classification model, for each tile, determining a classification from a plurality of classifications.

[0016] In some examples, the plurality of classifications include tumor epithelium, normal epithelium, immune, stroma, necrosis, blood, and fat.

[0017] In accordance with yet another example, a method for cell detection in a digital image of a slide includes: receiving the digital image; identifying cell objects in the digital image; generating a digital overlay drawing for the digital image; and displaying, in the digital overlay drawing, a polygon outlining each of the identified cell objects.

[0018] In some examples, the cell objects include lymphocyte cells and not lymphocyte cells.

[0019] In some examples, the cell objects comprise CD3, CD8, CD20, pancytokeratin, and smooth muscle actin.

[0020] In accordance with yet another example, a method for tissue classification of a digital image of a slide includes: receiving the digital image; segmenting the digital image into a plurality of tiles; determining a predicted class for each tile; identifying a plurality of cell objects in the digital image; determining a predicted class for each of the plurality of cell objects; and for

each of the plurality of tiles that corresponds to one of the plurality of cell objects, assigning the tile the predicted class of the corresponding cell object in place of the predicted class of the tile.

[0021] In some examples, the method further includes: storing, in a first file, for each tile, a tile position and the predicted class of the tile; and storing, in a second file, for each cell object, a polygon outlining the cell object and the predicted class of the cell object.

[0022] In some examples, the method further includes generating a digital overlay drawing for the digital image, where the digital overlay drawing is a cell mask displaying a polygon around each cell object.

[0023] In some examples, the method further includes generating a digital overlay drawing for the digital image, where the digital overlay drawing is a histology mask displaying the plurality of tiles and the predict class for each tile.

[0024] In some examples, the method further includes generating a digital overlay drawing for the digital image; and displaying, in the digital overlay drawing, the plurality of tiles and the predicted class for each tile that does not correspond to a cell object, and displaying, the plurality of cell objects and the predicted class of each cell object.

[0025] In some examples, the method further includes: generating a digital overlay drawing for the digital image, wherein the digital overlay drawing includes the digital image; and displaying the digital overlay drawing.

[0026] In some examples, the method further includes: generating a digital overlay drawing for the digital image, wherein the digital overlay drawing includes a generated version of the digital image; and displaying the digital overlay drawing.

[0027] In some examples, the method further includes: generating a plurality of digital overlay drawings for the digital image, where each digital overlay drawing corresponds to a different predicted class; and selectively displaying one of the plurality of digital overlay drawings.

[0028] In some examples, the method further includes generating a digital overlay drawing for the digital image, where the digital overlay drawing comprises percentages of predict classes corresponding to the digital image.

[0029] In some examples, the method further includes generating a digital overlay drawing for the digital image, where the digital overlay drawing comprises total counts of predict classes corresponding to the digital image.

[0030] In accordance with another example, a system includes a processor and a computer-readable memory is provided, the memory storing instructions that when executed by the processor cause the processor to perform the methods described herein.

[0031] In accordance with an example, a system includes a processor and a computer-readable memory storing instructions that when executed by the processor cause the processor to: receive the digital image; separate the digital image into a plurality of tiles; and identify the majority class of tissue visible within each tile in the plurality of tiles, based on a multi-tile analysis.

[0032] In accordance with an example, a system includes a processor and a computer-readable memory storing instructions that when executed by the processor cause the processor to: receive the digital image; generate a digital overlay drawing of a tissue region in the digital image; and display, in the digital overlay drawing, tiles over the tissue region and visually identify predicted content of each tile, to generate a classification map classifying the digital image.

[0033] In accordance with an example, a system includes a processor and a computer-readable memory storing instructions that when executed by the processor cause the processor to: receive the digital image; identify cell objects in the digital image; generate a digital overlay drawing for the digital image; and display, in the digital overlay drawing, a polygon outlining each of the identified cell objects.

[0034] In accordance with an example, a system includes a processor and a computer-readable memory storing instructions that when executed by the processor cause the processor to: receive the digital image; segment the digital image into a plurality of tiles; determine a predicted class for each tile; identify a plurality of cell objects in the digital image; determine a predicted class for each of the plurality of cell objects; and for each of the plurality of tiles that corresponds to one of the plurality of cell objects, assign the tile the predicted class of the corresponding cell object in place of the predicted class of the tile.

BRIEF DESCRIPTION OF THE DRAWINGS

[0035] This patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the United States Patent and Trademark Office upon request and payment of the necessary fee.

[0036] The figures described below depict various aspects of the system and methods disclosed herein. It should be understood that each figure depicts an example of aspects of the present systems and methods.

[0037] FIG. 1 illustrates an example of a digital image of a histology slide, in accordance with an example.

[0038] FIG. 2 is an overview of a digital tissue segmenter, in accordance with an example.

[0039] FIG. 3A illustrates a tissue segmentation overlay map created by the digital tissue segmenter, in accordance with an example.

[0040] FIG. 3B illustrates a cell outer edge overlay map created by the digital tissue segmenter, in accordance with an example.

[0041] FIG. 3C illustrates a tumor probability overlay heatmap map created by the digital tissue segmenter, in accordance with an example.

[0042] FIG. 4 is a flowchart of a method for preparing digital images of histology slides for tissue segmentation and mapping analysis, in accordance with an example.

[0043] FIGS. 5A, 5B, and 5C each illustrate an exemplary grid overlay used for tissue segmentation on a digital image of a histology slide, in accordance with an example.

[0044] FIG. 6A illustrates the layers of an exemplary tissue segmentation algorithm, in accordance with an example.

[0045] FIG. 6B compares the layers of a known image classification algorithm with an exemplary tissue segmentation algorithm, in accordance with an example.

[0046] FIG. 6C is a visualization of the layers of an exemplary 3-dimensional input image matrix being convoluted by two exemplary 3-dimensional filter matrices, in accordance with an example.

[0047] FIG. 7 illustrates an exemplary grid overlay showing lymphocyte detections (little red dots) and the tile classifications, in accordance with an example.

[0048] FIG. 8 illustrates exemplary training set images for the cell type locator, in accordance with an example.

[0049] FIG. 9 illustrates example overlay maps identifying a programmed death ligand 1 (PD-L1) tumor positive score (TPS) using different immunohistochemical (IHC) stain images, in accordance with an example.

[0050] FIG. 10 a schematic illustration of an example computer processing system having a digital tissue image segmenter for performing techniques and methods herein, in accordance with an example.

DETAILED DESCRIPTION OF THE INVENTION

[0051] FIG. 1 illustrates an example of a digital image of a histology slide, also known as a pathology slide. In one example, the digital image is created by a scanner that visually captures a histology slide. In an alternative example, the digital image is created by a digital camera attached to a microscope. The histology slide may be made of two transparent glass layers with a slice of tissue affixed between the two layers of glass. The thin slice of tissue may be very thin, with a thickness, for instance, of approximately 5 microns. Tissue may be preserved in a fixative, including formaldehyde, formalin, and paraffin. The tissue contains a combination of many individual biological cells that are visible on the slide. The scanner may include a Philips digital pathology slide scanner, or any scanner known in the art that can create a digital image file.

[0052] In one example, the tissue slice contains stain that attaches to certain types of cells or cell parts within the tissue. The stain may include hematoxylin and eosin (H & E) stain and any immunohistochemical (IHC) stain. Hematoxylin is a stain that will bind to DNA and cause the nucleus of a cell to appear blue or purple. Eosin is a stain that will bind to proteins and cause all of the remaining parts of the cell, namely the cytoplasm interior, to appear red or pink. An IHC stain is comprised of an antibody coupled with a molecule that displays one of many colors. The antibody may be designed to bind to any surface shape to target a specific molecule such as a protein or a sugar. The IHC stain will result in a concentration of dye of the selected color near any copies of a specific target molecule present on the slide. Some commonly monitored proteins in tumor samples include programmed death ligand 1 (PD-L1), whose presence in a

tumor region can indicate whether a tumor will respond to immunotherapy, and cluster of differentiation 3 (CD3), which is associated with T lymphocyte immune cells. The presence of CD3 in a tumor region may be associated with tumor infiltrating lymphocytes which can indicate that the tumor will be susceptible to anti-cancer immunotherapy.

[0053] In some examples, the slide may also contain additional control slices of tissue that are not from the tumor sample, which serve as a positive and/or negative control for the staining process. Control tissue slices are more common on slides that have IHC staining.

[0054] FIG. 2 is an overview of a digital tissue image segmenter **201**. The digital tissue segmenter **201** may comprise a computational method and apparatus that receives a digital image of a slide, displaying a slice of a tumor sample, and creates a high-density, grid-based digital overlay map that identifies the majority class of tissue visible within each grid tile in the digital image. The digital tissue segmenter **201** may also generate a digital overlay drawing of the outer edge of each cell in the slide image, at the resolution level of an individual pixel.

[0055] In another example, the digital tissue segmenter **201** is a computational method and apparatus that receives a digital radiology image and efficiently creates a high-density, grid-based digital overlay map that identifies the majority class of tissue visible within each grid tile in the digital image. The radiology image may depict a tumor within a patient's body. The radiology image may be 3-dimensional (3-D) and the digital tissue segmenter **201** may receive 2-dimensional slices of the 3-dimensional image as an input image. Radiology images include but are not limited to X-rays, CT scans, MRI's, ultrasounds, and PET.

[0056] The digital tissue segmenter **201** shown at FIG. 2 includes a tissue detector **204** for detecting the areas of a digital image that have tissue, and storing data that includes the locations of the areas detected to have tissue. The tissue detector **204** transfers tissue area location data to a tissue class tile grid projector **208** and a cell tile grid projector **212**. The tissue class tile grid projector **208** receives the tissue area location data, as described in further detail below and with reference to FIGS. 5A and 5C. For each of several tissue class labels, the tissue class locator **216** calculates a percentage that represents the likelihood that the tissue class label accurately describes the image within each tile to determine where each tissue class is located in the digital image. For each tile, the total of all of the percentages calculated for all tissue class labels will sum to 1, which reflects 100%. In one example, the tissue class locator **216** assigns one tissue class label to each tile to determine where each tissue class is located in

the digital image. The tissue class locator stores the calculated percentages and assigned tissue class labels associated with each tile.

[0057] Examples of tissue classes include but are not limited to tumor, stroma, normal, immune cluster, necrosis, hyperplasia/dysplasia, red blood cells, and tissue classes or cell types that are positive (contain a target molecule of an IHC stain) or negative for an IHC stain target molecule (do not contain that molecule). Examples also include tumor positive, tumor negative, lymphocyte positive, and lymphocyte negative. The grid-based digital overlay map or a separate digital overlay may also highlight individual immune cells, including lymphocytes, cytotoxic T cells, B cells, NK cells, macrophages, etc.

[0058] In one example, the digital tissue segmenter **201** includes a multi-tile algorithm that concurrently analyzes many tiles in an image, both individually and in conjunction with the portion of the image that surrounds each tile. The multi-tile algorithm may achieve a multiscale, multiresolution analysis that captures both the contents of the individual tile and the context of the portion of the image that surrounds the tile. The multi-tile algorithm is described further with reference to FIGS. 5A-5C and 6A-6B. Because the portions of the image that surround two neighboring tiles overlap, analyzing many tiles and their surroundings concurrently instead of separately analyzing each tile with its surroundings reduces computational redundancy and results in greater processing efficiency.

[0059] In one example, the digital tissue segmenter may store the analysis results in a 3-dimensional probability data array, which contains one 1-dimensional data vector for each analyzed tile. In one example, each data vector contains a list of percentages that sum to 100%, each indicating the probability that each grid tile contains one of the tissue classes analyzed. The position of each data vector in the orthogonal 2-dimensional plane of the data array, with respect to the other vectors, corresponds with the position of the tile associated with that data vector in the digital image, with respect to the other tiles.

[0060] The cell type tile grid projector receives the tissue area location data and projects a cell type tile grid onto the areas of an image with tissue, as described with further detail with respect to FIG. 8. The cell type locator may detect each biological cell in the digital image within each grid, prepare an outline on the outer edge of each cell, and classify each cell by cell type. The cell type locator stores data including the location of each cell and each pixel that contains a cell outer edge, and the cell type label assigned to each cell.

[0061] The overlay map generator and metric calculator may retrieve the stored 3-dimensional probability data array from the tissue class locator, and convert it into an overlay map that displays the assigned tissue class label for each tile. The assigned tissue class for each tile may be displayed as a transparent color that is unique for each tissue class. In one example, the tissue class overlay map displays the probabilities for each grid tile for the tissue class selected by the user. The overlay map generator and metric calculator also retrieves the stored cell location and type data from the cell type locator, and calculates metrics related to the number of cells in the entire image or in the tiles assigned to a specific tissue class.

[0062] FIGS. 3A and 3B illustrate examples of a digital overlay created by the digital tissue segmenter **201**. FIG. 3A illustrates a tissue class overlay map created by the overlay map generator of the digital tissue segmenter **201**. FIG. 3B illustrates a cell outer edge overlay map created by the overlay map generator of the digital tissue segmenter **201**. The overlay map generator may display the digital overlays as transparent or opaque layers that cover the slide image, aligned such that the slide location shown in the overlay and the slide image are in the same location on the display. The overlay map may have varying degrees of transparency. The degree of transparency may be adjustable by the user. The overlay map generator may report the percentage of the labeled tiles that are associated with each tissue class label, ratios of the number of tiles classified under each tissue class, the total area of all grid tiles classified as a single tissue class, and ratios of the areas of tiles classified under each tissue class. FIG. 3C illustrates a tumor probability overlay heatmap map created by the digital tissue segmenter, in accordance with another example.

[0063] The digital tissue segmenter **201** may also report the total number of cells or the percentage of cells that are located in an area defined by either a user, the entire slide, a single grid tile, by all grid tiles classified under each tissue class, or cells that are classified as immune cells. The digital tissue segmenter **201** may also report the number of cells classified as immune cells that are located within areas classified as tumor or any other tissue class.

[0064] In one example, the digital tissue segmenter **201** is capable of calculating the percentage of cells that are colored by an IHC stain to highlight particular cells containing the molecule targeted by the stain. The percentage of cells may be specific to a tissue class region or a cell type (or cell object). For example, if the IHC stain targets programmed death ligand 1 (PD-L1) protein, the digital tissue segmenter **201** may determine the percentage of cancer cells in the tumor tissue class that contain PD-L1 protein. If the IHC stain targets cluster of

differentiation 3 (CD3) protein, the digital tissue segmenter **201** may determine the percentage of lymphocytes or total cells that contain CD3. In other examples, the digital tissue segmenter **201** may determine the percentage of CD8, CD20, pancytokeratin, and/or smooth muscle actin in the digital image.

[0065] The map generator and metric calculator **224** may also create a digital overlay map, showing predicted IHC staining on a digital image of a slide that contains no IHC stain. In one example, the tissue class locator **216** can predict where IHC staining for a specific molecule would exist on a slide, or the percentage of cells that express a specific protein, based on input images that only contain H and E stain.

[0066] The digital overlays and reports generated by the digital tissue segmenter **201** can be used to assist medical professionals in more accurately estimating tumor purity, and in locating regions or diagnoses of interest, including invasive tumors having tumor cells that protrude into the non-tumor tissue region that surrounds the tumor. They can also assist medical professionals in prescribing treatments. For example, the number of lymphocytes in areas classified as tumor may predict whether immunotherapy will be successful in treating a patient's cancer.

[0067] The digital overlays and reports generated by the digital tissue segmenter **201** can also be used to determine whether the slide sample has enough high-quality tissue for successful genetic sequence analysis of the tissue. Genetic sequence analysis of the tissue on a slide is likely to be successful if the slide contains an amount of tissue and/or has a tumor purity value that exceeds a user-defined tissue amount and tumor purity thresholds. In one example, the digital tissue segmenter **201** may label a slide as accepted or rejected for sequence analysis, depending on the amount of tissue present on the slide and the tumor purity of the tissue on the slide. In some examples, the digital tissue segmenter **201** is configured to determine tumor-infiltrating lymphocytes (TILS). An example of a TILS process and engine is disclosed, for example, in U.S. Provisional Patent Application No. 62/889,521, titled "Determining Therapeutic Tumor-Infiltrating Lymphocytes (TILS) from Histopathology Slide Images," filed on August 20, 2019, which is incorporated herein by reference and in its entirety for all purposes.

[0068] The digital tissue segmenter **201** may also label a slide as uncertain, to recommend that it be manually reviewed by a trained analyst, who may be a member of a pathology team. In

one example, if the amount of tissue present on the slide is approximately equal to the user-defined tissue amount threshold or within a user-defined range, the digital tissue segmenter **201** may label the slide as uncertain. In one example, if the tumor purity of the tissue present on the slide is approximately equal to the user-defined tumor purity threshold or within a user-defined range, the digital tissue segmenter **201** may label the slide as uncertain.

[0069] In one example, the overlay map generator and metric calculator calculates the amount of tissue on a slide by measuring the total area covered by the tissue or by counting the number of cells on the slide. The number of cells on the slide may be determined by the number of cell nuclei visible on the slide. In one example, the digital tissue segmenter **201** calculates the proportion of tissue that is cancer cells by dividing the number of cell nuclei within grid areas that are labeled tumor by the total number of cell nuclei on the slide. The digital tissue segmenter **201** may exclude cell nuclei or outer edges of cells that are located in tumor areas but which belong to cells that are characterized as lymphocytes. The proportion of tissue that is cancer cells is known as the tumor purity of the sample. In some examples, the overlay map generator and metric calculator is configured to predict aligned tumor purity, for example, by calculating the fraction of tumor cells within a micro-dissected tissue area instead of within the whole tissue. These techniques herein may be configured into other tumor purity models based incorporated one or multiple definitions of tumor purity definition from pathologists.

[0070] In one example, the digital tissue segmenter **201** compares the tumor purity to the user-selected minimum tumor purity threshold and the number of cells in the digital image to a user-selected minimum cell threshold and approves the slide if both thresholds are exceeded. In one example, the user-selected minimum tumor purity threshold is 0.20, which is 20%.

[0071] In one example, the slide is given a composite tissue amount score that multiplies the total area covered by tissue detected on the slide by a first multiplier value, multiplies the number of cells counted on the slide by a second multiplier value, and sums the products of these multiplications.

[0072] The digital tissue segmenter **201** may calculate whether the grid areas that are labeled tumor are spatially consolidated or dispersed among non-tumor grid areas. If the digital tissue segmenter **201** determines that the tumor areas are spatially consolidated, the digital tissue segmenter **201** may produce a digital overlay of a recommended cutting boundary that separates the slide regions classified as tumor and the slide regions classified as non-tumor or

within the areas classified as non-tumor, proximal to the areas classified as tumor. This recommended cutting boundary can be a guide to assist a technician in dissecting a slide to isolate a maximum amount of tumor or non-tumor tissue from the slide, especially for genetic sequence analysis.

[0073] The digital tissue segmenter **201** may include clustering algorithms that calculate and report information about the spacing and density of type classified cells, tissue class classified tiles, or visually detectable features on the slide. For example, this may be accomplished by choosing a clustering algorithm and a set of objects to cluster, and the output can either be a qualitative description or a numerical value. An example clustering algorithm for cells would find spatial information of cells, such as, the center locations of tumor cells or lymphocytes, the size of each cell group, and the distance between the center of each cell group, where such spatial information is used for further outcome related analysis. The spacing information includes distribution patterns and heat maps for immune cells, tumor cells, or other cells. These patterns may include clustered, dispersed, dense, and non-existent. This information is useful to determine whether immune cells and tumor cells cluster together and what percentage of the cluster areas overlap, which may facilitate in predicting immune infiltration and patient response to immunotherapy.

[0074] The digital tissue segmenter **201** may also calculate and report average tumor cell roundness, average tumor cell perimeter length, and average tumor nuclei density.

[0075] The spacing information also includes mixture levels of tumor cells and immune cells. The clustering algorithms can calculate the probability that two adjacent cells on a given slide will be either two tumor cells, two immune cells, or one tumor cell and one immune cell.

[0076] The clustering algorithms can also measure the thickness of any stroma pattern located around an area classified as tumor. The thickness of this stroma surrounding the tumor region may be a predictor of a patient's response to treatment.

[0077] The digital tissue segmenter **201** may also calculate and report statistics including mean, standard deviation, sum, etc. for the following information in each grid tile of either a single slide image or aggregated from many slide images: red green blue (RGB) value, optical density, hue, saturation, grayscale, and stain deconvolution. Deconvolution includes the removal of the visual signal created by any individual stain or combination of stains, including hematoxylin, eosin, or IHC staining.

[0078] The digital tissue segmenter **201** may also incorporate known mathematical formulae from the fields of physics and image analysis to calculate visually detectable basic features for each grid tile. Visually detectable basic features, including lines, patterns of alternating brightness, and outlineable shapes, may be combined to create visually detectable complex features including cell size, cell roundness, cell shape, and staining patterns referred to as texture features.

[0079] The digital overlays, reports, statistics and estimates produced by the digital tissue segmenter **201** may be useful for predicting patient survival, PD-L1 status of a tumor or immune cluster, microsatellite instability, TILS%, and the origin of a tumor when the origin is unknown or the tumor is metastatic. Overlays of such status information can provide an illustrated estimate of how well a patient will respond to certain immunotherapies, in some examples. The overlay reports are not inherently biased towards any class of therapy, and may be used for indicating cell-based, vaccine, and cytokine therapies, for example. The visualized reports may be adjusted to better display tissue type, cell type, TILS%, etc. for the purpose of identifying certain immunotherapies. For example, the techniques herein can be adapted with requisite training data for specific cell- and tumor-type identification. For example, various models can be chosen to predict subtype of tumor (1 a model can predict DCIS, a subtype of tumor in breast cancer which is associated with higher risk of developing invasive breast cancer and 2 a model can predict squamous cell carcinoma vs adenocarcinoma in lung cancer). Another example is the model can be adjusted to predict subtypes of cells in lymphatic tissues, such as lymphocytes vs plasma cells.

[0080] The digital tissue segmenter **201** may calculate relative densities of each type of immune cell on an entire slide, in the areas designated as tumor or another tissue class. Immune tissue classes include lymphocytes, cytotoxic T cells, B cells, NK cells, macrophages, etc.

[0081] In one example, the act of scanning or otherwise digitally capturing a histology slide automatically triggers the digital tissue segmenter **201** to analyze the digital image of that histology slide.

[0082] In one example, the digital tissue segmenter **201** allows a user to edit a cell outer edge or a border between two tissue classes on a tissue class overlay map or a cell outer edge overlay map and saves the altered map as a new overlay.

[0083] FIG. 4 is a flowchart of a method for preparing digital images of histology slides for tissue classification and mapping analysis.

[0084] In one example, each digital image file received by the digital tissue segmenter **201** contains multiple versions of the same image content, and each version has a different resolution. The file stores these copies in stacked layers, arranged by resolution such that the highest resolution image containing the greatest number of bytes is the bottom layer. This is known as a pyramidal structure. In one example, the highest resolution image is the highest resolution achievable by the scanner or camera that created the digital image file.

[0085] In one example, each digital image file also contains metadata that indicates the resolution of each layer. The digital tissue segmenter **201** can detect the resolution of each layer in this metadata and compare it to user-selected resolution criteria to select a layer with optimal resolution for analysis. In one example, the optimal resolution is 1 pixel per micron (downsampled by 4).

[0086] In one example, the digital tissue segmenter **201** receives a Tagged Image File Format (TIFF) file with a bottom layer resolution of four pixels per micron. This resolution of 4 pixels per micron corresponds to the resolution achieved by a microscope objective lens with a magnification power of "40x". In one example, the area that may have tissue on the slide is up to 100,000 x 100,000 pixels in size.

[0087] In one example, the TIFF file has approximately 10 layers, and the resolution of each layer is half as high as the resolution of the layer below it. If the higher resolution layer had a resolution of four pixels per micron, the layer above it will have two pixels per micron. The area represented by one pixel in the upper layer will be the size of the area represented by four pixels in the lower layer, meaning that the length of each side of the area represented by one upper layer pixel will be twice the length of each side of the area represented by one lower layer pixel.

[0088] Each layer may be a 2x downsampling of the layer below it. Downsampling is a method by which a new version of an original image can be created with a lower resolution value than the original image. There are many methods known in the art for downsampling, including nearest-neighbor, bilinear, hermite, bell, Mitchell, bicubic, and Lanczos resampling.

[0089] In one example, 2x downsampling means that the red green blue (RGB) values from three of four pixels that are located in a square in the higher resolution layer are replaced by the RGB value from the fourth pixel to create a new, larger pixel in the layer above, which occupies the same space as the four averaged pixels.

[0090] In one example, the digital image file does not contain a layer or an image with the optimal resolution. In this case, the digital tissue segmenter **201** can receive an image from the file having a resolution that is higher than the optimal resolution and downsample the image at a ratio that achieves the optimal resolution.

[0091] In one example, the optimal resolution is 2 pixels per micron, or “20x” magnification, but the bottom layer of a TIFF file is 4 pixels per micron and each layer is downsampled 4x compared to the layer below it. In this case, the TIFF file has one layer at 40x and the next layer at 10x magnification, but does not have a layer at 20x magnification. In this example, the digital tissue segmenter **201** reads the metadata and compares the resolution of each layer to the optimal resolution and does not find a layer with the optimal resolution. Instead, the digital tissue segmenter **201** retrieves the 40x magnification layer, then downsamples the image in that layer at a 2x downsampling ratio to create an image with the optimal resolution of 20x magnification.

[0092] After the digital tissue segmenter **201** obtains an image with an optimal resolution, it transmits the image to the tissue detector **204**, which locates all parts of the image that depict tumor sample tissue and digitally eliminates debris, pen marks, and other non-tissue objects.

[0093] In one example, the tissue detector **204** differentiates between tissue and non-tissue regions of the image and uses gaussian blur removal to edit pixels with non-tissue objects. In one example, any control tissue on a slide that is not part of the tumor sample tissue can be detected and labeled as control tissue by the tissue detector or manually labeled by a human analyst as control tissue that should be excluded from the downstream tile grid projections.

[0094] Non-tissue objects include artifacts, markings, and debris in the image. Debris includes keratin, severely compressed or smashed tissue that cannot be visually analyzed, and any objects that were not collected with the sample.

[0095] In one example, a slide image contains marker ink or other writing that the tissue detector **204** detects and digitally deletes. Marker ink or other writing may be transparent over the tissue, meaning that the tissue on the slide may be visible through the ink. Because the ink

of each marking is one color, the ink causes a consistent shift in the RGB values of the pixels that contain stained tissue underneath the ink compared to pixels that contain stained tissue without ink.

[0096] In one example, the tissue detector **204** locates portions of the slide image that have ink by detecting portions that have RGB values that are different from the RGB values of the rest of the slide image, where the difference between the RGB values from the two portions is consistent. Then, the tissue detector may subtract the difference between the RGB values of the pixels in the ink portions and the pixels in the non-ink portions from the RGB values of the pixels in the ink portions to digitally delete the ink.

[0097] In one example, the tissue detector **204** eliminates pixels in the image that have low local variability. These pixels represent artifacts, markings, or blurred areas caused by the tissue slice being out of focus, an air bubble being trapped between the two glass layers of the slide, or pen marks on the slide.

[0098] In one example, the tissue detector **204** removes these pixels by converting the image to a grayscale image, passing the grayscale image through a gaussian blur filter that mathematically adjusts the original grayscale value of each pixel to a blurred grayscale value to create a blurred image. Other filters may be used to blur the image. Then, for each pixel, the tissue detector **204** subtracts the blurred grayscale value from the original grayscale value to create a difference grayscale value. In one example, if a difference grayscale value of a pixel is less than a user-defined threshold, it may indicate that the blur filter did not significantly alter the original grayscale value and the pixel in the original image was located in a blurred region. The difference grayscale values may be compared to a threshold to create a binary mask that indicates where the blurred regions are that may be designated as non-tissue regions. A mask may be a copy of an image, where the colors, RGB values, or other values in the pixels are adjusted to show the presence or absence of an object of a certain type to show the location of all objects of that type. For example, the binary mask may be generated by setting the binary value of each pixel to 0 if the pixel has a difference grayscale value less than a user-defined blur threshold and setting the binary value of each pixel to 1 if the pixel has a difference grayscale value higher than or equal to a user-defined blur threshold. The regions of the binary mask that have pixel binary values of 0 indicate blurred areas in the original image that may be designated as non-tissue.

[0099] The tissue detector **204** may also mute or remove extreme brightness or darkness in the image. In one example, the tissue detector **204** converts the input image to a grayscale image, and each pixel receives a numerical value according to how bright the pixel is. In one example, the grayscale values range from 0 to 255, where 0 represents black and 255 represents white. In pixels with a grayscale value above a brightness threshold value, the tissue detector will replace the grayscale value of those pixels with the brightness threshold value. For pixels with a grayscale value below a darkness threshold value, the tissue detector will replace the grayscale value of those pixels equal with the darkness threshold value. In one example, the brightness threshold value is approximately 210. In one example, the darkness threshold value is approximately 45. The tissue detector stores the image with the new grayscale values in a data file.

[0100] In one example, the tissue detector **204** analyzes the altered image for any artifacts, debris, or markings that remain after the first analysis. The tissue detector scans the image and categorizes any remaining groups of pixels with a certain color, size, or smoothness as non-tissue.

[0101] In one example, the slide has H and E staining and most tissue in the slide image will have a pink stain. In this example, the tissue detector **204** categorizes all objects without any pink or red hue, as determined by the RGB value of the pixels that represent the object, as non-tissue. The tissue detector **204** may interpret any color or the lack of any color in a pixel to indicate the presence or absence of tissue in that pixel.

[0102] In one example, the tissue detector **204** detects the contours of each object in the image in order to measure the size and smoothness of each object. Pixels that are very dark may be debris, and pixels that are very bright may be background, which are both non-tissue objects. Therefore, tissue detector **204** may detect the contours of each object by converting the image to grayscale, comparing the grayscale values of each pixel to a range of user-determined range of values that are not too bright or too dark, and determining whether the grayscale value is within the range to produce a binary image where each pixel is assigned one of two numerical values.

[0103] For example, to threshold an image, the tissue detector **204** may compare the grayscale values of each pixel to a user-defined range of values and replace each grayscale value outside of the user-defined range with the value 0 and each grayscale value within a user-

defined range with the value 1. Then, the tissue detector **204** draws all contours of all objects as the outer edge of each group of adjacent pixels having a value of 1. Closed contours indicate the presence of an object, and the tissue detector **204** measures the area within the contours of each object to measure the size of the object.

[0104] In one example, tissue objects on a slide are unlikely to make contact with the outer edges of the slide and the tissue detector **204** categorizes all objects that contact the edge of a slide as non-tissue.

[0105] In one example, after measuring the size of each object, the tissue detector **204** ranks the sizes of all objects and designates the largest value to be the size of the largest object. The tissue detector **204** divides the size of each object by the size of the largest object and compares the resulting size quotient to a user-defined size threshold value. If the size quotient for an object is smaller than the user-defined size threshold value, the tissue detector **204** designates that object as non-tissue. In one example, the user-defined size threshold value is 0.1.

[0106] Before measuring the size of each object, the tissue detector **204** may first downsample the input image to reduce the likelihood of designating portions of a tissue object as non-tissue. For example, a single tissue object may appear as a first tissue object portion surrounded by one or more additional tissue object portions having a smaller size. After thresholding, the additional tissue object portions may have a size quotient smaller than the user-defined size threshold value and may be erroneously designated as non-tissue. Downsampling before thresholding causes a small group of adjacent pixels having values of 1 surrounded by pixels having values of 0 in the original image to be included in a proximal, larger group of pixels having values of 1. The opposite may also be true, for small groups of adjacent pixels having values of 0 surrounded by pixels having values of 1 in the original image to be included in a proximal, larger group of pixels having values of 0.

[0107] In one example, the tissue detector **204** downsamples an image having 40x magnification by a ratio of 16x, so the magnification of the resulting downsampled image is 40/16x and each pixel in the downsampled image represents 16 pixels in the original image.

[0108] In one example, the tissue detector **204** detects the boundaries of each object on the slide as a cluster of pixels having binary or RGB values that do not equal zero, surrounded by pixels with RGB values equal to zero, indicating an object border. If the pixels forming the

boundaries lie on a relatively straight line, the tissue detector **204** classifies the object as non-tissue. For example, the tissue detector **204** outlines a shape with a closed polygon. If the number of vertices of the polygon is less than a user-defined minimum vertices threshold, the polygon is deemed to be a simple, inorganic shape that is too smooth, and marked as non-tissue.

[0109] FIGS. 5A, 5B, and 5C illustrate an exemplary grid overlay on a digital image of a histology slide for tissue classification. FIGS. 5A and 5B show examples of how a multi-field-of-view image tile input looks. FIG. 5A is the input image used to get prediction result of single central small tile. FIG. 5B is the input of image to get prediction results of a matrix of central small tiles. The grid overlays shown in FIGS. 5A and 5B, represent the tiling of the image that is performed by the PhiNet architecture, as described herein, e.g., at FIG 6A. In FIG. 5A, a single tile (teal) and its center (yellow) are shown. In FIG. 5B, shown is a grid of tiles, which is what is actually input into PhiNet architecture for inference. The PhiNet architecture processes each of the tiles in parallel, where each of the little yellow center squares receives a probability for each possible class (summing to 100% as described herein). FIG. 3A shows one possible output overlay map for a breast cancer model. There are as many overlay maps as there are models that may be displayed. For example, if the techniques include breast, colorectal, and lung models, then 3 overlay maps may be generated, and the techniques can also show the heat map (probability map) for a single class overlaid over the image.

[0110] After eliminating non-tissue pixels from the image, the tissue detector **204** stores data containing the locations of the remaining pixels representing tissue. In one example, the locations are stored as an image of the remaining pixels which could be a grayscale mask, a list of the locations of each vertex of the polygons that surround the remaining pixels, and/or other coordinates that indicate the locations of the tissue pixels.

[0111] The tissue class tile grid projector **208** receives the data containing the locations of the pixels representing tissue and overlays a tile grid over these pixels. In one example, the grid is composed of small square tiles that do not overlap.

[0112] The tissue class locator **216** includes a tissue classification algorithm (see FIGS. 6A through 6C) that assigns a tissue class label to the image represented in each small square tile. In one example, the digital tissue segmenter **201** may report the assigned tissue class label

associated with each small square tile by displaying a grid-based digital overlay map in which each tissue class is represented by a unique color (see FIG. 3A).

[0113] A smaller tile size may cause an increase in the amount of time required for the tissue class locator **216** to analyze the input image. Alternatively, a larger tile size may increase the likelihood that a tile will contain more than one tissue class and make it difficult to assign a single tissue class label to the tile. In this case, the tissue class locator **216** may calculate an equal probability for two or more tissue class labels being accurately assigned to a single small square tile instead of calculating that one of the tissue class labels has a higher probability of describing the image in the small square tile, compared to the other tissue class labels.

[0114] In one example, each side of each small square tile is approximately 32 microns long and approximately 5-10 cells fit in each small square tile. This small tile size allows the algorithm to create more spatially accurate borders when determining the boundary between two neighboring small square tile regions that depict two distinct tissue classes. In one example, each side of the small square tile can be as short as 1 micron.

[0115] In one example, the size of each tile is set by the user to contain a specific number of pixels. In this example, the resolution of the input image will determine the length of each side of the tile, as measured in microns. At different resolutions, the micron length of the tile side will vary and the number of cells in each tile may vary.

[0116] The algorithm recognizes various pixel data patterns in the portion of the digital image that is located within or near each small square tile and assigns a tissue class label to each small square tile based on those detected pixel data patterns. In one example, a medium square tile centered around a small square tile contains the area of a slide image that is close enough to the small square tile to contribute to the label assignment for that small square tile.

[0117] In one example, each side of a medium square tile is approximately 466 microns long, and each medium square tile contains approximately 225 (15x15) small square tiles. In one example, this medium tile size increases the likelihood that structural tissue features can fit within a single medium tile and provide context to the algorithm when labeling the central small square tile. Structural tissue features may include glands, ducts, vessels, immune clusters, etc.

[0118] In one example, this medium tile size is selected to counteract the shrinkage that occurs during convolution.

[0119] During convolution, an input image matrix is multiplied by a filter matrix to create a result matrix, and shrinkage refers to a case where the result matrix is smaller than the input image matrix. The dimensions of a filter matrix in a convolution layer affects the number of rows and columns lost to shrinkage. The total number of matrix entries that are lost to shrinkage by processing an image through a particular CNN can be calculated depending on the number of convolution layers in the CNN and the dimensions of the filter matrices in each convolution layer. (See FIGS. 6A through 6C)

[0120] In the example shown in FIG. 6B, the convolution layers in combination lose 217 total matrix rows or columns from the top, bottom, and two side edges of the matrix, so the medium square tile is set to equal the small square tile plus 217 pixels on each side of the small square tile.

[0121] In one example, two neighboring small square tiles share a side and are each at the center of a medium square tile. The two medium square tiles overlap. Of the 466×466 small pixels located in each medium square tile, the two medium square tiles will share all but 32×466 pixels. In one example, each convolution layer of the algorithm (see FIGS. 6A through 6B) analyzes both medium square areas simultaneously such that the algorithm produces two vectors of values (one for each of the two small square tiles).

[0122] The vector of values contains a probability value for each tissue class label, indicating the likelihood that the small square tile depicts that tissue class. The vectors of values will be arranged in a matrix, to form a 3-dimensional probability data array. The location of each vector in the 3-dimensional probability data array, relative to the other vectors, will correspond to the location of the associated small square tile, relative to the other small square tiles included in the algorithm analysis.

[0123] In the example, 434×434 (188,356) of the 466×466 (217,156) pixels in each medium square tile are common to both medium square tiles. By analyzing both medium square tiles simultaneously, the algorithm increases efficiency.

[0124] In one example, the algorithm can further increase efficiency by analyzing a large tile formed by multiple overlapping medium square tiles, each of which contains many small square tiles surrounding one center small square tile that receives a tissue class label. In this example, the algorithm still generates one data structure in the form of a 3-dimensional probability data array containing one vector of probabilities for each small square tile, wherein the location of the

vector within the 3-dimensional array corresponds to the location of the small tile within the large tile.

[0125] The tissue class locator **216** saves this 3-dimensional probability data array and the overlay map generator **224** converts the tissue class label probabilities for each small square tile into a tissue class overlay map. In one example, the overlay map generator **224** may compare the probabilities stored in each vector to determine the largest probability value associated with each small square tile. The tissue class label associated with that largest value may be assigned to that small square tile and only the assigned labels will be displayed in the tissue class overlay map.

[0126] In one example, matrices generated by each layer of the tissue class locator **216** for the large square tile are stored in graphics processing unit (GPU) memory. The capacity of the GPU memory and the amount of GPU memory required for each entry in the 3-dimensional probability data array may determine the maximum possible size of the large square tile. In one example, the GPU memory capacity is 250 MB and each entry in the matrices requires 4 bytes of GPU memory. This allows a large tile size of 4,530 pixels by 4,530 pixels, calculated as follows: 4 bytes/entry * 4530*4530*3 entries for each large tile = 246 (~250) MB of GPU memory required per large square tile. In another example, each entry in the matrices requires 8 bytes of GPU memory. In this example, a 16GB GPU can process 32 large tiles simultaneously, each large tile having dimensions of 4,530 pixels by 4,530 pixels, calculated as follows: 32 large tiles * 8 bytes/entry * 4530*4530*3 entries for each large tile = 14.7 (~16) GB of GPU memory required.

[0127] In one example, each entry in the 3-dimensional probability data array is a single precision floating-point format (float32) data entry.

[0128] In one example, there are 16,384 (128^2) non-overlapping small square tiles that form a large square tile. Each small square tile is the center of a medium square tile having sides that are each approximately 466 pixels long. The small square tiles form a center region of a large square tile having sides that are each approximately 4,096 pixels long. The medium square tiles all overlap and create a border around all four sides of the center region that is approximately 217 pixels wide. Including the border, each large square tile has sides that are each approximately 4,530 pixels long. (See FIG. 5C)

[0129] In this example, this large square tile size allows simultaneous calculations that reduce the redundant computation percentage by 99%. This may be calculated as follows: first, select a pixel on the interior of a large square tile (any pixel at least 434 pixels from the edge of the large square tile). Construct a region that is the size of a medium square tile (466 pixels per edge) with this model pixel at the center. Then for any small square tile centered within this constructed region, the model pixel is contained within that small square tile's corresponding medium square tile. There are $(466/32)^2 = \sim 217$ such small square tiles within the large square tile. For pixels not on the interior of the large square tile, the number of small square tiles that satisfy this condition is smaller. The number decreases linearly as the distance between the selected small square tile and the edge of the large square tile decreases, then again as the distance between the selected small square tile and the corner decreases, where a small number of pixels ($\sim 0.005\%$) only contribute towards the classification of a single small square tile. Performing classification on a single large square tile means the computations for each pixel are only performed once, instead of once per small square tile. Thus, the redundancy is reduced by nearly 217-fold. In one example, redundancy is not completely eliminated because a slide may contain several large square tiles, each of which may overlap slightly with its neighbors.

[0130] An upper bound on the redundant calculation percentage can be established (slight deviation from this upper bound depends on the number of large square tiles needed to cover the tissue and the relative arrangement of these tiles). The redundancy percentage is $1 - 1/r$ where r is the redundancy ratio, and r can be calculated as $(T/N + 1)(\sqrt{N} \cdot E + 434)^2 / (\sqrt{T} \cdot E + 434)^2$; T is the total number of small square tiles on the slide, N is the number of small square tiles per large square tile, and E is the edge size of the small square tiles.

[0131] FIG. 6A illustrates the layers of an example tissue segmentation algorithm. FIG. 6B compares the layers of a known image classification algorithm with an example tissue segmentation algorithm.

[0132] In one example, the tissue class locator **216** includes a convolutional neural network (CNN), a deep learning algorithm that has been trained by a training set to recognize pixel data patterns.

[0133] In one example, the tissue class locator **216** includes a tile-resolution fully convolutional network (FCN) black box deep learning model based on a known CNN ResNet-18

image recognition model. In one example, the tile-resolution FCN is a multi-tile algorithm known as PhiNet, the layers of which are shown in FIG. 6A.

[0134] FIG. 6B illustrates the differences between the ResNet-18 algorithm on the left, and the tile-resolution FCN PhiNet shown in FIG. 6A and on the right half of FIG. 6B. Compared to the ResNet-18 algorithm, the tile-resolution FCN PhiNet included in the digital tissue segmenter **201** has additional layers of 1x1 convolution in a skip connection, downsampling by a factor of 8 in a skip connection, and a confidence map layer, and replaces an average pooling layer with a concatenation layer, and a fully connected (fc N) layer with a 1x1 convolution and softmax layer.

[0135] The added layers convert a classification task into a classification-segmentation task. This means that instead of receiving and classifying a whole image as one tissue class label, the added layers allow the tile-resolution FCN to classify each small tile in the user-defined grid as a tissue class.

[0136] These added and replacement layers convert a CNN to a tile-resolution FCN without requiring the upsampling performed in the later layers of traditional pixel-resolution FCNs. Upsampling is a method by which a new version of an original image can be created with a higher resolution value than the original image. Upsampling is a time-consuming, computation-intensive process.

[0137] There are many methods known in the art for upsampling, including nearest-neighbor, bilinear, hermite, bell, Mitchell, bicubic, and Lanczos resampling. In one example, 2x upsampling means that a pixel with red green blue (RGB) values will be split into four pixels, and the RGB values for the three new pixels may be selected to match the RGB values of the original pixel. In another example, the RGB values for the three new pixels may be selected as the average of the RGB values from the original pixel and the pixels that are adjacent to the neighboring pixel.

[0138] Because the RGB values of the new pixels may not accurately reflect the visible tissue in the original slide that was captured by the digital slide image, upsampling can introduce errors into the final image overlay map produced by the overlay map generator **224**.

[0139] Additionally, traditional pixel-resolution FCNs require training data wherein each pixel is labeled, which is not feasible to produce from digital histology slides because it is too time

consuming and because each individual pixel does not contain enough contextual information to identify a tissue class.

[0140] In one example, instead of labeling individual pixels, the tile-resolution FCN is programmed to analyze a large square tile made of small square tiles (see FIG. 5C), producing a 3D array of values that each represent the probability that one tissue class classification label matches the tissue class depicted in each small tile.

[0141] A convolution layer, known in the art, performs the multiplication of at least one input image matrix by at least one filter matrix. In the first convolution later, the input image matrix has a value for every pixel in the large square tile input image, representing visual data in that pixel (for example, a value between 0 and 255 for each channel of RGB).

[0142] The filter matrix has dimensions selected by the user, and may contain weight values selected by the user or determined by backpropagation during CNN model training. In one example, in the first convolution layer, the filter matrix dimensions are 7x7 and there are 64 filters. The filter matrix may represent visual patterns that can distinguish one tissue class from another.

[0143] In an example where RGB values populate the input image matrix, the input image matrix and the filter matrices will be 3-dimensional. (See FIG. 6C)

[0144] Each filter matrix is multiplied by each input image matrix to produce a result matrix. All result matrices produced by the filters in one convolution layer may be stacked to create a 3-dimensional result matrix having dimensions such as rows, columns, and depth. The last dimension, depth, in the 3-D result matrix will have a depth equal to the number of filter matrices.

[0145] The result matrix from one convolution layer becomes the input image matrix for the next convolution layer.

[0146] A convolution layer title that includes “/n”, where n is a number, indicates that there is a downsampling (also known as pooling) of the result matrix produced by that layer. The n indicates the factor by which the downsampling occurs. Downsampling by a factor of 2 means that a downsampled result matrix with half as many rows and half as many columns as the original result matrix will be created by replacing a square of four values in the result matrix by

one of those values or a statistic calculated from those values. For example, the minimum, maximum, or average of the values may replace the original values.

[0147] The algorithm shown also adds skip connections (shown in FIG. 6A as black lines with arrows that connect blue convolution layers directly to the concatenation layer). The skip connection on the left includes downsampling by a factor of 8, and the skip connection on the right includes two convolution layers that multiply an input image matrix by filter matrices that each have dimensions of 1x1. Because of the 1x1 dimensions of the filter matrices in these layers, only an individual small square tile contributes to its corresponding probability vector in the result matrices created by the purple convolution layers. These result matrices represent a small focus of view.

[0148] In all of the other convolution layers, the larger dimensions of the filter matrices allow the pixels in each medium square tile, including the small square tile at the center of the medium square tile, to contribute to the probability vector in the result matrix that corresponds with that small square tile. These result matrices allow the contextual pixel data patterns surrounding the small square tile to influence the probability that each tissue class label applies to the small square tile. These result matrices represent a large focus of view.

[0149] The 1x1 convolution layers in the skip connection allow the algorithm to regard the pixel data patterns in the center small square tile as more or less important than pixel data patterns in the rest of the surrounding medium square tile. This is reflected by the weights that the trained model multiplies by the final result matrix from the skip connection layers (shown on the right side of FIG. 6A) compared to the weights that the trained model multiplies by the final result matrix from the medium tile convolution layers (shown in the center column of FIG. 6A) during the concatenation layer.

[0150] The downsampling skip connection shown on the left side of FIG. 6A creates a result matrix with a depth of 64. The 3x3 convolution layer having 512 filter matrices creates a result matrix with a depth of 512. The 1x1 convolution layer having 64 filter matrices creates a result matrix with a depth of 64. All three of these results matrices will have the same number of rows and the same number of columns. The concatenation layer concatenates these three results matrices to form a final result matrix with the same number of rows and the same number of columns as the three concatenated matrices, and a depth of $64+512+64$ (640). This final result matrix combines the large and small focus of view matrices.

[0151] The final result matrix may be flattened to 2 dimensions by multiplying a factor by every entry, and summing the products along each depth. Each factor may be selected by the user, or may be selected during model training by backpropagation. Flattening will not change the number of rows and columns of the final results matrix, but will change the depth to 1.

[0152] The 1x1 convolution layer receives the final result matrix and filters it with one or more filter matrices. The 1x1 convolution layer may include one filter matrix associated with each tissue class label in the trained algorithm. This convolution layer produces a 3-D result matrix that has a depth equal to the number of tissue class labels. Each depth corresponds to one filter matrix and along the depth of the result matrix there may be a probabilities vector for each small square tile. This 3-D result matrix is the 3-dimensional probability data array, and the 1x1 convolution layer stores this 3-D probability data array.

[0153] A softmax layer may create a 2-dimensional probability matrix from the 3-D probability data array by comparing every value in each probabilities vector and selecting the tissue class associated with the maximum value to assign that tissue class to the small square tile associated with that probabilities vector.

[0154] The stored 3-dimensional probability data array or the 2-D probability matrix may then be converted to a tissue class overlay map in the final confidence map layer in FIG. 6A, to efficiently assign a tissue class label to each tile.

[0155] In one example, to counteract shrinkage, input image matrices have added rows and columns on all four outer edges of the matrices, wherein each value entry in the added rows and columns is a zero. These rows and columns are referred to as padding. In this case, the training data input matrices will have the same number of added rows and columns with value entries equal to zero. A difference in the number of padding rows or columns in the training data input matrices would result in values in the filter matrices that do not cause the tissue class locator **216** to accurately label input images.

[0156] In the FCN shown in FIG. 6B, 217 total outer rows or columns on each side of the input image matrix will be lost to shrinkage before the skip connection, due to the gray and blue layers. Only the pixels located in the small square tiles will have a corresponding vector in the result matrices created by the green layers and beyond.

[0157] In one example, each medium square tile is not padded by adding rows and columns with value entries of zero around the input image matrix that corresponds to each medium square tile because the zeroes would replace image data values from neighboring medium square tiles that the tissue class locator **216** needs to analyze. In this case, the training data input matrices will not be padded either.

[0158] FIG. 6C is a visualization of each depth of an exemplary 3-dimensional input image matrix being convoluted by two exemplary 3-dimensional filter matrices.

[0159] In an example where an input image matrix contains RGB channels for each medium square tile, the input image matrix and filter matrices will be 3-dimensional. In one of the three dimensions, the input image matrix and each filter matrix will have three depths, one for red channel, one for green channel, and one for blue channel.

[0160] The red channel (first depth) of the input image matrix is multiplied by the corresponding first depth of the first filter matrix. The green channel (second depth) is multiplied in a similar fashion, and so on with the blue channel (third depth). Then, the red, green, and blue product matrices are summed to create a first depth of the 3-dimensional result matrix. This repeats for each filter matrix, to create an additional depth of the 3-dimensional result matrix that corresponds to each filter.

[0161] A variety of training sets may be used to train a CNN or FCN that is included in the tissue class locator **216**.

[0162] In one example, the training set may contain JPEG images of medium square tiles, each having a tissue class label assigned to its center small square tile, taken from at least 50 digital images of histology slides at a resolution of approximately 1 pixel per micron. In one example, a human analyst has outlined and labeled all relevant tissue classes or labeled each small square tile in each histology slide as non-tissue or as a specific type of cells. Classes of tissue may include tumor, stroma, normal, immune cluster, necrosis, hyperplasia/dysplasia, and red blood cells. In one example, each side of each center small square tile is approximately 32 pixels long.

[0163] In one example, the training set images are converted to input training image matrices and processed by the tissue class locator **216** to assign a tissue class label to each training set image. If the tissue class locator **216** does not accurately label the validation set of training

images to match the corresponding annotations added by a human analyst, the weights of each layer of the deep learning network may be adjusted automatically by stochastic gradient descent through backpropagation until the tissue class locator **216** accurately labels most of the validation set of training images.

[0164] In one example, the training data set has multiple classes where each class represents a tissue class. That training set will generate a unique model with specific hyperparameters (number of epochs, learning rate, etc.) that can recognize and classify the content in a digital slide image into different classes. Tissue classes may include tumor, stroma, immune cluster, normal epithelium, necrosis, hyperplasia/dysplasia, and red blood cells. In one example, the model can classify an unlimited number of tissue classes, provided each tissue class has a sufficient training set.

[0165] In one example, the training set images are converted into grayscale masks for annotation where different values (0-255) in the mask image represent different classes.

[0166] Each histology slide can exhibit large degrees of variation in visual features, including tumor appearance, so a training set may include digital slide images that are highly dissimilar to better train the model for the variety of slides that it may analyze.

[0167] A training set may also be specific to a cancer type. In this case, all of the histology slides that generated the digital images in a particular training set contain a tumor sample from the same cancer type. Cancer types may include breast, colorectal, lung, pancreatic, liver, stomach, skin, etc. Each training set may create a unique model specific to the cancer type. Each cancer type may also be split into cancer subtypes, known in the art.

[0168] In one example, a training set may be derived from histology slide pairs. A histology slide pair includes two histology slides that each have one slice of tissue, wherein the two slices of tissue were adjacent in the tumor sample. Therefore, the two slices of tissue are substantially similar. One of the slides in the pair is stained with H and E staining only, and the other slide in the pair is stained with IHC staining for a specific molecule target. The areas on the H and E stained slide that correspond to areas where IHC staining appears in the paired slide are annotated by a human analyst as containing a specific molecule target and the tissue class locator receives the annotated H and E slides as a training set.

[0169] In this case, the trained tissue class locator **216** receives digital images of H and E stained tissue to predict tiles that may contain a given molecule target and the overlay map generator **224** produces an overlay map showing which tiles contain a given molecule. In one example, the resolution of the overlay is at the level of an individual cell.

[0170] The overlay produced by a model trained by one or more training sets may be reviewed by a human analyst in order to annotate the digital slide image to add it to one of the training sets.

[0171] The pixel data patterns that the algorithm detects may represent visually detectable features. Some examples of those visually detectable features may include color, texture, cell size, shape, and spatial organization.

[0172] For example, color on a slide provides contextual information. An area on the slide that is purple has a higher density of cells and is more likely to be invasive tumor. Tumor also causes the surrounding stroma to become more fibrous in a desmoplastic reaction, which causes normally pink stroma to appear blue-grey. Color intensity also helps to identify individual cells of a certain type (for example, lymphocytes are uniformly very dark blue).

[0173] Texture refers to the distribution of stain within cells. Most tumor cells have a rough, heterogeneous appearance, with light pockets and dark nucleoli within their nuclei. A zoomed-out field of view with many tumor cells will have this rough appearance. Many non-tumor tissue classes each has distinguishing features. Furthermore, patterns of tissue classes that are present in a region can indicate the type of tissue or cell structures present in that region.

[0174] Additionally, cell size often indicates tissue class. If a cell is several times larger than normal cells elsewhere on the slide, the probability is high that it is a tumor cell.

[0175] The shape of individual cells, specifically how circular they are, can indicate what type of cell they are. Fibroblasts (stromal cells) are normally elongated and slim, while lymphocytes are very round. Tumor cells can be more irregularly shaped.

[0176] The organization of a group of cells can also indicate tissue class. Frequently, normal cells are organized in a structured and recognizable pattern, but tumor cells grow in denser, disorganized clusters. Each type and subtype of cancer can produce tumors with specific growth patterns, which include cell location relative to tissue features, the spacing of tumor cells relative to each other, formation of geometric elements, etc.

[0177] FIG. 7 illustrates an example of a first classification model for labeling an H&E Slide or IHC Slide using tile classifications and cell classifications. As shown, an exemplary model overlay may include red tiles for identifying clusters of lymphocytes, green tiles for identifying epithelium cells, and dark blue tiles for identifying tumor cells. The tile classifications are shown in the legend and generated using a PhiNet model architecture. The cell legend showing an example cell classification is shown, as generated used a UNet model architecture. In some examples, where a tile has a predicted tile classification and that tile corresponds to identified cell (i.e., cell object) with a classification, the cell classification may override, such that the segmenter **201** assigns the cell classification to that tile in place of the tile classification. In another example, an exemplary overlay may represent individual lymphocytes by coloring them red to distinguish between clusters of lymphocytes and individual lymphocytes infiltrating other tissues. In one example, individual lymphocytes detected within tumor tiles represent tumor-infiltrating lymphocytes which would be red dots within dark blue or green tiles.

[0178] In one example, the cell type locator **220** includes at least one cell outlining algorithm for detecting the outer edge of each cell and reporting the coordinates of every pixel that contains an outer edge of a cell to create a cell outer edge overlay map. (See FIG. 2B)

[0179] The cell type tile grid projector **212** may overlay a grid on the digital image of a histology slide. A smaller tile size may improve the training data set by increasing the number of tiles in the training data set, and may be easier for a human analyst to visually inspect before training the cell type locator **220**. The size of the tiles may be determined by the capacity of the memory.

[0180] The cell type locator **220** may recognize features that include changes in color and brightness between adjacent pixels, and the shape of areas with changes in color or brightness. The cell type locator **220** may also report the coordinates of the center of each detected cell and of all vertices of a polygon that encloses a single detected cell. The cell type locator **220** may include a radial symmetry filter or an edge detector such as a Laplacian edge filter or a Sobel edge filter. In one example, the cell type locator **220** is a UNet model, which is a three-class semantic segmentation FCN model.

[0181] In one example, the UNet is a three-class model, which means that the algorithm may classify each pixel as either background exterior to every cell, cell outer edge, or cell interior. This facilitates the counting of each individual cell, especially when two or more cells overlap

each other. In one example, tumor infiltrating lymphocytes will overlap tumor cells. In traditional two-class cell outlining models that only label whether a pixel contains a cell outer edge or not, each clump of two or more overlapping cells would be counted as one cell. Two-class UNet models with binary classification (cell vs background) are known in the art, but the three-class UNet model allows a type of classification that is not binary, which requires adaptation with the use of a different loss function. The classified results may be stored as matrix with different values to indicate cell exterior, cell outer edge, and cell interior. A training process is done by having an original image and a mask image together as input to a UNet model (e.g., the mask may be an image with pixel values at 0, 128 and 255, where 0 represent cell exterior, 128 represent cell outer edge, 255 represent cell interior).

[0182] In one example, to avoid the possibility that a cell that spans two tiles is counted twice, the algorithm may add a buffer around all four sides of each tile that is slightly wider than an average cell. The intention is to only count cells that appear in the center, non-buffered region for each tile. In this case, tiles will be placed so that the center, non-buffered region of neighboring tiles are adjacent and non-overlapping. Neighboring tiles will overlap in their respective buffer regions.

[0183] In one example, the cell segmentation algorithm includes two UNet models. One UNet model may be trained with images patches of mixed tissue classes, where a human analyst has highlighted a bounding polygon for each cell and classified each cell according to tissue class. In one example, training data includes digital slide images where every pixel has been labeled as either the interior of a cell, the outer edge of a cell, or the background which is exterior to every cell. In another example, the training data includes digital slide images where every pixel has been labeled with a yes or no to indicate whether it depicts the outer edge of a cell. The labels may be two-dimensional matrices with the same shapes as their corresponding image patches, where each matrix element takes on one of several values (the exact number of values is determined by the number of unique classes in the model).

[0184] Another UNet model may be trained with images of many cells of a single tissue class, or images of a diverse set of cells where cells of only one tissue class are outlined in a binary mask. In one example, the training set is labeled by associating a first value with all pixels showing a cell type of interest, a second value to all background pixels, and optionally a third value to pixels along the border of each cell (in the case of a three-class model instead of a two-class model). Visually, an image labeled in this way might appear as a black and white image

wherein all pixels showing a tissue class of interest would be white and all other pixels would be black, or vice versa. (See FIG. 8) For example, the images may have only labeled lymphocytes. This UNet model can recognize the outer edges of that particular cell type and assign a label to cells of that type in the digital image of the slide.

[0185] In one example, a non-UNet model may be used instead of a UNet model to label cells that are detected by a first UNet model. This non-UNet model could be another CNN that has a different architecture, or a non deep learning model like random forest or support vector machine.

[0186] In one example, the cell labels generated by the cell type locator **220** can be augmented by the tissue region labels generated by the tissue class locator **216**. If a cell is detected but not labeled a specific class by the cell type locator, it may be assigned a default label that is the label of the tissue region in which it is found.

[0187] In one example, the images in both training sets have a resolution of 2 pixels per micron.

[0188] The techniques herein, applying PhiNet architecture as described, may be used to predict other indicators, as well, such as microsatellite instability (MSI), consensus molecular subtypes (CMS), and other molecular biomarkers. Classification models may be trained to predict from H&E slides these biomarkers, using other additional information, such as other characteristics of the patient that may be included in the patient's electronic health record. Other biomarkers that can be predicted using trained PhiNet architecture include: tumor grade and/or stage, tumor origin (for TUO cases), histological subtypes (such as adenocarcinoma and squamous cell carcinoma), other biomarkers such as cell based, vaccine, or cytokine therapies).

[0189] Further, wherein exemplary embodiments disclosed herein are illustrated with H&E slides prepared by a pathologist, similar models may be generated for IHC slides as well. Exemplary IHC overlays may classify tiles according to a status of PD-L1 'presence' in the cell. In an example implementation, a digital tissue segmenter similar to that of segmenter **201** in FIG. 2 may be configured to as a digital PD-L1 tumor positive score (TPS) predictor. In an example, this PD-L1 TPS digital tissue segmenter includes a tissue detector, tissue class tile grid projector, and tissue class locator, like that of segmenter **201** in FIG. 2. The output of the tissue class locator 216 may be provided to a PD-L1 TPS metric calculator and overlay map generator (replacing the element **224** in the segmenter **201**) configured to count the number of

Tumor + tiles and divide by the total number of tumor tiles (i.e., the sum of Tumor + and Tumor - tiles), to determine a predicted PD-L1 TPS. Other quantities can also be calculated, such as the percent of stained lymphocytes or overall macrophage percentage. The PD-L1 tissue classifier in this example may be configured to predict any number of classes, including blood, epithelium, lymphocyte +, lymphocyte -, macrophage +, necrosis, stroma, tumor +, and tumor -, where “+” indicates positive PD-L1 expression and “-” indicates no PD-L1 expression. In this way, in some examples, a PD-L1 prediction process can be implemented by a digital tissue image segmenter implementing two main processes. First, the digital tissue image segmenter predicts different tissue classes overlaid onto an IHC image (i.e., a similar process as described in reference to H&E images described herein, but with different classes). Second, the digital tissue image segmenter obtains PD-L1 TPS by counting the number of predicted Tumor + tiles and dividing that by the total number of Tumor tiles (i.e. Tumor + added to Tumor -). An example overlay mapping process that may be performed by such a digital PD-L1 tumor positive score (TPS) predictor is shown in FIG. 9, showing different predicted tissue classifications. On the left are various IHC images and on the right are predicted TPS various of different tissue classifications overlaid on the IHC images.

[0190] FIG. 10 illustrates a system **1000** for performing the methods described herein may include a computing device **1002**, and more particularly may be implemented on one or more processing units **1004**, e.g., Central Processing Units (CPUs), and/or on one or more or Graphical Processing Units (GPUs) **1006**, including clusters of CPUs and/or GPUs. Features and functions described may be stored on and implemented from one or more non-transitory computer-readable media **1008** of the computing device. The computer-readable media **1008** may include, for example, an operating system **1010** and software modules, or “engines,” that implement the methods described herein, including a digital tissue image segmenter **1012** implementing methods and processing illustrated and described herein. Example processes include tissue detection, tissue class tile grid projection, cell type tile grid projection, tissue class locator, cell type locator, overlay map generator and metric calculator processes. More generally, the computer-readable media **1008** may store batch normalization process instructions for the engines for implementing the techniques herein. The computing device **1002** may be a distributed computing system, such as an Amazon Web Services cloud computing solution. The computing device **1002** may be implemented on one network accessible processing device **1050** or distributed across multiple such devices **1050**, **1052**, **1054**, etc.

[0191] The computing device **1002** includes a network interface **1014** communicatively coupled to network **1016**, for communicating to and/or from a portable personal computer, smart phone, electronic document, tablet, and/or desktop personal computer, or other computing devices for communicating overlay maps, predicted tile classifications and locations, predicted cell classifications and locations, etc. Such information may also be stored in a database **1024**. The computing device **1002** further includes an I/O interface **1018** connected to devices, such as digital displays **1020** for displaying generator overlay maps, user input devices **1022**, etc. A dashboard generator **1026** may be used to generate GUI and/or other digital displays allowing a user to review and interact with and adjust generated overlay maps, prediction classifications, etc.

[0192] The functions of the engines may be implemented across distributed computing devices, etc. connected to one another through a communication link. In other examples, functionality of the system may be distributed across any number of devices, including the portable personal computer, smart phone, electronic document, tablet, and desktop personal computer devices shown. The computing device **1002** may be communicatively coupled to the network **1016** and another network. The networks **1016** may be public networks such as the Internet, a private network such as that of a research institution or a corporation, or any combination thereof. Networks can include, local area network (LAN), wide area network (WAN), cellular, satellite, or other network infrastructure, whether wireless or wired. The networks can utilize communications protocols, including packet-based and/or datagram-based protocols such as Internet protocol (IP), transmission control protocol (TCP), user datagram protocol (UDP), or other types of protocols. Moreover, the networks can include a number of devices that facilitate network communications and/or form a hardware basis for the networks, such as switches, routers, gateways, access points (such as a wireless access point as shown), firewalls, base stations, repeaters, backbone devices, etc.

[0193] The computer-readable media **1008** may include executable computer-readable code stored thereon for programming a computer (e.g., comprising a processor(s) and GPU(s)) to the techniques herein. Examples of such computer-readable storage media include a hard disk, a CD-ROM, digital versatile disks (DVDs), an optical storage device, a magnetic storage device, a ROM (Read Only Memory), a PROM (Programmable Read Only Memory), an EPROM (Erasable Programmable Read Only Memory), an EEPROM (Electrically Erasable Programmable Read Only Memory) and a Flash memory. More generally, the processing units of the computing device may represent a CPU-type processing unit, a GPU-type processing

unit, a field-programmable gate array (FPGA), another class of digital signal processor (DSP), or other hardware logic components that can be driven by a CPU.

[0194] Throughout this specification, plural instances may implement components, operations, or structures described as a single instance. Although individual operations of one or more methods are illustrated and described as separate operations, one or more of the individual operations may be performed concurrently, and nothing requires that the operations be performed in the order illustrated. Structures and functionality presented as separate components in example configurations may be implemented as a combined structure or component. Similarly, structures and functionality presented as a single component may be implemented as separate components or multiple components. These and other variations, modifications, additions, and improvements fall within the scope of the subject matter herein.

[0195] Additionally, certain embodiments are described herein as including logic or a number of routines, subroutines, applications, or instructions. These may constitute either software (e.g., code embodied on a machine-readable medium or in a transmission signal) or hardware. In hardware, the routines, etc., are tangible units capable of performing certain operations and may be configured or arranged in a certain manner. In example embodiments, one or more computer systems (e.g., a standalone, client or server computer system) or one or more hardware modules of a computer system (e.g., a processor or a group of processors) may be configured by software (e.g., an application or application portion) as a hardware module that operates to perform certain operations as described herein.

[0196] In various embodiments, a hardware module may be implemented mechanically or electronically. For example, a hardware module may comprise dedicated circuitry or logic that is permanently configured (e.g., as a special-purpose processor, such as a microcontroller, field programmable gate array (FPGA) or an application-specific integrated circuit (ASIC)) to perform certain operations. A hardware module may also comprise programmable logic or circuitry (e.g., as encompassed within a processor or other programmable processor) that is temporarily configured by software to perform certain operations. It will be appreciated that the decision to implement a hardware module mechanically, in dedicated and permanently configured circuitry, or in temporarily configured circuitry (e.g., configured by software) may be driven by cost and time considerations.

[0197] Accordingly, the term "hardware module" should be understood to encompass a tangible entity, be that an entity that is physically constructed, permanently configured (e.g., hardwired), or temporarily configured (e.g., programmed) to operate in a certain manner or to perform certain operations described herein. Considering embodiments in which hardware modules are temporarily configured (e.g., programmed), each of the hardware modules need not be configured or instantiated at any one instance in time. For example, where the hardware modules comprise a processor configured using software, the processor may be configured as respective different hardware modules at different times. Software may accordingly configure a processor, for example, to constitute a particular hardware module at one instance of time and to constitute a different hardware module at a different instance of time.

[0198] Hardware modules can provide information to, and receive information from, other hardware modules. Accordingly, the described hardware modules may be regarded as being communicatively coupled. Where multiple of such hardware modules exist contemporaneously, communications may be achieved through signal transmission (e.g., over appropriate circuits and buses) that connects the hardware modules. In embodiments in which multiple hardware modules are configured or instantiated at different times, communications between such hardware modules may be achieved, for example, through the storage and retrieval of information in memory structures to which the multiple hardware modules have access. For example, one hardware module may perform an operation and store the output of that operation in a memory device to which it is communicatively coupled. A further hardware module may then, at a later time, access the memory device to retrieve and process the stored output. Hardware modules may also initiate communications with input or output devices, and can operate on a resource (e.g., a collection of information).

[0199] The various operations of the example methods described herein can be performed, at least partially, by one or more processors that are temporarily configured (e.g., by software) or permanently configured to perform the relevant operations. Whether temporarily or permanently configured, such processors may constitute processor-implemented modules that operate to perform one or more operations or functions. The modules referred to herein may, in some example embodiments, comprise processor-implemented modules.

[0200] Similarly, the methods or routines described herein may be at least partially processor-implemented. For example, at least some of the operations of a method can be performed by one or more processors or processor-implemented hardware modules. The

performance of certain of the operations may be distributed among the one or more processors, not only residing within a single machine, but also deployed across a number of machines. In some example embodiments, the processor or processors may be located in a single location (e.g., within a home environment, an office environment or as a server farm), while in other embodiments the processors may be distributed across a number of locations.

[0201] The performance of certain of the operations may be distributed among the one or more processors, not only residing within a single machine, but also deployed across a number of machines. In some example embodiments, the one or more processors or processor-implemented modules may be located in a single geographic location (e.g., within a home environment, an office environment, or a server farm). In other example embodiments, the one or more processors or processor-implemented modules may be distributed across a number of geographic locations.

[0202] Unless specifically stated otherwise, discussions herein using words such as "processing," "computing," "calculating," "determining," "presenting," "displaying," or the like may refer to actions or processes of a machine (e.g., a computer) that manipulates or transforms data represented as physical (e.g., electronic, magnetic, or optical) quantities within one or more memories (e.g., volatile memory, non-volatile memory, or a combination thereof), registers, or other machine components that receive, store, transmit, or display information.

[0203] As used herein any reference to "one embodiment" or "an embodiment" means that a particular element, feature, structure, or characteristic described in connection with the embodiment is included in at least one embodiment. The appearances of the phrase "in one embodiment" in various places in the specification are not necessarily all referring to the same embodiment.

[0204] Some embodiments may be described using the expression "coupled" and "connected" along with their derivatives. For example, some embodiments may be described using the term "coupled" to indicate that two or more elements are in direct physical or electrical contact. The term "coupled," however, may also mean that two or more elements are not in direct contact with each other, but yet still co-operate or interact with each other. The embodiments are not limited in this context.

[0205] As used herein, the terms "comprises," "comprising," "includes," "including," "has," "having" or any other variation thereof, are intended to cover a non-exclusive inclusion. For

example, a process, method, article, or apparatus that comprises a list of elements is not necessarily limited to only those elements but may include other elements not expressly listed or inherent to such process, method, article, or apparatus. Further, unless expressly stated to the contrary, "or" refers to an inclusive or and not to an exclusive or. For example, a condition A or B is satisfied by any one of the following: A is true (or present) and B is false (or not present), A is false (or not present) and B is true (or present), and both A and B are true (or present).

[0206] In addition, use of the "a" or "an" are employed to describe elements and components of the embodiments herein. This is done merely for convenience and to give a general sense of the description. This description, and the claims that follow, should be read to include one or at least one and the singular also includes the plural unless it is obvious that it is meant otherwise.

[0207] This detailed description is to be construed as an example only and does not describe every possible embodiment, as describing every possible embodiment would be impractical, if not impossible. One could implement numerous alternative embodiments, using either current technology or technology developed after the filing date of this application.

Claims

What is claimed is:

1. A method for creating an overlay map on a digital image of a slide, the method comprising:

receiving the digital image;
separating the digital image into a plurality of tiles;
identifying the majority class of tissue visible within each tile in the plurality of tiles,
based on a multi-tile analysis.

2. The method of claim 1, further comprising:

generating a digital overlay drawing of an outer edge of each cell in the image.

3. The method of claim 2, wherein the digital overlay drawing is prepared at the resolution level of an individual pixel.

4. A method for tissue classification of a digital image of a slide, the method comprising:

receiving the digital image;
generating a digital overlay drawing of a tissue region in the digital image; and
displaying, in the digital overlay drawing, tiles over the tissue region and visually identifying predicted content of each tile, to generate a classification map classifying the digital image.

5. The method of claim 4, the method further comprising:

determining the predicted content for each tile using a classification model configured as a multi-resolution fully convolutional network, the multi-resolution fully convolution network configured to perform classification on digital images of different zoom levels.

6. The method of claim 4, the method further comprising:

using a classification model, for each tile, determining a classification from a plurality of classifications.

7. The method of claim 6, wherein the plurality of classifications include tumor epithelium, normal epithelium, immune, stroma, necrosis, blood, and fat.

8. A method for cell detection in a digital image of a slide, the method comprising:
receiving the digital image;
identifying cell objects in the digital image;
generating a digital overlay drawing for the digital image; and
displaying, in the digital overlay drawing, a polygon outlining each of the identified cell objects.

9. The method of claim 8, wherein the cell objects include lymphocyte cells and not lymphocyte cells.

10. The method of claim 8, wherein the cell objects comprise CD3, CD8, CD20, pancytokeratin, and smooth muscle actin.

11. A method for tissue classification of a digital image of a slide, the method comprising:
receiving the digital image;
segmenting the digital image into a plurality of tiles;
determining a predicted class for each tile;
identifying a plurality of cell objects in the digital image;
determining a predicted class for each of the plurality of cell objects; and
for each of the plurality of tiles that corresponds to one of the plurality of cell objects, assigning the tile the predicted class of the corresponding cell object in place of the predicted class of the tile.

12. The method of claim 11, further comprising:
storing, in a first file, for each tile, a tile position and the predicted class of the tile; and
storing, in a second file, for each cell object, a polygon outlining the cell object and the predicted class of the cell object.

13. The method of claim 11, further comprising:
generating a digital overlay drawing for the digital image, where the digital overlay drawing is a cell mask displaying a polygon around each cell object.
14. The method of claim 11, further comprising:
generating a digital overlay drawing for the digital image, where the digital overlay drawing is a histology mask displaying the plurality of tiles and the predict class for each tile.
15. The method of claim 11, further comprising:
generating a digital overlay drawing for the digital image; and
displaying, in the digital overlay drawing, the plurality of tiles and the predicted class for each tile that does not correspond to a cell object, and displaying, the plurality of cell objects and the predicted class of each cell object.
16. The method of claim 11, further comprising:
generating a digital overlay drawing for the digital image, wherein the digital overlay drawing includes the digital image; and
displaying the digital overlay drawing.
17. The method of claim 11, further comprising:
generating a digital overlay drawing for the digital image, wherein the digital overlay drawing includes a generated version of the digital image; and
displaying the digital overlay drawing.
18. The method of claim 11, further comprising:
generating a plurality of digital overlay drawings for the digital image, where each digital overlay drawing corresponds to a different predicted class; and
selectively displaying one of the plurality of digital overlay drawings.
19. The method of claim 11, further comprising:
generating a digital overlay drawing for the digital image, where the digital overlay drawing comprises percentages of predict classes corresponding to the digital image.
20. The method of claim 11, further comprising:

generating a digital overlay drawing for the digital image, where the digital overlay drawing comprises total counts of predict classes corresponding to the digital image.

Abstract

Techniques for generating an overlay map on a digital medical image of a slide are provided, and include cell detection and tissue classification processes. Techniques include receiving a medical image, separating the image into tiles, and performing tile classifications and tissue classifications based on a multi-tile analysis. Techniques additionally include identifying cell objects in the image, separating the image into and displaying polygons identifying the cell objects and cell classifications. Generated displays may be overlays over the initial digital image.

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:
Stephen Yip et al.

Application No.: Not Yet Assigned

Confirmation No.: Not Yet Assigned

Filed: Concurrently Herewith

Art Unit: Not Yet Assigned

For: ARTIFICIAL INTELLIGENCE SEGMENTATION
OF TISSUE IMAGES

Examiner: Not Yet Assigned

PETITION TO ACCEPT COLOR DRAWINGS

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Sir:

Applicant submits this petition pursuant to 37 C.F.R. § 1.84 to request the acceptance of color drawings filed in connection with the above-identified application. These drawings are acceptable pursuant to 37 C.F.R. § 1.84 because color is the only practical medium by which to depict the subject matter set forth in some of the figures.

Pursuant to the requirements set forth in 37 C.F.R. § 1.84(a)(2)(i)–(iii), Applicant submits herewith the \$140.00 fee in accordance with 37 C.F.R. § 1.17(h), and have included in the application language identifying the specification as containing color drawings.

The requirement for three (3) sets of color drawings under 37 CFR § 1.84(a)(2)(ii) is not applicable to color drawings submitted via EFS-Web. Therefore, only one set of such color drawings is necessary.

Dated: December 31, 2019

Respectfully submitted,

Electronic signature: /Paul B. Stephens/
Paul B. Stephens

Registration No.: 47,970
MARSHALL, GERSTEIN & BORUN LLP
233 S. Wacker Drive
6300 Willis Tower
Chicago, Illinois 60606-6357
(312) 474-6300
Attorney for Applicant

Electronic Patent Application Fee Transmittal

Application Number:				
Filing Date:				
Title of Invention:	ARTIFICIAL INTELLIGENCE SEGMENTATION OF TISSUE IMAGES			
First Named Inventor/Applicant Name:	Stephen Yip			
Filer:	Paul Bryan Stephens/Jennie Mackowiak			
Attorney Docket Number:	32830/58872A			
Filed as Large Entity				
Filing Fees for Utility under 35 USC 111(a)				
Description	Fee Code	Quantity	Amount	Sub-Total in USD(\$)
Basic Filing:				
UTILITY APPLICATION FILING	1011	1	300	300
UTILITY SEARCH FEE	1111	1	660	660
UTILITY EXAMINATION FEE	1311	1	760	760
Pages:				
Claims:				
INDEPENDENT CLAIMS IN EXCESS OF 3	1201	1	460	460
Miscellaneous-Filing:				
LATE FILING FEE FOR OATH OR DECLARATION	1051	1	160	160

Description	Fee Code	Quantity	Amount	Sub-Total in USD(\$)
Petition:				
PETITION FEE- 37 CFR 1.17(H) (GROUP III)	1464	1	140	140
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				
Miscellaneous:				
Total in USD (\$)				2480

Electronic Acknowledgement Receipt

EFS ID:	38178289
Application Number:	16732242
International Application Number:	
Confirmation Number:	7185
Title of Invention:	ARTIFICIAL INTELLIGENCE SEGMENTATION OF TISSUE IMAGES
First Named Inventor/Applicant Name:	Stephen Yip
Customer Number:	04743
Filer:	Paul Bryan Stephens/Jennie Mackowiak
Filer Authorized By:	Paul Bryan Stephens
Attorney Docket Number:	32830/58872A
Receipt Date:	31-DEC-2019
Filing Date:	
Time Stamp:	20:08:12
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	CARD
Payment was successfully received in RAM	\$2480
RAM confirmation Number	E2019BUK08415678
Deposit Account	
Authorized User	

The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows:

File Listing:					
Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Application Data Sheet	Application_Data_Sheet.PDF	29180	no	6
			caf67df832d0c919d11fd2d90736e222cc3df6f4b		
Warnings:					
Information:					
This is not an USPTO supplied ADS fillable form					
2	Transmittal of New Application	Utility_Patent_Application_Transmittal.PDF	32943	no	1
			5ff82ee703ba3adef9d920eec283cf6399ee431b		
Warnings:					
Information:					
3	Drawings-other than black and white line drawings	Color_Drawing_as_filed.PDF	5984291	no	16
			46e65f4de7f1a09e4b6462e06276c9830fbbdcd08		
Warnings:					
Information:					
4		NON_PROV_APP_32830_58872A.PDF	198960	yes	45
			deeb1eea83e745ad45f621f24be4486c3e2cd9e0f		
	Multipart Description/PDF files in .zip description				
	Document Description	Start	End		
	Specification	1	40		
	Claims	41	44		
	Abstract	45	45		
Warnings:					
Information:					

5	A petition from Applicant to accept Color Drawings or Color Photographs	Petition_to_accept_color_drawings.PDF	15103	no	1
			d2a1b60aa684addb8790873067f288d09e201682		

Warnings:

Information:

6	Fee Worksheet (SB06)	fee-info.pdf	40201	no	2
			c9823b55de9ccb1f5f72d07f3fe8626c3eb3364f		

Warnings:

Information:

Total Files Size (in bytes):			6300678
-------------------------------------	--	--	---------

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

Application Data Sheet

Inventor Information

Inventor Number::	1
Given Name::	Stephen
Family Name::	Yip
City of Residence::	Chicago
State or Province of Residence::	IL
Country of Residence::	US
Street of mailing address::	3402 S. Oakley Ave.
City of mailing address::	Chicago
State or Province of mailing address::	IL
Country of mailing address::	US
Postal or Zip Code of mailing address::	60608
Inventor Number::	2
Given Name::	Irvin
Family Name::	Ho
City of Residence::	Wilmette
State or Province of Residence::	IL
Country of Residence::	US
Street of mailing address::	600 West Chicago Ave
City of mailing address::	Chicago
State or Province of mailing address::	IL
Country of mailing address::	US
Postal or Zip Code of mailing address::	60654

Inventor Number:: 3
Given Name:: Lingdao
Family Name:: Sha
City of Residence:: Chicago
State or Province of Residence:: IL
Country of Residence:: US
Street of mailing address:: 600 West Chicago Ave
City of mailing address:: Chicago
State or Province of mailing address:: IL
Country of mailing address:: US
Postal or Zip Code of mailing address:: 60654

Inventor Number:: 4
Given Name:: Boleslaw
Family Name:: Osinski
City of Residence:: Chicago
State or Province of Residence:: IL
Country of Residence:: US
Street of mailing address:: 600 West Chicago Ave
City of mailing address:: Chicago
State or Province of mailing address:: IL
Country of mailing address:: US
Postal or Zip Code of mailing address:: 60654

Correspondence Information

Correspondence Customer Number:: 04743

Application Information

Application Type:: Regular

Subject Matter:: Utility

CD-ROM or CD-R?: None

Sequence submission?: None

Computer Readable Form (CRF)?:: No

Title:: ARTIFICIAL INTELLIGENCE
SEGMENTATION OF TISSUE
IMAGES

Attorney Docket Number:: 32830/58872A

Request for Early Publication?: No

Request for Non-Publication?: No

Total Drawing Sheets:: 16

Small Entity?: No

Petition included?: No

Portions or all of the application associated with this No

Application Data Sheet may fall under a Secrecy
Order pursuant to 37 CFR 5.2::

Representative Information

Representative Customer Number:: 04743

Domestic Priority Information

Application::	Continuity Type::	Parent Application::	Parent Filing Date::	Prior Appl Status::
This application	Claims benefit of provisional	62/787,047	12/31/18	Pending

Applicant Information

Applicant Number:: 1

Applicant Type:: Assignee

Organization Name:: TEMPUS LABS, INC.

Street of mailing address:: 600 West Chicago Ave
Suite 510

City of mailing address:: Chicago

State or Province of mailing address:: IL

Country of mailing address:: US

Postal or Zip Code of mailing address:: 60654

Authorization or Opt-Out of Authorization to Permit Access

When this Application Data Sheet is properly signed and filed with the application, applicant has provided written authority to permit a participating foreign intellectual property (IP) office access to the instant application-as-filed (see paragraph A in subsection 1 below) and the European Patent Office (EPO) access to any search results from the instant application (see paragraph B in subsection 1 below).

Should applicant choose not to provide an authorization identified in subsection 1 below, applicant **must opt-out** of the authorization by checking the corresponding box A or B or both in subsection 2 below.

NOTE: This section of the Application Data Sheet is **ONLY** reviewed and processed with the **INITIAL** filing of an application. After the initial filing of an application, an Application Data Sheet cannot be used to provide or rescind authorization for access by a foreign IP office(s). Instead, Form PTO/SB/39 or PTO/SB/69 must be used as appropriate.

1. Authorization to Permit Access by a Foreign Intellectual Property Office(s)

A. Priority Document Exchange (PDX) - Unless box A in subsection 2 (opt-out of authorization) is checked, the undersigned hereby **grants the USPTO authority** to provide the European Patent Office (EPO), the Japan Patent Office (JPO), the Korean Intellectual Property Office (KIPO), the State Intellectual Property Office of the People's Republic of China (SIPO), the World Intellectual Property Organization (WIPO), and any other foreign intellectual property office participating with the USPTO in a bilateral or multilateral priority document exchange agreement in which a foreign application claiming priority to the instant patent application is filed, access to: (1) the instant patent application-as-filed and its related bibliographic data, (2) any foreign or domestic application to which priority or benefit is claimed by the instant application and its related bibliographic data, and (3) the date of filing of this Authorization. See 37 CFR 1.14(h)(1).

B. Search Results from U.S. Application to EPO - Unless box B in subsection 2 (opt-out of authorization) is checked, the undersigned hereby **grants the USPTO authority** to provide the EPO access to the bibliographic data and search results from the instant patent application when a European patent application claiming priority to the instant patent application is filed. See 37 CFR 1.14(h)(2).

The applicant is reminded that the EPO's Rule 141(1) EPC (European Patent Convention) requires applicants to submit a copy of search results from the instant application without delay in a European patent application that claims priority to the instant application.

2. Opt-Out of Authorizations to Permit Access by a Foreign Intellectual Property Office(s)

☐ A. Applicant **DOES NOT** authorize the USPTO to permit a participating foreign IP office access to the instant application-as-filed. If this box is checked, the USPTO will not be providing a participating foreign IP office with any documents and information identified in subsection 1A above.

☐ B. Applicant **DOES NOT** authorize the USPTO to transmit to the EPO any search results from the instant patent application. If this box is checked, the USPTO will not be providing the EPO with search results from the instant application.

NOTE: Once the application has published or is otherwise publicly available, the USPTO may provide access to the application in accordance with 37 CFR 1.14.

Signature:

NOTE: This Application Data Sheet must be signed in accordance with 37 CFR 1.33(b). **However, if this Application Data Sheet is submitted with the INITIAL filing of the application and either box A or B is not checked in subsection 2 of the “Authorization or Opt-Out of Authorization to Permit Access” section, then this form must also be signed in accordance with 37 CFR 1.14(c).**

This Application Data Sheet **must** be signed by a patent practitioner if one or more of the applicants is a **juristic entity** (e.g., corporation or association). If the applicant is two or more joint inventors, this form must be signed by a patent practitioner, **all** joint inventors who are the applicant, or one or more joint inventor-applicants who have been given power of attorney (e.g., see USPTO Form PTO/AIA/81) on behalf of **all** joint inventor-applicants.

See 37 CFR 1.4(d) for the manner of making signatures and certifications.

Signature	_/Paul B. Stephens/_	Date (YYYY-MM-DD)	2019-12-31
Name	Paul B. Stephens	Registration Number	47,970

SCORE Placeholder Sheet for IFW Content

Application Number: 16732242

Document Date: 12/31/2019

The presence of this form in the IFW record indicates that the following document type was received in electronic format on the date identified above. This content is stored in the SCORE database.

Since this was an electronic submission, there is no physical artifact folder, no artifact folder is recorded in PALM, and no paper documents or physical media exist. The TIFF images in the IFW record were created from the original documents that are stored in SCORE.

- Drawing

At the time of document entry (noted above):

- USPTO employees may access SCORE content via DAV or via the SCORE web page.
- External customers may access SCORE content via PAIR using the Supplemental Content tab.



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING or 371(c) DATE	GRP ART UNIT	FIL FEE REC'D	ATTY. DOCKET NO	TOT CLAIMS	IND CLAIMS
16/732,242	12/31/2019	2676	2340	32830/58872A	20	4

CONFIRMATION NO. 7185

4743

MARSHALL, GERSTEIN & BORUN LLP
233 SOUTH WACKER DRIVE
6300 WILLIS TOWER
CHICAGO, IL 60606-6357

FILING RECEIPT



0000000114331653

Date Mailed: 02/03/2020

Receipt is acknowledged of this non-provisional utility patent application. The application will be taken up for examination in due course. Applicant will be notified as to the results of the examination. Any correspondence concerning the application must include the following identification information: the U.S. APPLICATION NUMBER, FILING DATE, NAME OF FIRST INVENTOR, and TITLE OF INVENTION. Fees transmitted by check or draft are subject to collection.

Please verify the accuracy of the data presented on this receipt. If an error is noted on this Filing Receipt, please submit a written request for a corrected Filing Receipt, including a properly marked-up ADS showing the changes with strike-through for deletions and underlining for additions. If you received a "Notice to File Missing Parts" or other Notice requiring a response for this application, please submit any request for correction to this Filing Receipt with your reply to the Notice. When the USPTO processes the reply to the Notice, the USPTO will generate another Filing Receipt incorporating the requested corrections provided that the request is grantable.

Inventor(s)

Stephen Yip, Chicago, IL;
Irvin Ho, Wilmette, IL;
Lingdao Sha, Chicago, IL;
Boleslaw Osinski, Chicago, IL;

Applicant(s)

TEMPUS LABS, INC., Chicago, IL;

Power of Attorney: None

Domestic Priority data as claimed by applicant

This appln claims benefit of 62/787,047 12/31/2018

Foreign Applications for which priority is claimed (You may be eligible to benefit from the **Patent Prosecution Highway** program at the USPTO. Please see <http://www.uspto.gov> for more information.) - None.

Foreign application information must be provided in an Application Data Sheet in order to constitute a claim to foreign priority. See 37 CFR 1.55 and 1.76.

Permission to Access Application via Priority Document Exchange: Yes

Permission to Access Search Results: Yes

Applicant may provide or rescind an authorization for access using Form PTO/SB/39 or Form PTO/SB/69 as appropriate.

If Required, Foreign Filing License Granted: 01/30/2020

The country code and number of your priority application, to be used for filing abroad under the Paris Convention, is **US 16/732,242**

Projected Publication Date: 07/02/2020

Non-Publication Request: No

Early Publication Request: No
Title

ARTIFICIAL INTELLIGENCE SEGMENTATION OF TISSUE IMAGES

Preliminary Class

358

Statement under 37 CFR 1.55 or 1.78 for AIA (First Inventor to File) Transition Applications: No

PROTECTING YOUR INVENTION OUTSIDE THE UNITED STATES

Since the rights granted by a U.S. patent extend only throughout the territory of the United States and have no effect in a foreign country, an inventor who wishes patent protection in another country must apply for a patent in a specific country or in regional patent offices. Applicants may wish to consider the filing of an international application under the Patent Cooperation Treaty (PCT). An international (PCT) application generally has the same effect as a regular national patent application in each PCT-member country. The PCT process **simplifies** the filing of patent applications on the same invention in member countries, but **does not result** in a grant of "an international patent" and does not eliminate the need of applicants to file additional documents and fees in countries where patent protection is desired.

Almost every country has its own patent law, and a person desiring a patent in a particular country must make an application for patent in that country in accordance with its particular laws. Since the laws of many countries differ in various respects from the patent law of the United States, applicants are advised to seek guidance from specific foreign countries to ensure that patent rights are not lost prematurely.

Applicants also are advised that in the case of inventions made in the United States, the Director of the USPTO must issue a license before applicants can apply for a patent in a foreign country. The filing of a U.S. patent application serves as a request for a foreign filing license. The application's filing receipt contains further information and guidance as to the status of applicant's license for foreign filing.

Applicants may wish to consult the USPTO booklet, "General Information Concerning Patents" (specifically, the section entitled "Treaties and Foreign Patents") for more information on timeframes and deadlines for filing foreign patent applications. The guide is available either by contacting the USPTO Contact Center at 800-786-9199, or it can be viewed on the USPTO website at <http://www.uspto.gov/web/offices/pac/doc/general/index.html>.

For information on preventing theft of your intellectual property (patents, trademarks and copyrights), you may wish to consult the U.S. Government website, <http://www.stopfakes.gov>. Part of a Department of Commerce initiative, this website includes self-help "toolkits" giving innovators guidance on how to protect intellectual property in specific

countries such as China, Korea and Mexico. For questions regarding patent enforcement issues, applicants may call the U.S. Government hotline at 1-866-999-HALT (1-866-999-4258).

LICENSE FOR FOREIGN FILING UNDER
Title 35, United States Code, Section 184
Title 37, Code of Federal Regulations, 5.11 & 5.15

GRANTED

The applicant has been granted a license under 35 U.S.C. 184, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" followed by a date appears on this form. Such licenses are issued in all applications where the conditions for issuance of a license have been met, regardless of whether or not a license may be required as set forth in 37 CFR 5.15. The scope and limitations of this license are set forth in 37 CFR 5.15(a) unless an earlier license has been issued under 37 CFR 5.15(b). The license is subject to revocation upon written notification. The date indicated is the effective date of the license, unless an earlier license of similar scope has been granted under 37 CFR 5.13 or 5.14.

This license is to be retained by the licensee and may be used at any time on or after the effective date thereof unless it is revoked. This license is automatically transferred to any related applications(s) filed under 37 CFR 1.53(d). This license is not retroactive.

The grant of a license does not in any way lessen the responsibility of a licensee for the security of the subject matter as imposed by any Government contract or the provisions of existing laws relating to espionage and the national security or the export of technical data. Licensees should apprise themselves of current regulations especially with respect to certain countries, of other agencies, particularly the Office of Defense Trade Controls, Department of State (with respect to Arms, Munitions and Implements of War (22 CFR 121-128)); the Bureau of Industry and Security, Department of Commerce (15 CFR parts 730-774); the Office of Foreign Assets Control, Department of Treasury (31 CFR Parts 500+) and the Department of Energy.

NOT GRANTED

No license under 35 U.S.C. 184 has been granted at this time, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" DOES NOT appear on this form. Applicant may still petition for a license under 37 CFR 5.12, if a license is desired before the expiration of 6 months from the filing date of the application. If 6 months has lapsed from the filing date of this application and the licensee has not received any indication of a secrecy order under 35 U.S.C. 181, the licensee may foreign file the application pursuant to 37 CFR 5.15(b).

SelectUSA

The United States represents the largest, most dynamic marketplace in the world and is an unparalleled location for business investment, innovation, and commercialization of new technologies. The U.S. offers tremendous resources and advantages for those who invest and manufacture goods here. Through SelectUSA, our nation works to promote and facilitate business investment. SelectUSA provides information assistance to the international investor community; serves as an ombudsman for existing and potential investors; advocates on behalf of U.S. cities, states, and regions competing for global investment; and counsels U.S. economic development organizations on investment attraction best practices. To learn more about why the United States is the best country in the world to develop

technology, manufacture products, deliver services, and grow your business, visit <http://www.SelectUSA.gov> or call +1-202-482-6800.

PATENT APPLICATION FEE DETERMINATION RECORD

Substitute for Form PTO-875

Application or Docket Number
16/732,242

APPLICATION AS FILED - PART I

(Column 1)		(Column 2)	SMALL ENTITY		OR	OTHER THAN SMALL ENTITY	
FOR	NUMBER FILED	NUMBER EXTRA	RATE(\$)	FEE(\$)		RATE(\$)	FEE(\$)
BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A	N/A	N/A			N/A	300
SEARCH FEE (37 CFR 1.16(k), (l), or (m))	N/A	N/A	N/A			N/A	660
EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A	N/A			N/A	760
TOTAL CLAIMS (37 CFR 1.16(j))	20 minus 20 =	*			OR	x 100 =	0.00
INDEPENDENT CLAIMS (37 CFR 1.16(h))	4 minus 3 =	* 1				x 460 =	460
APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is \$310 (\$155 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).						0.00
MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))							0.00
* If the difference in column 1 is less than zero, enter "0" in column 2.			TOTAL			TOTAL	2180

APPLICATION AS AMENDED - PART II

(Column 1)		(Column 2)	(Column 3)	SMALL ENTITY		OR	OTHER THAN SMALL ENTITY	
AMENDMENT A	CLAIMS REMAINING AFTER AMENDMENT	HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA	RATE(\$)	ADDITIONAL FEE(\$)		RATE(\$)	ADDITIONAL FEE(\$)
Total (37 CFR 1.16(i))	*	Minus **	=	x	=	OR	x	=
Independent (37 CFR 1.16(h))	*	Minus ***	=	x	=	OR	x	=
Application Size Fee (37 CFR 1.16(s))						OR		
FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))						OR		
				TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	

(Column 1)		(Column 2)	(Column 3)	SMALL ENTITY		OR	OTHER THAN SMALL ENTITY	
AMENDMENT B	CLAIMS REMAINING AFTER AMENDMENT	HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA	RATE(\$)	ADDITIONAL FEE(\$)		RATE(\$)	ADDITIONAL FEE(\$)
Total (37 CFR 1.16(i))	*	Minus **	=	x	=	OR	x	=
Independent (37 CFR 1.16(h))	*	Minus ***	=	x	=	OR	x	=
Application Size Fee (37 CFR 1.16(s))						OR		
FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))						OR		
				TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	

* If the entry in column 1 is less than the entry in column 2, write "0" in column 3.

** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20".

*** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, enter "3".

The "Highest Number Previously Paid For" (Total or Independent) is the highest found in the appropriate box in column 1.



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING OR 371(C) DATE	FIRST NAMED APPLICANT	ATTY. DOCKET NO./TITLE
16/732,242	12/31/2019	Stephen Yip	32830/58872A

CONFIRMATION NO. 7185

INFORMAL NOTICE



4743
MARSHALL, GERSTEIN & BORUN LLP
233 SOUTH WACKER DRIVE
6300 WILLIS TOWER
CHICAGO, IL 60606-6357

Date Mailed: 02/03/2020

INFORMATIONAL NOTICE TO APPLICANT

Applicant is notified that the above-identified application contains the deficiencies noted below. No period for reply is set forth in this notice for correction of these deficiencies. However, if a deficiency relates to the inventor's oath or declaration, the applicant must file an oath or declaration in compliance with 37 CFR 1.63, or a substitute statement in compliance with 37 CFR 1.64, executed by or with respect to each actual inventor no later than the expiration of the time period set in the "Notice of Allowability" to avoid abandonment. See 37 CFR 1.53(f).

The item(s) indicated below are also required and should be submitted with any reply to this notice to avoid further processing delays.

- A properly executed inventor's oath or declaration has not been received for the following inventor(s):
Stephen Yip
Irvin Ho
Lingdao Sha
Boleslaw Osinski

Questions about the contents of this notice and the requirements it sets forth should be directed to the Office of Data Management, Application Assistance Unit, at (571) 272-4000 or (571) 272-4200 or 1-888-786-0101.

/qtran/



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
16/732,242	12/31/2019	Stephen Yip	32830/58872A	7185
4743 7590 03/23/2020 MARSHALL, GERSTEIN & BORUN LLP 233 SOUTH WACKER DRIVE 6300 WILLIS TOWER CHICAGO, IL 60606-6357			EXAMINER BELLA, MATTHEW C	
			ART UNIT 2667	PAPER NUMBER
			NOTIFICATION DATE 03/23/2020	DELIVERY MODE ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

mgbdoCKET@marshallip.com



UNITED STATES PATENT AND TRADEMARK OFFICE

Commissioner for Patents
United States Patent and Trademark Office
P.O. Box 1450
Alexandria, VA 22313-1450
www.uspto.gov

MARSHALL, GERSTEIN & BORUN LLP
233 SOUTH WACKER DRIVE
6300 WILLIS TOWER
CHICAGO IL 60606-6357

In Re Application of:
YIP, STEPHEN et al.
Application Serial No. **16/732,242**
Filed: **December 31, 2019**
Attorney Docket Number: **32830/58872A**
For: **ARTIFICIAL INTELLIGENCE SEGMENTATION**
OF TISSUE IMAGES

:
:
:
:
:
:
:

**DECISION ON PETITION
TO ACCEPT COLOR
DRAWINGS**

This is a decision on the petition under 37 C.F.R. § 1.84(a)(2), filed December 31, 2019, requesting acceptance of color drawings.

The petition requests that the color drawings, noted as figures 1, 3A-3C, and 5A-9 be accepted.

A grantable petition under 37 C.F.R. § 1.84(a)(2) must be accompanied by a fee set forth under 37 C.F.R. § 1.17(h), one (1) set of color drawings if submitted via the Office electronic filing system or otherwise, three (3) sets of the color drawings in question, a black and white photocopy of said drawings, and the specification must contain, or be amended to contain, the following language as the first paragraph in that portion of the specification relating to the brief description of the drawings:

" The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee."

Petitioner has met the requirements set forth above. Accordingly, the petition is **Granted**.

Any inquiry regarding this decision should be directed to Doris To, Quality Assurance Specialist, at (571) 272-7629.

/Doris To/

Doris To
Quality Assurance Specialist
Technology Center 2600
Communications

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)	Application Number		16732242
	Filing Date		2019-12-31
	First Named Inventor	Stephen Yip	
	Art Unit	2667	
	Examiner Name	BELLA, MATTHEW C	
	Attorney Docket Number	32830/58872A	

U.S.PATENTS						
Examiner Initial*	Cite No	Patent Number	Kind Code ¹	Issue Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear
	1	6463438	B1	2002-10-08	VELTRI et al.	
If you wish to add additional U.S. Patent citation information please click the Add button.						
U.S.PATENT APPLICATION PUBLICATIONS						
Examiner Initial*	Cite No	Publication Number	Kind Code ¹	Publication Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear
	1	20030215936	A1	2003-11-20	KALLIONIEMI et al.	
	2	20060064248	A1	2006-03-23	SAIDI et al.	
	3	20070216909	A1	2007-09-20	EVERETT et al.	
	4	20100111396	A1	2010-05-06	BOUCHERON LAURA E	
	5	20140073922	A1	2014-03-13	MAMMONE RICHARD J	

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number		16732242
Filing Date		2019-12-31
First Named Inventor	Stephen Yip	
Art Unit		2667
Examiner Name	BELLA, MATTHEW C	
Attorney Docket Number		32830/58872A

6	20140140607	A1	2014-05-22	ERJEFALT JONAS	
7	20140337052	A1	2014-11-13	PELLINI et al.	
8	20150100246	A1	2015-04-09	REMZI et al.	
9	20160063724	A1	2016-03-03	TUNSTALL et al.	
10	20160103973	A1	2016-04-14	SINGAL et al.	
11	20160253466	A1	2016-09-01	AGAIAN et al.	
12	20170076046	A1	2017-03-16	BARNES et al.	
13	20170337682	A1	2017-11-23	LIAO et al.	
14	20180005136	A1	2018-01-04	GAI et al.	
15	20180089373	A1	2018-03-29	MATSUGUCHI et al.	
16	20180121759	A1	2018-05-03	GABRANI et al.	

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number	16732242
Filing Date	2019-12-31
First Named Inventor	Stephen Yip
Art Unit	2667
Examiner Name	BELLA, MATTHEW C
Attorney Docket Number	32830/58872A

17	20180129911	A1	2018-05-10	MADABHUSHI et al.	
18	20180137394	A1	2018-05-17	WENZEL et al.	
19	20190169685	A1	2019-06-06	GEORGIADIS et al.	

If you wish to add additional U.S. Published Application citation information please click the Add button.

FOREIGN PATENT DOCUMENTS

Examiner Initial*	Cite No	Foreign Document Number ³	Country Code ²	Kind Code ⁴	Publication Date	Name of Patentee or Applicant of cited Document	Pages, Columns, Lines where Relevant Passages or Relevant Figures Appear	T ⁵
	1	107369151	CN	A	2017-11-21	WAN XIANGBO	English Abstract Submitted	☐
	2	2017/198790	WO	A1	2017-11-23	F HOFFMANN-LA ROCHE AG		☐
	3	2018/065434	WO	A1	2018-04-12	VENTANA MED SYST INC		☐
	4	2018/067937	WO	A1	2018-04-12	OMNISEQ INC		☐
	5	2018/136664	WO	A1	2018-07-26	ICHAN SCHOOL OF MEDICINE AT MOUNT SINAI		☐

If you wish to add additional Foreign Patent Document citation information please click the Add button

NON-PATENT LITERATURE DOCUMENTS

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number	16732242
Filing Date	2019-12-31
First Named Inventor	Stephen Yip
Art Unit	2667
Examiner Name	BELLA, MATTHEW C
Attorney Docket Number	32830/58872A

Examiner Initials*	Cite No	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc), date, pages(s), volume-issue number(s), publisher, city and/or country where published.	T ⁵
	1	"John Hopkins University Study Finds Laboratory-Developed Liquid Biopsy Tests Can Give Different Results; Call for 'Improved Certification' of Medical Laboratories That Develop These LDTs", Dark Daily, 6 pages (2018).	☐
	2	"Technical Specifications," FoundationOne (Registered) Liquid, Foundation Medicine, Inc. (2018).	☐
	3	ABDEL-RAHMAN, Correlation between PD-L1 expression and outcome of NSCLC patients treated with anti-PD-1/PD-L1 agents: A meta-analysis, Crit. Rev. Oncol. Hematol., 101:75-85 (2016).	☐
	4	ABELS et al., Current state of the regulatory trajectory for whole slide imaging devices in the USA, Journal of Pathology Informatics, 1:23 (2017).	☐
	5	ADERI, Pathology services in developing countries-the West African experience, Arch. Pathol. Lab. Med., 135:183-6 (2011).	☐
	6	AERTS et al., Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nat. Commun., 5:4006 (2014).	☐
	7	ASH et al., Joint analysis of gene expression levels and histological images identifies genes associated with tissue morphology, 1-28 (2018).	☐
	8	AYERS et al., IFN-g-related mRNA profile predicts clinical response to PD-1 blockade, The Journal of Clinical Investigation, 127(8):2930-2940 (2017).	☐
	9	BAAS et al., Relationship between level of PD-L1 expression and outcomes in the KEYNOTE-010 study of pembrolizumab vs docetaxel for previously treated, PD-L1-Positive NSCLC., J. Clin. Oncol. American Society of Clinical Oncology, 34:9015-9015 (2016).	☐
	10	BANDI et al., From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge, IEEE Transactions on Medical Imaging, 38:550-560 (2019).	☐

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number	16732242
Filing Date	2019-12-31
First Named Inventor	Stephen Yip
Art Unit	2667
Examiner Name	BELLA, MATTHEW C
Attorney Docket Number	32830/58872A

11	BANKHEAD et al., QuPath: Open source software for digital pathology image analysis, Sci. Rep., 7:16878 (2017).	☐
12	BEAUBIER et al., Clinical validation of the tempus xT next-generation targeted oncology sequencing assay, Oncotarget, 10:2384-2396 (2019).	☐
13	BEJNORDI et al., Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer, Jama, 318:2199-2210 (2017).	☐
14	BERGADO et al., Recurrent multiresolution convolutional networks for vhr image classification, (2018), <Downloaded from the Internet https://arxiv.org/pdf/1806.05793.pdf > <Downloaded on: 15 March 2020> < DOI: 10.1109/TGRS.2018.2837357 > entire document, especially.	☐
15	BORGHAEI et al., Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer, N. Engl. J. Med., 373:1627-39 (2015).	☐
16	BRAHMER et al., Safety and activity of anti-PD-L1 antibody in patients with advanced cancer, N. Engl. J. Med., 366:2455-65 (2012).	☐
17	BRAY et al., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin., 68:394-424 (2018).	☐
18	BRUGNARA et al., Management training for pathology residents: i. results of a national survey, American journal of clinical pathology, 101:559-563 (1994).	☐
19	BUNESCU et al., Multiple instance learning for sparse positive bags, Proceedings of the 24th Annual Intl. Conf. on Machine Learning (ICML-2007).	☐
20	BUTTNER et al., Programmed death-ligand 1 immunohistochemistry testing: A review of analytical assays and clinical implementation in non-small-cell lung cancer, J. Clin. Oncol., 35:3867-76 (2017).	☐
21	BYCHKOV et al., Deep learning for tissue microarray image-based outcome prediction in patients with colorectal cancer, Proc. of SPIE, 9791:979115 (2016).	☐

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number	16732242
Filing Date	2019-12-31
First Named Inventor	Stephen Yip
Art Unit	2667
Examiner Name	BELLA, MATTHEW C
Attorney Docket Number	32830/58872A

22	CAMPANELLA et al., Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nature Medicine, 25:1301-1309 (2019).	☐
23	CARAVAGNA et al., Algorithmic methods to infer the evolutionary trajectories in cancer progression, PNAS., 113(28):E4025-E4034 (2016).	☐
24	CARDIFF et al., Manual hematoxylin and eosin staining of mouse tissue sections, Cold Spring Harb Protoc., 655-8 (2014).	☐
25	CARTER et al., Absolute quantification of somatic DNA alterations in human cancer, Nat. Biotechnol., 30:413-421 (2012).	☐
26	CHANG, Chemotherapy, chemoresistance and the changing treatment landscape for NSCLC, Lung Cancer, 71:3-10 (2011).	☐
27	CLARK et al., Analytical validation of a hybrid capture - based next-generation sequencing clinical assay for genomic profiling of cell-free circulating tumorDNA, The Journal of Molecular Diagnostics, 20(5):686-702 (2018).	☐
28	CLUNIE, HIMA_2017_DICOMWSI_Clunie [PowerPoint slides]. Retrieved from https://www.dclunie.com/papers/HIMA_2017_DICOMWSI_Clunie.pdf , (2017).	☐
29	COOPER et al., PanCancer insights from the cancer genome atlas: the pathologist's perspective, J. Pathol., 244:512-524 (2018).	☐
30	CORTES-CIRIANO et al., A molecular portrait of microsatellite instability across multiple cancers, Nat. Commun., 8:12 (2017).	☐
31	COUDRAY et al., Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning, bioRxiv, 23 (2017).	☐
32	COUTURE et al., Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype, NPJ. Breast Cancer, 4:30 (2018).	☐

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number	16732242
Filing Date	2019-12-31
First Named Inventor	Stephen Yip
Art Unit	2667
Examiner Name	BELLA, MATTHEW C
Attorney Docket Number	32830/58872A

33	CRUZ-ROA et al., Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent, Scientific Reports, 7:46450, 14 (2017).	☐
34	D'INCECCO et al., PD-1 and PD-L1 expression in molecularly selected non-small-cell lung cancer patients, Br. J. Cancer, 112:95-102 (2015).	☐
35	DAUME et al., Domain adaptation for statistical classifiers, Journal of Artificial Intelligence Research, 26:101-126 (2006).	☐
36	DENKERT et al., Standardized evaluation of tumor-infiltrating lymphocytes in breast cancer: results of the ring studies of the international immuno-oncology biomarker working group, Mod. Pathol., 29:1155-64 (2016).	☐
37	DIENSTMANN et al., Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer, Nat. Rev. Cancer, 17:14 (2017).	☐
38	ERTOSUN et al., Automated grading of gliomas using deep learning in digital pathology images: A modular approach with ensemble of convolutional neural networks, AMIA. Annu. Symp. Proceedings AMIA. Symp., 1899-908 (2015).	☐
39	ESTEVA et al., Dermatologist-level classification of skin cancer with deep neural networks, Nature, 542:115-118 (2017).	☐
40	Evaluation of Automatic Class III Designation for MSK-Impact (Integrated Mutation Profiling of Actionable Cancer Targets), Decision Summary, 57 pp. (publicly available before May 14, 2018).	☐
41	FABRIZIO et al., Beyond microsatellite testing: assessment of tumor mutational burden identifies subsets of colorectal cancer who may respond to immune checkpoint inhibition, Journal of Gastrointestinal Oncology, 9(4):610-617 (2018).	☐
42	FELDMAN et al., Tissue processing and hematoxylin and eosin staining, Methods Mol. Biol., 1180:31-43 (2014).	☐
43	FoundationOne CDx (Trademark) Technical Information, Foundation Medicine, Inc., Cambridge, Massachusetts, 43 pp. (November 2017).	☐

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number	16732242
Filing Date	2019-12-31
First Named Inventor	Stephen Yip
Art Unit	2667
Examiner Name	BELLA, MATTHEW C
Attorney Docket Number	32830/58872A

44	FoundationOne Liquid, Genomic Testing 6 pages (2019).	☐
45	GAN et al., Applicability of Next generation sequencing technology in microsatellite instability testing, Genes, 6:46-59 (2015).	☐
46	GANIN et al., Domain-adversarial training of neural networks, The journal of machine learning research, 17:2096-2030 (2016).	☐
47	GARG et al., Patterns of locoregional failure in stage III non-small cell lung cancer treated with definitive chemoradiation therapy, Pract. Radiat. Oncol., 4:342-8 (2014).	☐
48	GERTYCH et al., Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides, Sci. Rep., nature.com, 9:1483 (2019).	☐
49	GitHub, "HistomicsTK", available online at < https://github.com/DigitalSlideArchive/HistomicsTK >, retrieved on April 17, 2020, 3 pages.	☐
50	GROSSMAN et al., Toward a shared vision for cancer genomic data, The New England Journal of Medicine, 375:1109-1112 (2016).	☐

If you wish to add additional non-patent literature document citation information please click the Add button

EXAMINER SIGNATURE

Examiner Signature		Date Considered	
--------------------	--	-----------------	--

*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

¹ See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. ² Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). ³ For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document. ⁴ Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. ⁵ Applicant is to place a check mark here if English language translation is attached.

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number	16732242
Filing Date	2019-12-31
First Named Inventor	Stephen Yip
Art Unit	2667
Examiner Name	BELLA, MATTHEW C
Attorney Docket Number	32830/58872A

CERTIFICATION STATEMENT

Please see 37 CFR 1.97 and 1.98 to make the appropriate selection(s):

☐ That each item of information contained in the information disclosure statement was first cited in any communication from a foreign patent office in a counterpart foreign application not more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(1).

OR

☐ That no item of information contained in the information disclosure statement was cited in a communication from a foreign patent office in a counterpart foreign application, and, to the knowledge of the person signing the certification after making reasonable inquiry, no item of information contained in the information disclosure statement was known to any individual designated in 37 CFR 1.56(c) more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(2).

☐ See attached certification statement.

☐ The fee set forth in 37 CFR 1.17 (p) has been submitted herewith.

☒ A certification statement is not submitted herewith.

SIGNATURE

A signature of the applicant or representative is required in accordance with CFR 1.33, 10.18. Please see CFR 1.4(d) for the form of the signature.

Signature	_/Paul B. Stephens/_	Date (YYYY-MM-DD)	2020-07-01
Name/Print	Paul B. Stephens	Registration Number	47,970

This collection of information is required by 37 CFR 1.97 and 1.98. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 1 hour to complete, including gathering, preparing and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether the Freedom of Information Act requires disclosure of these records.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspections or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)	Application Number		16732242
	Filing Date		2019-12-31
	First Named Inventor	Stephen Yip	
	Art Unit	2667	
	Examiner Name	BELLA, MATTHEW C	
	Attorney Docket Number	32830/58872A	

U.S.PATENTS						
Examiner Initial*	Cite No	Patent Number	Kind Code ¹	Issue Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear
	1					

If you wish to add additional U.S. Patent citation information please click the Add button.

U.S.PATENT APPLICATION PUBLICATIONS						
Examiner Initial*	Cite No	Publication Number	Kind Code ¹	Publication Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear
	1					

If you wish to add additional U.S. Published Application citation information please click the Add button.

FOREIGN PATENT DOCUMENTS								
Examiner Initial*	Cite No	Foreign Document Number ³	Country Code ²	Kind Code ⁴	Publication Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear	T ⁵
	1							☐

If you wish to add additional Foreign Patent Document citation information please click the Add button

NON-PATENT LITERATURE DOCUMENTS			
Examiner Initials*	Cite No	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc), date, pages(s), volume-issue number(s), publisher, city and/or country where published.	T ⁵

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number	16732242
Filing Date	2019-12-31
First Named Inventor	Stephen Yip
Art Unit	2667
Examiner Name	BELLA, MATTHEW C
Attorney Docket Number	32830/58872A

1	GROSSMANN et al., Defining the biological basis of radiomic phenotypes in lung cancer, Elife, 6:e23421 (2017).	☐
2	GUTMAN et al., The digital slide archive: a software platform for management, integration, and analysis of histology for cancer research, Cancer Research, 77:e75-e78 (2017).	☐
3	HAUSE et al., Classification and characterization of microsatellite instability across 18 cancer types, Nature Medicine, 22:1342-1350 (2016).	☐
4	HE et al., Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778 (2016).	☐
5	HEINDL et al., Microenvironmental niche divergence shapes BRCA1-dysregulated ovarian cancer morphological plasticity, Nat. Commun., (2018), doi:10.1038/s41467-018-06130-3.	☐
6	HERBST et al., Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial Lancet, 387:1540-50 (2016).	☐
7	HERBST et al., Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients, Nature, 515:563-7 (2014).	☐
8	ILSE et al., Attention-based deep multiple instance learning, Proceedings of the 35th International Conference on Machine Learning, PMLR 80:2127-2136 (2018).	☐
9	ING et al., A novel machine learning approach reveals latent vascular phenotypes predictive of renal cancer outcome, Sci. Rep., 7:13190 (2017).	☐
10	International Application No. PCT/US2019/032313, International Search Report and Written Opinion, mailed July 26, 2019.	☐
11	International Application No. PCT/US2019/069164, International Search Report and Written Opinion, mailed April 09, 2020.	☐

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number	16732242
Filing Date	2019-12-31
First Named Inventor	Stephen Yip
Art Unit	2667
Examiner Name	BELLA, MATTHEW C
Attorney Docket Number	32830/58872A

12	IOFFE et al., Batch normalization: Accelerating deep network training by reducing internal covariate shift [Internet], arXiv [cs.LG], (2015), Available from: http://arxiv.org/abs/1502.03167 .	☐
13	JANOWCZYK et al., Deep learning for digital pathology imageanalysis: A comprehensive tutorial with selected use cases, J. Pathol. Inform., 7:29 (2016).	☐
14	JANZIC et al., PD-L1 Expression in Squamous-cell Carcinoma and Adenocarcinoma of the lung, Radiol. Oncol., 51:357-62 (2017).	☐
15	KAPIL et al., Deep semi supervised generative learning for automated PD-L1 tumor cell scoring on NSCLC tissue needle biopsies, 1-10 (2018).	☐
16	KATHER et al., Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer, Nat. Med., 25:1054-1054 (2019).	☐
17	KATHER et al., Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study, PLoS medicine 16:e1002730 (2019).	☐
18	KAZANDJIAN et al., FDA approval summary: Nivolumab for the treatment of metastatic non-small cell lung cancer with progression on or after platinum-based chemotherapy, Oncologist., 21:634-42 (2016).	☐
19	KERR et al., Programmed death ligand-1 immunohistochemistry: Friend or foe?, Arch. Pathol. Lab. Med., 140:326-31 (2016).	☐
20	KIM et al., The landscape of microsatellite instability in colorectal and endometrial cancer genomes, Cell, 155(4):21 (2013).	☐
21	KITANO et al., Tumour-infiltrating lymphocytes are correlated with higher expression levels of PD-1 and PD-L1 in early breast cancer, ESMO Open, 2:e000150 (2017).	☐
22	KOMURA et al., Machine learning methods for histopathological image analysis, Comput. Struct. Biotechnol. J., 16:34-42 (2018).	☐

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number	16732242
Filing Date	2019-12-31
First Named Inventor	Stephen Yip
Art Unit	2667
Examiner Name	BELLA, MATTHEW C
Attorney Docket Number	32830/58872A

23	KRISTENSEN et al., Principles and methods of integrative genomic analyses in cancer, Nat. Rev., 14:299-313 (2014).	☐
24	KRUPINSKI et al., A new software platform to improve multidisciplinary tumor board workflows and user satisfaction: A pilot study, J. Pathol. Inform., 9:26 (2018).	☐
25	KURAKIN et al., Adversarial Examples in the Physical World, Technical report, (2016).	☐
26	LAN et al., Quantitative histology analysis of the ovarian tumour microenvironment, Sci. Rep., 5:16317 (2015).	☐
27	LECUN et al., Deep learning, Nature, 521:436-444 (2015).	☐
28	LEEK et al., Association of macrophage infiltration with angiogenesis and prognosis in invasive breast carcinoma, Cancer Res., 56:4625-9 (1996).	☐
29	LEMERY et al., First FDA approval agnostic of cancer site-when a biomarker defines the indication, N. Engl. J. Med., 377:1409-1412 (2017).	☐
30	LEVENSON et al., Pigeons (Columba livia) as trainable observers of pathology and radiology breast cancer images, PLoS One, 10:e0141357 (2015).	☐
31	LEVINE et al., Rise of the machines: Advances in deep learning for cancer diagnosis, Trends in Cancer, 5(3):157-169 (2019).	☐
32	LITJENS et al., A survey on deep learning in medical image analysis, Med. Image Anal. Elsevier, 42:60-88 (2017).	☐
33	LITJENS et al., Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis, Sci. Rep. Nature.com, 6:26286 (2016).	☐

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number	16732242
Filing Date	2019-12-31
First Named Inventor	Stephen Yip
Art Unit	2667
Examiner Name	BELLA, MATTHEW C
Attorney Docket Number	32830/58872A

34	LIU et al., Artificial intelligence-based breast cancer nodal metastasis detection, Arch. Pathol. Lab. Med. arpa., 2018-0147-OA (2018). doi:10.5858/arpa.2018-0147-OA.	☐
35	LIU et al., Detecting cancer metastases on gigapixel pathology images, 13 (2017).	☐
36	LONG et al., Fully convolutional networks for semantic segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 39:640-651 (2017), https://ieeexplore.ieee.org/document/7298965 .	☐
37	LOWE, Distinctive image features from scale-invariant keypoints, Int. J. Computer Vision, 28 (2004).	☐
38	LUO et al., Comprehensive computational pathological image analysis predicts lung cancer prognosis, J. Thorac. Oncol., 12:501-509 (2017).	☐
39	MADABHUSHI et al., Image analysis and machine learning in digital pathology: Challenges and opportunities, Med. Image Anal., 33:170-175 (2016).	☐
40	MCLAUGHLIN et al., Quantitative assessment of the heterogeneity of PD-L1 expression in non-small-cell lung cancer, JAMA Oncol. American Medical Association, 2:46-54 (2016).	☐
41	MENG et al., Predictive biomarkers in PD-1/PD-L1 checkpoint blockade immunotherapy, Cancer Treat. Rev., 41:868-76 (2015).	☐
42	MOBADERSANY et al., Predicting cancer outcomes from histology and genomics using convolutional networks, Proc. Natl. Acad. Sci. U. S. A., 115:E2970-9 (2018).	☐
43	MURPHY et al., Comparison of the microsatellite instability analysis system and the bethesda panel for the determination of microsatellite instability in colorectal cancers, Journal of Molecular Diagnostics, 8(3):305-311 (2006).	☐
44	NALISNIK et al., Interactive phenotyping of large-scale histology imaging data with HistomicsML, Sci. Rep., 1-12 (2017).	☐

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number		16732242
Filing Date		2019-12-31
First Named Inventor	Stephen Yip	
Art Unit	2667	
Examiner Name	BELLA, MATTHEW C	
Attorney Docket Number	32830/58872A	

45	NAWAZ et al., Beyond immune density: critical role of spatial heterogeneity in oestrogen receptor-negative breast cancer, doi:10.1038/modpathol., 37 (2015).	☐
46	NEWBURN, Understanding the role of Human Leukocyte Antigen in Cancer Immunotherapy, Personalis, (2019).	☐
47	O'ROURKE et al., Is concurrent chemoradiation the standard of care for locally advanced non-small cell lung cancer?, A review of guidelines and evidence, Clin. Oncol., 22:347-55 (2010).	☐
48	OpenSeadragon API: Getting Started, available online at <https://openseadragon.github.io/docs/>, 3 (2020).	☐
49	ORANGE et al., Identification of three rheumatoid arthritis disease subtypes by machine learning integration of synovial histologic features and RNA sequencing data, Arthritis Rheumatol., 70:690-701 (2018).	☐
50	PARDOLL, The blockade of immune checkpoints in cancer immunotherapy, Nat. Rev. Cancer, 12:252-264 (2012).	☐

If you wish to add additional non-patent literature document citation information please click the Add button

EXAMINER SIGNATURE

Examiner Signature		Date Considered	
--------------------	--	-----------------	--

*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

¹ See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. ² Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). ³ For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document. ⁴ Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. ⁵ Applicant is to place a check mark here if English language translation is attached.

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number	16732242
Filing Date	2019-12-31
First Named Inventor	Stephen Yip
Art Unit	2667
Examiner Name	BELLA, MATTHEW C
Attorney Docket Number	32830/58872A

CERTIFICATION STATEMENT

Please see 37 CFR 1.97 and 1.98 to make the appropriate selection(s):

☐ That each item of information contained in the information disclosure statement was first cited in any communication from a foreign patent office in a counterpart foreign application not more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(1).

OR

☐ That no item of information contained in the information disclosure statement was cited in a communication from a foreign patent office in a counterpart foreign application, and, to the knowledge of the person signing the certification after making reasonable inquiry, no item of information contained in the information disclosure statement was known to any individual designated in 37 CFR 1.56(c) more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(2).

☐ See attached certification statement.

☐ The fee set forth in 37 CFR 1.17 (p) has been submitted herewith.

☒ A certification statement is not submitted herewith.

SIGNATURE

A signature of the applicant or representative is required in accordance with CFR 1.33, 10.18. Please see CFR 1.4(d) for the form of the signature.

Signature	_/Paul B. Stephens/_	Date (YYYY-MM-DD)	2020-07-01
Name/Print	Paul B. Stephens	Registration Number	47,970

This collection of information is required by 37 CFR 1.97 and 1.98. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 1 hour to complete, including gathering, preparing and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether the Freedom of Information Act requires disclosure of these records.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspections or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

INFORMATION DISCLOSURE STATEMENT BY APPLICANT (Not for submission under 37 CFR 1.99)	Application Number		16732242
	Filing Date		2019-12-31
	First Named Inventor	Stephen Yip	
	Art Unit	2667	
	Examiner Name	BELLA, MATTHEW C	
	Attorney Docket Number	32830/58872A	

U.S.PATENTS						
Examiner Initial*	Cite No	Patent Number	Kind Code ¹	Issue Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear
	1					

If you wish to add additional U.S. Patent citation information please click the Add button.

U.S.PATENT APPLICATION PUBLICATIONS						
Examiner Initial*	Cite No	Publication Number	Kind Code ¹	Publication Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear
	1					

If you wish to add additional U.S. Published Application citation information please click the Add button.

FOREIGN PATENT DOCUMENTS								
Examiner Initial*	Cite No	Foreign Document Number ³	Country Code ²	Kind Code ⁴	Publication Date	Name of Patentee or Applicant of cited Document	Pages,Columns,Lines where Relevant Passages or Relevant Figures Appear	T ⁵
	1							☐

If you wish to add additional Foreign Patent Document citation information please click the Add button

NON-PATENT LITERATURE DOCUMENTS			
Examiner Initials*	Cite No	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc), date, pages(s), volume-issue number(s), publisher, city and/or country where published.	T ⁵

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number	16732242
Filing Date	2019-12-31
First Named Inventor	Stephen Yip
Art Unit	2667
Examiner Name	BELLA, MATTHEW C
Attorney Docket Number	32830/58872A

1	PATEL et al., Development of immunohistochemistry services for cancer care in western Kenya: Implications for low-and middle-income countries, S. Afr. J. Lab. Clin. Med., 5:187 (2016).	☐
2	PATEL et al., PD-L1 expression as a predictive biomarker in cancer immunotherapy, Mol. Cancer Ther., 14:847-56 (2015).	☐
3	PMA P170019: FDA Summary of Safety and Effectiveness of Data, FoundationOne CDx (Trademark), 58 pp. (November 30, 2017).	☐
4	POPOVICI et al., Image-based surrogate biomarkers for molecular subtypes of colorectal cancer, Bioinformatics, 33(13):2002-2009 (2017).	☐
5	Press Release - Guardant Health Presents Data on Immunotherapy-related Biomarkers MSI and TMB in blood of advanced cancer patients at European Society of Medical Oncology Annual Meeting," 4 pages, (2018).	☐
6	RATCLIFFE et al., Agreement between programmed cell death ligand-1 diagnostic assays across multiple protein expression cutoffs in non-small cell lung cancer. Clin. Cancer Res., 23:3585-91 (2017).	☐
7	RIMM et al., A prospective, Multi-institutional, Pathologist-based assessment of 4 immunohistochemistry assays for PD-L1 expression in non-small cell lung cancer, JAMA. Oncol., 3:1051-8 (2017).	☐
8	RIVENSON et al., Deep learning-based virtual histology staining using auto-fluorescence of label-free tissue, Nat. Biomed. Eng., (2019)	☐
9	ROACH et al., Development of a companion diagnostic PD-L1 immunohistochemistry assay for pembrolizumab therapy in non-small-cell lung cancer, Appl. Immunohistochem. Mol. Morphol., 24:392-7 (2016).	☐
10	Roche Molecular Systems, Inc., NAVIFY Tumor Board, available online at < https://www.navify.com/tumorboard/ >, retrieved on 9 (2020).	☐
11	RONNEBERGER et al., U-Net: Convolutional networks for biomedical image segmentation, arXiv:1505.04597 [cs] (2015), https://arxiv.org/abs/1505.04597 .	☐

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number	16732242
Filing Date	2019-12-31
First Named Inventor	Stephen Yip
Art Unit	2667
Examiner Name	BELLA, MATTHEW C
Attorney Docket Number	32830/58872A

12	RUSSAKOVSKY et al., ImageNet large scale visual recognition challenge [Internet], arXiv [cs.CV], (2014), Available from: http://arxiv.org/abs/1409.0575 .	☐
13	SALTZ et al., Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images, Cell Reports, 23:181-193 (2018).	☐
14	SCHAUMBERG et al., H&E-stained whole slide image deep learning predicts SPOP mutation state in prostate cancer [Internet], bioRxiv, [cited 2019 Feb 13], 064279 (2018), Available from: https://www.biorxiv.org/content/10.1101/064279v9 .	☐
15	SCHMIDHUBER, Deep learning in neural networks: an overview, Neural Netw. Elsevier, 61:85-117 (2015).	☐
16	SIMONYAN et al., Very deep convolutional networks for large-scale image recognition, 10 (2014).	☐
17	SPRINGENBERG et al., Striving for simplicity: The All Convolutional Net., ICLR (2014).	☐
18	SUL et al., FDA approval summary: Pembrolizumab for the treatment of patients with metastatic non-small cell lung cancer whose tumors express programmed death-ligand 1, Oncologist., 21:643-50 (2016).	☐
19	SUNDAR et al., Nivolumab in NSCLC: latest evidence and clinical potential, Ther. Adv. Med. Oncol., 7:85-96 (2015).	☐
20	SZEGEDY et al., Going deeper with convolutions, IEEE, 9 (2015).	☐
21	TENG et al., Classifying cancers based on T-cell infiltration and PD-L1, Cancer Res., 75:2139-45 (2015).	☐
22	THORSSON et al, The immune landscape of cancer, Immunity, 48:812-830 (2018).	☐

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number	16732242
Filing Date	2019-12-31
First Named Inventor	Stephen Yip
Art Unit	2667
Examiner Name	BELLA, MATTHEW C
Attorney Docket Number	32830/58872A

23	THRANE et al., Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma, Cancer Res., 78(20):5970-5979, (2018).	☐
24	TIZHOOSH et al., Artificial intelligence and digital pathology: Challenges and opportunities, J. Pathol. Inform, 9:38 (2018).	☐
25	TSUJIKAWA et al., Quantitative multiplex immunohistochemistry reveals myeloid-inflamed tumor-immune complexity associated with poor prognosis. Cell Reports, 19:203-217 (2017).	☐
26	TZENG et al., Adversarial discriminative domain adaptation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7167-7176 (2017).	☐
27	U.S. Food & Drug Administration (FDA), KEYTRUDA label [Internet], [cited Last accessed Feb 12, 2019], Available from: https://www.accessdata.fda.gov/drugsatfda_docs/label/2017/125514s024lbl.pdf#page=46	☐
28	U.S. Food & Drug Administration (FDA), PD-L1 IHC 22C3 pharmDx. SUMMARY OF SAFETY AND EFFECTIVENESS DATA [Internet], [cited Last accessed Feb 12, 2019], Available from: https://www.accessdata.fda.gov/cdrh_docs/pdf15/P150013S006b.pdf	☐
29	U.S. Food & Drug Administration, Letter of Review of FoundationOne CDx (Trademark), 5 pp. (November 30, 2017).	☐
30	UMAR et al., Revised bethesda guidelines for hereditary nonpolyposis colorectal cancer (lynch syndrome) and microsatellite instability, Journal of the National Cancer Institute, 96:261-268 (2004).	☐
31	University Hospital of Liège and Osimis S.A., REST API of Orthanc, available online at < ">https://book.orthanc-server.com/users/rest.html#> , 17 (2015).	☐
32	VASCONCELOS et al., Increasing deep learning melanoma classification by classical and expert knowledge based image transforms, CoRR, (2017).	☐
33	VASSILAKOPOULOU et al., Evaluation of PD-L1 expression and associated tumor-infiltrating lymphocytes in laryngeal squamous cell carcinoma, Clin. Cancer Res., 22:704-13 (2016).	☐

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number		16732242
Filing Date		2019-12-31
First Named Inventor	Stephen Yip	
Art Unit		2667
Examiner Name	BELLA, MATTHEW C	
Attorney Docket Number		32830/58872A

34	VELCHETI et al., Programmed death ligand-1 expression in non-small cell lung cancer, Lab. Invest., 94:107-16 (2014).	☐
35	VILAR et al., Microsatellite instability in colorectal cancer-the stable evidence, Nat. Rev. Clin. Oncol., 7:153-162 (2010).	☐
36	WANG et al., Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome, Scientific Reports, 8:10393 (2018).	☐
37	WANG et al., Deep learning assessment of tumor proliferation in breast cancer histological images, IEEE Int. Conf. Bioinfo. Biomed. (BIBM), 600-603 (2017).	☐
38	WEINSTEIN et al., The cancer genome atlas pan-cancer analysis project, Nat. Gene., 45:1113-1120 (2013).	☐
39	WIMBERLY et al., PD-L1 expression correlates with tumor-infiltrating lymphocytes and response to neoadjuvant chemotherapy in breast cancer, Cancer Immunol. Res., 3(4):326-332 (2015).	☐
40	XU et al., A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images, Neurocomputing, 191:214-223 (2016).	☐
41	YI et al., Microvessel prediction in H&E stained pathology images using fully convolutional neural networks, BMC Bioinformatics, 19:64 (2018).	☐
42	YU et al., Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features, Nat. Comm., 7:12474 (2016).	☐
43	YUAN et al. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling, Science Translational Medicine, 4(157):157ra143 (2012).	☐
44	YUAN et al., Supplementary materials for quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling, (2012), doi:10.1126/scitranslmed.3004330.	☐

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number		16732242
Filing Date		2019-12-31
First Named Inventor	Stephen Yip	
Art Unit	2667	
Examiner Name	BELLA, MATTHEW C	
Attorney Docket Number	32830/58872A	

45	ZHANG et al., Adversarial deep learning for microsatellite instability prediction from histopathology, slides, First Conference on Medical Imaging with Deep Learning (MIDL, 2018), Amsterdam, The Netherlands.	☐
46	ZHU et al., Deciphering genomic underpinnings of quantitative MRI-based radiomic phenotypes of invasive breast carcinoma, Sci. Rep., 5:17787 (2016).	☐

If you wish to add additional non-patent literature document citation information please click the Add button

EXAMINER SIGNATURE

Examiner Signature		Date Considered	
--------------------	--	-----------------	--

*EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609. Draw line through a citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

¹ See Kind Codes of USPTO Patent Documents at www.USPTO.GOV or MPEP 901.04. ² Enter office that issued the document, by the two-letter code (WIPO Standard ST.3). ³ For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document. ⁴ Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. ⁵ Applicant is to place a check mark here if English language translation is attached.

**INFORMATION DISCLOSURE
STATEMENT BY APPLICANT**
(Not for submission under 37 CFR 1.99)

Application Number	16732242
Filing Date	2019-12-31
First Named Inventor	Stephen Yip
Art Unit	2667
Examiner Name	BELLA, MATTHEW C
Attorney Docket Number	32830/58872A

CERTIFICATION STATEMENT

Please see 37 CFR 1.97 and 1.98 to make the appropriate selection(s):

☐ That each item of information contained in the information disclosure statement was first cited in any communication from a foreign patent office in a counterpart foreign application not more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(1).

OR

☐ That no item of information contained in the information disclosure statement was cited in a communication from a foreign patent office in a counterpart foreign application, and, to the knowledge of the person signing the certification after making reasonable inquiry, no item of information contained in the information disclosure statement was known to any individual designated in 37 CFR 1.56(c) more than three months prior to the filing of the information disclosure statement. See 37 CFR 1.97(e)(2).

☐ See attached certification statement.

☐ The fee set forth in 37 CFR 1.17 (p) has been submitted herewith.

☒ A certification statement is not submitted herewith.

SIGNATURE

A signature of the applicant or representative is required in accordance with CFR 1.33, 10.18. Please see CFR 1.4(d) for the form of the signature.

Signature	_/Paul B. Stephens/_	Date (YYYY-MM-DD)	2020-07-01
Name/Print	Paul B. Stephens	Registration Number	47,970

This collection of information is required by 37 CFR 1.97 and 1.98. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 1 hour to complete, including gathering, preparing and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether the Freedom of Information Act requires disclosure of these records.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspections or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

Electronic Acknowledgement Receipt

EFS ID:	39885432
Application Number:	16732242
International Application Number:	
Confirmation Number:	7185
Title of Invention:	ARTIFICIAL INTELLIGENCE SEGMENTATION OF TISSUE IMAGES
First Named Inventor/Applicant Name:	Stephen Yip
Customer Number:	4743
Filer:	Paul Bryan Stephens/Audrey Nagelberg
Filer Authorized By:	Paul Bryan Stephens
Attorney Docket Number:	32830/58872A
Receipt Date:	01-JUL-2020
Filing Date:	31-DEC-2019
Time Stamp:	13:43:05
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no
------------------------	----

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Non Patent Literature	73_KRISTENSEN_433193.pdf	1067987	no	15
			f3e98b590dbf202a54ed46cc51ac7f4994073e0e		

Warnings:

Information:					
2	Non Patent Literature	74_Krupinski1_433201.pdf	851767	no	7
			2f6b267767d25f39061c26b10e3c3ff65f87b495		
Warnings:					
Information:					
3	Non Patent Literature	75_Kurakin2016_354983.pdf	2483234	no	15
			f534a71a9b165f91be381ca43b023b776a57ebc6		
Warnings:					
Information:					
4	Non Patent Literature	76_LAN_433209.pdf	4711094	no	12
			69e7d13fc03971598eab366cc58cc415265ba467		
Warnings:					
Information:					
5	Non Patent Literature	77_LeCun436_354984.pdf	12828846	no	9
			2977455ac3539abe5bed470291b784726ff0d585		
Warnings:					
Information:					
6	Non Patent Literature	78_Leek4625_433226.pdf	1052844	no	6
			0805a9249ecc9059f9f92ecaae5e4f5b18fc2eaf		
Warnings:					
Information:					
7	Non Patent Literature	79_Lemery1409_354985.pdf	3933281	no	4
			f9445c6b8fc2ed3ce0f35a7e6dbfdc38c3907a36		
Warnings:					
Information:					
8	Non Patent Literature	7_ASH_433179.pdf	5129459	no	28
			5961e9e675d29fd4d9b3f9936ef7e772d1b8a7f9		
Warnings:					
Information:					

9	Non Patent Literature	80_Levensone014135_354986.pdf	20199142	no	21
			b91d4f0dad78cd9d9ae81d3fa1203f83cbe70f3d		
Warnings:					
Information:					
10	Non Patent Literature	81_Levine2019_354987.pdf	10717477	no	13
			95cfd8a6642d8d6573bf0f994a16584af9866347		
Warnings:					
Information:					
11	Non Patent Literature	82_a_survey_on_dee_433234.pdf	1769691	no	29
			9107889a85dbd4f77ee7eefc275b2ba51cb0bc24		
Warnings:					
Information:					
12	Non Patent Literature	83_Litjens1_433218.pdf	4042689	no	11
			d89f4dd7b0997ae771dd906bce742b3b5347d7b9		
Warnings:					
Information:					
13	Non Patent Literature	84_Liu859_433185.pdf	2816101	no	10
			2880ab01b884545f7606c0bd77b3bd6425cd1e51		
Warnings:					
Information:					
14	Non Patent Literature	85_Liu1_354988.pdf	3155925	no	13
			1e91ae66bfda17942c3b3073662a97d934cd47171		
Warnings:					
Information:					
15	Non Patent Literature	86_Long_Fully_Conv_433256.pdf	1524045	no	10
			7b48cfc86be7792f7cba892f27c779540cd32f40		
Warnings:					
Information:					

16	Non Patent Literature	87_Lowe_354989.pdf	1854691	no	28
			910fb1096bdf9fdce7a988ec7de7d5bb95086c73		
Warnings:					
Information:					
17	Non Patent Literature	88_Luo501_433194.pdf	1606589	no	9
			89869fe7c16856ca138ebe57600ef28f0f4a9682		
Warnings:					
Information:					
18	Non Patent Literature	89_Madabhushi170_433202.pdf	847083	no	14
			7f7b13ef27ea636f01c9846c234569f669f3bbeed		
Warnings:					
Information:					
19	Non Patent Literature	8_Ayers2930_354955.pdf	1671942	no	11
			9727eaa0ba83430aedc78e41a7e48d0fcb0b604		
Warnings:					
Information:					
20	Non Patent Literature	90_McLaughlinE1_433210.pdf	1713344	no	19
			881fa466f550de42d8aaf51f67d43d5ed5697530		
Warnings:					
Information:					
21	Non Patent Literature	91_Meng868_433243.pdf	388963	no	9
			0946ccd0cac2179f28f06fad241dbc2a02c980cd		
Warnings:					
Information:					
22	Non Patent Literature	92_MOBADERSANY_433227.pdf	1832165	no	10
			95f45f4bdaa35e003844441effb6d37eab59298a		
Warnings:					
Information:					

23	Non Patent Literature	93_Murphy305_354991.pdf	729397	no	7
			12f700da0158291309e23c785bba994cb483a1b2		
Warnings:					
Information:					
24	Non Patent Literature	94_Interactive_phe_433235.pdf	4101701	no	12
			35d036bc12501ea3417736a6bd6e446261b2080a		
Warnings:					
Information:					
25	Non Patent Literature	95_Nawaz766_433219.pdf	6932303	no	12
			7e9c0df9e9f1954ba42bb221099eb2537e08c44c		
Warnings:					
Information:					
26	Non Patent Literature	96_Newburn1_354992.pdf	262199	no	4
			342dc0fafe3f87a81ef92080db5c7bdab31fd1ff		
Warnings:					
Information:					
27	Non Patent Literature	97_NPL_ORourke347_433186.pdf	9680096	no	9
			206d81aec704ba63de993522447ce5c909179013		
Warnings:					
Information:					
28	Non Patent Literature	98_OpenSeadragon_A_433195.pdf	122408	no	3
			a49a6cab5a0fb684301772f38b3d628e0d8d1c52		
Warnings:					
Information:					
29	Non Patent Literature	99_Orange690_433203.pdf	3493044	no	12
			a7a6964a22eed7171d3abd2b69b0349790b3fa64		
Warnings:					
Information:					

30	Non Patent Literature	9_Baas9015_433188.pdf	3615178	no	6
			a5af09f4a89f6a48b8330442b40f09edc97e1b4		
Warnings:					
Information:					
31	Foreign Reference	1_CN107369151A_433276.pdf	1138092	no	15
			1a7240459d9d5ee5f1bd8e866677f50065ed24d9		
Warnings:					
Information:					
32	Foreign Reference	2_WO2017198790A1_433742.pdf	6108841	no	63
			f60de8e28ec3bb8e00195893d9281506c9f027ad		
Warnings:					
Information:					
33	Foreign Reference	3_WO2018065434A1_433282.pdf	6766157	no	57
			ab24ee2f00fbd854065ddb2fe2c1240b59d0a5c3		
Warnings:					
Information:					
34	Foreign Reference	4_WO2018067937A1_355037.pdf	2994495	no	56
			f0c84e7a1dc236bc28ed2d2f58ae9e6ea3797527		
Warnings:					
Information:					
35	Foreign Reference	5_WO2018136664A1_355038.pdf	18981370	no	339
			4331553e617d8ece28dd484149634347bbd93ea1		
Warnings:					
Information:					
36	Information Disclosure Statement (IDS) Form (SB08)	IDS_16732242_1.pdf	126307	no	10
			3a7671b68826d9c9b166086c43b04362eeaa89564		
Warnings:					
Information:					
This is not an USPTO supplied IDS fillable form					

37	Information Disclosure Statement (IDS) Form (SB08)	IDS_16732242_2.pdf	110920	no	8
			e4d2e3802208cc9e5c5cdbe8e4d94eafcbcc3b88		

Warnings:

Information:

This is not an USPTO supplied IDS fillable form

38	Information Disclosure Statement (IDS) Form (SB08)	IDS_16732242_3.pdf	111518	no	8
			37095467685e18ee1c6ce2c6b1fe8adae6ccf1bf		

Warnings:

Information:

This is not an USPTO supplied IDS fillable form

Total Files Size (in bytes):			151472385
-------------------------------------	--	--	-----------

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

Publication Number: CN 107369151 A

Publication Date: Nov 21, 2017

Abstract

The invention discloses a gastrointestinal stromal tumor pathological diagnosis support system and method based on big-data deep learning, and the system comprises an image data obtaining unit which is used for obtaining a normal stomach intestine tissue section image and a pathological section image of a diagnosed gastrointestinal stromal tumor case as the input image data; an image data marking unit which is used for marking the inputted image data; an image database construction unit which is used for carrying out the classification and arrangement of the marked image data provided by the image data marking unit, and constructing a pathological image database; a CNN (convolution neural network) construction unit which is used for constructing a first CNN model; and a CNN model training unit which is used for obtaining an ideal CNN model. According to the invention, the system and method can achieve the precise and high-efficiency intelligent reading of a film, so as to assist the pathological diagnosis of the gastrointestinal stromal tumor clinically, thereby improving the accuracy of pathological diagnosis, work efficiency and working persistent state.



(12)发明专利申请

(10)申请公布号 CN 107369151 A

(43)申请公布日 2017. 11. 21

(21)申请号 201710425297.X

G06K 9/62(2006.01)

(22)申请日 2017.06.07

G06F 17/30(2006.01)

(71)申请人 万香波

地址 510000 广东省广州市天河区员村二
横路26号

申请人 陈浩 王磊 范新娟 林黄靖
冯莉莉 窦琪 王平安 朱亚希
黄艳 中山大学附属第六医院
深圳视见医疗科技有限公司

(72)发明人 万香波

(74)专利代理机构 广州三环专利商标代理有限
公司 44202

代理人 宋静娜 郝传鑫

(51)Int. Cl.

G06T 7/00(2017.01)

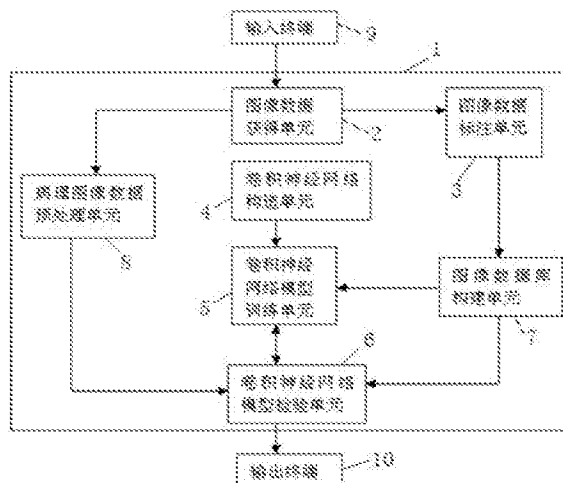
权利要求书2页 说明书7页 附图4页

(54)发明名称

基于大数据深度学习的胃肠间质瘤病理诊
断支持系统和方法

(57)摘要

本发明公开了一种基于大数据深度学习的胃肠间质瘤病理诊断支持系统和方法,该系统包括:图像数据获得单元,用于获得正常胃肠组织切片图像和已确诊的胃肠间质瘤病例的病理切片图像作为已输入图像数据;图像数据标注单元,用于对已输入图像数据进行标注;图像数据库构建单元,用于对图像数据标注单元提供的已标注图像数据分类、整理,构建病理图像数据库;卷积神经网络(CNN)构造单元,用于构造第一卷积神经网络模型;以及卷积神经网络模型训练单元,获得理想的卷积神经网络模型。通过本发明的胃肠间质瘤病理诊断支持系统和方法可实现精准和高效的智能读片,以辅助临床上胃肠间质瘤的病理诊断工作,提高其准确率、工作效率及工作持续状态。



1. 胃肠间质瘤病理诊断支持系统,其特征在於,所述支持系统包括:

图像数据获得单元,用于获得正常胃肠组织切片图像和已确诊的胃肠间质瘤病例的病理切片图像作为已输入图像数据;

图像数据标注单元,用于对所述已输入图像数据进行标注,以及保证图像的标签和图像的真实病理诊断结果一致;

图像数据库构建单元,用于对所述图像数据标注单元提供的已标注图像数据分类、整理,构建病理图像数据库;

卷积神经网络构造单元,用于构造第一卷积神经网络模型;以及

卷积神经网络模型训练单元,利用所述病理图像数据库的图像数据对所述第一卷积神经网络模型的参数进行调节,以及训练所述第一卷积神经网络模型,获得可用于检测患者病理图像数据的第二卷积神经网络模型。

2. 根据权利要求1所述的支持系统,其特征在於,所述支持系统还包括卷积神经网络模型检验单元,用于获得理想的卷积神经网络模型。

3. 根据权利要求2所述的支持系统,其特征在於,所述卷积神经网络模型检验单元包括卷积神经网络模型校验单元和卷积神经网络模型测试单元,所述卷积神经网络模型校验单元用于检测所述第二卷积神经网络模型的准确率;所述卷积神经网络模型测试单元,用于检测所述第二卷积神经网络模型是否过拟合,以筛选出鲁棒的第三卷积神经网络模型。

4. 根据权利要求1所述的支持系统,其特征在於,所述支持系统还包括卷积神经网络模型预训练单元,用于当所述图像数据获得单元收集到的已输入图像数据不足时,对所述第一卷积神经网络模型进行预训练。

5. 根据权利要求1所述的支持系统,其特征在於,所述支持系统还包括病理图像数据处理单元,用于筛选和示出患者病理图像中的待检测区域。

6. 根据权利要求5所述的支持系统,其特征在於,所述预处理单元采用自适应阈值法筛选出所述待检测区域。

7. 根据权利要求1所述的支持系统,其特征在於,所述卷积神经网络训练单元采用微调法训练所述第一卷积神经网络模型。

8. 一种胃肠间质瘤病理诊断的支持方法,其特征在於,所述支持方法包括以下步骤:

图像数据获得:获得正常胃肠组织切片图像和已确诊的胃肠间质瘤病例的病理切片图像作为已输入图像数据;

图像数据标注:对所述已输入图像数据进行标注,以及保证图像的标签和图像的真实病理诊断结果一致;

图像数据库构建:对所述图像数据标注单元提供的已标注图像数据分类、整理,构建病理图像数据库;

卷积神经网络构造:构造第一卷积神经网络模型;以及

卷积神经网络模型训练:利用所述病理图像数据库的图像数据对所述第一卷积神经网络模型的参数进行调节,以及训练所述第一卷积神经网络模型,获得可用于检测患者病理图像数据的第二卷积神经网络模型。

9. 根据权利要求8所述的支持方法,其特征在於,所述支持方法还包括卷积神经网络模型检验步骤:获得理想的卷积神经网络模型;所述卷积神经网络模型检验步骤包括卷积神

神经网络模型校验和卷积神经网络模型测试,所述卷积神经网络模型校验用于检测所述第二卷积神经网络模型的准确率;所述卷积神经网络模型测试,用于检测所述第二卷积神经网络模型是否过拟合,以筛选出鲁棒的第三卷积神经网络模型。

10.根据权利要求1-6任一所述的支持设备在诊断胃肠间质瘤中的应用。

基于大数据深度学习的胃肠间质瘤病理诊断支持系统和方法

技术领域

[0001] 本发明涉及一种基于大数据深度学习的胃肠间质瘤病理诊断支持系统和方法。

背景技术

[0002] 深度学习是目前人工智能领域用于图像识别和语音分析的最契合、应用最广泛的算法,其灵感来自于人类大脑的工作机制,是通过建立卷积神经网络对外部输入的数据进行自动化特征提取,从而能够使机器理解学习数据,获得信息并输出。目前,基于深度学习的人工智能已应用于各个行业领域,包括语音识别、人脸识别、车标识别、手写汉字识别等。近年来人工智能辅助医疗技术的产品研发也取得重大进展,如由谷歌大脑及Verily公司研发的用于乳腺癌病理诊断的人工智能产品可达到89%的肿瘤定位准确率;浙江大学附属第一医院利用人工智能实现快速分析甲状腺B超中结节区域的位置和良恶性。

[0003] 在医疗诊断过程中,病理组织切片检查需要极高的规范化和精确度。目前绝大部分的病理组织切片是由人工制作,并由病理医生结合自身长期积累的临床诊断经验来分析和判断。胃肠间质瘤(Gastrointestinal stromal tumors,GIST)是好发于胃肠道的间叶源性的肿瘤。GIST常伴随着胃肠道出血、吞咽困难,严重影响患者正常生活,因此GIST患者需要得到及时有效的治疗以防病情持续恶化。但由于缺乏特异性临床表现,确诊GIST必须依据病理组织学和免疫组化检测结果。将病变组织制备成病理切片标本,置于光学显微镜下观察,可见GIST瘤细胞形态多变,主要是上皮样细胞和梭形细胞,前者呈圆形、椭圆形或多边形,胞浆丰富,核仁明显,嗜酸性较弱,多成弥漫、巢状或围绕血管成器官样排列;后者核仁不明显,两侧呈钝圆,嗜酸性可较弱也可为中度,染色质不集中,主要呈交叉束状、栅栏状、漩涡状排列。根据组织良恶性程度可见不同程度的细胞异型性和核分裂像。

[0004] 因此,病理组织学方法诊断GIST的缺点主要是:病理玻片的结果判断是由病理医生肉眼观察所得,这种人工阅片方法与病理医生自身经验、工作状态等主观因素密切相关,容易产生误差。临床上也往往因为胃肠道内环境复杂,GIST常被误诊为平滑肌源性肿瘤或神经源性肿瘤,这对于经验不足的病理医生更容易漏诊、误诊。同时,病理医生要负责检查切片上的所有可见的生物组织,而每个病人都会有许多切片,在进行40倍放大时每个切片都有100多亿的像素,因此人工阅片工作量极大,容易受阅片者主观情绪和疲劳阅片等因素的影响。而且,不同的病理学家对同一病人可能给出相当不同的诊断结论。因此,这种高度依赖人为因素的病理组织学诊断方法具有主观差异性,加上其工作强度大、时间成本高且诊断不一致性等缺点,很大程度上会影响GIST的早期诊治从而影响患者预后。另外,培养合格专业的病理医生需要进行长期的专业培训和实践过程,培养周期长,且易受当前社会经济、文化等社会因素的影响,意味着我国乃至全世界病理医生数量“供不应求”、职业缺口大的严峻形势亟需解决。

发明内容

[0005] 针对病理组织学人工阅片的缺点,本发明拟通过计算机对大量胃肠间质瘤病理图

像进行深度学习,来建立智能化的胃肠间质瘤病理诊断数学模型,搭建基于大数据及深度学习算法的胃肠间质瘤辅助病理诊断人工智能平台,从而实现高准确率和高效率的智能读片,以辅助临床上胃肠间质瘤的病理诊断工作,提高其准确率、工作效率及工作持续状态。

[0006] 基于此,本发明的目的在于克服上述现有技术的不足之处而提供一种可提高临床诊断胃肠间质瘤时的效率、降低医疗成本的胃肠间质瘤病理诊断支持系统。

[0007] 为实现上述目的,本发明采取的技术方案为:一种基于大数据深度学习的胃肠间质瘤病理诊断支持系统,所述支持系统包括:图像数据获得单元,用于获得正常胃肠组织切片图像和已确诊的胃肠间质瘤病例的病理切片图像作为已输入图像数据;图像数据标注单元,用于对所述已输入图像数据进行标注,以及保证图像的标签和图像的真实病理诊断结果一致;图像数据库构建单元,用于对所述图像数据标注单元提供的已标注图像数据分类、整理,构建病理图像数据库;卷积神经网络构造单元,用于构造第一卷积神经网络模型;以及卷积神经网络模型训练单元,利用所述病理图像数据库的图像数据对所述第一卷积神经网络模型的参数进行调节,以及训练所述第一卷积神经网络模型,获得可用于检测患者病理图像数据的第二卷积神经网络模型。

[0008] 由此,医生可结合所述支持设备针对输入的患者病理图像给出的分类结果和相应的概率,以及医生的专业知识和经验快速地诊断出该患者的是否患有胃肠间质瘤,显著提高临床诊断的效率,从而降低医疗成本;其中,为了保证收集到的图像数据准确无误,可以利用图像标注工具ASAP,对每张病理切片图像进行标注,以保证图像的标签和真实值一致;为了加快训练网络模型的速度,可以使用具有高速并行计算的GPU来代替CPU进行训练;为了加快卷积神经网络模型的检测速度,基于卷积神经网络训练单元,可以将训练好的网络模型重新建模为可变步长的CNN分类模型结构,以用于实际操作中的检测方法;该模型将对巨大的全切片图像进行分块化处理,将预选出的活体组织区域分割成大小相同的ROI分块,由于分块间的检测可以高度并行化,使得检测的速度在多GPU的并行下得到显著提升,然后通过可变步长的CNN分类模型的检测,生成预测概率分布图像;图像数据库将病理图像数据分为训练集、测试集和校验集等;第一卷积神经网络模型的参数包括学习速率、训练次数和多少层网络等网络参数,训练指在寻求最优解时,自动调节参数的过程。

[0009] 优选地,所述支持系统还包括卷积神经网络模型检验单元,用于获得理想的卷积神经网络模型。应当说明的是,此处“理想”是指卷积神经网络模型的准确率高,以及“鲁棒”。

[0010] 优选地,所述卷积神经网络模型检验单元包括卷积神经网络模型校验单元和卷积神经网络模型测试单元,所述卷积神经网络模型校验单元用于检测所述第二卷积神经网络模型的准确率;所述卷积神经网络模型测试单元,用于检测所述第二卷积神经网络模型是否过拟合,以筛选出鲁棒的第三卷积神经网络模型;应当说明的是,如若模型在测试集上的准确率与校验集训练中的准确率相差较大,则说明模型过拟合,此时,可返回卷积神经网络训练单元中,调节网络结构或参数,进行再次训练以获得更好的网络模型;如若在测试集上的准确率与校验集训练中的准确率十分接近,则说明该模型较为鲁棒。

[0011] 优选地,所述支持系统还包括卷积神经网络模型预训练单元,用于当所述图像数据获得单元收集到的已输入图像数据不足时,对所述第一卷积神经网络模型进行预训练。

[0012] 优选地,所述支持系统还包括病理图像数据预处理单元,用于筛选和示出患者病

理图像中的待检测区域。

[0013] 优选地,为了确保检测的有效性,所述预处理单元采用自适应阈值法筛选出所述待检测区域。

[0014] 优选地,所述卷积神经网络训练单元采用微调法训练所述第一卷积神经网络模型。

[0015] 作为本发明的另一方面,本发明还提供了一种胃肠间质瘤的病理诊断支持方法,所述支持方法包括以下步骤:

[0016] 图像数据获得:获得正常胃肠组织切片图像和已确诊的胃肠间质瘤病例的病理切片图像作为已输入图像数据;

[0017] 图像数据标注:对所述已输入图像数据进行标注,以及保证图像的标签和图像的真实病理诊断结果一致;

[0018] 图像数据库构建:对所述图像数据标注单元提供的已标注图像数据分类、整理,构建病理图像数据库;

[0019] 卷积神经网络构造:构造第一卷积神经网络模型;以及

[0020] 卷积神经网络模型训练:利用所述病理图像数据库的图像数据对所述第一卷积神经网络模型的参数进行调节,以及训练所述第一卷积神经网络模型,获得可用于检测患者病理图像数据的第二卷积神经网络模型。

[0021] 应当说明的是,图像数据标注和图像数据库构建可以看作病理图像数据库构建阶段。优选地,所述支持方法还包括卷积神经网络模型检验步骤:获得理想的卷积神经网络模型;所述卷积神经网络模型检验步骤包括卷积神经网络模型校验和卷积神经网络模型测试,所述卷积神经网络模型校验用于检测所述第二卷积神经网络模型的准确率;所述卷积神经网络模型测试,用于检测所述第二卷积神经网络模型是否过拟合,以筛选出鲁棒的第三卷积神经网络模型。应当说明的是,卷积神经网络构造、卷积神经网络模型训练和卷积神经网络模型检验可看作卷积神经网络的训练阶段,用于获得理想的卷积神经网络模型。

[0022] 作为本发明的第三个方面,本发明还涉及上述支持系统在病理诊断胃肠间质瘤中的临床应用。

[0023] 综上所述,本发明的有益效果为:

[0024] 与现有的病理医生人工阅片相比较,本发明基于大数据和深度学习算法的胃肠间质瘤病理诊断的支持设备具有准确率高、耗时短且工作持续时间长的优点,且此发明在各大医院包括三甲、基层医院的推广和云端服务化将有助于解决医疗资源分配不均匀、实现远程优质医疗等,为广大患者提供更便捷、更准确的病理诊断服务;上述优点的实现是因为本发明的设备和方法利用深度学习算法在图像识别的优势,让计算机进行大数据级别的胃肠间质瘤病理切片的深度学习,从而训练出能模拟病理医生阅片并与之媲美的智能化神经网络模型,经过不断学习和验证,该神经网络模型可以实现对胃肠间质瘤病理切片的智能阅片、快速识别并得出科学结论。

附图说明

[0025] 图1为本发明的胃肠间质瘤病理诊断支持系统的结构框图;

[0026] 图2为本发明实施例二的胃肠间质瘤病理诊断支持方法的流程图;

[0027] 图3为对胃肠间质瘤切片图；

[0028] 图4为本发明的实施例二的快速检测模型的原理图；

[0029] 图5为本发明的胃肠间质瘤病理诊断支持系统应用时的流程图；

[0030] 其中,1、胃肠间质瘤病理诊断支持系统,2、图像数据获得单元,3、图像数据标注单元,4、卷积神经网络构造单元,5、卷积神经网络模型训练单元,6、卷积神经网络模型检验单元,7、图像数据库构建单元,8、病理图像数据预处理单元,9、输入终端,10、输出终端。

具体实施方式

[0031] 为更好的说明本发明的目的、技术方案和优点,下面将结合附图和具体实施例对本发明作进一步说明。

[0032] 实施例1

[0033] 参见图1,本发明的胃肠间质瘤病理诊断支持系统1的一种实施例,其包括:

[0034] 图像数据获得单元2,用于获得正常胃肠组织切片图像和已确诊的胃肠间质瘤病例的病理切片图像作为已输入图像数据;

[0035] 图像数据标注单元3,用于对已输入图像数据进行标注,以及保证图像的标签和图像的真实病理诊断结果一致;

[0036] 图像数据库构建单元7,用于对图像数据标注单元提供的已标注图像数据分类、整理,构建病理图像数据库;

[0037] 卷积神经网络构造单元4,用于构造第一卷积神经网络模型;

[0038] 卷积神经网络模型训练单元5,利用病理图像数据库的图像数据对第一卷积神经网络模型的参数进行调节,以及训练第一卷积神经网络模型,获得可用于检测患者病理图像数据的第二卷积神经网络模型;

[0039] 卷积神经网络模型检验单元6,用于获得理想的卷积神经网络模型,包括卷积神经网络模型校验单元(图中未显示)和卷积神经网络模型测试单元(图中未显示),卷积神经网络模型校验单元用于检测第二卷积神经网络模型的准确率;卷积神经网络模型测试单元,用于检测第二卷积神经网络模型是否过拟合,以筛选出鲁棒的第三卷积神经网络模型。

[0040] 卷积神经网络模型预训练单元(图中未显示),用于当图像数据获得单元收集到的已输入图像数据不足时,对第一卷积神经网络模型进行预训练;以及

[0041] 病理图像数据预处理单元8,用于筛选和示出患者病理图像中的待检测区域。

[0042] 其中,病理图像数据预处理单元8采用自适应阈值法筛选出待检测区域;卷积神经网络模型训练单元5采用微调(fine-tune)法训练第一卷积神经网络模型;数据库中包括以下四类数据集:训练集、消验集、测试集以及常规的已公开病理图像数据集。

[0043] 另外,输入终端9用于将现有的正常胃肠组织切片图像和已确诊的胃肠间质瘤病例的病理切片图像输入图像数据获得单元2,并且,这些输入的数据最后将由图像数据库构建单元7分类收集,用于支持后续的临床诊断工作;

[0044] 以及将待检测的患者的病理切片图像输入病理图像数据预处理单元8;

[0045] 输出终端10,用于将由卷积神经网络模型训练单元5获得的鲁棒的卷积神经网络模型对输入病理图像数据预处理单元8的待检测的患者的病理切片图像分类检测的结果(病理类型和相应的概率)呈现给医生,以供临床诊断参考。

[0046] 实施例2

[0047] 参见图2,本发明的胃肠间质瘤诊断支持方法的一种实施例,其包括如下步骤:

[0048] (1) 采集图像数据

[0049] 以中山大学附属第六医院医学生物库数据为数据来源,采集12000张病理切片图像,包括6000张正常组织切片图像和6000张胃肠间质瘤组织切片,并分别按照训练集:校验集:测试集=3:1:1的数量比例进行随机分组。如下表1所示:

[0050] 表1病理切片图像的具体数据。

[0051]		胃肠正常组	胃肠道间质瘤	总共(张)
		组织切片图像	组织切片图像	
	训练集(张)	3600	3600	7200
[0052]	校验集(张)	1200	1200	2400
	测试集(张)	1200	1200	2400
	总共(张)	6000	6000	12000

[0053] (2) 标注图像信息

[0054] 利用现有的ASAP图像标注软件对步骤(1)所收集的训练集、校验集和测试集的病理切片图像进行数据标记。为保证信息标注的准确性,需在标注前对图像进行优化处理。对图像的标注工作主要包括:用不同颜色/粗细/虚实线条勾勒出图像中各种病理结构区域,例如,GIST瘤细胞形态多变,主要是上皮样细胞和梭形细胞,前者呈圆形、椭圆形或多边形,胞浆丰富,核仁明显,嗜酸性较弱,多成弥漫、巢状或围绕血管成器官样排列;后者核仁不明显,两侧呈钝圆,嗜酸性可较弱也可为中度,染色质不集中,主要呈交叉束状、栅栏状、漩涡状排列。根据组织良恶性程度可见不同程度的细胞异型性和核分裂像,然后对图像分类并赋予分值,并将所勾勒的区域进行标签命名。将正确标注后的病理图像进行数字化存储,以进行下一步的网络模型训练和校验。图3为对胃肠间质瘤中不典型增生区域的标注图。

[0055] (3) 训练卷积神经网络

[0056] ①设计模型

[0057] (a) 按照卷积层、最大采样层、非线性函数、全连接层级联的方式来构造卷积神经网络;

[0058] (b) 采用多层网络来增强网络的拟合能力;

[0059] (c) 网络最后的全连接层的输出的通道数设为2,分别代表该图像是正常胃肠组织切片图像、胃肠间质瘤组织切片图像。

[0060] ②训练网络

[0061] (a) 根据步骤(1)、(2)中收集到的图像数据,对卷积神经网络模型的参数进行调节,在校验集上观察分类的准确率;

[0062] (b) 为了加快训练网络的速度,使用具有高速并行计算的GPU来代替CPU进行训练;

[0063] (c) 卷积神经网络权值参数的更新的方法使用SGD来求解,如若收敛速度较慢,使用Adadel ta,Adam等优化方法来求解;

[0064] (d) 如若步骤(1)收集到的训练数据(即图像数据)数目太少,卷积神经网络模型采用先在常规的公开病理图像数据集预训练,再将收集到的图像数据采用fine-tune(微调)的方法来训练卷积神经网络模型;

[0065] (e) 如在现有卷积神经网络模型上训练,分类的准确率无法上升,可以通过增加卷积神经网络模型的深度来增加卷积神经网络模型的拟合能力。

[0066] ③设计快速检测模型(如图4所示)

[0067] 1.为了提高检测效率,在预处理阶段采用自适应阈值法,从全切片图像中预选出活体组织区域,作为卷积神经网络的检测对象(如图4箭头101所示,代表预处理过程)。

[0068] 2.为了提高检测的准确度以及灵活性,基于步骤(3),可以将训练好的卷积神经网络重新建模为可变步长的CNN分类模型,以用于实际操作中的检测方法;该模型将对巨大的全切片图像进行分块化处理,将预选出的活体组织区域分割成大小相同的ROI分块;由于分块间的检测可以高度并行化,使得检测的速度在多GPU的情况下得到有效的提升(如图4箭头102所示,代表模型快速检测过程)。通过卷积神经网络模型的检测,生成预测概率分布图像。

[0069] 3.基于第2步的预测概率分布图像,在后处理中,筛除零散点后,分析预测概率分布图,以得到最终全切片图像的预测结果(如图4箭头103所示,代表后处理过程)。

[0070] (4) 测试集验证

[0071] (a) 基于步骤③的可变步长的分类模型结构,使用步骤(3)中训练好的卷积神经网络模型来对测试集测试,观察模型在测试集上的准确率。

[0072] (b) 如若步骤(3)中训练好的卷积神经网络模型在测试集上的准确率与训练中的校验集的准确率相差较大,则说明模型过拟合;此时,可返回步骤(3)中,调节卷积神经网络模型网络结构或参数,获得更好的网络模型。

[0073] (c) 如若步骤(3)中训练好的卷积神经网络模型在测试集上的准确率与训练中的校验集的准确率十分接近,则说明该训练所得的卷积神经网络模型较为鲁棒,可作为合适的检测病患的病理图像的网络模型。

[0074] 实施例3

[0075] 本发明的胃肠间质瘤病理诊断支持系统的一种应用例,将待检测的病理图像通过输入终端9输入本发明的胃肠间质瘤诊断支持设备中的病理图像数据预处理单元8,之后的运行流程参见图5,其中,

[0076] (a) 为了确保检测的有效性,在预处理阶段采用自适应阈值法,从患者组织全切片图像中预选出活体组织区域,然后以组织重心为中心提取或框选出待检测区域(即患者病理组织区域);

[0077] (b) 之后,对患者病理组织区域图片进行预处理,预处理包括去噪、直方图均衡、归一化等步骤;

[0078] (c) 用先前训练好的卷积神经网络模型(即实施例1中第二卷积神经网络模型)对预处理后的图片中待检测区域进行分类检测,从而得出胃肠间质瘤的预测结果,包括该病理切片所属的胃肠间质瘤类别以及相应的概率。

[0079] 实施例4本发明的胃肠间质瘤病理诊断支持方法与现有方法的比较

[0080] 目前临床上病理诊断工作是由经过规范化培训的病理医生人工阅读病理组织切

片,结合自身长期积累的临床诊断经验来作出分析和诊断。由于这种人工肉眼阅片方法与病理医生自身经验、工作状态、主观情绪等因素密切相关,故准确率不高,但耗时长,工作持续时间有限,容易产生漏诊、误诊及诊断不一致等情况。本发明则利用计算机对标准化的大量胃肠间质瘤病理图像的深度学习,对卷积神经网络进行参数调节和拟合训练,从而获得较为鲁棒的网络模型。这种基于大数据和深度学习的神经网络能模拟人工阅片,根据输入的新病理图像得出对应的输出值即病理诊断结论。再者,通过模型重建,在不影响检测准确度的情况下,极大提高了检测速度。

[0081] 选取35名具有3年以上胃肠间质瘤诊断与治疗经验的医生,分别每人提供40张疑似胃肠间质瘤的病理图像,判断其类型,然后计算准确率和平均时间,统计医生诊断状态,与本发明的诊断支持方法相比较,其结果如下表2所示。

[0082] 表2胃肠间质瘤诊断结果的比较

[0083]

	病理医师阅片		人工智能阅片	
诊断准确率	图像测试正确例数(1400例)	测试准确率	图像测试正确例数(1400例)	测试准确率

[0084]

	1224	高(87.4%)	1348	更高(96.3%)
诊断速度(平均时间)	慢 (15-30min)		快 (3-5min)	
诊断稳定性	不稳定,易受主观影响		稳定	
工作持续状态	有限,易受主观影响		无限	

[0085] 从上表2可知,采用本发明的方法阅读病理组织切片,其准确率比专业病理医生的水平更高,且耗时更短,工作持续时间长。

[0086] 最后所应当说明的是,以上实施例仅用以说明本发明的技术方案而非对本发明保护范围的限制,尽管参照较佳实施例对本发明作了详细说明,本领域的普通技术人员应当理解,可以对本发明的技术方案进行修改或者等同替换,而不脱离本发明技术方案的实质和范围。

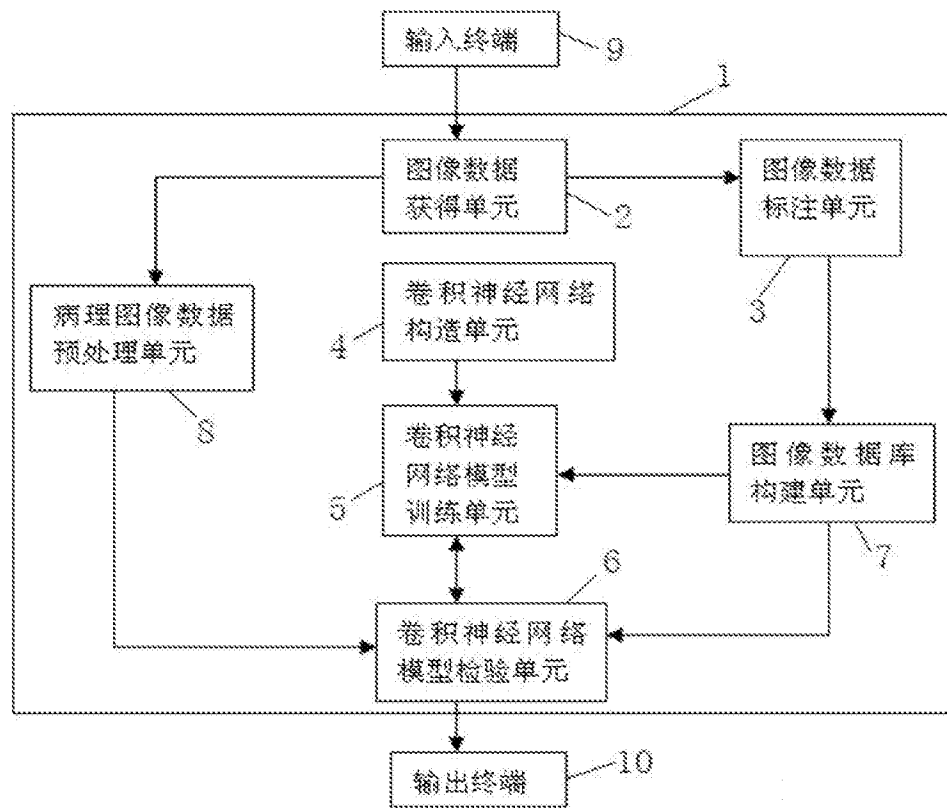


图1

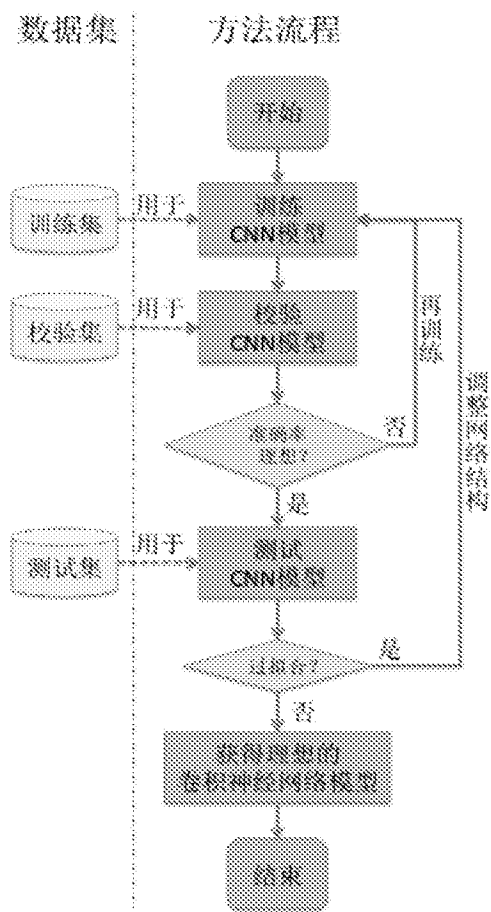


图2

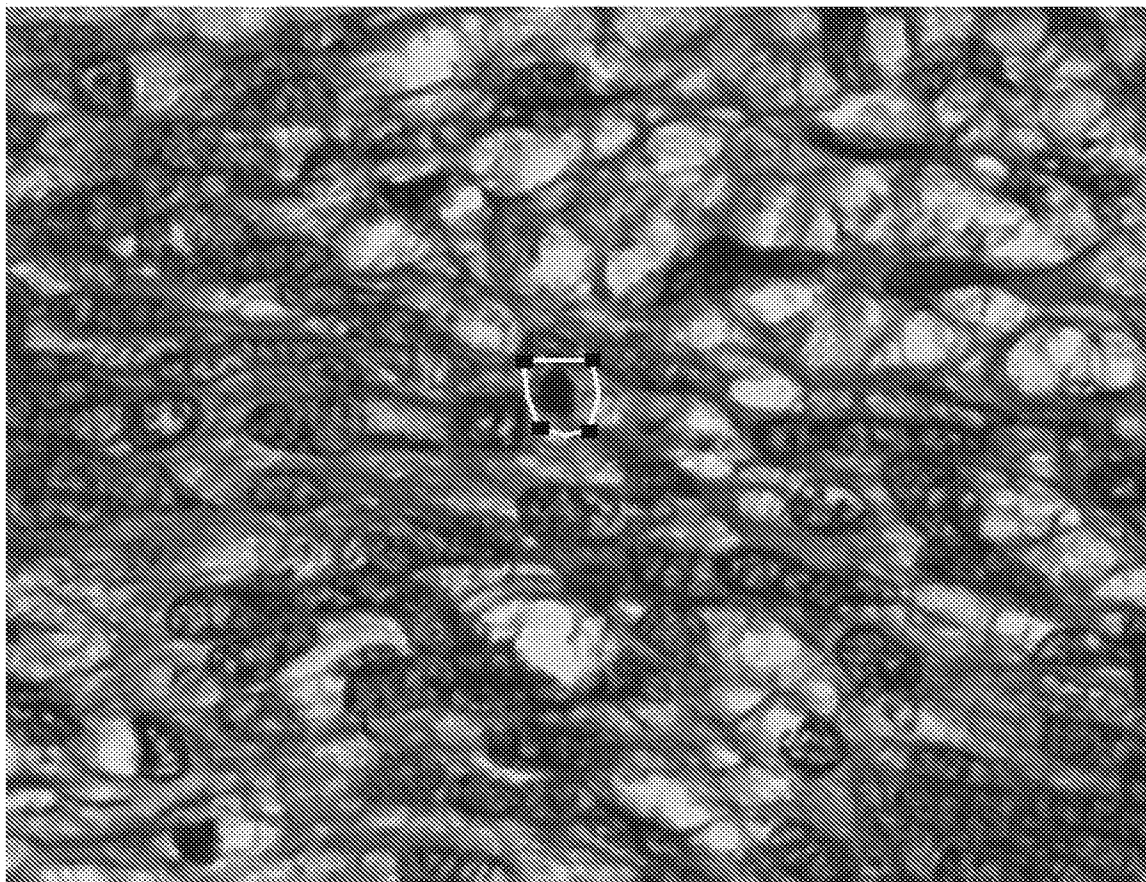


图3

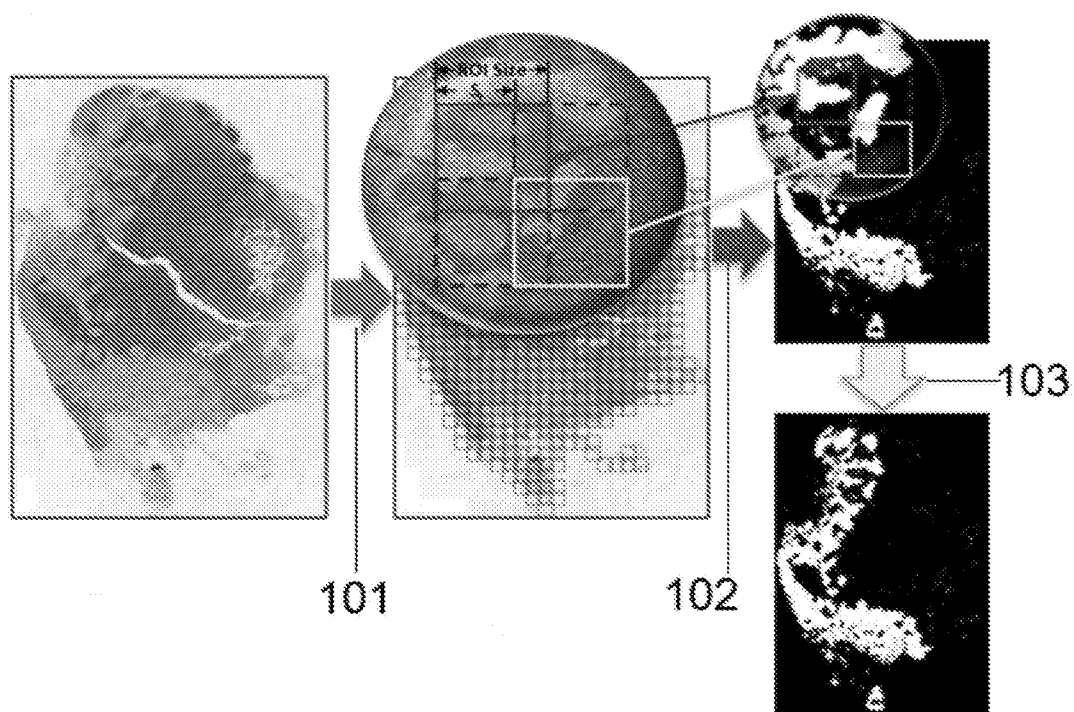


图4

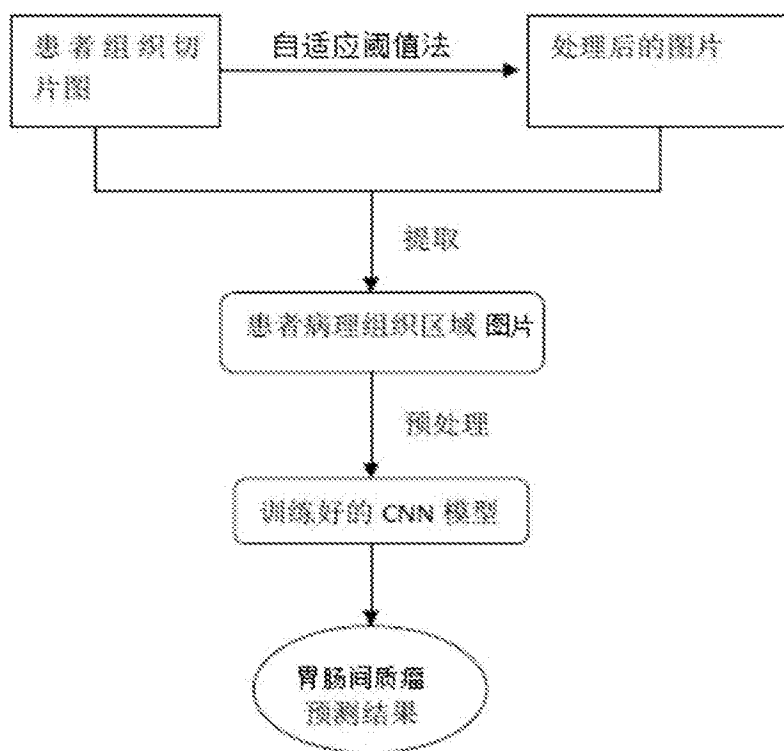


图5

Publication Number: WO 2017/198790 A1

Publication Date: Nov 23, 2017

Abstract

The invention relates to an image analysis system (100) for tumor classification. The system is configured for: receiving (202) at least one digital image (118) of a tissue sample; analyzing (204) the at least one received image for identifying immune cells and tumor cells in the at least one received image; for each of the identified tumor cells, determining (206) the distance of said tumor cell to the nearest immune cell; computing (208) a proximity measure as a function of the determined distances; in dependence on the proximity measure, classifying (210) the identified tumor cells into tumor cells of an inflammatory tumor or as tumor cells of a non-inflammatory tumor; and storing (212) the classification result on a storage medium and/or displaying (214) the classification result on a display device (102).



(51) International Patent Classification:

G06T 7/00 (2017.01)

(21) International Application Number:

PCT/EP2017/062037

(22) International Filing Date:

18 May 2017 (18.05.2017)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

16170224.6 18 May 2016 (18.05.2016) EP

16201272.8 29 November 2016 (29.11.2016) EP

(71) Applicant (for all designated States except US): **F. HOFFMANN-LA ROCHE AG** [CH/CH]; Grenzacherstraße 124, 4070 Basel (CH).

(71) Applicant (for US only): **HOFFMANN-LA ROCHE INC.** [US/US]; Great Notch, 150 Clove Road 8th Floor, Little Falls, New Jersey 07424 (US).

(72) Inventors: **GAIRE, Fabien**; Sonnenbichlweg 1, 82319 Starnberg (DE). **KLAIMAN, Eldad**; St.-Valentins-Weg 4, Starnberg 82319 (DE). **KORSKI, Konstanty**; Ul. Kormorana 21, 61-058 Poznan (PL).

(74) Agent: **RICHARDT PATENTANWÄLTE PARTG MBB**; Wilhelmstraße 7, 65185 Wiesbaden (DE).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM,

(54) Title: TUMOR PROXIMITY MEASURE

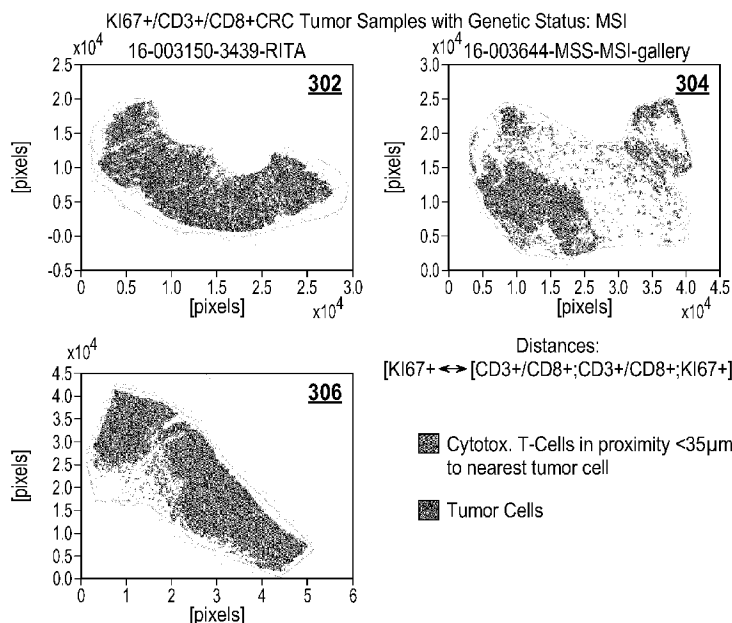


Fig. 3

(57) Abstract: The invention relates to an image analysis system (100) for tumor classification. The system is configured for: receiving (202) at least one digital image (118) of a tissue sample; analyzing (204) the at least one received image for identifying immune cells and tumor cells in the at least one received image; for each of the identified tumor cells, determining (206) the distance of said tumor cell to the nearest immune cell; computing (208) a proximity measure as a function of the determined distances; in dependence on the proximity measure, classifying (210) the identified tumor cells into tumor cells of an inflammatory tumor or as tumor cells of a non-inflammatory tumor; and storing (212) the classification result on a storage medium and/or displaying (214) the classification result on a display device (102).

PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

TUMOR PROXIMITY MEASURE

Description

Field of the invention

The invention relates to the field of image analysis, and more particularly to the field of image-based tumor classification.

Background and related art

Microsatellite instability (MSI) is the condition of genetic hypermutability that results from impaired DNA mismatch repair (MMR). The presence of MSI represents phenotypic evidence that MMR is not functioning normally and thus is an indicator of the genetic state of a cell considered as a potential tumor cell.

Microsatellite instability is known to be associated with colon cancer, gastric cancer, endometrium cancer, ovarian cancer, hepatobiliary tract cancer, urinary tract cancer, brain cancer, and skin cancers. MSI is most prevalent as the cause of colon cancers. Each year, there are over 500,000 colon cancer cases worldwide. Based on findings from over 7,000 patients stratified for MSI-High (MSI-H), MSI-Low (MSI-L), or Microsatellite Stable (MSS) colon cancers, those with MSI-H had a more positive prognosis by 15% compared to MSI-L or MSS tumors.

MSI is a good marker for determining a prognosis for cancer treatments. Additional approaches which are based on genetic information of a tumor exist. For example, extensive research is ongoing to decipher the underlying genetic patterns of tumor tissue cells with the hope to improve early cancer diagnosis and treatment. The recent progress in next generation sequencing technologies has revolutionized the field of cancer genomics. However, one caveat of these studies remains the large amount of genetic variations identified and their interpretation.

Summary

It is an objective of the present invention to provide for an improved method and image analysis system for classifying tumor cells as specified in the independent claims. Embodiments of the invention are given in the dependent claims.

Embodiments of the present invention can be freely combined with each other if they are not mutually exclusive.

In one aspect, the invention relates to an image analysis method for tumor classification. The method comprises:

- receiving, by an image analysis system, at least one digital image of a tissue sample;
- 5 – analyzing, by the image analysis system, the at least one received image for identifying immune cells and tumor cells in the at least one received image;
- for each of the identified tumor cells, determining, by the image analysis system, the distance of said tumor cell to the nearest immune cell;
- computing, by the image analysis system, a proximity measure as a function of
10 the determined distances;
- in dependence on the proximity measure, classifying, by the image analysis system, the identified tumor cells into tumor cells of an inflammatory tumor or as tumor cells of a non-inflammatory tumor; and
- storing, by the image analysis system, the classification result on a storage
15 medium and/or displaying the classification result on a display device.

Said approach for classifying the identified tumor cells into cells of an inflammatory vs. non-inflammatory tumor may be advantageous for multiple reasons:

- 20 The distance of immune cells and tumor cells appears to more accurately reflect cellular processes related to the immune response of an organism against a tumor than, for example, the MSS/MSI status. Thus, the proximity measure appear to more accurately reflect the inflammatory status of a tumor (an inflammation is— very generally speaking — the body's immune system's response to a stimulus) than a
25 genetic approach. It has been observed that at least in some cases inflammatory versus non-inflammatory tumor were correctly distinguished based on the proximity measure which were not correctly distinguished based on their MSS/MSI status.

- 30 The high mutational load in MSI tumors creates many tumor-specific neoantigens, typically 10–50 times that of MSS tumors. Some of these neoantigens will be processed, presented on MHC, and recognized as foreign by T cells. The high neoantigen burden might be one explanation for the high level of tumor-infiltrating

lymphocytes ("inflammatory tumor state") in MSI tumors. MSI tumors have a better prognosis than MSS tumors and respond better to several classes of drugs which boost or modulate the immune response. By determining a proximity measure based on the individual distances of tumor cells and immune cells, a more accurate
5 classification of inflammatory vs. non-inflammatory tumors can be made and/or a more accurate prognosis and treatment recommendation (whether or not to prescribe a drug that boosts or modulates the immune response) can be given.

According to a further beneficial aspect, the proximity measure can quickly be
10 computed from a digital image of a tumor tissue sample having been stained with an immunofluorescence assay. Thus, an additional time consuming and expensive (and-as it appears-less accurate) MSS/MSI status analysis of tumor cells may be avoided.

15 In a further beneficial aspect (compared to e.g. to various immune score computations), and increased accuracy of tumor classification and prognosis may be achieved, because distances between individual immune cells and tumor cells are determined. Instead of roughly determining the outline of a tumor region and then determining whether an immune cell is contained in said tumor region or not,
20 embodiments of the invention use individual immune cell-tumor cell distances as a basis for computing a proximity measure from all said determined distances. This proximity measure is a more accurate indication of whether immune cells and tumor cells are located sufficiently close to each other to allow for an effective immune response against the tumor cells. In other words, the proximity measure for example
25 indicates whether a sufficiently large number of immune-promoting immune cells such as cytotoxic T cells, B cells, memory cells, T-helper cells and/or macrophages are within an "immunologically effective distance" from the tumor cells.

An "immunologically effective distance" as used herein is a distance between an immune cell and its nearest tumor cell which is sufficiently small to allow for the
30 killing of said tumor cell by said immune cell. The killing of the tumor cell may be performed directly or indirectly and may be performed by different pathways which may depend on the type of immune cell. For example, the killing of the tumor cell

may be achieved via direct cell-cell interaction between the immune cell and the tumor cell, via a direct or signal-molecule based interaction of said immune cell with a further immune cell in the proximity of the tumor cell whereby the further immune cells kills or triggers the apoptosis of the tumor cell, or by triggering the apoptosis of the tumor cell (e.g. via direct cell-cell interaction or via signal molecules secreted by the immune cell).

In a further beneficial aspect, the distance information may not only be used as an input for accurately classifying tumor as an inflammatory versus a non-inflammatory tumor, the distance information may also be used for generating a graphical output that may allow a pathologist to easily and accurately identify the location and distribution of inflammatory regions within a tumor tissue.

In a further beneficial aspect, once the individual immune cells and tumor cells have been identified, the distance computation for individual pairs of neighboring tumor cells and in immune cells may be performed very quickly and efficiently. As the computed distances on numerical values, also the proximity measure computation as a function of a plurality of numerical values may be performed quickly and with comparatively low computational effort even in case in the depicted tissue sample comprises many thousand cells.

In a further aspect, the invention relates to an image analysis method for tumor classification. The method comprises:

- receiving , by an image analysis system, at least one digital image of a tissue sample;
- analyzing , by the image analysis system, the at least one received image for identifying tumor cells in the at least one received image;
- analyzing , by the image analysis system, the identified tumor cells for identifying tumor regions, wherein a tumor region is a tissue region whose majority of cells consist of tumor cells;
- analyzing , by the image analysis system, the at least one received image for identifying immune cells lying within one of the identified tumor regions or lying

within a boundary region around one of the tumor regions, the width of the boundary being smaller than 200 μ m;

- for each of the identified immune cells, determining, by the image analysis system, the distance of said immune cell to the nearest tumor cell;
- 5 – computing , by the image analysis system, a proximity measure as a function of the determined distances;
- in dependence on the proximity measure, classifying, by the image analysis system, the identified tumor cells into tumor cells of an inflammatory tumor or as tumor cells of a non-inflammatory tumor; and
- 10 – storing, by the image analysis system, the classification result on a storage medium and/or displaying the classification result on a display device.

Thus, according to said alternative approach, at first the tumor cells and corresponding tumor regions are identified in the digital image. Then, immune cells
15 in said tumor regions or within close spatial proximity to said tumor regions are identified and the immune cell – tumor cell distances are identified. In this scenario, the distances may be computed starting from the individual identified immune cells (“for each identified immune cell”) or may be computed starting from the individual identified tumor cells (for each identified tumor cell”). For example, gray scale and
20 color segmentation techniques, edge detection, voting and radial symmetry based image analysis techniques can be used for identifying and/or classifying cells and for identifying regions consisting of or mainly comprising similarly colored or shaped cells having been identified as tumor cells. Moreover, various machine learning techniques with supervised learning methods such as SVM (Support vector
25 Machines), DNN (Deep Neural Networks), Random Forest, etc. can be used to classify cells into types and/or to identify tumor regions, non-tumor tissue regions and glass regions on a slide. Limiting the distance determination on the tumor and its boundary may ensure that in case a whole tissue slide comprises only a small tumor region and the majority of immune cells is located far away from the tumor,
30 the effect of said distant immune cells (which may be present in said far-away regions irrespective of whether the tumor is inflammatory or not) does not modify the proximity measure.

According to embodiments, the image analysis method further comprises analyzing the received digital image for determining a tumor cell density of the identified tumor cells. The classification of the identified tumor cells into tumor cells of an inflammatory tumor or as tumor cells of a non-inflammatory tumor comprises

- 5 inputting the proximity measure and the tumor cell density into a classifier configured to perform the classification.

According to embodiments, the image analysis method further comprises analyzing the received digital image for determining an immune cell density of the identified immune cells or of a particular type of immune cells (e.g. T-cells, B-cells,

- 10 macrophages, etc.). The classification of the identified tumor cells into tumor cells of an inflammatory tumor or as tumor cells of a non-inflammatory tumor comprises inputting the proximity measure and the immune cell density into a classifier configured to perform the classification.

According to embodiments, the classification of the identified tumor cells into tumor cells of an inflammatory tumor or as tumor cells of a non-inflammatory tumor

- 15 comprises inputting the proximity measure, the tumor cell density and the immune cell density into a classifier configured to perform the classification.
- According some embodiments, the determination of the tumor cell density comprises determining the tumor cell density for each tumor cell individually, e.g. by
- 20 determining the tumor cell density in a predefined area centered around said tumor cell. In this case, each of the determined tumor cell density values can be fed as input into the classifier. Alternatively, an average value is computed from the individual density values and is input as a single value into the classifier. Likewise, the determination of the immune cell density comprises determining the immune cell
- 25 density for each immune cell individually, e.g. by determining the immune cell density in a predefined area centered around said immune cell. In this case, each of the determined immune cell density values can be fed as input into the classifier. Alternatively, an average value is computed from the individual density values and is input as a single value into the classifier.

Using cell density information as an additional input parameter of the classifier may be advantageous as the accuracy of the classification may significantly be increased.

According to embodiments, the displaying of the classification result comprises
5 representing all identified tumor cells as first pixel blobs having a first color and representing all identified immune cells whose distance to its nearest tumor cells is below a threshold as second pixel blobs having a second color. Then, the image analysis system displays the second pixel blobs as an overlay of the first pixel blobs.

For example, the tissue sample may be a biopsy sample of colorectal cancer, the
10 identified immune cells may be cytotoxic T cells and the distance threshold may be 35 μm . All cells having been identified as tumor cells may be colorized with blue color and may correspond to a first image layer. All cells having been identified as immune cells whose distance to its nearest tumor cell is below 35 μm are colorized with orange color and may correspond to a second image layer. The second image
15 layer is superimposed on the first image layer with the consequence that the orange “tumor affine” immune cells cover and hide at least some of the tumor cells. In this case, orange regions in the tumor tissue indicate inflammatory regions. If the tissue section represented in the generated overlay-image comprises only few orange regions, the immune response is predicted to be only weak or is predicted not to significantly affect tumor development. The patient in this case is predicted not
20 profit from the application of drugs boosting or modulating the immune response because the immune cells are too far away from the tumor cells as to be able to kill them. Thus, a user of pathologist can immediately and easily derive from such an overlay image if a particular tumor tissue is inflammatory or not and/or whether the
25 treatment of the patient with substances known to boost or modulate an immune reaction will be effective in treating this tumor or not.

According to embodiments, the tissue sample is a whole slide tissue sample and the digital image is a whole slide image. Typically, such whole slide tissue samples comprise several thousand or even 10,000s of cells. Thus, the proximity measure
30 may be a function of the distances of many thousand neighboring immune cell-tumor cell pairs.

According to embodiments, the identification of the tumor cells comprises identifying proliferating non-lymphoid cells and using said identified cells as the tumor cells.

For example, proliferating cells may be identified by using generic proliferation markers such as fluorescent stained antibodies binding to the KI67 protein or to the PCNA protein. Non-lymphoid cells are identified, for example, as cells which do not express typical immune cell markers such as CD3. Thus, any cell having the expression profile KI67+/CD3- is identified as a proliferating non-lymphoid cells and thus as a tumor cell. This approach may be beneficial as it provides a very generic way of identifying tumor cells that is applicable for many different types of tumors and for many different types of tissues.

In addition, or alternatively, the identification of the tumor cells comprises identifying cells expressing a (specific) set of one or more tumor-specific biomarkers and using said identified cells as the tumor cells.

For example, a cytokeratin profile (data indicating whether a particular set of cytokeratins is expressed in a cell or not) of the cells in the tissue sample is determined, e.g. by using fluorescence stains selectively staining the respective cytokeratins. The obtained cytokeratin profile is then compared by the image analysis system against known cytokeratin profiles of various tissues such as liver, lung, colon and so on. In case the cytokeratin profile of a cell in the tissue sample depicted in the image differs from the typical cytokeratin profile of the tissue from which the tissue sample was taken, the cells having this "deviant/untypical" cytokeratin profile are automatically or semi-automatically identified as tumor cells.

Cytokeratins are proteins of keratin-containing intermediate filaments found in the intracytoplasmic cytoskeleton of epithelial tissue. For example, basic cytokeratins CK1, CK2, CK3, CK4, CK5, CK6, CK7, CK8 and acidic cytokeratins CK9, CK10, CK12, CK13, CK14, CK16, CK17, CK18, CK19, and CK20 are known. Expression of these cytokeratins is frequently organ or tissue specific. Therefore, cytokeratin expression profiles can be used by anatomic pathologists to detect the presence and/or cell of origin of various tumors.

Depending on the type of tissue of the tissue sample and/or the type of the tumor suspected to be present in a patient, the presence (sufficiently high expression level) of other tumor-specific biomarkers may be determined, e.g. the presence of the Her2 protein for breast cancer.

- 5 According to embodiments, the identification of the immune cells comprises or consists of identifying cytotoxic T-cells and using the identified cytotoxic T-cells as the identified immune cells. For example, the CD8 protein that is specifically expressed in cytotoxic T-cells is used as a biomarker. A stain that selectively binds to the CD8 protein in the tissue sample is used and an image whose pixel intensity
- 10 values are indicative of the presence of the CD8 protein is taken from the stained tissue sample. An image analysis operation is performed on the CD8 biomarker related staining and intensity signal, e.g. a threshold-based blob detection operation, and the corresponding pixel blob is identified as CD8+ cell and thus as a cytotoxic T-cell.
- 15 According to other embodiments, CD8+ and CD3 + cells (i.e., cells expressing both the CD8 protein and the CD3 protein) are identified as cytotoxic T-cells. The CD3 protein is a protein expressed on many different immune cells and is used as a further biomarker that helps to ensure that a particular blob is really an immune cell and not, for example, an artifact created by an unspecific staining reaction of the
- 20 antibody used for staining the CD8 protein.

- In addition, or alternatively, the identification of the immune cells comprises or consists of identifying T-helper-cells and using the identified T-helper-cells as the identified immune cells. For example, CD4 expressing cells ("CD4+ cells") are identified as T-helper cells. According to other examples, cells expressing both the
- 25 CD4 and the CD3 protein ("CD4+/CD3+ cells") are identified as T-helper-cells. Again, CD3 protein expression was used as a further indicator to make sure a blob is really an immune cell, not a blob of an unspecific stain. Instead of or in addition to the CD4 marker, the T-bet marker may likewise be used for identifying T-helper cells.

In addition, or alternatively, the identification of the immune cells comprises or consists of identifying memory cells and using the identified memory cells as the identified immune cells. For example, CD45RO expressing cells ("CD45RO+ cells") are identified as memory cells.

- 5 In addition, or alternatively, the identification of the immune cells comprises or consists of identifying B- cells and using the identified B cells as the identified immune cells. For example, CD20 expressing cells ("CD20 + cells") are identified as B-cells.

- 10 In addition, or alternatively, the identification of the immune cells comprises or consists of identifying activated T- cells and using the identified activated T- cells as the identified immune cells. For example, CD137 expressing cells ("CD137+ cells") are identified as activated T-cells.

- 15 In addition, or alternatively, the identification of the immune cells comprises identifying PD1+ immune cells and using the identified PD1+ immune cells as the identified immune cells. In addition, the identification of the tumor cells comprises identifying PDL1+ tumor cells and using the identified PDL1+ tumor cells as the identified tumor cells. A PDL1+ tumor cell is a tumor cell that expresses the PDL1 protein. For example, as mentioned before, a tumor cell can be identified e.g. as a cell expressing the KI67 protein and/or expressing the PCNA protein and not
20 expressing a CD3 protein. Thus, a PDL1+ tumor cell would be a cell having the expression profile PDL1+/KI67+/CD3-. Alternatively, any other biomarker indicating that a cell is a tumor cell (e.g. panCK+ or a particular cytokine profile that deviates from the profile of the surrounding tissue) can be used for identifying tumor cells.

- 25 A cell that "expresses" a biomarker BM ("BM+ cell") as used herein refers to a cell expressing said biomarker at least in a minimum amount. For example, the question if said minimum amount is generated can be determined by comparing an intensity value on a respective digital image with an intensity threshold value. Said threshold value may be specific for the biomarker or stain used. Depending on the type of biomarker examined, a cell that "expresses" a biomarker can also be a cell that
30 expresses a greater amount of said biomarker than a "typical" cell of said cell type

or than another reference cell type. In this case, the label "BM+ cell" refers to a cell that over-expresses the biomarker BM (relative to a reference expression level).

Programmed death-ligand 1 (PDL1) is a 40kDa type 1 transmembrane protein that plays a major role in suppressing the immune system during particular events such as pregnancy, tissue allografts, autoimmune disease and other disease states such as hepatitis. Normally the immune system reacts to foreign antigens by triggering proliferation of antigen-specific CD8+ T cells.

Programmed cell death protein 1 (PD1) is a cell surface receptor that belongs to the immunoglobulin superfamily and is expressed on T cells and pro-B cells. PD1 binds two ligands, PDL1 and PD-L2. PD1 is an immune checkpoint and plays an important role in down regulating the immune system by preventing the activation of T-cells. PD1 is expressed on the surface of activated T cells, B cells, and macrophages.

The formation of PD1 / PDL1 ligand complex transmits an inhibitory signal on the immune response. It appears that some cancers upregulate PDL1 expression, thereby managing to evade the host immune system. High expression of PDL1 by tumor cells has been observed to be associated with increased tumor aggressiveness and an increased risk of death.

By computing the proximity measure in by using the above mentioned PDL1/PD1 biomarkers, it is possible to determine if the immune cells lying within an immunologically effective distance from the tumor cells express PD1 receptor protein or not and to determine if the tumor cells (over)express the PDL1 protein.

In addition, or alternatively, the identification of the immune cells comprises or consists of identifying macrophages and using the identified macrophages as the identified immune cells. For example, cells expressing both the CD163 and the CD68 protein ("CD163+/CD68+ cells") are identified as macrophages. Instead of or in addition to the CD163 marker, the CSF1R marker may likewise be used for identifying macrophages.

It has been observed that automated full slide analysis of a cohort of slides stained with an immunofluorescence assay targeting the biomarkers CD3, CD4, CD8 and KI67 reveals that metrics of CD3+/CD8+ and CD3+/CD4+ immune cell distribution and proximity to KI67+ proliferating non-lymphoid cells (CD3-/KI67+) in the tumor area can serve as a strong indicator for immune response in the tumor area. By computing a proximity score that is based on individual tumor cell – immune cell distances, a particularly accurate tumor classification method is provided that is more accurate than state of the art MSS/MSI state based or immunoscore based classification schemes. Moreover, the proposed computation of a cell distance based proximity score and the use of the score for tumor classification has been observed in colorectal cancers to have a higher prognostic value than the AJCC/UICC TNM-classification.

Cytotoxic T-cells, memory cells, T-helper cells and macrophages are immune cells which are responsible or are able to initiate and/or boost an immune response. To the contrary, regulatory T-cells are often involved in the downregulation of the immune response. Thus, by selectively identifying those immune cells having a positive or stimulatory effect on the immune response, and by selectively identifying the distances of said positively stimulating immune cells, an even higher accuracy of the proximity score derived from said distances can be achieved: the presence of regulatory T-cells potentially having a negative impact on the immune response are ignored, thereby preventing the computation of a proximity score that indicates that many immune cells are in close proximity to tumor cells in a situation when predominantly regulatory T-cells but not macrophages, cytotoxic T-cells and/or T-helper cells are in close proximity to individual tumor cells.

According to embodiments, the identification of the immune cells comprises or consists of identifying immune cell of one or more immune cell types that boost an immune reaction (e.g. macrophages, cytotoxic T-cells, memory cells, B cells, and T-helper cells) and ignoring ("filtering out") immune cells of an immune cell type that suppresses or downregulates an immune reaction (e.g. regulatory T-cells). Thus, the totality of identified immune cells used for computing the proximity measure may

consist of immune cells of the same or of multiple different immune cell types. In this case, for computing the proximity measure as a function (e.g. a ratio or a histogram slope) of the distances, immune cell type specific functions (e.g. with immune cell type specific distance thresholds corresponding to immune cell type specific

5 immunologically effective distances of said immune cell type) are used. For example, the immune cell ratios may be computed for each immune cell type individually and immune cell type specific ratios may be computed for immune cell type specific distance thresholds being identical to the immunologically effective distance of said immune cell type. Then, an average, e.g. the arithmetic mean of the
10 immune cell type specific ratios may be computed as a final ratio and used as the proximity measure, for example. In case the immunologically effective distance of the different immune response boosting immune cell types is identical or very similar, the same function, e.g. the same distance threshold for ratio computation may be used for multiple or all immune cell types considered.

15 Computing a single proximity measure from multiple different immune cell types (using the same or immune cell type specific function of the tumor cell – immune cell distances) may be advantageous as a better and more complete data basis of the spatial relationship between relevant immune response promoting immune cells and their potential targets, the tumor cells, is provided and a more accurate classification
20 into inflammatory and non-inflammatory tumor cells may be provided.

In addition, or alternatively, multiple individual proximity measures may be computed for multiple respective immune cell types and/or tumor cell types, e.g. for determining if a tumor of a tissue sample is a PDL1+ tumor and if a sufficient portion of the tumor cells is within an immunologically effective distance of a PD1+ immune
25 cell. Said additional proximity measures may allow for a further sub-classification of the tumor and for a better prediction of the treatment schema.

According to embodiments, the identification of the immune cells comprises identifying regulatory T-cells and identifying immune cells of at least one type of immune cells that boosts the immune response and selectively using the identified
30 immune cells that boost the immune response but not the identified regulatory T-cells as the identified immune cells. For example, cells expressing the FoxP3

protein ("FoxP3+ cells") are identified as regulatory T-cells and are ignored when the tumor cell – immune cell distances are determined. According to other embodiments, FoxP3+/CD3 + cells are identified as regulatory T-cells, whereby the expression of the CD3 biomarker is used as a further indicator to make sure a blob is really an immune cell, not a blob of an unspecific stain.

According to embodiments, the proximity measure is determined multiple times as described herein for embodiments of the invention, thereby respectively analyzing different digital images whose pixel intensities represent different biomarkers, the different biomarkers being indicative of immune cells of different types (e.g.: CD8 for cytotoxic T-cells, FoxP3 for regulatory T-cells, CD163 or CSF1R for macrophages, CD20 for B-cells, CD137 for activated T-cells, PD1 for some sub-types of T- and pro-B cells) and/or being indicative of tumor cells of different types (e.g. PDL1+/PDL1- tumor cells). Each of said multiple proximity measures may be used as a predictor of tumor cell class and/or prognosis and/or of a suitable treatment scheme.

Computing the proximity measure multiple times for different types of immune cell types and/or cancer cell types may allow to more accurately predict which ones of different classes of immune response modulating drugs will be effective in treating the tumor and which ones will not. By combining information contained in the spatial proximity of different immune cell and tumor cell types, the accuracy of classifying the tumor into inflammatory or non-inflammatory tumors may be increased and a classification into further sub-classes of tumors (e.g. PDL1+ or PDL1- tumors) may be enabled. In addition, the accuracy of a prediction if a particular drug class will be effective in treating the tumor may increase.

When computing multiple proximity measures for different immune cell sub-types and/or different tumor cell sub-types, also immune cells having an inhibitory effect on the immune system are taken into account. Although their presence within an immunologically effective distance from tumor cells may not indicate that the tumor is an inflammatory tumor, information on their spatial proximity to tumor cells may allow predicting whether applying a drug suppressing said inhibitory immune cell types may boost the immune response against the tumor.

According to embodiment, the image analysis system automatically outputs a prediction whether or not a particular class of drugs, will be effective in treating the cancer or not.

For example, one of said classes of drugs can be PD1 inhibitors. PD1 inhibitors
5 activate the immune system to attack tumors and are used with varying success to treat some types of cancer by blocking or otherwise inhibiting the activity of the PD1 receptor. According to embodiment, the image analysis system is configured to predict that the PDL1 inhibitors will be effective only in case the proximity measure indicates that a sufficiently large fraction of PDL1+ tumor cells comprises at least
10 one PD1+ immune cell within the predefined distance (immunologically effective distance). If the immune cells are not within an immunologically effective distance from the tumor cells or if the immune cells are located within said distance next a PDL1+ tumor cell but are not expressing PD1, the image analysis system is configured to predict that an anti-PDL1 drug will not to be effective in treating the
15 tumor.

Likewise, other drug classes exist which boost or modulate an immune response and which specifically modulate the activity of individual immune cell types such as macrophages, T helper cells, B-cells, cytotoxic T-cells (T-cell engaging agents) and others. It may also be possible that said immune response modulating or boosting
20 drugs are only effective in respect to specific tumor types. By automatically computing the proximity measure for multiple different immune cell types and/or tumor cell types, in a single step quantitative location information (cell-cell distance smaller than immunologically effective distance or not) as well as qualitative information on the cell classes is used as a basis for predicting whether or not a
25 particular drug class will be effective in treating the tumor or not.

According to embodiments of the invention wherein the distances are determined for each of the identified tumor cells, the computation of the proximity measure as a function of the determined distances comprises:

- 30 – identifying a first and a second sub-set of the identified tumor cells; the first sub-set ("tumor cells with a near immune cell") selectively comprising tumor cells whose nearest immune cell is less than a predefined distance away; the second

sub-set (“tumor cell with far immune cells”) selectively comprising tumor cells whose nearest immune cell is at least the predefined distance away from the tumor cell; for example, the predefined distance can be the “immunologically effective distance” of the identified immune cells;

- 5 – computing a ratio of the number of tumor cells contained in the first subset and the number of tumor cells in the second sub-set; and
- using the ratio as the proximity measure, wherein the higher the ratio, the higher the probability that the classification result indicates that the identified tumor cells belong to the inflammatory tumor.

10

According to alternative embodiments of the invention wherein the distances are determined for each of the identified immune cells, the computation of the proximity measure as a function of the determined distances comprises:

- identifying a first and a second sub-set of the identified immune cells, the first
15 sub-set selectively comprising immune cells whose nearest tumor cell is less than a predefined distance away, the second sub-set selectively comprising immune cells whose nearest tumor cell is at least the predefined distance away from the tumor cell;
- computing a ratio of the number of immune cells contained in the first subset and
20 the number of immune cells in the second sub-set; and
- using the ratio as the proximity measure, wherein the higher the ratio, the higher the probability that the classification result indicates that the identified tumor cells belong to the inflammatory tumor.

- 25 According to embodiments, the predefined distance is the “immunologically effective distance of the immune cell. The immunologically effective distance is a maximum distance within which the identified immune cell is able to directly or indirectly trigger the killing of or the performing of apoptosis by the tumor cell.

According to embodiments, the predefined distance is larger than 15 μm .

30

According to embodiments, the predefined distance is in a range of 20 μm to 50 μm , e.g. 35 μm .

For example, the tumor type may be colorectal cancer, the distance may be 35 μ m. For other tumor types, the distance may slightly vary, but typically lies in the range of 20 μ m to 50 μ m.

- 5 According to embodiments the classification result indicates that the identified tumor cells belong to the inflammatory tumor in case the ratio exceeds a predefined percentage or threshold, typically 50%. Thus, according to one embodiment, the classification result indicates that the identified tumor cells belong to the inflammatory tumor in case the number of tumor cells contained in the first subset is
- 10 larger than the number of tumor cells in the second sub-set (and thus, the fraction of the tumor cells in the first subset exceeds 50% of all tumor cells).

According to embodiments related to the determination of the distances for each of the tumor cells, the computation of the proximity measure as a function of the

15 determined distances comprises:

- generating a histogram of the distances of the tumor cells to their respective nearest one of the identified immune cells; the histogram comprises at least two distance bins; the histogram covers a distance range of 0 μ m to at least 50 μ m (typically up to 80 μ m), each of the bins corresponding to a bar of the histogram,
- 20 each of the bars indicating a count of the identified tumor cells having a distance to their nearest immune cell that falls into said bin;
- connecting the upper end of the first one of the bars with the upper end of the last one of the bars with a line; the line may be straight or curved; for example, the line may be a balance line; the first bar corresponds to the one of the bins
- 25 covering the smallest distances; the last bar corresponds to the one of the bins covering the largest distances of the distance range;
- determining the slope of the line; for example, the first bar may be located at the left boundary of the histogram, the last bar may be located at the right boundary of the histogram, and the slope may be determined of the line connecting the left-
- 30 side, first bar with the right-side last bar;
- using the slope as the proximity measure, wherein in case the slope indicates that the tumor cell count of the first bar is higher than the tumor cell count of the

last bar (i.e., is negative if the first bar is plotted to on the left side of the histogram), the classification result is that the identified tumor cells belong to an inflammatory tumor.

- 5 According to embodiments related to the determination of the distances for each of the immune cells, the computation of the proximity measure as a function of the determined distances comprises:
- generating a histogram of the distances of the immune cells to their respective nearest one of the identified tumor cells, the histogram comprising at least two
 - 10 distance bins, the histogram covering a distance range of 0 μm to at least 50 μm , each of the bins corresponding to a bar of the histogram, each of the bars indicating a count of the identified immune cells having a distance to their nearest tumor cell that falls into said bin;
 - connecting the upper end of the first one of the bars with the upper end of the last
 - 15 one of the bars with a line, the first bar corresponding to the one of the bins covering the smallest distances, the last bar corresponding to the one of the bins covering the largest distances of the distance range;
 - determining the slope of the line;
 - using the slope as the proximity measure, wherein in case the slope indicates
 - 20 that the immune cell count of the first bar is higher than the immune cell count of the last bar, the classification result is that the identified tumor cells belong to an inflammatory tumor.

Computing a histogram slope as the proximity measure may have the advantage

25 that the histogram that needs to be computed for determining the slope can be used as graphical output that visually represents the relative immune cell – tumor cell distribution in a tissue sample. The histogram may be output via a display device, thereby providing a user, e.g. a pathologist, with valuable information on the distance distribution of immune and tumor cells.

30

According to embodiments, the image analysis system outputs a signal selectively in case the classification result indicates that the identified tumor cells belong to an

inflammatory tumor, wherein the signal is indicative of a treatment recommendation to use a substance that boosts or modulates an immune response as a drug for treating the tumor.

- 5 According to embodiments, the method comprises using the proximity measure in addition to or in replacement of an MSS-MSI status indicator for computing a prognosis of the responsiveness of the tumor to a substance that boosts or modulates an immune response.
- 10 According to embodiments, the identified tumor cells are colorectal cancer cells. However, the method and system according to embodiments of the invention are also applicable to other tumor types, e.g. breast cancer, pancreatic cancer, lung cancer, and the like.
- 15 According to embodiments, the identification of the immune cells comprises identifying immune cells of multiple different immune cell types by analyzing pixel intensity values in the at least one digital image, whereby the pixel intensities represent the presence of different immune cell type specific biomarkers, and using the totality of the identified immune cells of the multiple different immune cell types
- 20 as the identified immune cells for computing the proximity measure.

According to embodiments, the tumor cells are identified as cells expressing at least a first tumor-type specific biomarker. The immune cells being identified as cells expressing at least a first immune cell type specific biomarker. The method further

- 25 comprises:
- analyzing the at least one received image or a further image of the tissue sample for identifying further immune cells as cells expressing a second immune cell type specific biomarker; and
 - analyzing the at least one received image or the further image of the tissue
 - 30 sample for identifying further tumor cells as cells expressing a second tumor type specific biomarker or using the identified immune cells as further tumor cells;
 - for each of the identified further tumor cells, determining the distance of said further tumor cell to the nearest further immune cell; alternatively, for each of the

identified further immune cells, the distance of said further immune cell to the nearest further tumor cell is determined;

- computing a further proximity measure as a function of said determined distances;
- 5 – in dependence on the further proximity measure, sub-classifying the identified tumor cells and/or predicting the treatability of the tumor by a class of substances.

In a further aspect, the invention relates to an image analysis system for tumor classification, the system being configured for:

- 10 – receiving at least one digital image of a tissue sample;
- analyzing the at least one received image for identifying immune cells and tumor cells in the at least one received image;
- for each of the identified tumor cells, determining the distance of said tumor cell to the nearest immune cell;
- 15 – computing a proximity measure as a function of the determined distances;
- in dependence on the proximity measure, classifying the identified tumor cells into tumor cells of an inflammatory tumor or as tumor cells of a non-inflammatory tumor; and
- storing the classification result on a storage medium and/or displaying the
- 20 classification result on a display device.

In a further aspect, the invention relates to a computer readable medium comprising instructions that when executed by a processor causes the processor to execute a method according to any one of the embodiments described herein.

25

In a further aspect, the invention relates to an image analysis system for tumor classification, the system being configured for:

- receiving at least one digital image of a tissue sample;
- analyzing the at least one received image for identifying tumor cells in the at
- 30 least one received image;
- analyzing the identified tumor cells for identifying tumor regions, wherein a tumor region is a tissue region whose majority of cells consist of tumor cells;

- analyzing the at least one received image for identifying immune cells lying within one of the identified tumor regions or lying within a boundary region around one of the tumor regions, the width of the boundary being smaller than 200 μ m; preferentially, the width of the boundary region is between 120 μ m and 200 μ m; thus, the boundary region may have the shape of a belt surrounding a tumor region, the belt width being the boundary region width.
- for each of the identified immune cells, determining the distance of said immune cell to the nearest tumor cell;
- computing a proximity measure as a function of the determined distances;
- in dependence on the proximity measure, classifying the identified tumor cells into tumor cells of an inflammatory tumor or as tumor cells of a non-inflammatory tumor; and
- storing the classification result on a storage medium and/or displaying the classification result on a display device.

Thus, the width of the boundary region around the tumor regions defines the area (tumor region(s) and tumor boundary region(s)) which shall be used for counting individual immune cell - tumor cell distances. Selectively identifying immune cells within the exterior outline of a boundary region around an identified tumor region may allow to ensure that the accuracy of the proximity measure is not negatively affected by including the immune cells that are normally distributed in tissue regions far away from a tumor which would never engage in an immune response against the tumor anyway. In addition, this may ensure that the accuracy of the proximity measure is not negatively affected by a too narrow boundary that would exclude immune cells in close proximity of the tumor whose position is key for predicting the inflammatory status of the tumor.

According to some embodiments, the width of the boundary region is determined for a particular cancer type empirically. The determination comprises: choosing an array of adjacent regions (e.g. circles, squares or rectangles having a predefined diameter, width and/or height e.g. in the range of 20 μ m to 40 μ m) in the image, the array of adjacent regions starting at a pixel on the tumor region outline ("starting

region”) and extending at least 200µm radially away from the tumor center into the non-tumor tissue region (the last one of the adjacent regions being referred to as “terminal region”; counting the number of immune cells in each of said adjacent regions; scanning the immune cell counts of the adjacent regions from the starting

5 region in direction of the terminal region, thereby identifying the first one of the regions lying within a range of 120µm to 200µm away from the boundary of the tumor region and basically having the same immune cell count as all subsequently scanned regions. This region is considered to comprise a number of immune cells that is typical for the “normal” non-tumor tissue far away from any tumor tissue; and
10 using the position of the identified region (e.g. a distance of any one of the pixel within said identified region from the pixel on the tumor region outline covered by the “starting region”) as the boundary region width.

A “tumor” or “neoplasm” as used herein is an abnormal growth of tissue, and when
15 also forming a mass is commonly referred to as a tumor. This abnormal growth usually but not always forms a mass.

An “inflammatory tumor” as used herein is a tumor against which a localized protective response by the immune system is observable. Said response may
20 comprise the infiltration of plasma cells, lymphocytes, and eosinophils in the tumor.

A “non-inflammatory tumor” as used herein is a tumor where no or no significant response of the immune system against said tumor is observable.

25 An “immune cell” as used herein is a type of cell formed in the myelopoietic, lymphoid, and reticular portions of the reticuloendothelial system in various parts of the body, and normally present in those sites and in the circulating blood (rarely in other tissues). In particular, the immune cell can be a lymphocyte, e.g. T lymphocytes, B lymphocytes, and macrophages.

30

A proximity measure is a data value computed from a plurality of immune cell- tumor cell distances. For example, it can be a numerical data value, e.g. a ratio (a value

between 0 and 1) of different groups of distances or a slope of a line connecting histogram bins or a data value derived therefrom. The proximity measure may indicate whether the majority of tumor cell – immune cell distances examined is shorter than a predefined distance threshold, e.g. the immunologically effective distance of an immune cell.

An “image analysis system” as used herein is an electronic system, e.g. a computer, configured for extracting meaningful information from digital images by means of digital image processing techniques. Image analysis tasks can comprise color deconvolution, connected component analysis and/or edge detection for identifying cells, for identifying the type of the cells (tumor or stroma cell, different types of immune cells) and distance measurements.

Brief description of the drawings

In the following embodiments of the invention are explained in greater detail, by way of example only, making reference to the drawings in which:

Detailed description

Figure 1 is a block diagram of an image analysis system;
Figure 2 is a flow chart of an image analysis method for classifying tumors;
Figure 3 depicts three overlay images derived from respective MSI tumor samples, the image comprising a first layer with tumor cells and a second layer with differently colored immune cells lying within a maximum distance from the nearest tumor cell;
Figure 4 depicts 8 overlay images derived from respective MSS tumor samples;
Figure 5 depicts three histograms derived from the three MSI tissue samples depicted in Fig. 3;
Figure 6 depicts 8 histograms derived from the 8 MSS tissue samples depicted in Fig. 4;
Figure 7 depicts three IHC fluorescent images of the three MSI tissue samples depicted in Fig. 3;

Figure 8 depicts 8 fluorescent images of the 8 MSS tissue samples depicted in Fig. 4;

Figure 9 depicts a comparison of the overlay image, the IHC fluorescent image and the histogram generated for a MSI sample (left column) and a
5 MSS sample (right column); and

Figure 10 depicts a plot indicating the probability density that a particular immune cell lies within a given distance from its nearest tumor cell.

Figure 11 depicts the combination of proximity information and cell density information for capturing tumor context information for classifying the
10 tumor.

Figure 12 depicts stained tissue samples of two patients having similar T-cell densities but different proximity measures.

Figure 1 is a block diagram of an image analysis system 100 according to an
15 embodiment of the invention. The system comprises one or more processors 104, a main memory 106 and a non-volatile storage medium 108. The storage medium comprises one or more application programs or modules 110, 114, 112, 116 configured for performing one or more image processing tasks. For example, a first module 110 may perform a connected component analysis and edge detection
20 routines in order to identify pixel blobs representing cells. The identification may be performed on different monochromatic fluorescent or brightfield microscopy images having been derived from the same tissue sample whereby the pixel intensities of the different monochromatic images respectively are indicative of a particular biomarker. Multiple monochromatic fluorescent images can be derived from a multi-
25 spectral fluorescent image of a particular tissue sample by applying a color deconvolution algorithm.

The storage medium 108 may further comprise one or more digital images 118 of a tissue sample having been stained with one or more biomarker specific stains.

30 Moreover, the system 100 is coupled to or comprises a display 102, e.g. an LCD display. The system uses the display 102 for displaying the digital images 118 of tissue samples of various patients (see for example figures 7 and 8), for displaying

overlay images (as shown for example in figures 3 and 4) having been derived from various tissue images 118, for displaying a graphical representation of proximity measures (for example histograms depicted in figures 5 and 6) and/or for displaying a tumor classification result or treatment suggestion.

5

For example, a tissue sample, e.g. a biopsy sample, is used that has been stained by a plurality of biomarker specific stains. One or more of that stains selectively bind to strongly proliferating non-lymphoid cells or to proteins selectively expressed by tumor cells. Intensity signals of said stains and the corresponding monochromatic
10 images can be analyzed by module 110 for identifying tumor cells. One or more further ones of said stains selectively bind to immune cells or specific subtypes of immune cells, e.g. cytotoxic t cells or T-helper cells. Intensity signals of said stains and the corresponding monochromatic images can be analyzed by module 110 for identifying immune cells. After having executed application program or module 110,
15 tumor cells and immune cells within a tissue sample and the corresponding digital images have been identified.

Then, a distance measurements application program or module 112 uses the location of the pixel blobs identified as immune cells and tumor cells as input for
20 determining the distance of tumor cells and their closest neighboring immune cells. For example, module 112 identifies the center of each cell and determines the distance between the respective cell centers. Alternatively, module 112 determines the distance as the distance between the outer boundaries of the two neighboring cells.

25

The totality of distances measured by module 112 is provided as input to the application program or module 114 which computes a proximity measure as a function of the totality of distances provided by module 112 for a particular tissue sample. The module 114 computes, for example, a ratio between tumor cells for
30 which at least one immune cell within a predefined distance threshold was identified and tumor cells for which no immune cells within that predefined distance threshold was identified. In addition, or alternatively, module 114 may compute a histogram for

all identified distances between any of the tumor cells and its nearest immune cell neighbor and may compute a slope of the histogram bins as illustrated by figures 5, 6 and 9.

5 In addition, the totality of distances measured by module 112 is provided as input to the application program or module 116 which generates an overlay image as depicted, for example, in figures 3 and 4. In an overlay image tumor cells are represented in a base image layer in a first color and immune cells having been identified as being the next neighbor to at least one of the tumor cells and having in
10 addition been identified as lying within a distance from said tumor cell that is below a distance threshold (e.g. a predefined "immunologically effective distance") are represented in a further image layer in a second color. The further layer is presented on top of the base layer. The images 118, the colorized overlay image is generated by module 116 and/or the histograms are presented to a user, e.g. a pathologist, via
15 display 102.

Figure 2 is a flow chart of an image analysis method for classifying tumors. The methods depicted in figure 2 can be implemented, for example, by an image analysis system 100 as depicted in figure 1.

20

In a first step 202, the image analysis system receives a digital image 118 of a tissue sample. The digital image can be a multispectral fluorescent immunohistochemistry (IHC) image that is decomposed by the image analysis system into a plurality of monochromatic images by applying a color deconvolution
25 algorithm. Alternatively, the image analysis system may receive a plurality of monochromatic digital images of the tissue sample. The multispectral digital image and/or the plurality of monochromatic digital images of the tissue sample can be provided via an interface with an image acquisition system. Alternatively, the image analysis system 100 may receive the digital images of the tissue sample by reading
30 the images from a storage medium, e.g. a CD-ROM or flash drive.

For example, the tissue sample from which the images were derived can be a biopsy of a colorectal cancer tissue sample having been stained by a plurality of fluorescent stains selectively binding the following biomarkers: KI67, CD3 and CD8.

Pixel regions for which a KI67 signal but no CD3 or CD8 signal was obtained are

5 identified as proliferating non-lymphoid cells and thus are identified as tumor cells.

Pixel regions for which a CD3 and a CD8 signal was obtained are identified as cytotoxic T-cells and thus are identified as immune cells. A "KI67 signal" as used

herein is a light signal emitted by a stain that selectively stains the KI67 protein, whereby said light signal is represented in the form of pixel intensity values in a

10 respective monochromatic image that selectively captures the emission spectrum of the stain used for selectively staining the KI67 protein. This definition applies analogously also to the other "biomarker signals".

For example, digital images of IHC-stained glass slides can be acquired using a

15 Ventana iScan HT Slide Scanner. Images were viewed and organized using the Roche IRIS Platform. Ventana image analysis software VDP-SKD and the Ventana Digital Pathology Software Development Kit was used for performing most of the

image analysis methods for identifying the cell boundaries and for identifying separate tissue and glass regions, tumor cells and immune cells and tumor cell

20 regions mainly (e.g. to more than 50%) consisting of tumor cells.

In a second step 204, module 110 of the image analysis system performs one or

more image analysis routines for identifying immune cells and tumor cells in the

received digital image or images of the tissue sample. Image analysis routines for

25 detecting cells, e.g. via a connected component analysis, gray scale and color

segmentation techniques, intensity thresholding and the like are known in the art.

In a further step 206, module 112 of the image analysis system determines, for each

of the identified tumor cells, the distance of the tumor cell to the nearest immune

30 cell, e.g. by measuring the distance between cell boundaries.

After step 206 has completed, module 114 computes a proximity measure as a function of the determined distances in step 208. For example, the proximity measure can be a ratio of tumor cells having at least one “near immune cell” and of tumor cells not having such a “near immune cell”, whereby a “near immune cell” is a
5 immune cell lying within a predefined (e.g. immunologically effective) maximum distance from a tumor cell. In other words, this ratio describes the ratio of tumor cells that can potentially be attacked by a nearby immune cell and those tumor cells which appear not to be attackable by an immune cell. In addition, or alternatively a slope of two or more bins of a distance histogram can be computed, whereby a
10 negative slope (from short distance bins to long distance bins) indicates that the majority of tumor cells can potentially be attacked by a nearby immune cell and whereby a positive slope (from short distance bins to long distance bins) indicates that the majority of tumor cells can probably not be attacked successfully by an immune cell.

15 In step 210, module 114 classifies the identified tumor cells into tumor cells of an inflammatory tumor or as tumor cells of a non-inflammatory tumor in dependence on the proximity measure. For example, if the ratio is larger than 1 and/or if the slope is negative, the tumor cells are classified as inflammatory tumor cells. If the ratio is smaller than 1 and/or if the slope is positive, the tumor cells are classified as non-
20 inflammatory tumor cells.

In step 212, the classification result is stored on a storage medium 108. In addition, or alternatively, the classification result is displayed on a display device 102. For example, an overlay image is depicted in figures 3 and 4 can be used as graphical
25 representation of the classification result: if the majority of tumor cells are covered by an immune cell being within the “immunologically effective distance”, this is an indication that the tumor cell is depicted in the overlay image is a tumor cell of an inflammatory tumor.

30 **Figures 3 and 4** depict overlay images generated for a group of colorectal cancer (CRC) patients. Using an immunofluorescence assay (CD3/CD4/CD8/Ki67/DAPI), surgical samples obtained from primary tumors of colorectal cancer (CRC) patients

(n=23) was stained. The MSS/MSI status of said patients was known. KI67+/CD3- cells were identified as tumor cells. CD3+/CD8+ cells were identified as immune cells.

- 5 **Figure 3** depicts three overlay images 302-306 computed by the image analysis system for three of the CRC patients whose microsatellite stability status was identified as “instable” (“MSI”).

Each of the images comprises a first layer wherein all tumor cells are colorized with
10 a first color, e.g. blue, irrespective of the presence or location of any immune cell. In each of said images 302-306, said first layer is overlaid with (and covered by) a second layer with pixel blobs representing immune cells having been identified as lying within a predefined maximum distance from at least one tumor cell. For example, these “potentially immunologically effective” immune cells are colorized in
15 orange. It is important to note that the second layer does not indicate the presence of immune cells not lying within the predefined maximum distance. Thus, the second layer is not simply an overlay of layers representing immune cells and tumor cells in different colors, but rather is a computational result being indicative of potentially immunologically effective immune cells within the tumor. It has been observed that
20 this particular form of overlay image provides a better and more accurate indication regarding the inflammatory state of a tumor than a mere superposition of a generic tumor cell image layer and a generic immune cell image layer.

As can be seen from figure 3, the majority of tumor cells in each of the overlay
25 images 302, 304, 306 is covered by “potentially immunologically effective immune cells”. The most complete coverage is observed in image 306. Thus, the tumor cells of the tissue samples corresponding to overlay images 302-306 are classified as cells of an inflammatory tumor. Thus, for those three patients, the status “MSI” is “in line” with the distance-based classification result.

30

Figure 4 depicts 8 overlay images derived for 8 other patients whose microsatellite stability status is “stable” (MSS). The overlay images were generated as already

described for figure 3. The overlay images 402, 406-416 clearly show that the majority of tumor cells is not covered by pixels representing a “potentially immunologically effective” immune cell, i.e., an immune cell lying within a predefined maximum distance from at least one tumor cell. Thus, the tumor cells of the tissue samples corresponding to overlay images 402, 406-416 are classified as cells of a non-inflammatory tumor. For the corresponding seven patients, the status “MSS” is “in-line” with the distance-based classification result.

However, in the overlay image 404 the tumor cells are almost completely covered by “potentially immunologically effective” immune cells. Thus, the image analysis system according to embodiments of the invention classifies the tumor cells of the tissue sample depicted in overlay image 404 as tumor cells of an inflammatory tumor. This is in contradiction to the known genetic MSS status of the corresponding patient. An in-depth analysis by a pathologist revealed that indeed the tumor depicted in overlay image 404 was an inflammatory tumor. If the MSS/MSI status would be used as a basis for a treatment decision, the corresponding patient would not have been prescribed a drug that stimulates or modulates an immune response. However, using a proximity measure as described herein for embodiments of the invention as a basis for a treatment suggestion, the patient corresponding to overlay image 404 is identified by the image analysis system as a patient benefiting from the subscription of a drug that stimulates the immune system.

Figure 5 depicts three histograms 502, 504, 506 derived from the same three MSI tissue samples whose computed overlay images 302, 304, 306 are depicted in Fig. 3.

Each histogram depicted in figures 5, 6 and 9 comprises four distance bins. The first bin covers a distance of 0 μm to 20 μm , the second bin covers a distance from 21 μm to 40 μm , the third bin covers a distance of 41 μm to 60 μm and the last bin covers a distance from 61 μm to 80 μm . The term “distance” relates to the distance of anyone of the identified tumor cells to its nearest identified immune cell. The distance refers to the “real” distance between respective cells in the tissue sample.

The image analysis system uses information on the resolution and/or zoom factor of the image as input for computing the cell-cell distance in μm from the numbers of pixels separating the boundaries of two cells within one image. For each bin, a corresponding bar is plotted in the histogram, whereby the height of each bar reflects the number of tumor cells whose distance to its nearest immune cell falls within said bin.

As can be inferred from the histograms 502, 504, 506, the majority of tumor cells is contained in the first and second bin representing distances within the immunologically effective distance of immune cells. By plotting a straight or curved line that connects the top of the bar of the first bin with the top of the bar of the last bin and by determining the slope of that line, a negative (falling) slope will be identified. A negative slope in the distance histogram is used as a proximity measure that indicates that the tumor cells of the tumor belong to an inflammatory tumor. This proximity measure based classification result is in line with the microsatellite instability status information "MSI".

Figure 6 depicts 8 histograms derived from the 8 MSS tissue samples whose overlay images are depicted in Fig. 4. Thus, the histograms 602-616 correspond to the 8 patients whose microsatellite stability status is "stable" (MSS). The histograms were generated as already described for figure 5. The histograms, 602, 606-616 all have a positive slope, thereby clearly showing that the majority of tumor cells does not comprise a neighboring immune cell within the predefined maximum distance that reflects the "immunologically effective" immune cell distance. Thus, the tumor cells of the tissue samples corresponding to histograms 602, 606-616 are classified as tumor cells of a non-inflammatory tumor. For the corresponding seven patients, the status "MSS" is "in-line" with the slope-based classification result.

However, in histogram 604 (corresponding to the same patient as overlay image 404), a negative slope is observed and the image analysis system classifies the tumor cells of the tumor corresponding to the distance histogram 604 as an inflammatory tumor. This is in contradiction to the known genetic MSS status of the

corresponding patient. As was explained above, the in-depth analysis revealed that indeed the tumor corresponding to histogram 604 was an inflammatory tumor.

In other words, the proximity measure based tumor classification, e.g. the tumor cell ratio based classification illustrated in figures 3 and four as well as the histogram slope based classification illustrated in figures 5 and six a more accurate than state-of-the-art MSS/MSI status based tumor classification.

Figure 7 depicts three IHC fluorescent images 702, 704, 806 of the MSI tissue samples of the three patients for which respective overlay images were generated as depicted in Fig. 3 and for which respective distance histograms were generated as depicted in Fig. 5.

Figure 8 depicts 8 IHC fluorescent images 802-816 of the MSI tissue samples of the 8 patients for which respective overlay images were generated as depicted in Fig. 4 and for which respective distance histograms were generated as depicted in Fig. 6. IHC image 804 shows the inflammatory tumor whose IHC image looks more similar to the IHC images 702-706 than to IHC images 802, 806-816.

Figure 9 depicts a comparison of two groups of images, wherein the group depicted on the left corresponds to an inflammatory CRC tumor of a first patient and the group to the right corresponds to a non-inflammatory CRC tumor of a second patient. Tumor cells were identified as KI67+/CD3- cells. Immune cells (here: cytotoxic T-cells) were identified as CD8+/CD3+ cells. The first (topmost) row depicts overlay images computed in the same way as described for figures 3 and 4. The second row depicts the IHC fluorescent images of the respective tissue samples and the third row depicts the respectively computed distance histograms generated as described for figures 5 and 6.

For inflammatory tumor tissue (left column), the majority of tumor cells are covered in the overlay image 902 by immune cells having been identified to lie within a predefined maximum distance (immunologically effective distance) from a tumor cell. A corresponding field of view (FOV) 902 from the corresponding tissue below shows a significant degree of tumor immune infiltration. The slope of the

corresponding distance histogram determined as described for figures 5 and 6 is negative.

For non-inflammatory tumor tissue (right column), the majority of tumor cells are not covered in the overlay image 904 by immune cells having been identified to lie within a predefined maximum distance (immunologically effective distance) from a tumor cell. A corresponding field of view (FOV) 908 from the corresponding tissue below shows that there is almost no tumor infiltration by immune cells will. The slope of the corresponding distance histogram determined is positive.

The difference in distribution of cytotoxic t cells engaging cancer cells between different tumor categories as expressed graphically in the overlay images 902, 904 and as expressed in the slope of the histograms 910, 912 is apparent.

Figure 10 depicts a plot indicating the probability density that a particular immune cell lies within a given distance from its nearest tumor cell.

A probability density function (PDF) is a function that describes the relative likelihood for a variable to take on a given value. The probability of the variable falling within a particular range of values is given by the integral of this variable's density over that range—that is, it is given by the area under the density function but above the horizontal axis and between the lowest and greatest values of the range. The probability density function is nonnegative everywhere, and its integral over the entire space is equal to one.

The plot depicts multiple thin MSS curves respectively representing the probability function in dependence on the distance of a tumor cell to its nearest immune cell obtained from an image of a tissue sample having been classified as MSS sample. For the MSS samples, the probability density strongly increases at a distance larger than 90µm. A bold MSS curve represents the average of all MSS curves in the plot.

The plot further depicts multiple thin MSI curves respectively representing the probability function in dependence on the distance of a tumor cell to its nearest immune cell obtained from an image of a tissue sample having been classified as

MSI sample. For the MSI samples, the probability density is highest at a distance smaller than 35 μm . A bold MSI curve represents the average of all MSI curves in the plot.

In addition, the plot comprises some curves obtained for tissue samples with
5 unknown MMS/MSI status.

The automated evaluation of tumor cell-immune cell distances in a tissue sample on a whole slide level provides a more comprehensive and accurate insight into tumor
10 biology.

Figure 11 depicts a digital image of a stained colorectal cancer sample comprising annotations of tumor context information. A combination of proximity information and cell density information is captured as tumor context information for classifying the tumor. As was described for several embodiments of the invention, proximity
15 measures between tumor cells and immune cells have been observed to represent valuable predictive parameters for classifying tumor cells into tumor cells of an inflammatory tumor or as tumor cells of a non-inflammatory tumor. The classification result may also be used for predicting whether or not the tumor will be treatable by a drug that boosts the immune system. Thus, the distances between immune cells
20 and tumor cells and the distance measure derived therefrom represent tumor context information that provides valuable predictive information on the likely tumor type, tumor progression and treatability.

It has been observed that the accuracy of the classification can significantly be increased by determining and using as input also the density of the tumor cells
25 and/or the density of immune cells. Thus, according to embodiments, the tumor cell density and/or the immune cell density (measured e.g. in number of cells per 100 $\mu\text{m} \times 100 \mu\text{m}$) is automatically determined and is fed as an additional input to the classifier. According to embodiments, the density of the immune cells is determined in an immune cell type specific manner (resulting in the computation of a density of
30 cytotoxic t-cells, a density for macrophages, a density for B-cells, etc.).

Accordingly, in the training phase of the classifier, density information can be assigned to each tumor cell and/or to each immune cell contained in any of the training images. For example, for each tumor cell in a training image, a square-shaped or circular area of about $100\mu\text{m} \times 100\mu\text{m}$ comprising said tumor cell in its geometrical center may be determined. Then, the number of tumor cells in said area is automatically determined and the tumor cell density is computed for said area. Then, the computed density is assigned to the tumor cell in the center of the area. This procedure may be repeated for each tumor cell. Likewise, for each immune cell in the training image, a square-shaped or circular area of about $100\mu\text{m} \times 100\mu\text{m}$ comprising said immune cell in its geometrical center may be determined. Then, the number of immune cells in said area is automatically determined and the immune cell density is computed for said area. Then, the computed density is assigned to the immune cell in the center of the area. According to embodiments, the density of the immune cells is determined in the training phase in an immune cell type specific manner (resulting in the computation of a density of cytotoxic t-cells, a density for macrophages, a density for B-cells, etc.).

In the depicted image, several tumor cells (Ki67 positive cells) and several immune cells (CD8 positive T-cells) were identified and the distance between the tumor cells and the T-cells was determined (see the large two circles around the two small tumor cells): each of the large circle represents a distance threshold of e.g. about $35\mu\text{m}$. For each of the identified tumor cells, the image analysis system determines the distance of said tumor cell to the nearest immune cell and determined if said distance is below the predefined distance threshold indicated by the large circles. The ratio of tumor cells for which a "close distance" smaller than the distance threshold and tumor cells for which a "large distance" larger than the distance threshold was measured is fed in addition to e.g. immune cell density information of the respective tumor cells into the classifier for classifying a tumor as "inflammatory" or "non-inflammatory".

According to embodiments, the classifier is configured such that the classification of the identified tumor cells into "tumor cells of an inflammatory tumor" or as "tumor cells of a non-inflammatory tumor" comprising checking if the tumor cell densities of

the majority of the tumor cells in the tumor depicted in the received image are within a tumor cell density range that is expected for a tumor cell of an inflammatory tumor or if the tumor cell densities of the majority of the tumor cells in the tumor depicted in the received image are within a tumor cell density range that is expected for a tumor cell of a non-inflammatory tumor. For example, the expected density ranges can be specified manually for different types of tumors (inflammatory/non-inflammatory breast cancer/colon cancer/lung cancer etc) or can be specified automatically and implicitly during the training phase by using the tumor cell densities as additional training parameters.

- 10 In addition or alternatively, the classifier is configured such that the classification of the identified tumor cells into "tumor cells of an inflammatory tumor" or as "tumor cells of a non-inflammatory tumor" comprising checking if the immune cell densities of the majority of the immune cells in spatial proximity to the tumor are within an immune cell density range that is expected in the context of an inflammatory tumor or if the immune cell densities of the majority of the immune cells in spatial proximity to the tumor are within an immune cell density range that is expected in the context of a tumor cell of a non-inflammatory tumor. As mentioned above, the expected density ranges for the immune cell densities can be specified manually for different types of tumors (inflammatory/non-inflammatory breast cancer/colon cancer/lung cancer etc) or can be specified automatically and implicitly during the training phase by using the immune cell densities as additional training parameters.

- 25 According to embodiments, the proximity measures and the cell densities are determined multiple times before, during and/or after the treatment of a patient for predicting the treatability of a tumor and for determining whether or not an applied drug had an effect on the tumor or the immune cells.

- Moreover, according to embodiments, the dynamic interplay of any cell population (e.g. tumor cells or immune cells) with a given reference point (e.g. a cell of another cell type than that of the cell population like immune cells, tumor cells, stroma cells or other tissues, e.g. vessels) is determined and used as an additional predictive input parameter for training an untrained classifier and/or for feeding a trained classifier.

Figure 12 depicts stained tissue samples of two patients A, B having similar T-cell densities but different proximity measures. Different degrees of immune cell infiltration in tumor nests have been observed to represent important signs of tumor cell recognition and/or eradication by immune cells. In the two compared tissue samples, the same T-cell density but strongly deviating proximity measures were observed: while in sample A only 4% of the tumor cells were observed in close proximity to a T-cell, in sample B more than 20% of the tumor cells had at least one T-cell in close proximity, e.g. within a distance of about 35µm. It has been observed that a low degree of tumor infiltration by immune cells correlates with a microsatellite status "MSS" while a high degree of tumor infiltration (sample B) correlates with a microsatellite status "MSI". Moreover, it has been observed that while already the computation of the proximity measure provides good or even better inflammatory/non-inflammatory tumor classification results than the conventional and expensive determination of MSS/MSI state, it has been observed that the additional determination and use of density information provides significantly more accurate classification results than the use of the proximity measure alone. This is because the density of a tissue may also have an impact on the proximity measure which however, may be more related to the type of the tissue from which a tumor stems than to the status of immune cell infiltration of the tumor.

C l a i m s

1. An image analysis method for tumor classification, the method comprising:
- receiving (202), by an image analysis system (100), at least one digital
5 image (118) of a tissue sample;
 - analyzing (204), by the image analysis system, the at least one received
image for identifying immune cells and tumor cells in the at least one
received image;
 - for each of the identified tumor cells, determining (206), by the image
10 analysis system, the distance of said tumor cell to the nearest immune cell;
 - computing (208), by the image analysis system, a proximity measure as a
function of the determined distances;
 - in dependence on the proximity measure, classifying (210), by the image
analysis system, the identified tumor cells into tumor cells of an
15 inflammatory tumor or as tumor cells of a non-inflammatory tumor; and
 - storing (212), by the image analysis system, the classification result on a
storage medium and/or displaying (214) the classification result on a
display device.
- 20 2. An image analysis method for tumor classification, the method comprising:
- receiving, by an image analysis system (100), a digital image (118) of a
tissue sample;
 - analyzing, by the image analysis system, the at least one received image
for identifying tumor cells in the at least one received image;
 - 25 – analyzing, by the image analysis system, the identified tumor cells for
identifying tumor regions, wherein a tumor region is a tissue region whose
majority of cells consist of tumor cells;
 - analyzing, by the image analysis system, the at least one received image
for identifying immune cells lying within one of the identified tumor regions
30 or lying within a boundary region around one of the tumor regions, the
width of the boundary being smaller than 200µm;

- for each of the identified immune cells, determining, by the image analysis system, the distance of said immune cell to the nearest tumor cell;
 - computing, by the image analysis system, a proximity measure as a function of the determined distances;
 - 5 – in dependence on the proximity measure, classifying, by the image analysis system, the identified tumor cells into tumor cells of an inflammatory tumor or as tumor cells of a non-inflammatory tumor; and
 - storing , by the image analysis system, the classification result on a storage medium and/or displaying the classification result on a display device.
 - 10
3. The method of claim 1 or 2, the displaying of the classification result comprising:
- representing all identified tumor cells as first pixel blobs having a first color;
 - 15 and
 - representing all identified immune cells whose distance to its nearest tumor cells is below a threshold as second pixel blobs having a second color;
 - displaying the second pixel blobs as an overlay of the first pixel blobs.
 - 20
4. The image analysis method of any one of the previous claims, the tissue sample being a whole slide tissue sample and the digital image being a whole slide image.
- 25 5. The image analysis method of any one of the previous claims, further comprising:
- analyzing the received digital image for determining a tumor cell density of the identified tumor cells;
 - wherein the classification of the identified tumor cells into tumor cells of an inflammatory tumor or as tumor cells of a non-inflammatory tumor
 - 30 comprises inputting the proximity measure and the tumor cell density into a classifier configured to perform the classification.

6. The image analysis method of any one of the previous claims, further comprising:
- 5 -- analyzing the received digital image for determining an immune cell density of the identified immune cells or of a particular type of immune cells;
- wherein the classification of the identified tumor cells into tumor cells of an inflammatory tumor or as tumor cells of a non-inflammatory tumor comprises inputting the proximity measure and the immune cell density
- 10 into a classifier configured to perform the classification.
7. The image analysis method of any one of the previous claims, the identification of the tumor cells comprising:
- 15 -- identifying proliferating non-lymphoid cells and using said identified cells as the tumor cells; and/or
- identifying cells expressing a set of one or more tumor-specific biomarkers and using said identified cells as the tumor cells.
8. The image analysis method of any one of the previous claims,
- 20 -- the identification of the immune cells comprising identifying cytotoxic T-cells and using the identified cytotoxic T-cells as the identified immune cells; and/or
- the identification of the immune cells comprising identifying T-helper-cells and using the identified T-helper-cells as the identified immune cells;
- 25 and/or
- the identification of the immune cells comprising identifying macrophages and using the identified macrophages as the identified immune cells; and/or
- 30 -- the identification of the immune cells comprising identifying memory cells and using the identified memory cells as the identified immune cells; and/or

- the identification of the immune cells comprising identifying B- cells and using the identified memory cells as the identified immune cells; and/or
- the identification of the immune cells comprising identifying activated T-cells and using the identified activated T cells as the identified immune cells; and/or
- the identification of the immune cells comprising identifying PD1+ immune cells and using the identified PD1+ immune cells as the identified immune cells, the identification of the tumor cells comprising identifying PDL1+ tumor cells and using the identified PDL1+ tumor cells as the identified tumor cells .

9. The image analysis method of any one of the previous claims,

- the identification of the immune cells comprising identifying regulatory T-cells and identifying immune cells of at least one type of immune cells that boosts an immune response; and
- selectively using the identified immune cells that boost the immune response but not the identified regulatory T-cells as the identified immune cells.

10. The image analysis method of any one of the previous claims 1, 3-9, the computation of the proximity measure as a function of the determined distances comprising:

- identifying a first and a second sub-set of the identified tumor cells, the first sub-set selectively comprising tumor cells whose nearest immune cell is less than a predefined distance away, the second sub-set selectively comprising tumor cells whose nearest immune cell is at least the predefined distance away from the immune cell;
- computing a ratio of the number of tumor cells contained in the first subset and the number of tumor cells in the second sub-set; and
- using the ratio as the proximity measure, wherein the higher the ratio, the higher the probability that the classification result indicates that the identified tumor cells belong to the inflammatory tumor.

11. The image analysis method of any one of the previous claims 2-9, the computation of the proximity measure as a function of the determined distances comprising:
- 5 – identifying a first and a second sub-set of the identified immune cells, the first sub-set selectively comprising immune cells whose nearest tumor cell is less than a predefined distance away, the second sub-set selectively comprising immune cells whose nearest tumor cell is at least the predefined distance away from the tumor cell;
- 10 – computing a ratio of the number of immune cells contained in the first subset and the number of immune cells in the second sub-set; and
- using the ratio as the proximity measure, wherein the higher the ratio, the higher the probability that the classification result indicates that the identified tumor cells belong to the inflammatory tumor.
- 15
12. The image analysis method of any one of claims 10-11, the predefined distance being the immunologically effective distance of the identified immune cells, the immunologically effective distance being a maximum distance within which the identified immune cell is able to directly or indirectly trigger the killing
- 20 of or the performing of apoptosis by the tumor cell.
13. The image analysis method of any one of claims 10-12, the predefined distance is in a range of 20 μm to 50 μm .
- 25 14. The image analysis method of any one of claims 10-13, wherein in case the ratio exceeds a predefined percentage, e.g. 50%, the classification result indicates that the identified tumor cells belong to the inflammatory tumor.
- 30 15. The image analysis method of any one of the previous claims 1, 3-9, the computation of the proximity measure as a function of the determined distances comprising:

- generating a histogram (502-506; 602-616; 910, 912) of the distances of the tumor cells to their respective nearest one of the identified immune cells, the histogram comprising at least two distance bins, the histogram covering a distance range of 0 μm to at least 50 μm , each of the bins corresponding to a bar of the histogram, each of the bars indicating a count of the identified tumor cells having a distance to their nearest immune cell that falls into said bin;
- connecting the upper end of the first one of the bars with the upper end of the last one of the bars with a line, the first bar corresponding to the one of the bins covering the smallest distances, the last bar corresponding to the one of the bins covering the largest distances of the distance range;
- determining the slope of the line;
- using the slope as the proximity measure, wherein in case the slope indicates that the tumor cell count of the first bar is higher than the tumor cell count of the last bar, the classification result is that the identified tumor cells belong to an inflammatory tumor.

16. The image analysis method of any one of the previous claims 2-9, the computation of the proximity measure as a function of the determined distances comprising:

- generating a histogram of the distances of the immune cells to their respective nearest one of the identified tumor cells, the histogram comprising at least two distance bins, the histogram covering a distance range of 0 μm to at least 50 μm , each of the bins corresponding to a bar of the histogram, each of the bars indicating a count of the identified immune cells having a distance to their nearest tumor cell that falls into said bin;
- connecting the upper end of the first one of the bars with the upper end of the last one of the bars with a line, the first bar corresponding to the one of the bins covering the smallest distances, the last bar corresponding to the one of the bins covering the largest distances of the distance range;
- determining the slope of the line;

- using the slope as the proximity measure, wherein in case the slope indicates that the immune cell count of the first bar is higher than the immune cell count of the last bar, the classification result is that the identified tumor cells belong to an inflammatory tumor.

5

17. The image analysis method of any one of the previous claims, the method further comprising:

- selectively in case the classification result indicates that the identified tumor cells belong to an inflammatory tumor, outputting a signal being indicative of a treatment recommendation to use a substance that boosts or modulates an immune response as a drug for treating the tumor.

10

18. The image analysis method of any one of the previous claims, the method further comprising:

- using the proximity measure in addition to or in replacement of an MSS-MSI status indicator for computing a prognosis of the responsiveness of the tumor to a substance that boosts or modulates an immune response.

15

19. The image analysis method of any one of the previous claims, the identified tumor cells being colorectal cancer cells.

20

20. The image analysis method of any one of claims, the identification of the immune cells comprising identifying immune cells of multiple different immune cell types by analyzing pixel intensity values in the at least one digital image representing the presence of different immune cell type specific biomarkers, and using the totality of the identified immune cells of the multiple different immune cell types as the identified immune cells for computing the proximity measure.

25

21. The image analysis method of any one of claims, the tumor cells being identified as cells expressing at least a first tumor-type specific biomarker, the

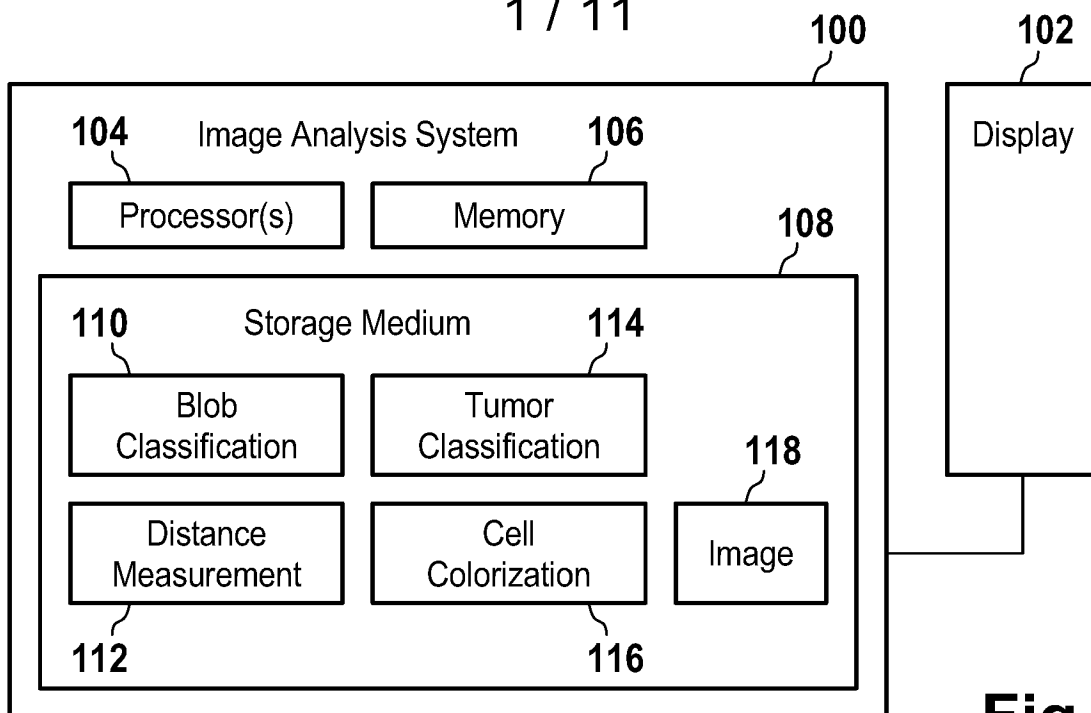
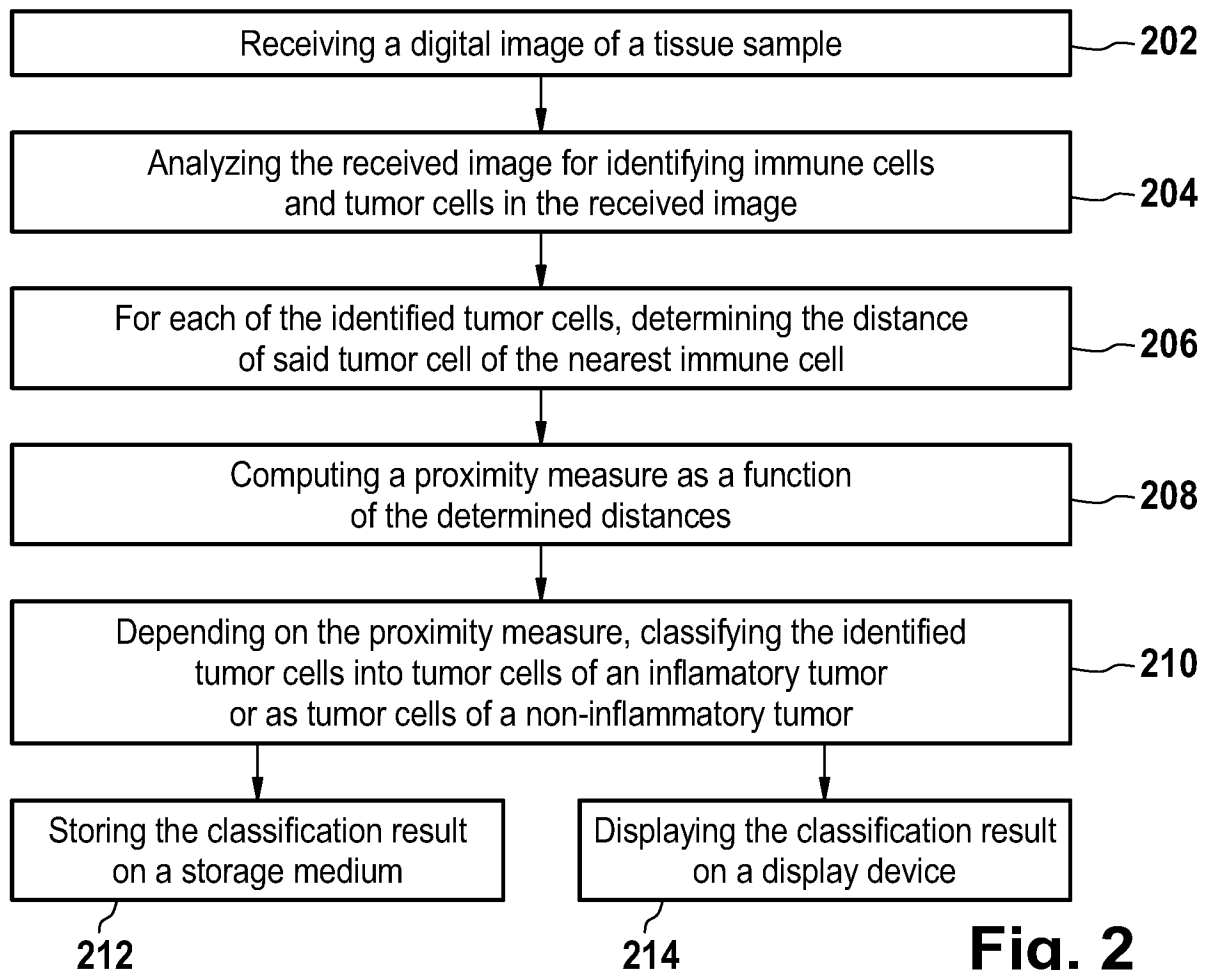
30

immune cells being identified as cells expressing at least a first immune cell type specific biomarker, the method further comprising:

- 5 — analyzing the at least one received image or a further image of the tissue sample for identifying further immune cells as cells expressing a second immune cell type specific biomarker; and
- analyzing the at least one received image or the further image of the tissue sample for identifying further tumor cells as cells expressing a second tumor type specific biomarker or using the identified immune cells as further tumor cells;
- 10 — for each of the identified further tumor cells, determining the distance of said further tumor cell to the nearest further immune cell;
- computing a further proximity measure as a function of said determined distances;
- 15 — in dependence on the further proximity measure, sub-classifying the identified tumor cells and/or predicting the treatability of the tumor by a class of substances.

22. An image analysis system (100) for tumor classification, the system being configured for:

- 20 — receiving (202) a digital image (118) of a tissue sample;
- analyzing (204) the at least one received image for identifying immune cells and tumor cells in the at least one received image;
- for each of the identified tumor cells, determining (206) the distance of said tumor cell to the nearest immune cell;
- 25 — computing (208) a proximity measure as a function of the determined distances;
- in dependence on the proximity measure, classifying (210) the identified tumor cells into tumor cells of an inflammatory tumor or as tumor cells of a non-inflammatory tumor; and
- 30 — storing (212) the classification result on a storage medium and/or displaying (214) the classification result on a display device (102).

**Fig. 1****Fig. 2**

2 / 11

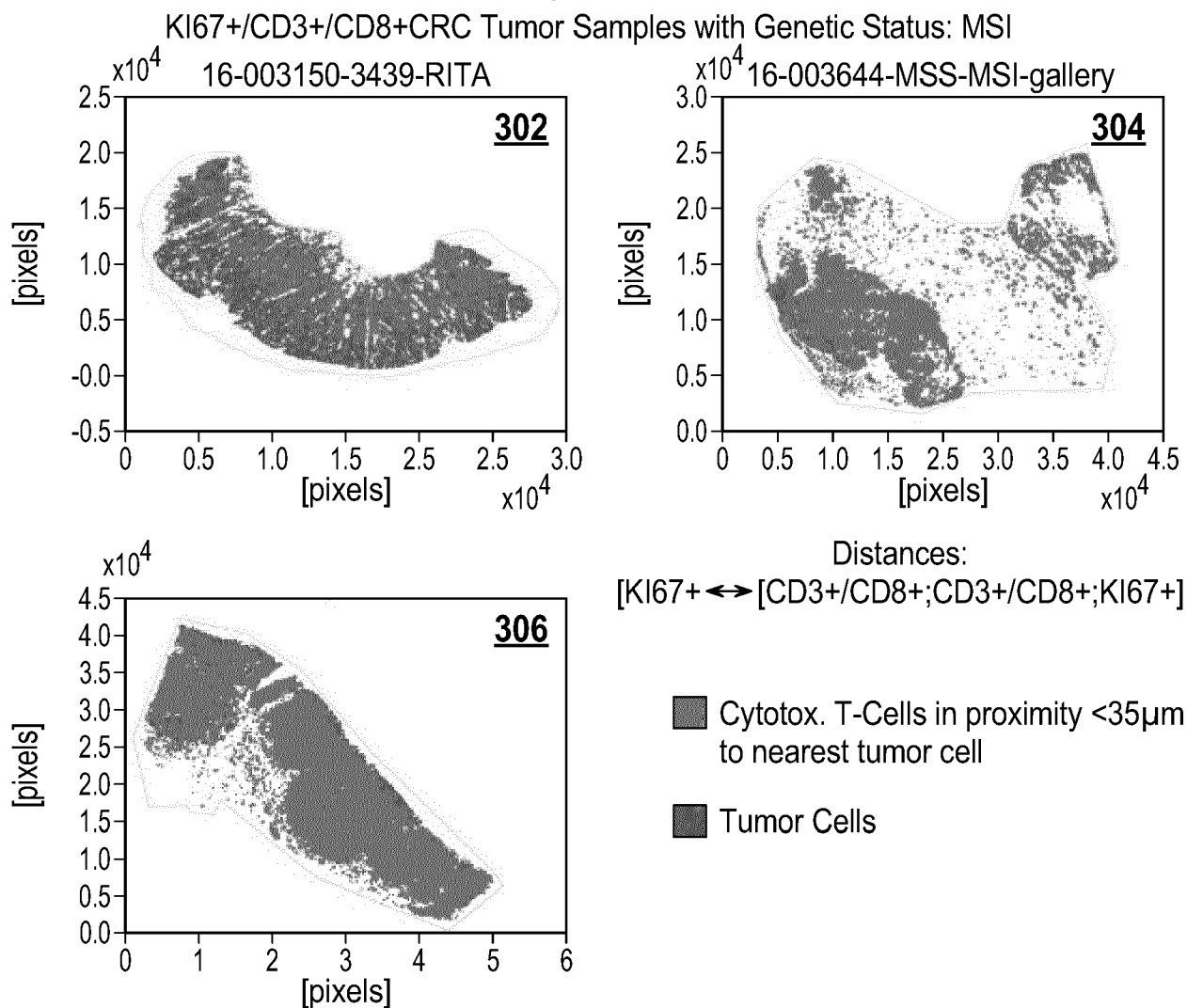


Fig. 3

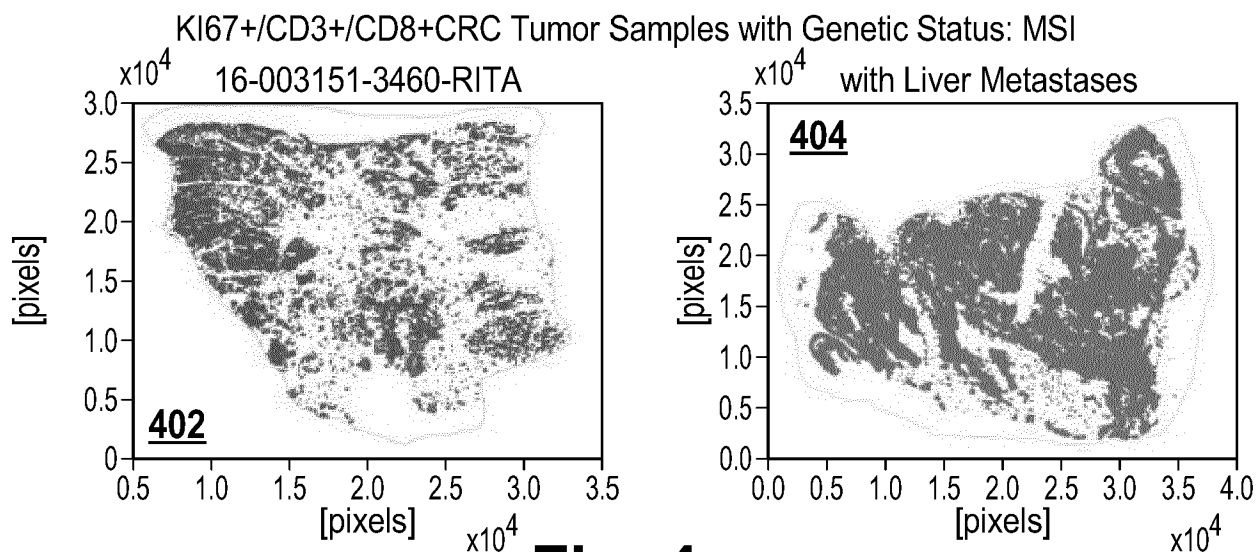
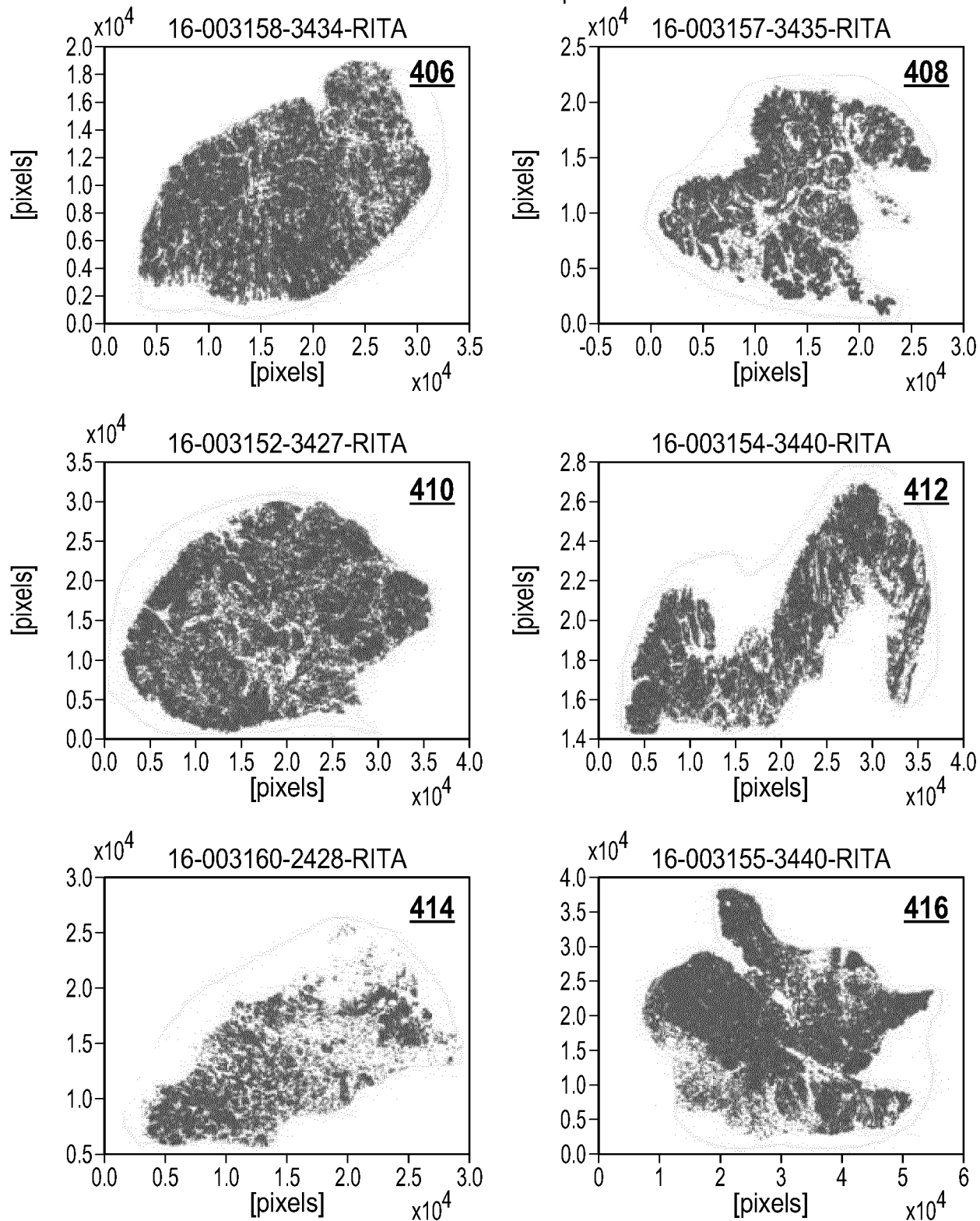


Fig. 4a

3 / 11

Further KI67+/CD3+/CD8+CRC Tumor Samples with Genetic Status: MSS

**Fig. 4b**

SUBSTITUTE SHEET (RULE 26)

Distance Histograms of KI67+/CD3+/CD8+ CRC Tumor Samples with Genetic Status: MSI

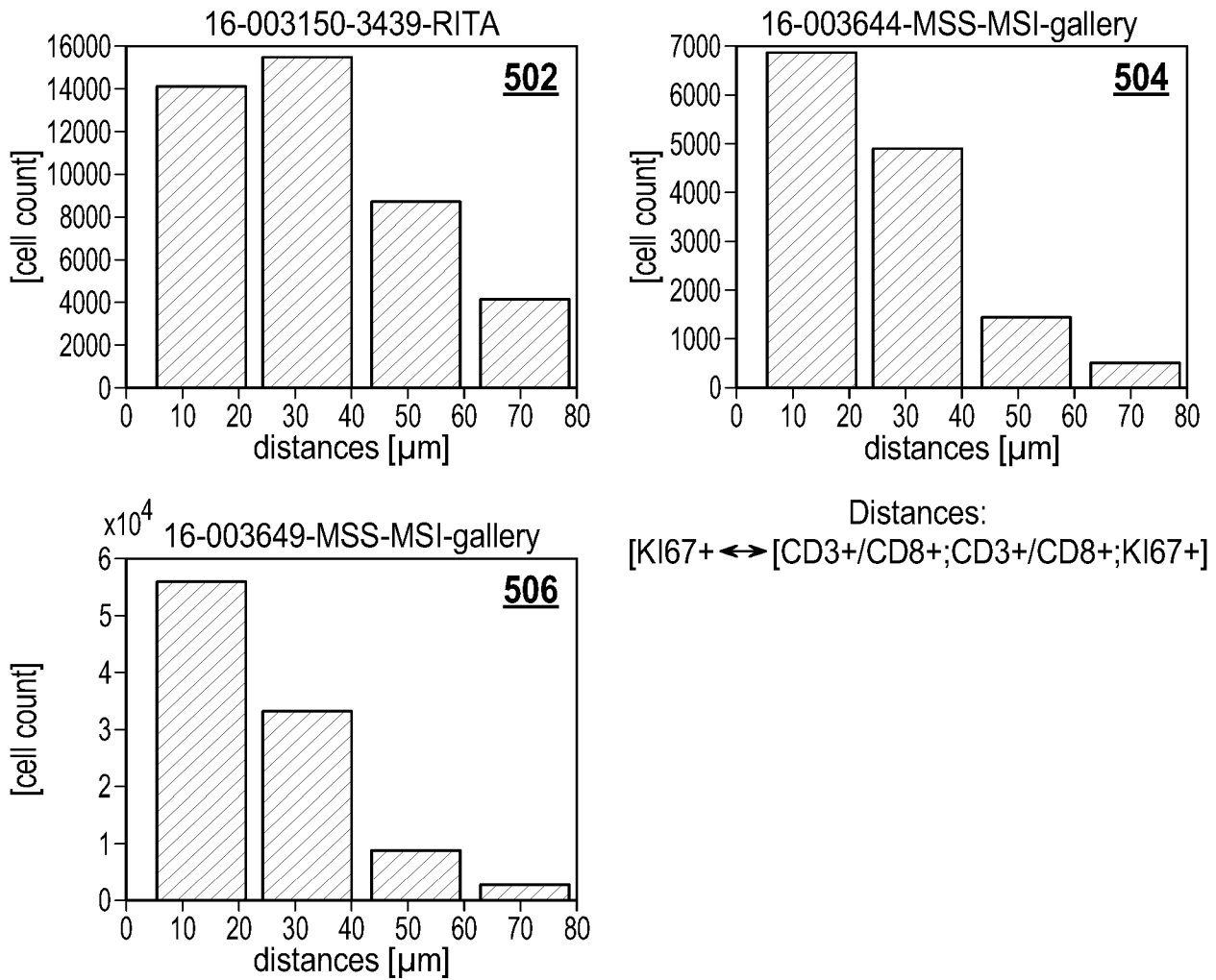


Fig. 5

Distance Histograms of KI67+/CD3+/CD8+ CRC Tumor Samples with Genetic Status: MSS

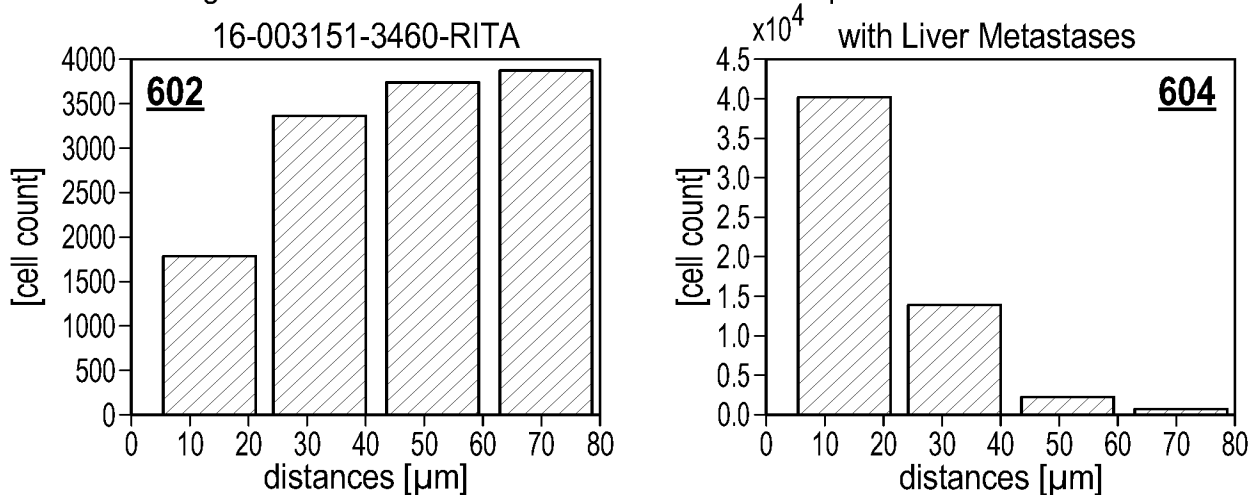
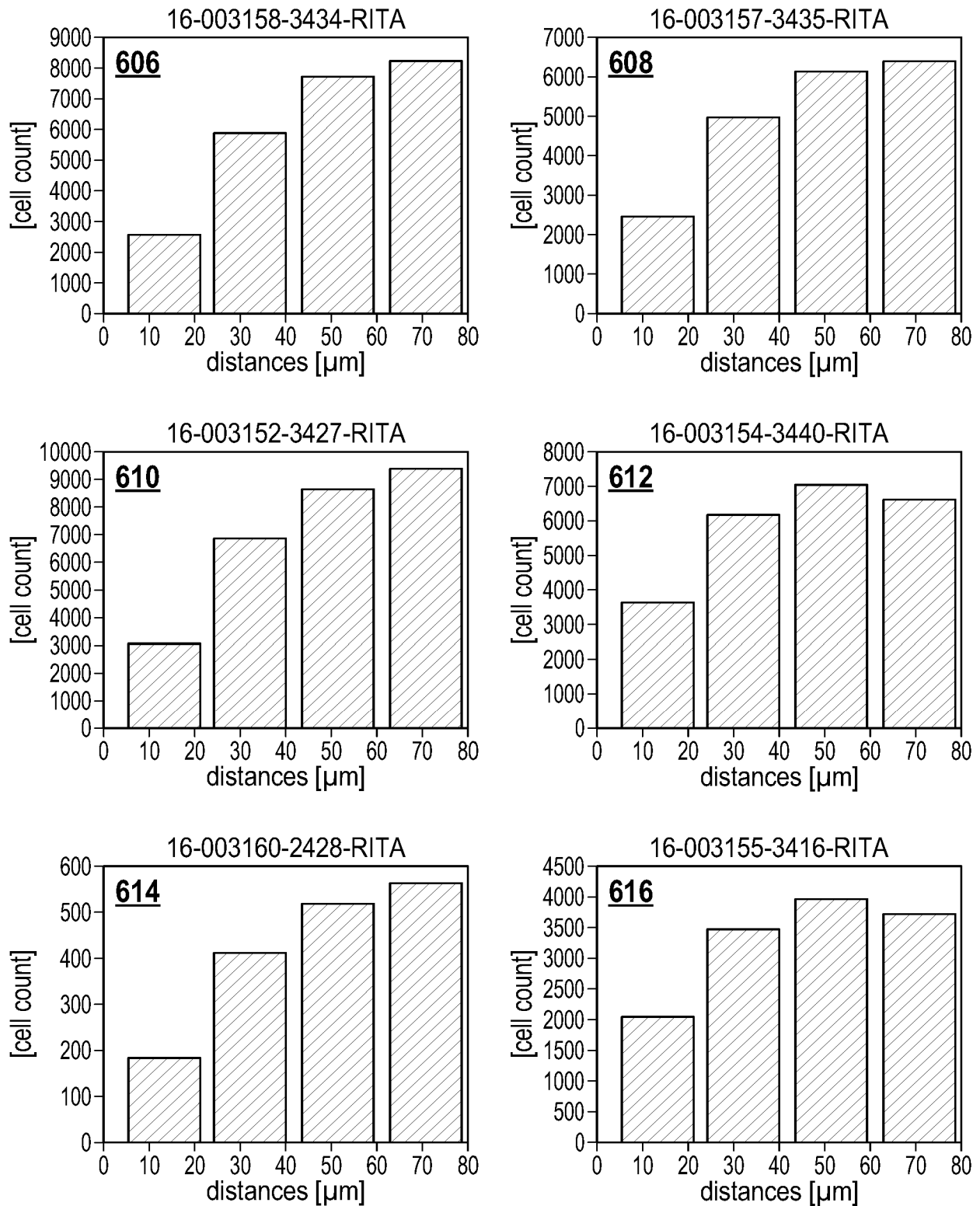


Fig. 6a

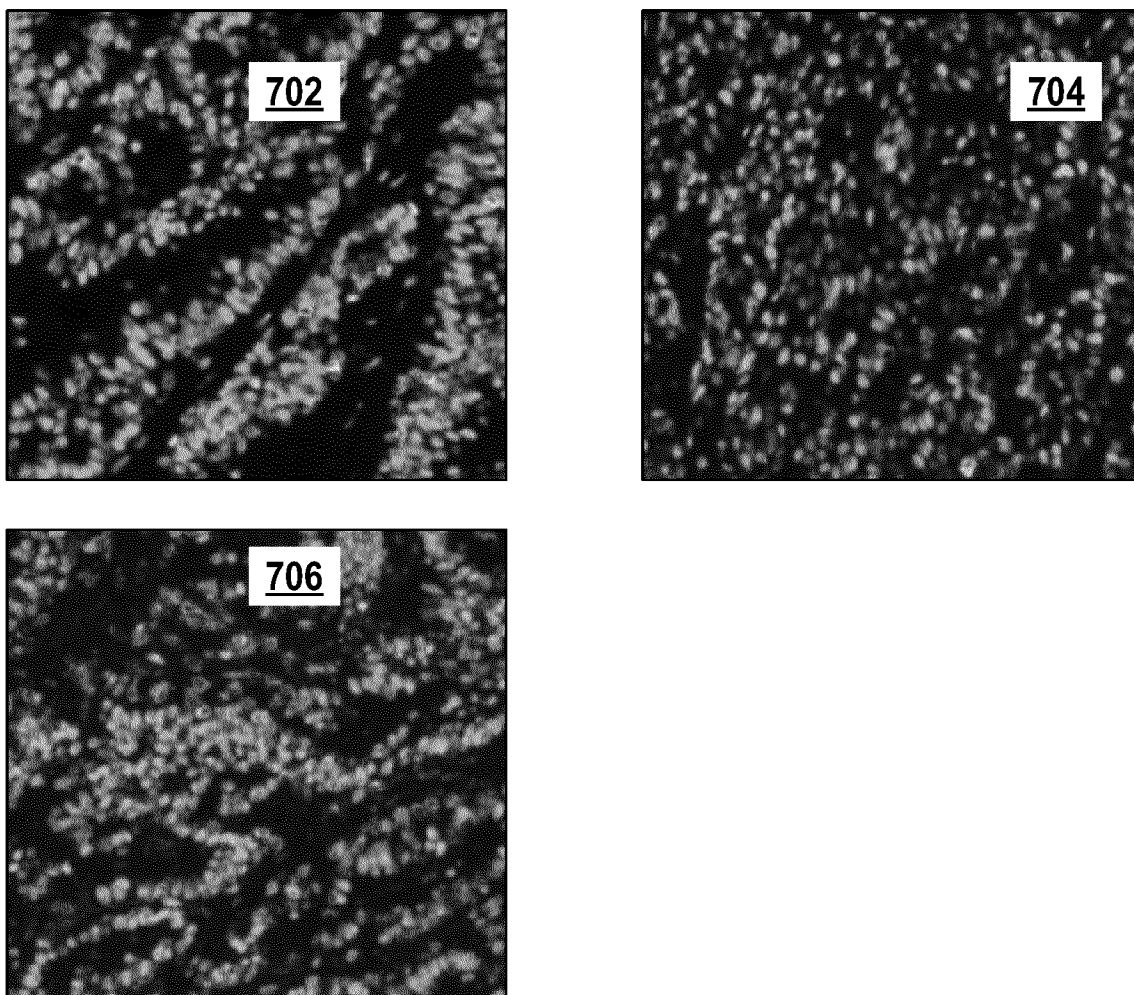
Further Distance Histograms of KI67+/CD3+/CD8+ CRC Tumor Samples with Genetic Status: MSS



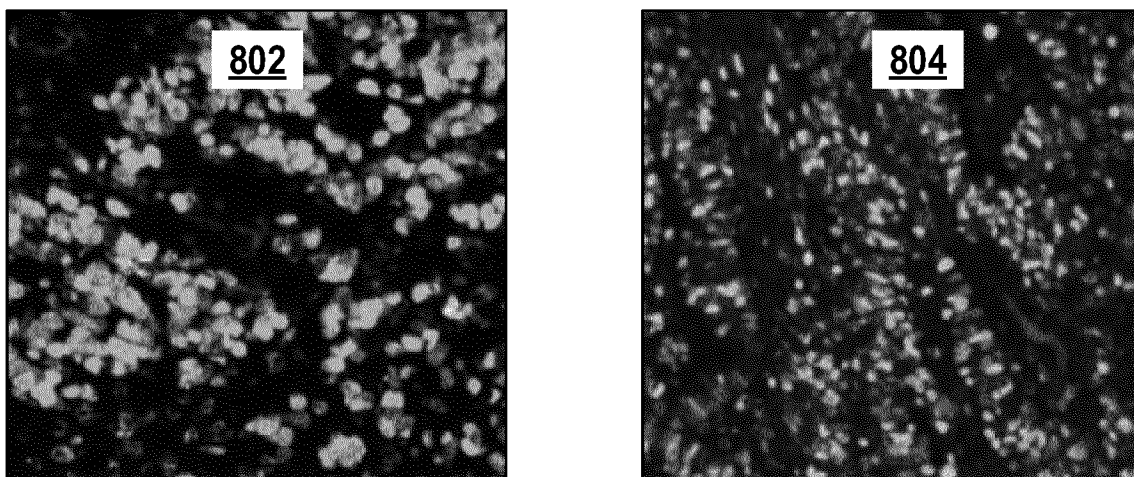
Distances: [KI67+ ↔ [CD3+/CD8+;CD3+/CD8+;KI67+]

Fig. 6b

CRC Tumor Samples with Genetic Status: MSS

**Fig. 7**

Tumor Samples with Genetic Status: MSS

**Fig. 8a**

Tumor Samples with Genetic Status: MSS

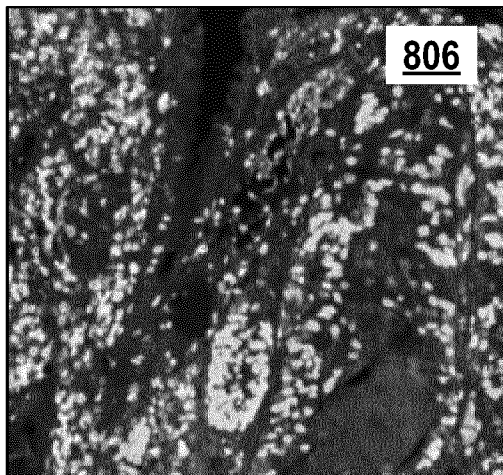
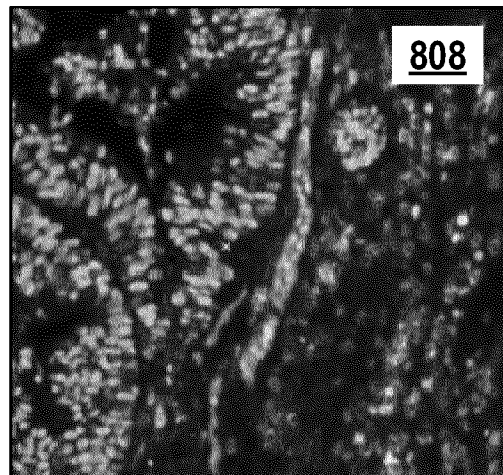
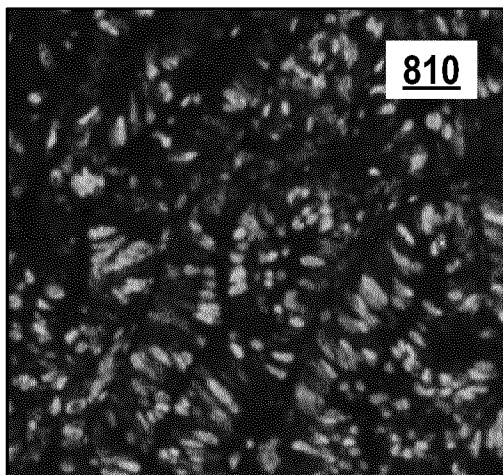
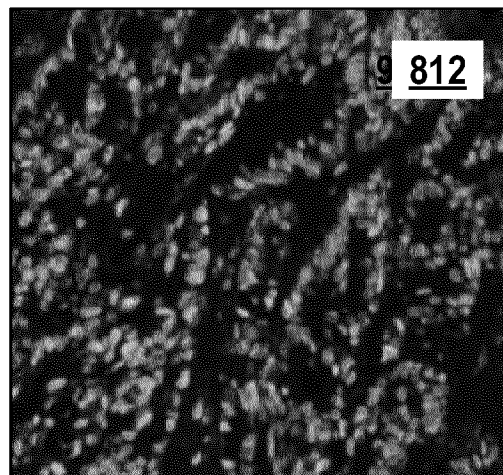
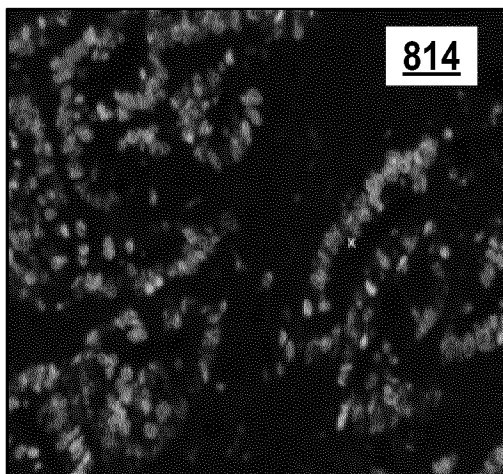
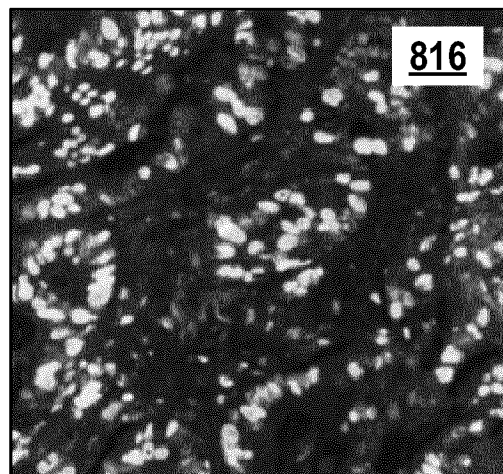
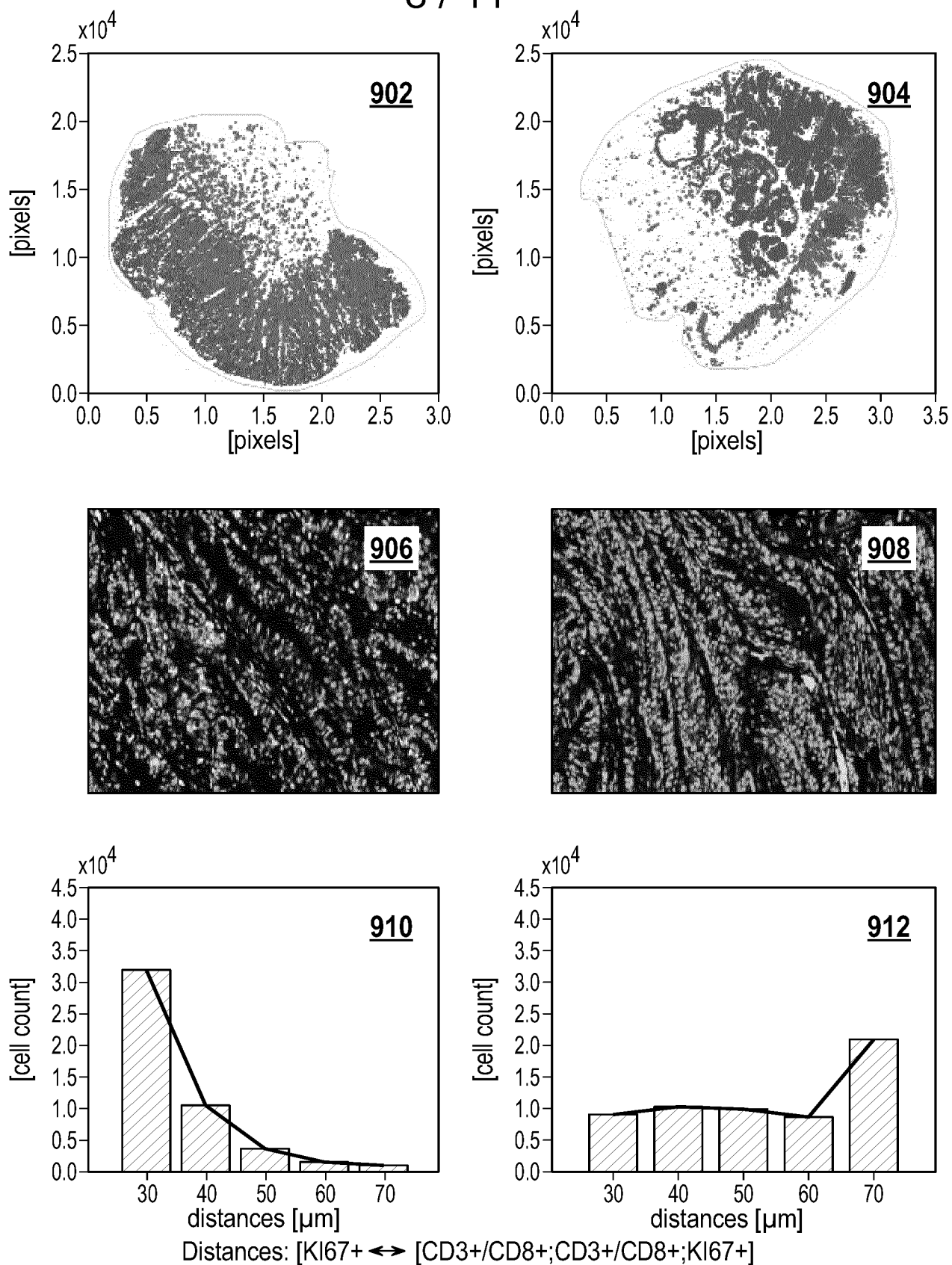


Fig. 8b

**Fig. 9**

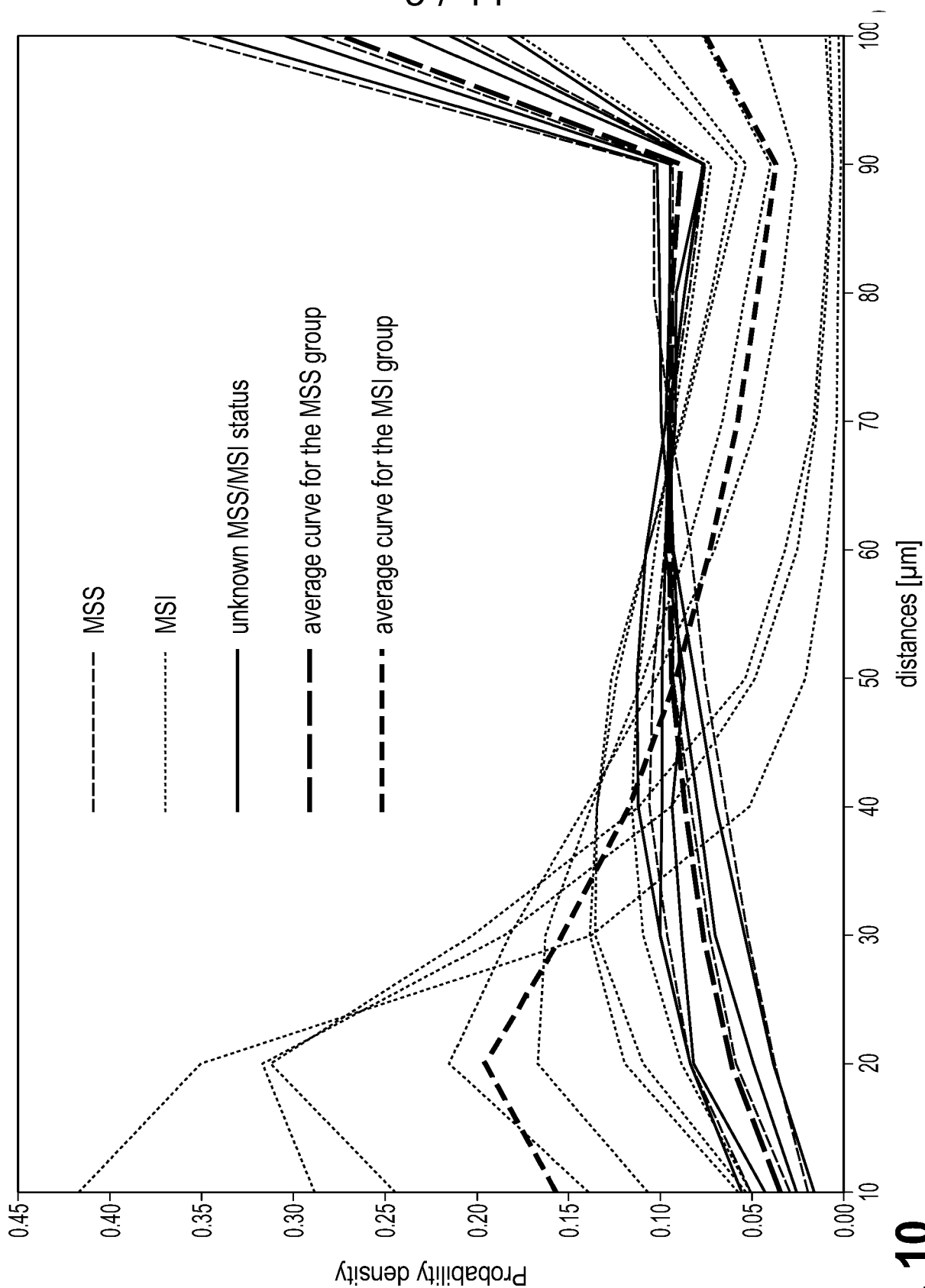
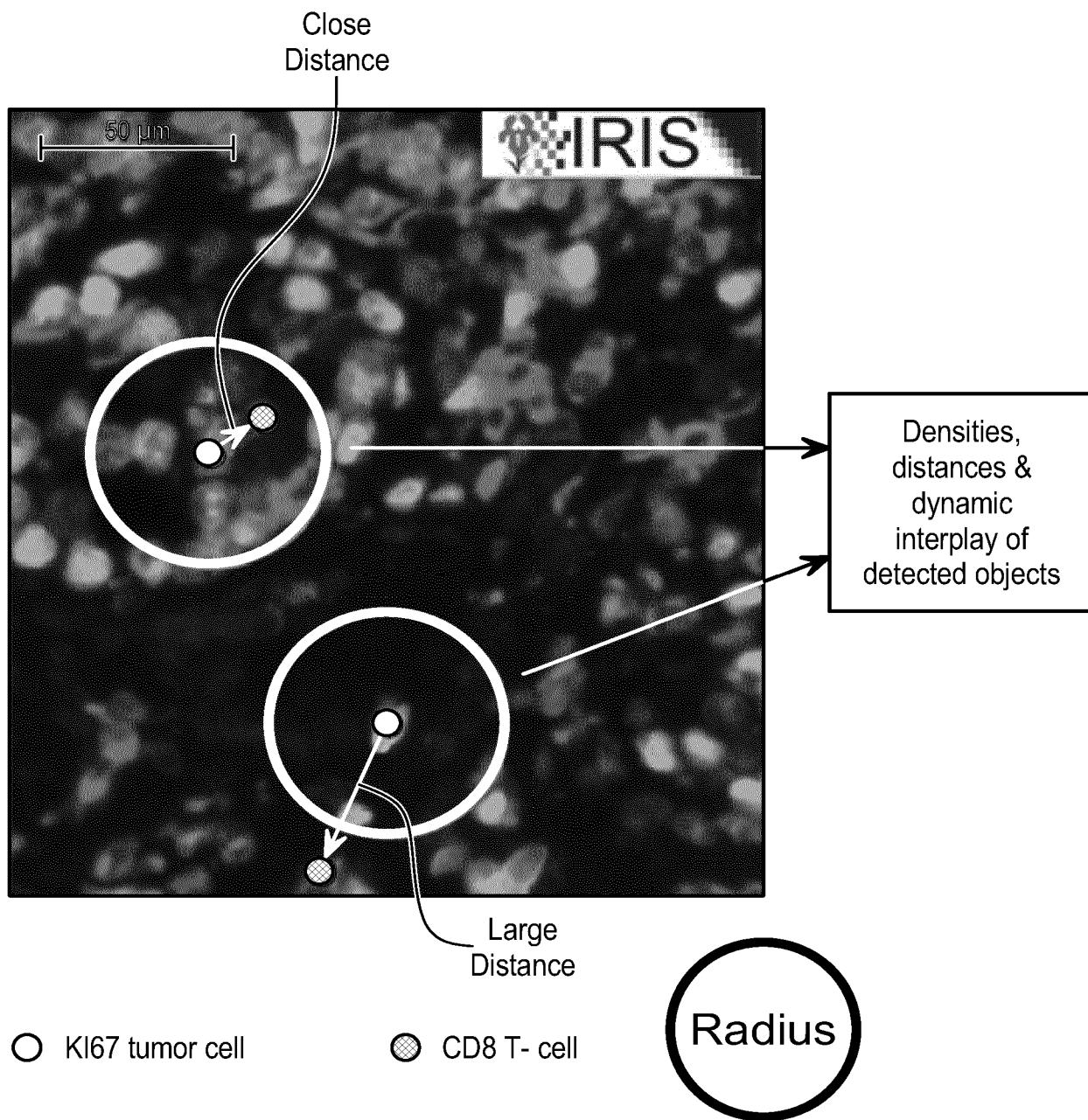


Fig. 10

**Fig. 11**

Two cases with similar T-cell densities

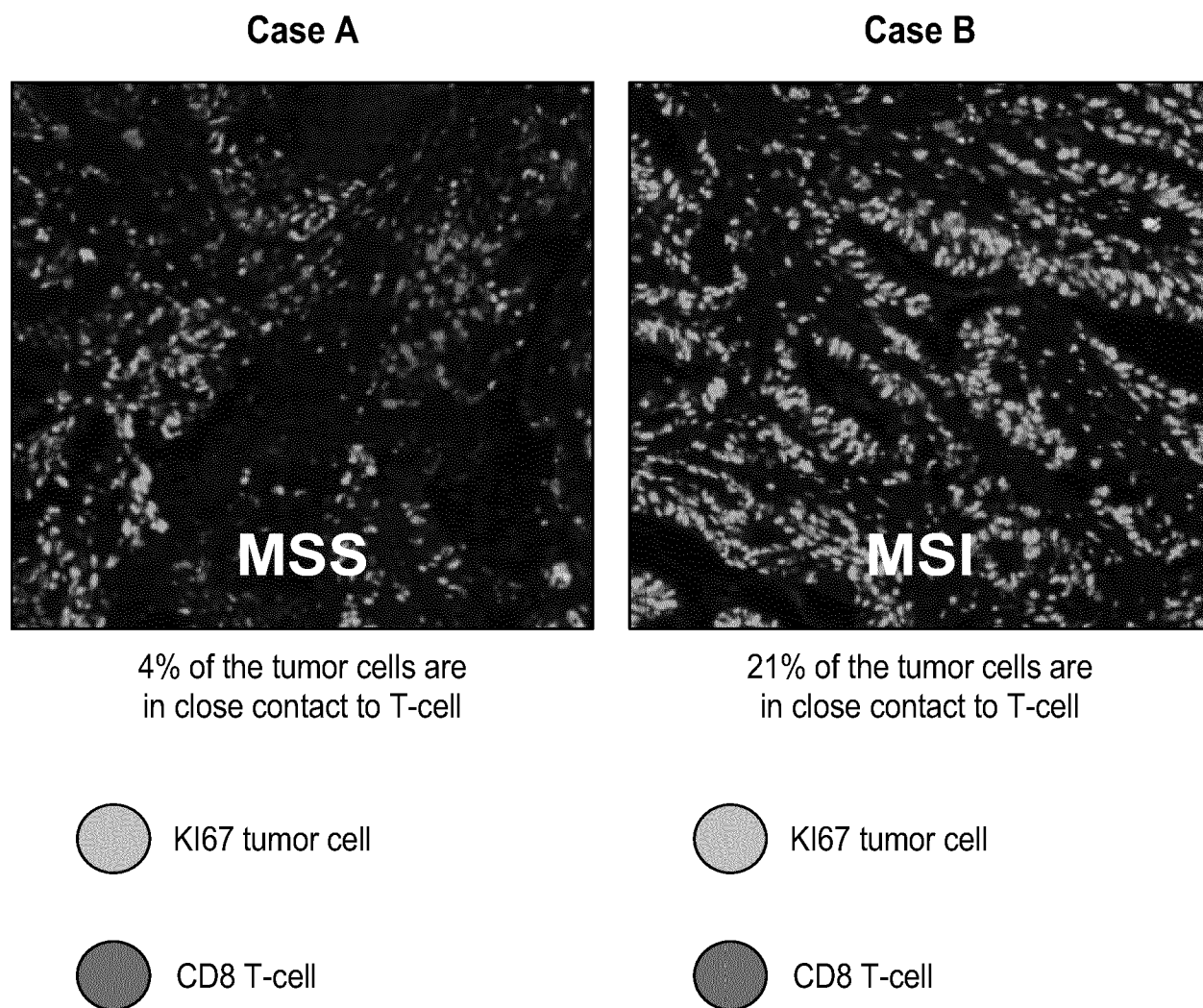


Fig. 12

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2017/062037

A. CLASSIFICATION OF SUBJECT MATTER INV. G06T7/00 ADD.		
According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) G06T		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) EP0-Internal		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2015/189264 A1 (VENTANA MED SYST INC [US]; HOFFMANN LA ROCHE [CH]; UNIV LELAND STANFORD) 17 December 2015 (2015-12-17) abstract paragraphs [0012], [0014], [0029], [0032], [0037], [0038], [0040], [0045], [0048], [0065], [0066], [0071], [0072], [0125], [0128], [0130] -----	1-22
X	US 2016/123964 A1 (TUMEH PAUL C [US] ET AL) 5 May 2016 (2016-05-05) abstract paragraphs [0016], [0017], [0026], [0027], [0045], [0047], [0048], [0096] - [0102], [0114], [0116], [0122] - [0124], [0126] paragraphs [0138], [0165], [0220], [0221], [0305], [0306], [0308], [0309] ----- - / - -	1-22
☒ Further documents are listed in the continuation of Box C. ☒ See patent family annex.		
* Special categories of cited documents : <div style="display: flex; justify-content: space-between;"> <div style="width: 48%;"> <p>"A" document defining the general state of the art which is not considered to be of particular relevance</p> <p>"E" earlier application or patent but published on or after the international filing date</p> <p>"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)</p> <p>"O" document referring to an oral disclosure, use, exhibition or other means</p> <p>"P" document published prior to the international filing date but later than the priority date claimed</p> </div> <div style="width: 48%;"> <p>"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention</p> <p>"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone</p> <p>"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art</p> <p>"&" document member of the same patent family</p> </div> </div>		
Date of the actual completion of the international search 19 September 2017		Date of mailing of the international search report 27/09/2017
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016		Authorized officer Casteller, Maurizio

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2017/062037

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2015/133321 A1 (BHAUMIK SRABANI [US] ET AL) 14 May 2015 (2015-05-14) abstract; figures 1,2 paragraphs [0026], [0029], [0032], [0035], [0037], [0046], [0063], [0086], [0087] -----	1-22
X	JULIANE M KRÜGER ET AL: "Combat or surveillance? Evaluation of the heterogeneous inflammatory breast cancer microenvironment", THE JOURNAL OF PATHOLOGY, vol. 229, no. 4, 15 February 2013 (2013-02-15), pages 569-578, XP055181126, ISSN: 0022-3417, DOI: 10.1002/path.4150 abstract; figures 1,4,5 page 570, paragraphs "Image acquisition", "Colour deconvolution" page 571, paragraph "Statistical analysis" -----	1-22

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2017/062037

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2015189264 A1	17-12-2015	EP 3155592 A1 US 2017091937 A1 WO 2015189264 A1	19-04-2017 30-03-2017 17-12-2015
US 2016123964 A1	05-05-2016	CN 107109459 A EP 3215852 A1 US 2016123964 A1 WO 2016073760 A1	29-08-2017 13-09-2017 05-05-2016 12-05-2016
US 2015133321 A1	14-05-2015	US 2015133321 A1 WO 2015073516 A1	14-05-2015 21-05-2015

Publication Number: WO 2018/065434 A1

Publication Date: Apr 12, 2018

Abstract

A digital pathology system and associated method and computer program product provide a quantitative analysis of entire tissue slides as well as intuitive, effective, fast, and precise quantification of biomarker expressions across relevant areas of the entire tissue slides. The digital pathology system enables a novel workflow that allows the user to efficiently outline clinically relevant morphology in its entirety, including solid tumor areas. Quantitative analysis results are then efficiently and intuitively provided to the user for all tissue content (i.e., millions of cells) within seconds. This efficiency is made possible by a pre-computation step that computes and stores image analysis results for later retrieval. Visualizing vast amount of data effectively is another component of the system that provides information and confidence to the user about the biomarker expression levels.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property
Organization
International Bureau

(43) International Publication Date
12 April 2018 (12.04.2018)



(10) International Publication Number
WO 2018/065434 A1

(51) International Patent Classification:
G06T 7/00 (2017.01)

(21) International Application Number:
PCT/EP2017/075113

(22) International Filing Date:
04 October 2017 (04.10.2017)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/405,808 07 October 2016 (07.10.2016) US

(71) Applicant: VENTANA MEDICAL SYSTEMS, INC.
[US/US]; 1910 E. Innovation Park Dr., Tucson, Arizona
85755 (US).

(71) Applicant (for CL only): F. HOFFMANN-LA ROCHE
AG [CH/CH]; Grenzacherstrasse 124, 4070 Basel (CH).

(72) Inventors: BARNES, Michael; 1500 E Pusch Wilderness
Dr., Unit 2206, Oro Valley, Arizona 85737 (US). BRED-
NO, Joerg; 397 27th Street, San Francisco, CA 94131 (US).
CHUKKA, Srinivas; 1045 Oak Tree Drive, San Jose, CA
95129 (US). GUETTER, Christoph; 3221 Fernside Boule-
vard, Unit #A, Alameda, CA 94501 (US). LORSAKUL,
Auranuch; 3485 Monroe Street Apt 421, Santa Clara, CA
95051 (US). SARKAR, Anindya; 794 Coyote St, Milpitas,
CA 95035 (US). SUZUE, Ellen; 806 Fife Way, Sunnyvale,
CA 94087 (US).

(74) Agent: GOEHRING, Frank et al.; Roche Diagnostics
GmbH, Patent Department (LPP.....6164), P.O.Box 11 52,
82372 Penzberg (DE).

(81) Designated States (unless otherwise indicated, for every
kind of national protection available): AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ,
CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO,
DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN,
HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP,
KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME,
MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ,
OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA,
SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN,
TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every
kind of regional protection available): ARIPO (BW, GH,
GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ,
UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ,
TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK,
EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV,
MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,
TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW,
KM, ML, MR, NE, SN, TD, TG).

Published:

- with international search report (Art. 21(3))
- before the expiration of the time limit for amending the
claims and to be republished in the event of receipt of
amendments (Rule 48.2(h))

(54) Title: DIGITAL PATHOLOGY SYSTEM AND ASSOCIATED WORKFLOW FOR PROVIDING VISUALIZED WHOLE-
SLIDE IMAGE ANALYSIS

(57) Abstract: A digital pathology system and associated method and computer program product provide a quantitative analysis of
entire tissue slides as well as intuitive, effective, fast, and precise quantification of biomarker expressions across relevant areas of the
entire tissue slides. The digital pathology system enables a novel workflow that allows the user to efficiently outline clinically relevant
morphology in its entirety, including solid tumor areas. Quantitative analysis results are then efficiently and intuitively provided to
the user for all tissue content (i.e., millions of cells) within seconds. This efficiency is made possible by a pre-computation step that
computes and stores image analysis results for later retrieval. Visualizing vast amount of data effectively is another component of the
system that provides information and confidence to the user about the biomarker expression levels.



WO 2018/065434 A1

DIGITAL PATHOLOGY SYSTEM AND ASSOCIATED WORKFLOW FOR PROVIDING VISUALIZED WHOLE-SLIDE IMAGE ANALYSIS

FIELD OF THE INVENTION

5 [0001] The present invention generally relates to a digital pathology system that hosts a workflow for providing a quantitative analysis of whole-slide images as well as intuitive visualization thereof. More particularly, the present workflow provides precise quantification of biomarker expressions to aid in in-vitro diagnostics.

BACKGROUND OF THE INVENTION

10 [0002] Digital pathology equipment is often used to produce digital images of microscope slides. Pathologists and histotechnologists often visually examine the digital images to obtain information about tissue samples and to identify the most appropriate treatment to improve a clinical outcome.

15 [0003] Current digital pathology image analysis systems are increasingly expected to analyze large amounts of high-resolution digital images of tissue slides or "virtual slides." The digital image analysis has proven to be computationally intensive, requiring significant storage, and lessening the ability for real time diagnostic and interaction execution.

20 [0004] In addition, a conventional digital pathology system typically analyzes one or more small areas of the digital slide, generally referred to as fields of view (FOVs). Such partitioning of the digital slide forces the user or pathologist to contribute significant mental effort into approximating the correct solution by carefully selecting those FOVs that are more likely to represent the information of the entire tissue affixed to the slide.

25 [0005] Despite their potential advantages, current digital pathology systems have not achieved the desired efficiency in providing an expeditious, real time quantitative analysis of entire tissue slides as well as intuitive visualization thereof.

[0006] For example, conventional digital imaging systems with limited processing power restrict the size of each field of view to approximately 1,000 to 2,500 pixels, a small fraction of the whole slide image that can contain, for example, 60,000 to 110,000 pixels, and occupy, for example, between 0.5 GB and 1.5 GB of memory space.

SUMMARY OF THE INVENTION

[0007] It is anticipated that whole-slide image analyses will become widely adapted in automated image analyses for therapeutic, prognostic, or diagnostic decision support. As a result, the present disclosure describes an efficient whole-slide image analysis workflow for use in a digital pathology system to perform quantitative analysis of the entire tissue slides as well as intuitive visualization.

[0008] More particularly, the present workflow effectively provides precise, visual quantification of biomarker expressions to aid pathologists in providing accurate in-vitro diagnostics.

[0009] The digital pathology system provides automated digital slide creation, management, analysis, and viewing. It can be used for in vitro diagnostic as an aid to the pathologist in the display, detection, counting, review, and classification of tissues and cells of clinical interest based on particular morphology, color, intensity, size, pattern, and shape.

[0010] The workflow of the digital pathology system can automatically outline, or enable the user, such as a pathologist, to efficiently outline clinically relevant morphology (e.g., one or more solid tumor areas) over the entire slide containing the tissue sample. The quantitative analysis result can then efficiently and intuitively be provided to the user for the entire outlined clinically relevant morphology (which in some cases may contain thousands or millions of cells) within seconds. Furthermore, the present workflow enables the visualization of this vast amount of data effectively, which significantly improves the diagnostic and which provides information and confidence to the user about the biomarker expression levels.

[0011] The present workflow may reside on a workflow module. It may be executed on a remote or local pathology workstation, and may act as an interface with the user. The workflow may start by preparing a whole-slide image (or images) before the user starts
5 analyzing a case. To this end, the digital pathology system may acquire a scanned digital image of the entire slide that can be viewed and assessed by the user at various magnification levels.

[0012] In one embodiment, the user reviews the whole-slide image and manually separates relevant from non-relevant areas in
10 the whole-slide image. To this end, the user resorts to the digital pathology system to manually select one or more relevant areas, for example, by outlining the relevant areas using an input device such as a mouse or using a finger in case of a touch-sensitive display. Each outlined relevant area may be viewed at different
15 magnification levels. The selected relevant areas are also referred to herein as "whole tumor areas" or "clinically relevant morphology," to signify the ability to include within the selected relevant areas entire tumor areas along with optional surrounding regions of interest.

[0013] It should be amply clear that the user is no longer
20 restricted by specific geometric shapes, such as a rectangle, for defining and selecting the relevant areas of interest in the tissue sample. Rather, the present digital pathology system enables the user to perform a freehand selection of large relevant areas in the tissue sample, by outlining the relevant areas, such as by manually
25 sketching a boundary around the relevant areas of interest.

[0014] It should also be understood that other manual selection methods might be used. As an example, the user is enabled to annotate the whole-slide digital image in order to specify the areas of interest that require scoring. Alternatively, the selection process of
30 the relevant areas could be performed automatically by the digital pathology system.

[0015] Once the step of selecting the relevant areas is completed, the user invokes an image analysis module that initiates an image analysis algorithm of the selected relevant areas. The

5 image analysis module provides clinical scores based on cell detection and classification, for example, in the form of a percent positive immune cell score. In one embodiment, the clinical scores range between 0 and 100, with "0" indicating no expression of one or more biomarkers of interest expression and "100" indicating full expression of one or more biomarkers of interest. The scoring results represent large amounts of biological information, such as diagnostics (positive/negative) or informing treatment decisions.

10 [0016] A visualization module prepares computer-generated overlays of the image analysis results, and provides the user with a user-friendly visual rendering of the scoring results atop the originally selected whole-slide image, as a representation of density. The overlays can represent one or more maps, such as a heat map, a location map, a correlation map, etc., and can show variations of information using color and/or lines depending on the application.

20 [0017] Upon reviewing the visual rendering by the visualization module, the user can optionally invoke an artefact extraction module for removing undesired artefacts, such as folds, stroma separation artefact, speckling, blank areas (i.e., areas where substantially no tissue is present), etc. The artefacts removal might be implemented either manually by outlining the artefacts, or by resorting to an automated method that detects, visualizes, and excludes the artefact areas from the visual rendering.

25 [0018] The user can further optionally invoke a tissue content extraction module for removing undesired tissue or staining content, such as anthracotic pigment, necrosis, intra-vascular neutrophils, etc. The tissue or staining content removal might be implemented by either manually outlining the areas of interest, or by employing an automated method of tissue content detection and extraction.

30 [0019] The visual rendering will therefore represent large, accurate, and specific biological information pertaining to the clinically relevant areas that are selected by the user, such as solid tumor areas, but excludes data considered by the user to be extraneous, such as necrosis, and various artefacts and tissue

content. The visual rendering is expected to assist a qualified pathologist in providing a more accurate diagnostic interpretation.

[0020] The user then decides whether to accept the visual rendering or to manually override it either partially or in its entirety. If
5 override is selected, the results of any one or more of the image analysis module, the artefact extraction module, and the tissue content extraction module, can be selectively disregarded, and the user is provided with the ability to manually enter the final scores the
10 he or she derives from the assessment of the visual rendering provided by the visualization module.

BRIEF DESCRIPTION OF THE DRAWINGS

[0021] The patent or application file contains at least one drawing executed in color. Copies of this patent or patent
15 application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

[0022] The various features of the present invention and the manner of attaining them will be described in greater detail with reference to the following description, claims, and drawings wherein:

[0023] FIG. 1 illustrates an exemplary computer-based digital
20 pathology system that operates in a network environment and that hosts a workflow residing on a workflow module, for providing a visual quantitative analysis of a whole-slide image as well as intuitive visualization of the quantification of biomarker expressions, in accordance with one embodiment of the present invention;

25 [0024] FIG. 2 is an exemplary overview block diagram of the workflow module of FIG. 1;

[0025] FIG. 3 is an exemplary high level flow chart that illustrates the overall operation of the workflow module of FIGS. 1 and 2, for
30 enabling users to efficiently outline clinically relevant morphology, including all solid tumor areas, across the whole-slide image, in accordance with one embodiment of the present invention;

[0026] FIG. 4 is comprised of FIGS. 4A and 4B, and represents an exemplary flow chart of the overall operation of the workflow module of FIGS. 1 and 2, in accordance with an alternative embodiment of the present invention;

5 [0027] FIG. 5 illustrates an exemplary whole-slide image generated by a scanner that forms part of the digital pathology system of FIG. 1;

[0028] FIGS. 6 through 9 are exemplary screen shots of the whole-slide image shown in FIG. 5, illustrating the viewing environment, using different magnification powers, prior to invoking the workflow module of FIGS. 1 and 2;

10

[0029] FIG. 10 is a screen shot illustrating exemplary annotations added by the user to capture clinically relevant morphology or relevant areas within the whole-slide image of FIGS. 6 through 9, in preparation for invoking the workflow module of FIGS. 1 and 2;

15

[0030] FIG. 11 is an exemplary screen shot illustrating a visual rendering of the analysis performed by the workflow module of FIGS. 1 and 2, which provides the user with overlays that assist the user with the understanding of the whole slide pattern at low power magnification, as well as the understanding of the individual cell results at high power magnification, and which further provides the user with a quantitative scoring pattern of the relevant areas of FIG. 10;

20

[0031] FIGS. 12 through 14 are exemplary screen shots illustrating the ability provided to the user by the workflow module, to toggle between the visual rendering of FIG. 11 and the annotated view of FIG. 10, at varying power magnification levels, in order to provide the user with the ability to compare the views and to better assess the analysis provided by the workflow module of FIGS. 1 and 2;

25

[0032] FIG. 15 is an exemplary screen shot illustrating a high power magnification view of a designated clinically relevant morphology within the visual rendering of FIG. 11, wherein individual dots in the overlays mark individual classified cells, thus enabling the

30

- 7 -

user to understand the whole slide pattern at a low magnification power and the individual cell results at a high magnification power for a very granular comparison;

5 [0033] FIG. 16 is an exemplary screen shot of the same clinically relevant morphology of FIG. 15, with the overlays being toggled OFF, to allow for a better appreciation of the visual rendering by the workflow module of FIGS. 1 and 2;

10 [0034] FIG. 17 is an exemplary screen shot of the analysis rendering of FIG. 11, providing the user with the ability to either accept, override, or modify the scoring assessment by the workflow module of FIGS. 1 and 2;

15 [0035] FIG. 18 is an exemplary side-by-side comparison view of H&E (A) and PD-L1 (B) screen shot images, visualizing biomarker expressions in tumor exemplary samples, with image (B) showing an overlay on a PD-L1 bladder, wherein positive areas are represented by colored dots;

[0036] FIGS. 19 and 20 are exemplary screen shots of an exemplary colored overlay at a 2X magnification power level (FIG. 19) and at a 10X magnification power level (FIG. 20); and

20 [0037] FIGS. 21 through 26 illustrate exemplary views of the whole-slide heat map visualization process of the present invention.

[0038] It will be appreciated that for simplicity and clarity of illustration, reference numbers may be reused among the figures to indicate corresponding or analogous features.

25 **DETAILED DESCRIPTION**

30 [0039] FIG. 1 illustrates a computer-based digital pathology system 100 that operates in a network environment for providing a visual quantitative analysis of a whole-slide image as well as intuitive visualization of the quantification of biomarker expressions, in accordance with one embodiment of the present invention. The digital pathology system 100 interfaces with a plurality of client computer systems (or user stations) 140, 142 over a network 130.

[0040] The digital pathology system 100 may include, among other things, a stainer 110, a scanner 111, a workflow module 200 and a processor or computer 125. The users of the client computer systems 140, 142, such as pathologists, histotechnologists, or like professionals, may be able to access, view, and interface with the outputs of the scanner 111 and workflow module 200 on a real time basis, either remotely or locally. These outputs may alternatively be stored and accessed on networked databases 150.

[0041] As further detailed in FIG. 2, the workflow module 200 may generally include an image analysis module 210, a visualization module 220, an artefact extraction module 230, and a tissue content extraction module 240, the operation of which will be described in more detail.

[0042] The workflow module 200 may host a workflow 300 (FIG. 3) that is executed on the processor 125 and on one or more user stations 140, 142. The workflow 300 starts at step 305 by preparing a digital pathology slide (or slides) pursuant to the appropriate assay guidelines, including staining the tissue sample. In an exemplary implementation, images are acquired of one or more microscope glass slides containing tissue sections (e.g., formalin-fixed, paraffin-embedded (FFPE) tissue sections) that have been stained using, for example, immunohistochemical (IHC) staining, for the presence of PD-L1 protein, by employing the PD-L1 (SP142) assay developed by Ventana Medical Systems, Inc. The IHC slide can be stained, for example, on a BenchMark ULTRA stainer 110, also a product of Ventana Medical Systems, Inc. In addition to the IHC slide, in some embodiments, additional digital images may be acquired of slides containing adjacent tissue sections of the same tissue block can be stained with other IHC stains, a hematoxylin and eosin (H&E) stain, or using other types of staining techniques.

[0043] At step 310, the scanner 111 acquires a whole-slide image (or images) 500 (FIG.5) of the sample tissue, before the user starts analyzing a case. The scanned digital image 500 can be viewed and assessed by the user at various magnification levels. For example, the slide can be loaded into a VENTANA iScan HT whole-

slide scanner 111 that can generate the digital image 500 of the specimen in monochrome or multiple color channels. This digital image 500 may be stored, for example, in the databases 150. The batch control can also be scanned and available for review.

5 [0044] At step 315, the user may review the digital image of a relevant area (or clinically relevant morphology) 610 (FIG. 6), without the benefit of the image analysis provided by the workflow module 200. FIGS. 6 through 9 illustrate various screen shots 600, 700, 800, 900 of this relevant area 610, using different magnification levels, prior to
10 the invocation of the workflow module 200 of FIGS. 1 and 2. While the user reviews the digital images 600, he or she will begin to formulate overall positivity of the slide and to assess the relevancy of the stained cells. The user then defines (or identifies) relevant areas, such as tumor areas, and annotates the whole-slide image 500
15 accordingly. In some embodiments, the relevant areas may be identified automatically, without the user's input, as described in other parts of the present disclosure.

[0045] The image analysis module 210 of the workflow module 200 can independently analyze, at step 320, the content over the
20 entire slide. In some embodiments, analyzing "the entire slide" can mean analyzing all areas of the slide containing at least some tissue. That is, in some embodiments, analyzing the entire slide can include automatically detecting areas, within the whole-slide image, that contain no tissue (i.e., no biological material), excluding such areas
25 from the analysis, and analyzing only those areas that contain at least some tissue.

[0046] The image analysis module 210 can provide initial clinical scores based on cell detection and classification in the form of a percent positive immune cell score. In a preferred embodiment, the
30 clinical scores range between 0 and 100, with "0" indicating no expression of one or more biomarkers of interest expression and "100" indicating full expression of one or more biomarkers of interest. The scoring results represent large amounts of biological information, such as cell type, location, density, formation, etc.

[0047] As it will be explained later in greater detail, in connection with FIGS. 21 through 26, the visualization module 220 prepares computer-generated overlays of the image analysis results, and provides the user with a user-friendly visual rendering (or overlay) 1100 (FIG. 11) of the scoring results atop the originally selected whole-slide image, as a representation of the density of the biomarker expressions across the selected clinically relevant morphology. The overlays can represent one or more maps, such as a heat map, a location map, a correlation map, etc., and can show variations of information depending on the application. More specifically, the workflow module 200 provides a quantitative analysis of entire tissue slides as well as intuitive visualization thereof, including precise, visual quantification of biomarker expressions.

[0048] The image analysis module 210 analyses the entire slide offline, typically immediately after the image is scanned and becomes available to the image analysis module 210. The analysis may include determining locations of all objects of interest (e.g., cells) within the slide, classifying them (i.e., determining their type), and storing the location and classification data, along with any other information related to the objects of interest, in databases 150.

[0049] Depending on the size of the entire slide, the amount information that it contains, and the processing power of the analysis module 210, the above-described analysis of the entire slide may take significant time, sometimes as long as 10 minutes or longer. Accordingly, doing this analysis "offline" (i.e., beginning the analysis as soon as the slide image is scanned and before the slide is requested by the user) may enable the analysis to be complete by the time the user accesses the analysis for the first time for review and annotation. The image analysis module 210 anticipates the user's interaction with the resulting analysis, such as by adding annotations, e.g., drawing a boundary 1010 around the relevant clinically relevant morphology 1020 (FIG. 10).

[0050] Consequently, by the time that the user starts interacting with the resulting analysis, all computationally intensive steps such as locating and classifying cells can be completed with the results

being stored in databases 150, in which case the image analysis module 210 may only need to load the pre-calculated information from databases 150, and perform quantification operations, e.g., quantifying the cells of certain types within the selected clinically relevant morphology. Such quantification operations may be performed quickly enough to provide the user with "real-time" experience where the quantification results become visible within a very short time (e.g., within a second or within a few seconds).

[0051] In some embodiments, as part of the offline analysis, image analysis module 210 may also generate and store in databases 150 one or more whole-slide overlay images corresponding to the identified objects (e.g., cells) of different types. This way, when the user selects a region of clinically relevant morphology, image analysis module 210, instead of generating the overlay image(s), may load pre-generated overlay image(s) from databases 150, and display the portion of the overlay image(s) that correspond to the selected region of clinically relevant morphology.

[0052] At step 325 of the workflow 300, the user compares the annotated whole-slide image 500 (step 315) with the scored visual rendering 1100 (step 320) by, for example, toggling the heat map overlays ON and OFF and by comparing the overall slide scores of the visual rendering 1100 at multiple magnifications. As shown in the screen shot 1000 of FIG. 10, the user then annotates relevant areas of interest 1020 on the visual rendering 1000, such as by drawing a contour 1010 around the relevant areas 1020 (e.g., solid tumor areas and surrounding regions).

[0053] The contour 1010 represents an instruction to the workflow module 200 to exclusively limit the analysis to the relevant areas 1020 within the contour 1010. Since the workflow 300 requires assessment on a whole tumor level, the user's annotations are drawn to capture all relevant or suspect tumor areas. The user may need to refer back to the negative control or H&E in order to continue refining his or her understanding of the relevant areas to encompass within the contour 1010. If certain areas are irrelevant to the whole slide score, such as large areas of necrosis, the user can annotate the rendering

1000, or alternatively use one or more exclusion fields of view, to remove those nonrelevant areas from the analysis. Alternatively, the original contour 1010 can be redrawn.

5 [0054] At step 330, the user invokes the image analysis algorithm that is embedded on the image analysis module 210 of the workflow module 200, and forwards the user annotations 1010 thereto, in order to instruct the image analysis module 210 and the visualization module 220 to limit the analysis and rendering exclusively to the annotated relevant areas 1020, thus minimizing processing and rendering time, and expediting the analysis process.

10 [0055] At step 335, the workflow module 200 responds to the user's invocation by generating two main pieces of information for user review against the digital image rendering of a relevant area 610 (FIG. 6). The first piece of information includes an overlay 1120 (FIG. 11) wherein the workflow module 200 shows the locations of the positive PD-L1 immune cells. This data is based on the single cell classification level, so when any particular area has a high density of positive IC, the color coalesces to form the area visualized at low magnification power. As the user views the annotation overlay 1120 on higher power, individual dots mark the individual classified cells, giving the user the ability to understand the whole slide pattern at low magnification power and also to understand the individual cell results on high magnification power, for a very granular comparison.

20 [0056] As further illustrated in the screen shot 1200 of FIG. 12, this annotation overlay 1120 can be toggled ON and OFF, allowing the user to compare the digital image rendering 1200 which is devoid of image analysis, with the digital image rendering 1100 that contains the image analysis of the present invention, at low and high magnification levels. Reference is made to screen shots 1300, 1400 of FIGS. 13, 14, respectively. This allows the user to understand how the workflow module 200 classifies cells at high magnification power as well as the overall pattern of staining at low magnification power.

30 [0057] The second piece of information that is provided by the workflow module 200, enables the user to compare the image

analysis results against an overall slide metrics (or scores) 1130 (shown in the left margin of the screen shot 1100 of FIG. 11). The image analysis module 210 generates the overall quantitative score which is the cumulative proportion of positive immune cell area to the total relevant tumor area, and which is represented as a percentage. This calculation is the same process as one performs for the assay. In addition, the image analysis module 210 also indicates two other metrics including the total count of positive immune cells and the total tumor area. The immune cells count is converted to an area concept based on the average size of lymphocytes as well as other relevant immune cell metrics (i.e., macrophages).

[0058] With reference to the screen shot 1500 of FIG. 15, as the user views the annotation overlay 1120 of FIG. 11, on a higher magnification power, individual dots 1555 mark the individual classified cells, giving the user the ability to understand the whole slide pattern at a low magnification power, and also to understand the individual cell results on a high magnification power, for a very granular comparison.

[0059] As further illustrated in the screen shot 1600 FIG. 16, the annotation overlay 1120, including the individual dots 1555, can be toggled ON and OFF, allowing the user to compare the digital image rendering 1200 which is devoid of image analysis, with the digital image rendering 1100 that contains the image analysis of the present invention, at low and high magnification levels. This enables the user to understand how the image analysis module 210 classifies cells at higher magnification levels as well as the overall pattern of staining at lower magnification levels.

[0060] At step 340, following the foregoing user comparative review, the user determines whether to accept the results generated by the workflow module 200, at step 335, or whether to manually revise or override these results.

[0061] If override or revision is chosen and actually performed by the user, the workflow 300 proceeds to step 325, and continues to iterate the user input until the user is fully satisfied with the final digital

rendering. The final digital rendering can then be used by a professional practitioner for the intended purpose. The user will then sign out the case, and the selected results provided by the workflow module 200 are incorporated into a report.

5 [0062] FIG. 18 is a side-by-side exemplary comparison view of H&E (A) and PD-L1 (B) screen shot images 1800(A) and 1800(B), visualizing biomarker expressions in tumor exemplary samples, with image (B) showing an overlay on a PD-L1 bladder, wherein positive areas are represented by colored dots.

10 [0063] FIGS. 19 and 20 are exemplary screen shots 1900, 2000, respectively, of an exemplary colored overlay at a 2X magnification power level (FIG. 19) and at a 10X magnification power level (FIG. 20).

15 [0064] Considering now FIG. 4, it is comprised of FIGS. 4A and 4B, and represents a workflow 400 that is hosted on the workflow module 200 of FIGS. 1 and 2, in accordance with an alternative embodiment of the present invention. Similarly, to step 305, the workflow 400 starts at step 405 by preparing a digital pathology slide (or slides) pursuant to the appropriate assay guidelines, including staining the tissue sample.

20 [0065] At step 410, and similarly to step 310, the scanner 111 acquires a whole-slide image (or images) 500 (FIG.5) of the sample tissue, before the user starts analyzing a case. The scanned digital image 500 can be viewed and assessed by the user at various magnification levels. This digital image 500 may be stored and accessed on the databases 150. The batch control is also scanned and available for review.

25 [0066] At step 415, the user logs into the digital pathology system 100 and reviews the stored control images. In addition, at step 420, the user reviews the digital image 500 of the stained slide without the benefit of the image analysis that is provided by the workflow module 200.

30

[0067] At step 425, the user separates the relevant areas from the non-relevant areas in the whole-slide digital image 500. To this end, the user uses the digital pathology system 100 to manually select the relevant areas of interest by, for example, outlining the relevant areas. Each relevant area may selectively be viewed at a different magnification level.

[0068] It should be amply clear that the user is no longer restricted by specific geometric shapes, such as a rectangle, for defining and selecting the relevant areas of interest in the tissue sample. Rather, the present digital pathology system enables the user to perform a freehand selection of large relevant areas in the tissue sample, by outlining these relevant areas, such as by manually drawing a boundary 1010 around the relevant areas of interest 1020, as explained earlier in connection with FIG. 10.

[0069] It should also be understood that other manual selection methods might be used. As an example, the user is enabled to annotate the whole-slide digital image in order to specify the areas of interest that require scoring. Alternatively, the selection process of the relevant areas could be performed automatically by the digital pathology system 100. For example, digital pathology system 100 may automatically detect one or more areas within the whole-slide digital image containing tumor (e.g., based on any known machine-learning or deep-learning technique, such as using a neural network trained to detect tumor areas) and designate such area(s) as relevant areas.

[0070] Once the selection of the relevant areas is completed at step 420, the workflow 400 proceeds to step 430, to enable the user to invoke the image analysis module 210 of the workflow module 200, in order to initiate an image analysis algorithm of the selected relevant areas 1010. The image analysis module 210 provides clinical scores based on cell detection and classification in the form of a percent positive immune cell score, ranging between 0 and 100. The scoring results represent large amounts of biological information, such as cell type, location, density, formation, etc.

[0071] At step 435 of FIG. 4B, the visualization module 220 responds to the user's instructions and prepares computer-generated overlays 1120 (FIG. 11) of the image analysis results, and provides the user with a user-friendly visual rendering 1100 of the scoring results atop the originally selected whole-slide image, as a representation of density. The overlays can represent one or more maps, such as a heat map, a location map, a correlation map etc., and can show variations of information depending on the application, as explained earlier in connection with step 335.

[0072] At step 440, the user verifies the resulting scores and annotates the visual rendering 1100, as needed.

[0073] At step 445, the user invokes the artefact extraction module 230 (FIG. 2) which enables the user to select undesired image artefacts, such as folds, stroma separation artefact, speckling, etc. To this end, the user may either perform the selection by manually outlining the undesired artefacts (or artefact areas), or by implementing an automated method that detects, visualizes these undesired artefacts. The artefact extraction module 230 then excludes the undesired artefacts from the resulting image analysis and visualization by the image analysis module 210 and the visualization module 220.

[0074] At step 450, the user invokes the tissue content extraction module 240 (FIG. 2) which enables the removal of undesired tissue and staining content, such as anthracotic pigment, necrosis, intra-vascular neutrophils, etc. To this end, the user may either perform the selection by manually outlining the undesired content areas, or by implementing an automated method that detects, visualizes these undesired content areas. The tissue content extraction module 240 then excludes the undesired artefacts from the resulting image analysis and visualization by the image analysis module 210 and the visualization module 220.

[0075] At decision step 455, the user determines whether or not to accept the scoring results generated by the workflow module 200. If the scoring results are accepted, then the workflow module 200

finalizes the digital rendering at step 460. The user will then sign out the case, and the selected results provided by the workflow module 200 are incorporated into a report.

5 [0076] If, however, the user selects override or revision of the scoring results at step 455, the workflow 400 proceeds to step 465, and enters the revised scores. The workflow module 200 continues to iterate the user input until the user is fully satisfied with the final digital rendering. The final digital rendering is then finalized and the final results provided by the workflow module 200 are incorporated into a
10 report at step 460. The user will then sign out the case, and the report will be made available to a professional practitioner for the intended purpose.

[0077] It should be clear that the present invention is not limited
15 to the detection and visualization of one type of cells, but it can rather analyze entire tumor areas containing various types of cells. Each type of cell may be visualized with a different color.

[0078] The whole-slide heat map visualization process of the present invention will now be described in more detail, in connection with FIGS. 21 through 26. In general, the whole-slide heat map
20 overlay is a method of visualizing digital pathology images in order to show areas where a higher density or cluster of biomarker expression occurs. The heat map visualization process provides the user with the ability to visually perceive the density of biomarker expression independently of the zoom factor at low image magnification (e.g.,
25 4X). The whole-slide heat maps are generated at low magnification to indicate the density of the cell detection results from the whole-slide image. A color gradient is used to indicate areas of increasingly higher density, wherein each cell is assigned a density value and the entire layer is visualized using a gradient. The interpolation of discrete
30 points creates a continuous surface that is referred to as a density surface. The end visualization which affects how the data is interpreted by the viewer is subjective.

[0079] FIGS. 21 through 24 illustrate examples of a digital pathology image 2100 (FIG. 21) that comprises a clinically relevant

morphology (or area of interest to the user) 2111. FIG. 21 shows the clinically relevant morphology 2111 at a low magnification level, while FIG. 22 illustrates a highly magnified view of this clinically relevant morphology 2111. While the clinically relevant morphology 2111 is illustrated as having a generally square shape, it should be understood that the user may select the morphology of interest using either regular or irregular contours.

[0080] The image analysis algorithm that is embedded on the image analysis module 210 of the workflow module 200 is performed onto the clinically relevant morphology 2111 of FIG. 22, to provide the cell detection results (or map) 2300, which is shown in FIG. 23 as green dots 2323. After obtaining the cell detection results for the whole-slide image, the cell detection results are used to generate the whole-slide heat map (or heat map overlay) 2400 (FIG. 24) corresponding to the density of the biomarker expression. FIG. 24 shows the whole-slide heat map overlay 2400 on the original whole-slide image 2100 (FIG. 21). The color gradient is used to indicate areas of increasingly higher density.

[0081] In addition to heat map concept of visualizing the density of biomarker expression, the present invention describes a visualization process that enables the detection and identification of multiple cell types within the whole-slide 2100. Rather than using color to provide an indication of the quantity of a biomarker expression in the tissue, color will now be used to represent cell types, such as tumor or immune cells, and whether they express for a certain biomarker or not, i.e., positive or negative.

[0082] FIGS. 25, 26 visualize examples for a bladder and lung cancer tissue where PD-L1 positive immune cells as well as PD-L1 positive/negative tumor cells are visualized as detected by the image analysis algorithm of the present invention. In this case, each cell type is assigned a different color and the density of each cell type is visualized by using a color intensity gradient that smoothly assigns a lower color intensity value to locations further away from the center location of the respective cell type. Consequently, high density areas will be represented by a continuously strong color

region and areas of mixed cell types will be represented by new colors that result from contributions of all different cell types in the surrounding area.

5 [0083] FIG. 25 illustrates an example of positive immune cells 2550 detected on a PD-L1 bladder slide 2500. These positive immune cells 2550 are represented by the green areas. FIG. 26 illustrates an example of PD-L1 positive immune cells 2650 (green areas), PD-L1 positive tumor cells 2655 (red areas), and PD-L1 negative immune cells 2660 (blue areas). Higher density areas of a certain cell type
10 appear grouped and mixed areas appear in mixed color.

[0084] In each of the flow charts described herein, one or more of the methods may be embodied in a computer readable medium containing computer readable code such that a series of steps are performed when the computer readable code is executed on a
15 computing device. In some implementations, certain steps of the methods are combined, performed simultaneously or in a different order, or perhaps omitted, without deviating from the spirit and scope of the invention. Thus, while the method steps are described and illustrated in a particular sequence, the use of a specific
20 sequence of steps is not meant to imply any limitations on the invention. Changes may be made with regards to the sequence of steps without departing from the spirit or scope of the present invention. The use of a particular sequence is therefore, not to be taken in a limiting sense, and the scope of the present invention is
25 defined only by the appended claims.

[0085] As it will be appreciated by one skilled in the art, aspects of the present invention may be embodied as a system, method, or computer program product. Accordingly, aspects of the present invention may take the form of an entirely hardware embodiment, an entirely software embodiment (including firmware, resident
30 software, micro-code, etc.) or an embodiment combining software and hardware aspects that may all generally be referred to herein as a "circuit," "module" or "system." Furthermore, aspects of the present invention may take the form of a computer program product

embodied in one or more computer readable medium(s) having computer readable program code embodied thereon.

[0086] As it will be further appreciated, the processes in embodiments of the present invention may be implemented using any combination of software, firmware or hardware. As a preparatory step to practicing the invention in software, the programming code (whether software or firmware) will typically be stored in one or more computer readable storage mediums for example, but not limited to, an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, or device, or any suitable combination of the foregoing. More specific examples (a non-exhaustive list) of the computer readable storage medium would include the following: an electrical connection having one or more wires, a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), a portable compact disc read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any suitable combination of the foregoing. In the context of this document, a computer readable storage medium may be any tangible medium that can contain, or store a program for use by or in connection with an instruction execution system, apparatus, or device.

[0087] The article of manufacture containing the programming code is used by either executing the code directly from the storage device, by copying the code from the storage device into another storage device such as a hard disk, RAM, etc., or by transmitting the code for remote execution using transmission type media such as digital and analog communication links. The methods of the invention may be practiced by combining one or more machine-readable storage devices containing the code according to the present invention with appropriate processing hardware to execute the code contained therein. An apparatus for practicing the invention could be one or more processing devices and storage systems containing or having network access to program(s) coded in accordance with the invention.

[0088] A computer readable signal medium may include a propagated data signal with computer readable program code embodied therein, for example, in baseband or as part of a carrier wave. Such a propagated signal may take any of a variety of forms, including, but not limited to, electro-magnetic, optical, or any suitable combination thereof. A computer readable signal medium may be any computer readable medium that is not a computer readable storage medium and that can communicate, propagate, or transport a program for use by or in connection with an instruction execution system, apparatus, or device.

[0089] Program code embodied on a computer readable medium may be transmitted using any appropriate medium, including but not limited to wireless, wireline, optical fiber cable, R.F., etc., or any suitable combination of the foregoing. Computer program code for carrying out operations for aspects of the present invention may be written in any combination of one or more programming languages, including an object oriented programming language such as Java, Smalltalk, C++ or the like and conventional procedural programming languages, such as the "C" programming language or similar programming languages. The program code may execute entirely on the user's computer, partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider).

[0090] Thus, it is important that while illustrative embodiments of the present invention are described in the context of a fully functional computer (server) system with installed (or executed) software, those skilled in the art will appreciate that the software aspects of the illustrative embodiments of the present invention are capable of being distributed as a program product in a variety of forms, and that an illustrative embodiment of the present invention

applies equally regardless of the particular type of media used to actually carry out the distribution.

[0091] In addition, while the present invention has been described with reference to exemplary embodiments, it will be understood by those skilled in the art that various changes may be made and equivalents may be substituted for elements thereof without departing from the scope of the invention. Furthermore, many modifications may be made to adapt a particular system, device or component thereof to the teachings of the invention without departing from the essential scope thereof. Therefore, it is intended that the invention not be limited to the particular embodiments disclosed for carrying out this invention, but that the invention will include all embodiments falling within the scope of the appended claims.

[0092] As used herein, the singular forms "a", "an", and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will be further understood that the terms "comprises" and/or "comprising," when used in this specification, specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof. Moreover, the use of the terms first, second, etc. do not denote any order or importance, but rather the terms first, second, etc. are used to distinguish one element from another. In addition, listing terms such as "a", "b", "c", "first", "second", and "third" are used herein and in the appended claims for purposes of description and are not intended to indicate or imply relative importance or significance.

[0093] The corresponding structures, materials, acts, and equivalents of all means or step plus function elements in the claims below are intended to include any structure, material, or act for performing the function in combination with other claimed elements as specifically claimed. The description of the present invention has been presented for purposes of illustration and description, but is not intended to be exhaustive or limited to the invention in the form

disclosed. Many modifications and variations will be apparent to those of ordinary skill in the art without departing from the scope and spirit of the invention. The embodiment was chosen and described in order to best explain the principles of the invention and the practical application, and to enable others of ordinary skill in the art to understand the invention for various embodiments with various modifications as are suited to the particular use contemplated.

CLAIMS

What is claimed is:

1. A digital pathology system that hosts a workflow for providing a quantitative analysis of an entire tissue slide and of biomarker expressions across a clinically relevant morphology contained within the entire tissue slide, the workflow comprising:

an image analysis module that generates an image analysis of the entire tissue slide, in preparation for access by the user;

- a visualization module that enables a user to select any of a part or the entirety of the clinically relevant morphology; and

wherein, in response to a user selection of the clinically relevant morphology, the visualization module limits the image analysis exclusively to the selected clinically relevant morphology, and generates a visual overlay that includes a set of quantifiable scores across the selected clinically relevant morphology, as a representation of the biomarker expressions across the selected clinically relevant morphology.

2. The digital pathology system of claim 1, wherein the set of quantifiable scores is a representation of a density of the biomarker expressions across the selected clinically relevant morphology.

3. The digital pathology system of claim 1 or 2, wherein the clinically relevant morphology includes at least one tumor area.

4. The digital pathology system of claim 1 or 2, wherein the clinically relevant morphology includes a plurality of different solid tumor areas.

5. The digital pathology system of any of claims 1 to 4, wherein the visual overlay represents an overlay map.

6. The digital pathology system of claim 5, wherein the overlay map includes at least one of a heat map, a location map, and a correlation map.

7. The digital pathology system of any of claims 1 to 6, wherein the image analysis module accepts user inputs; and

the visualization module revises the visual overlay accordingly.

8. The digital pathology system of any of claims 1 to 7, further comprising a scanner;

wherein the scanner scans the entire tissue slide to generate a scanned image of the entire tissue slide; and

wherein the image analysis module generates the image analysis within a predetermined time after the generation of the scanned image, and before the user selection of the clinically relevant morphology.

9. The digital pathology system of claim 8, wherein generating the image analysis comprises determining locations and types of a plurality of cells within the scanned image of the entire tissue slide.

10. The digital pathology system of claim 9, wherein the digital pathology system further comprises a database, and wherein the image analysis module stores the determined locations and types of the plurality of cells in the database to be accessed after the user selection of the clinically relevant morphology.

11. The digital pathology system of claim 9, wherein generating the image analysis further comprises generating at least one overlay image based on the determined locations and types of the plurality of cells, and storing the at least one overlay image.

12. The digital pathology system of any of claims 1 to 11, further comprising an artefact extraction module that enables the user to select undesired image artefacts for exclusion from the analysis of the visualization module, wherein the undesired image artefacts include at least one of a fold, a stroma separation artefact, and a speckling.

13. The digital pathology system of claim 11, further comprising a tissue content extraction module that enables the user to select

undesired tissue and staining content for exclusion from the analysis of the visualization module.

14. The digital pathology system of claim 13, wherein the undesired tissue and staining content includes at least an
5 anthracotic pigment, a necrosis, and an intra-vascular neutrophils.

15. A digital pathology method for providing a quantitative analysis of an entire tissue slide, and of biomarker expressions across a clinically relevant morphology contained within the entire tissue slide, the method comprising:

10 generating an image analysis of the entire tissue slide, in preparation for access by the user;

automatically selecting or enabling a user to select the clinically relevant morphology;

15 performing an image analysis of the selected clinically relevant morphology, wherein the image analysis is substantially limited to the selected clinically relevant morphology; and

20 generating a visual overlay that includes a set of quantifiable scores across the selected clinically relevant morphology, as a representation of the biomarker expressions across the selected clinically relevant morphology.

16. The digital pathology method of claim 15, wherein the set of quantifiable scores is a representation of a density of the biomarker expressions across the selected clinically relevant morphology.

17. The digital pathology method of claim 15 or 16, wherein the
25 clinically relevant morphology includes at least one tumor area.

18. The digital pathology method of any of claims 15 to 17, wherein automatically selecting the clinically relevant morphology includes processing the entire tissue slide by a neural network trained to detect tumor areas.

19. The digital pathology method of any of claims 15 to 18, wherein the visual overlay represents an overlay map.

20. The digital pathology method of claim 19, wherein the overlay map includes any one of a heat map, a location map, or a correlation map.

21. The digital pathology method of any of claims 15 to 20, further including accepting user inputs; and

revising the visual overlay accordingly.

22. A computer program product located on a processor readable medium for providing a quantitative analysis of an entire tissue slide, and of biomarker expressions across a clinically relevant morphology contained within the entire tissue slide, the computer program product including processor-executable instruction codes and comprising:

a first set of instruction codes for generating an image analysis of the entire tissue slide, in preparation for access by the user;

a second set of instruction codes for enabling a user to select any of a part or the entirety of the clinically relevant morphology; and

wherein, in response to a user selection of the clinically relevant morphology, a third set of instruction codes for limiting the image analysis exclusively to the selected clinically relevant morphology, and for generating a visual overlay that includes a set of quantifiable scores across the selected clinically relevant morphology, as a representation of the biomarker expressions across the selected clinically relevant morphology.

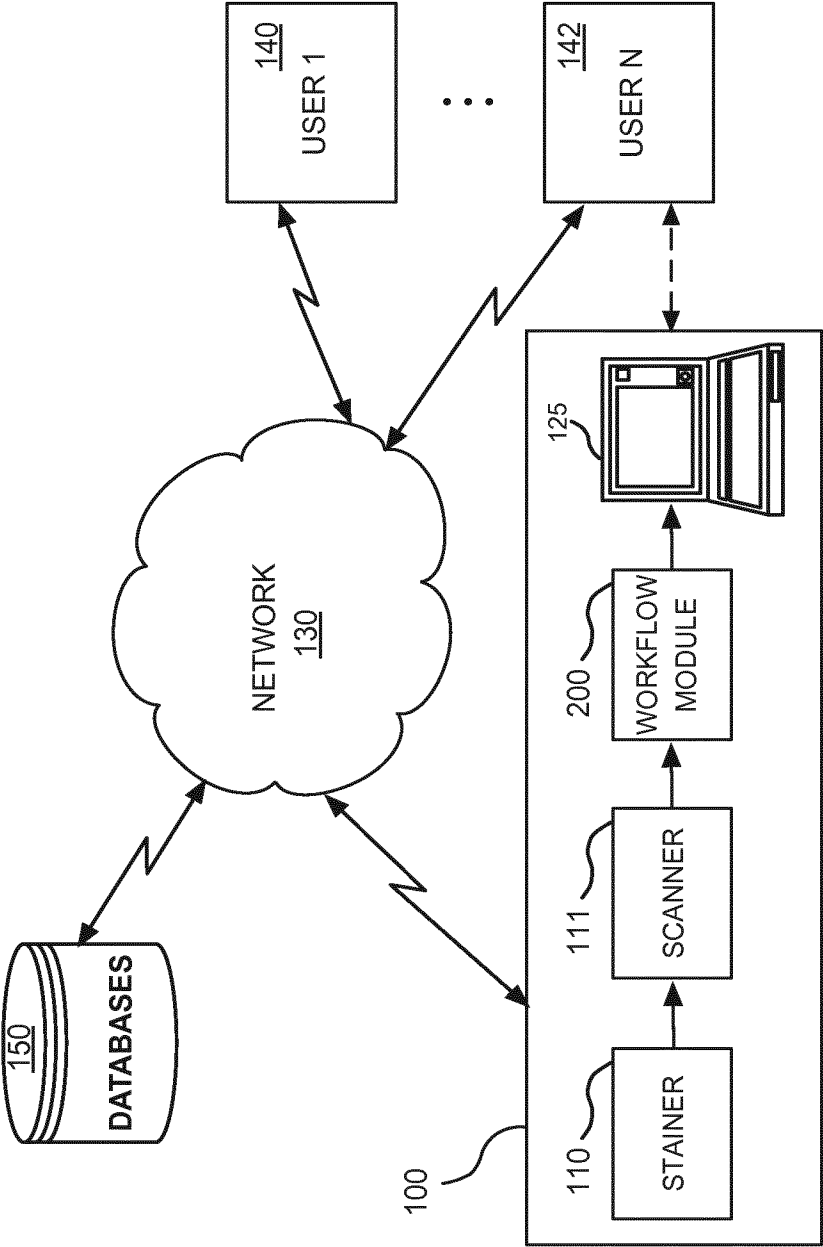


FIG. 1

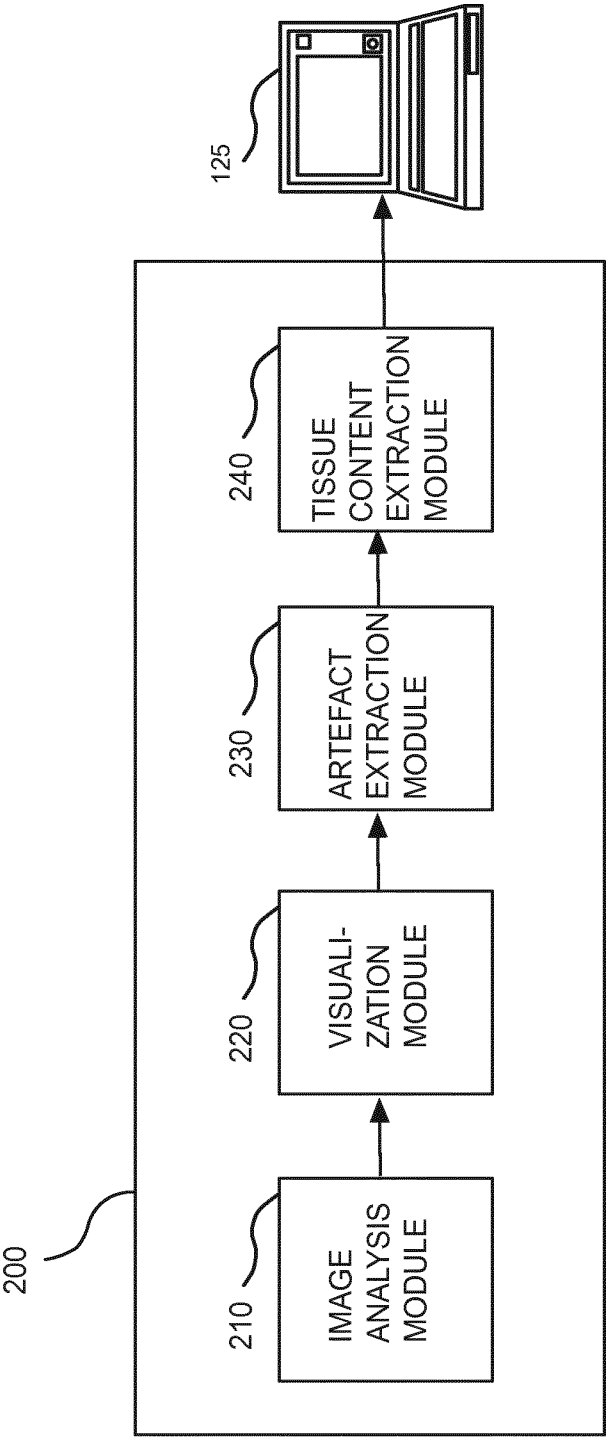


FIG. 2

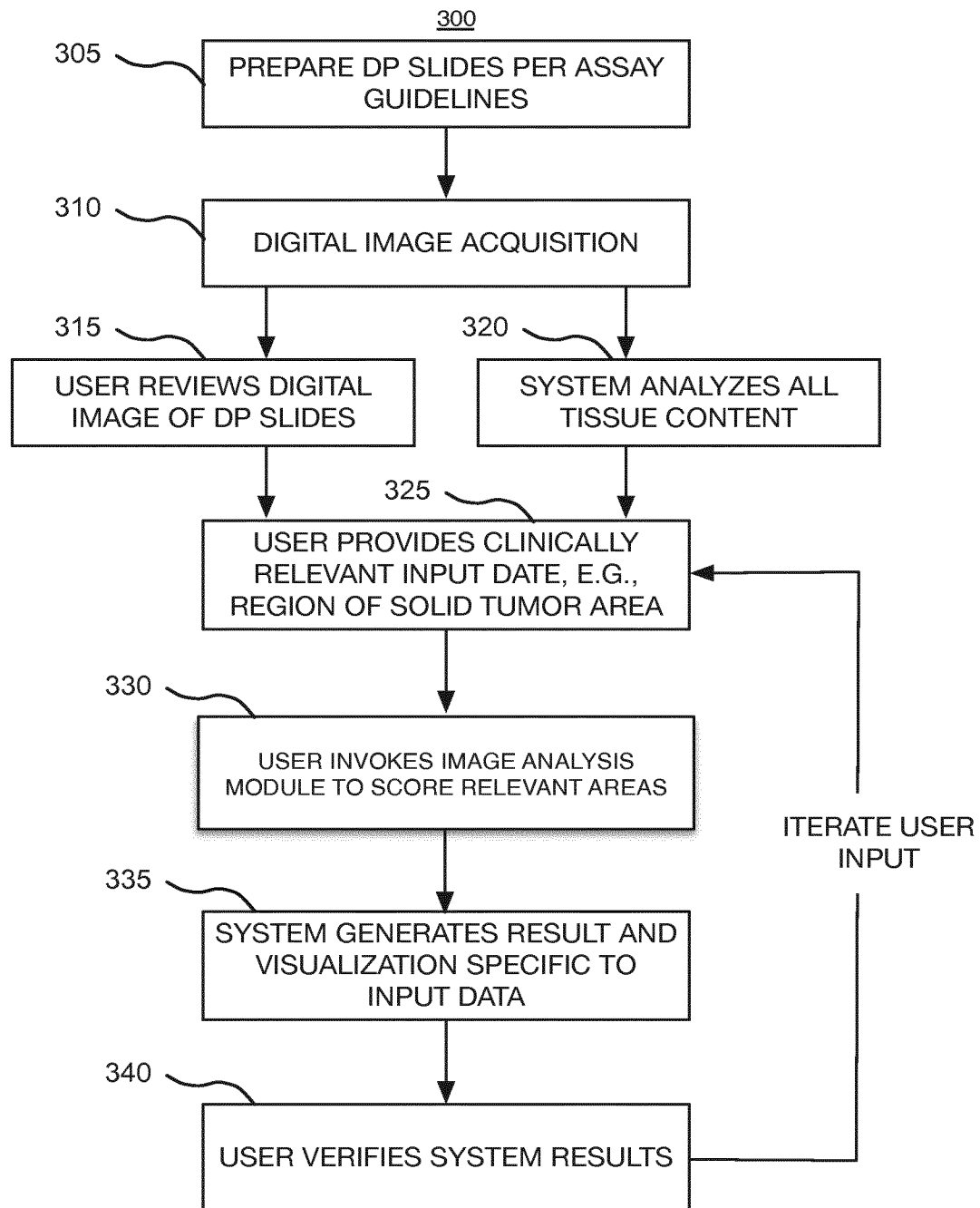


FIG. 3

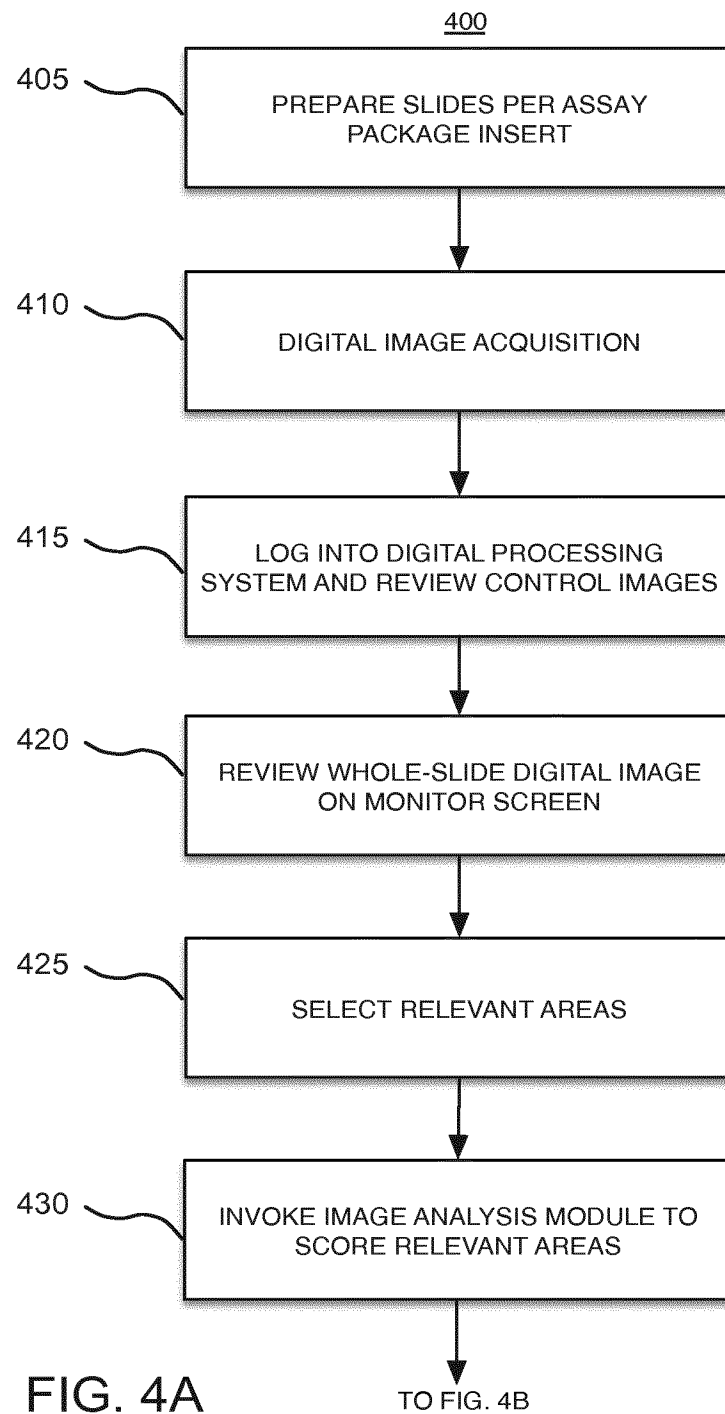


FIG. 4A

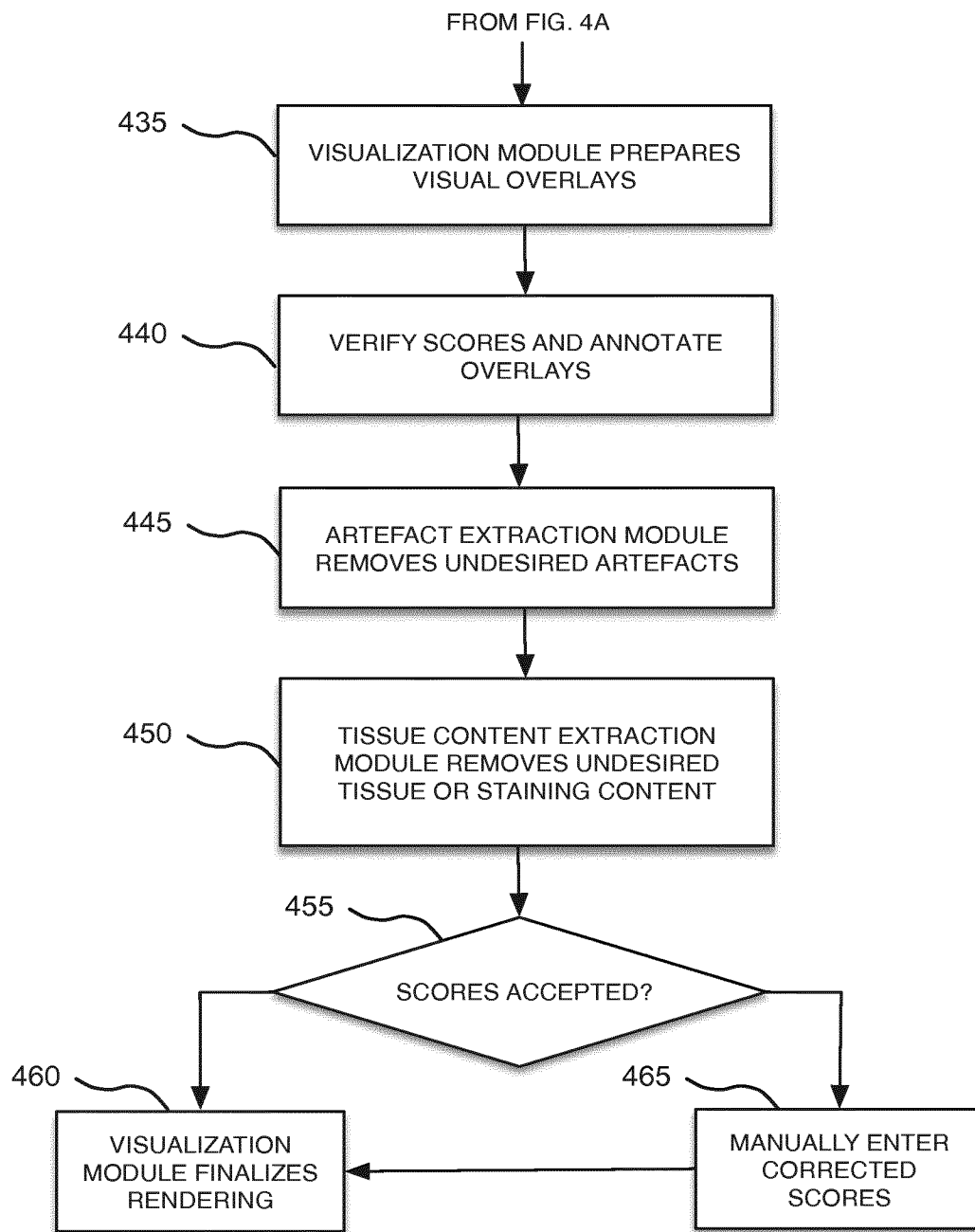


FIG. 4B

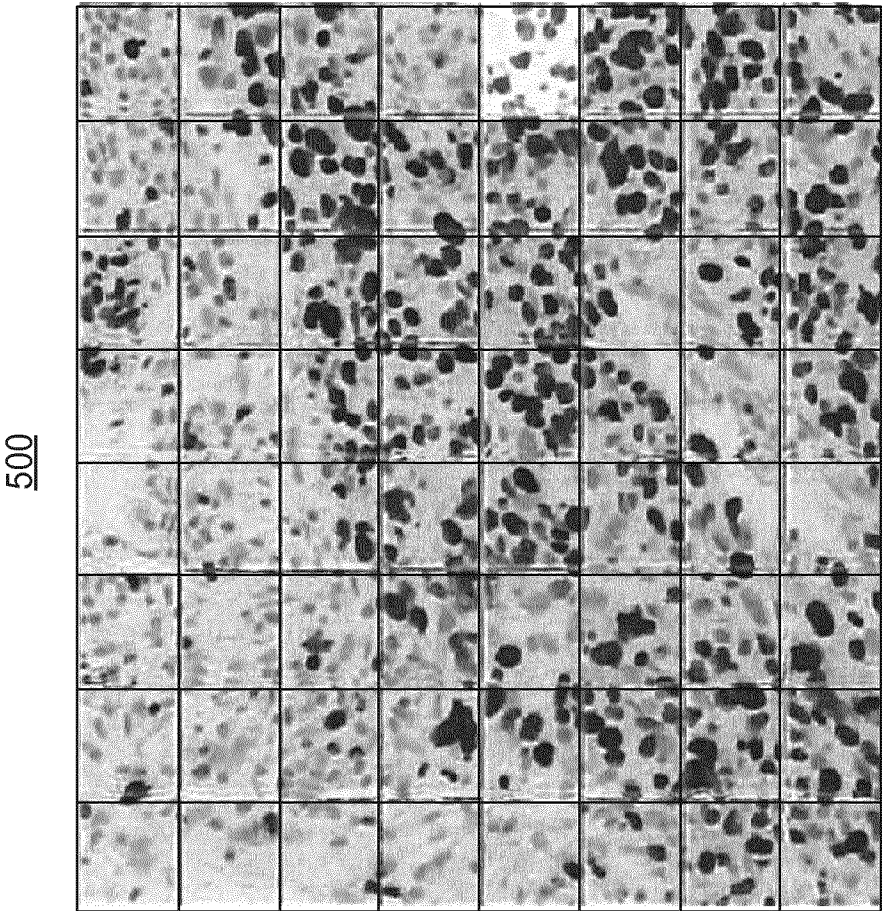


FIG. 5

600

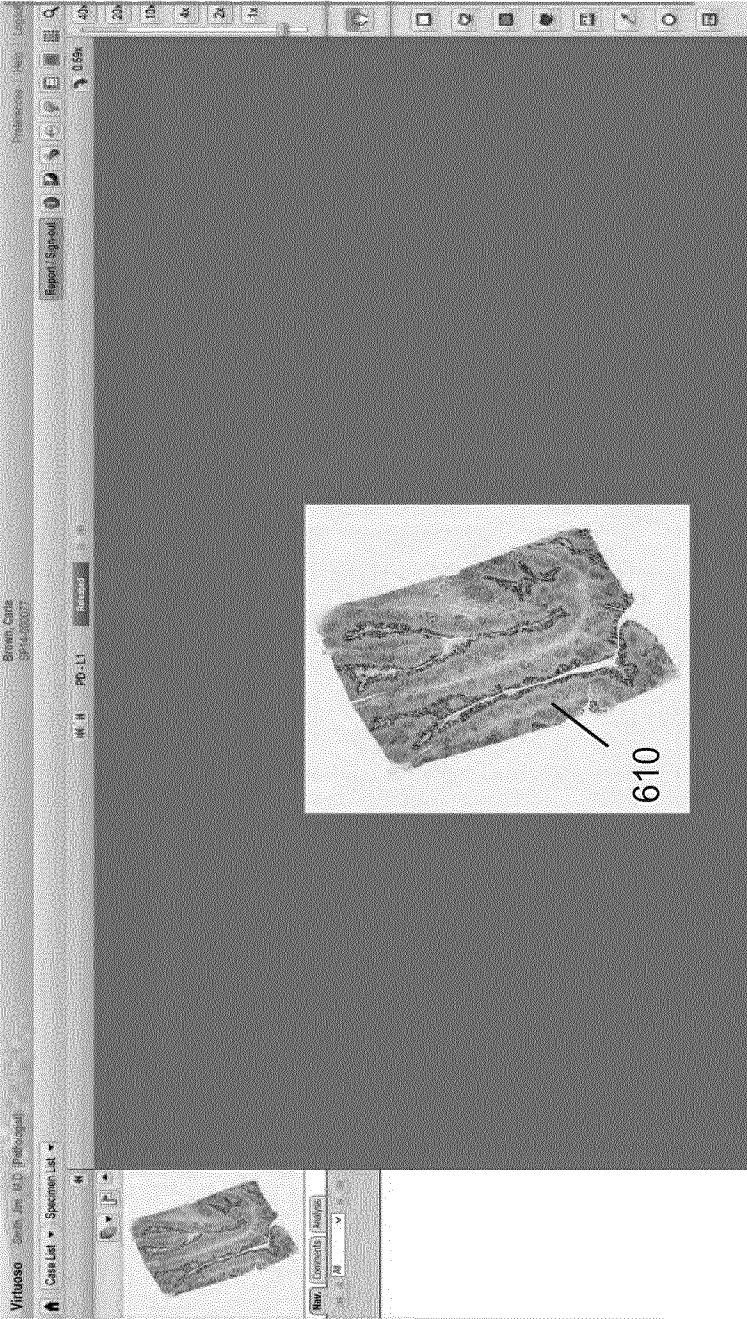


FIG. 6

700

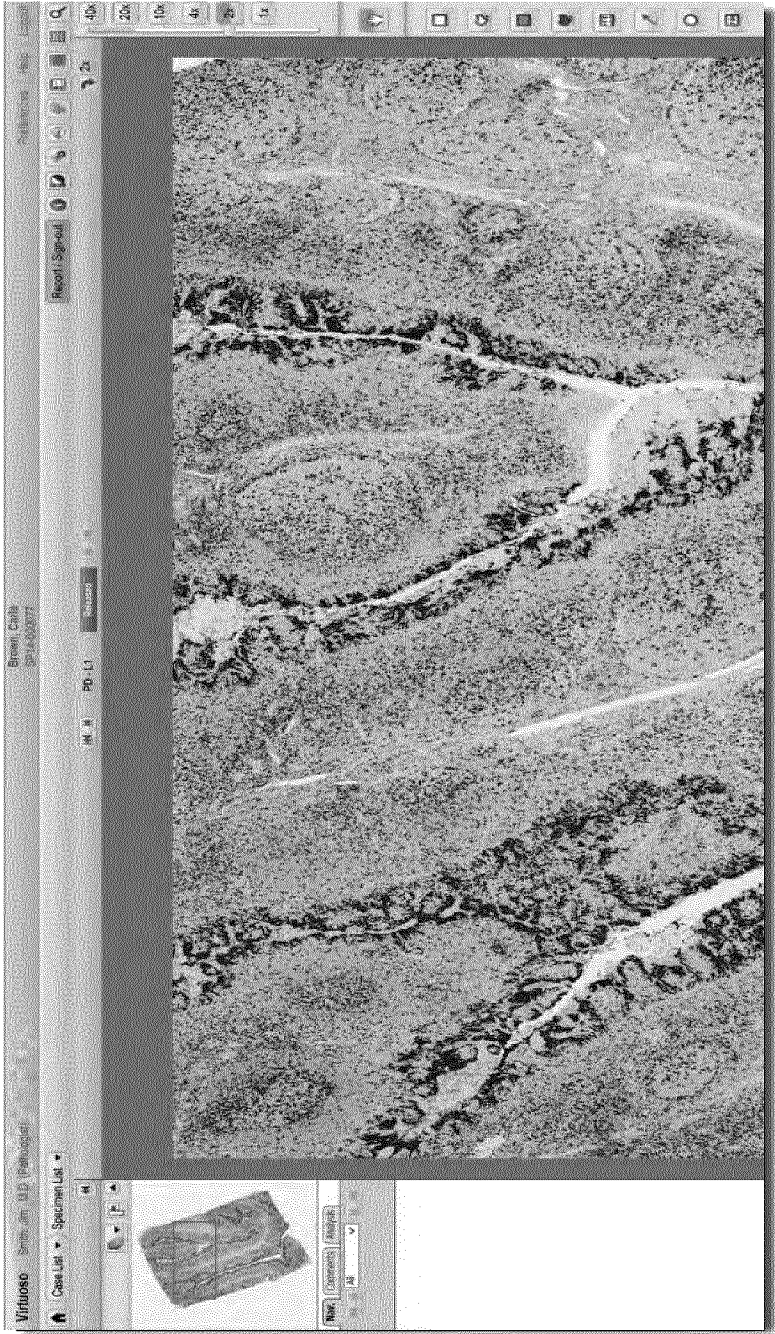


FIG. 7

800

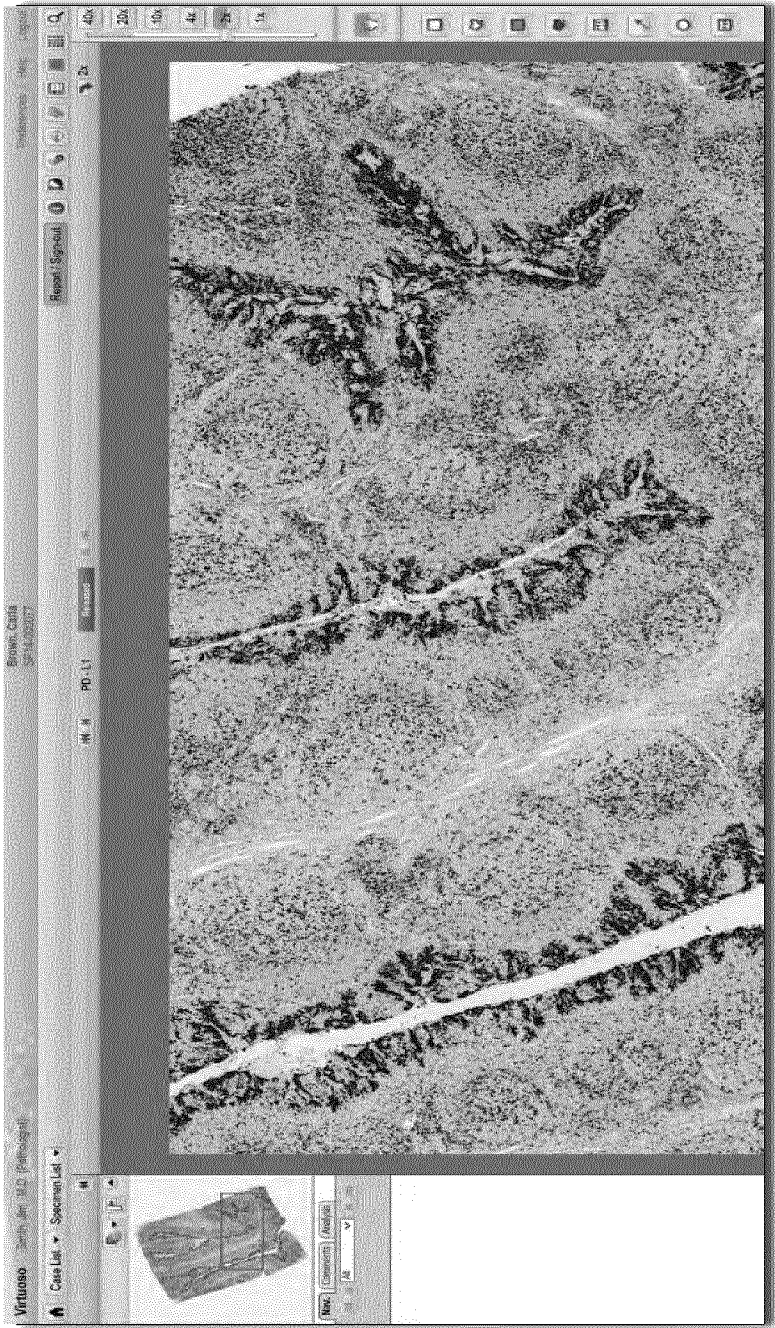


FIG. 8

900

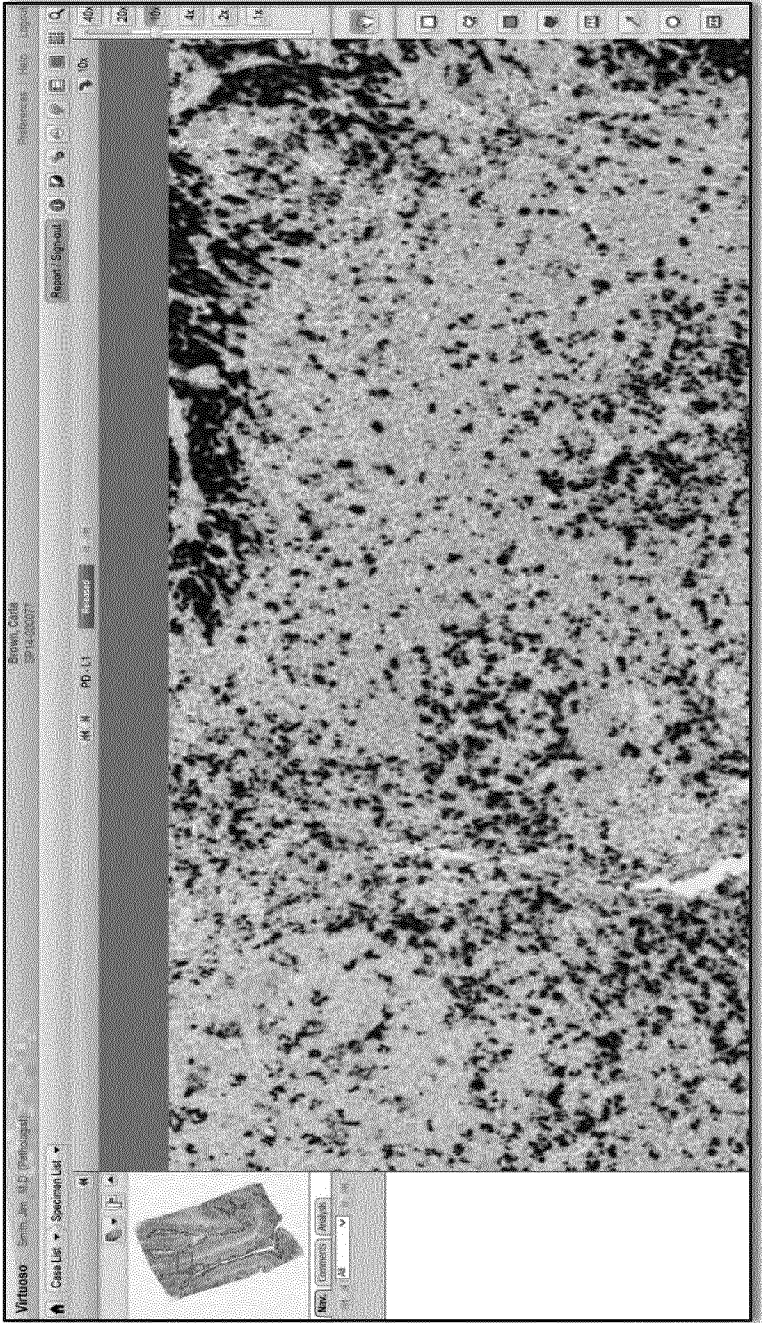


FIG. 9

1000

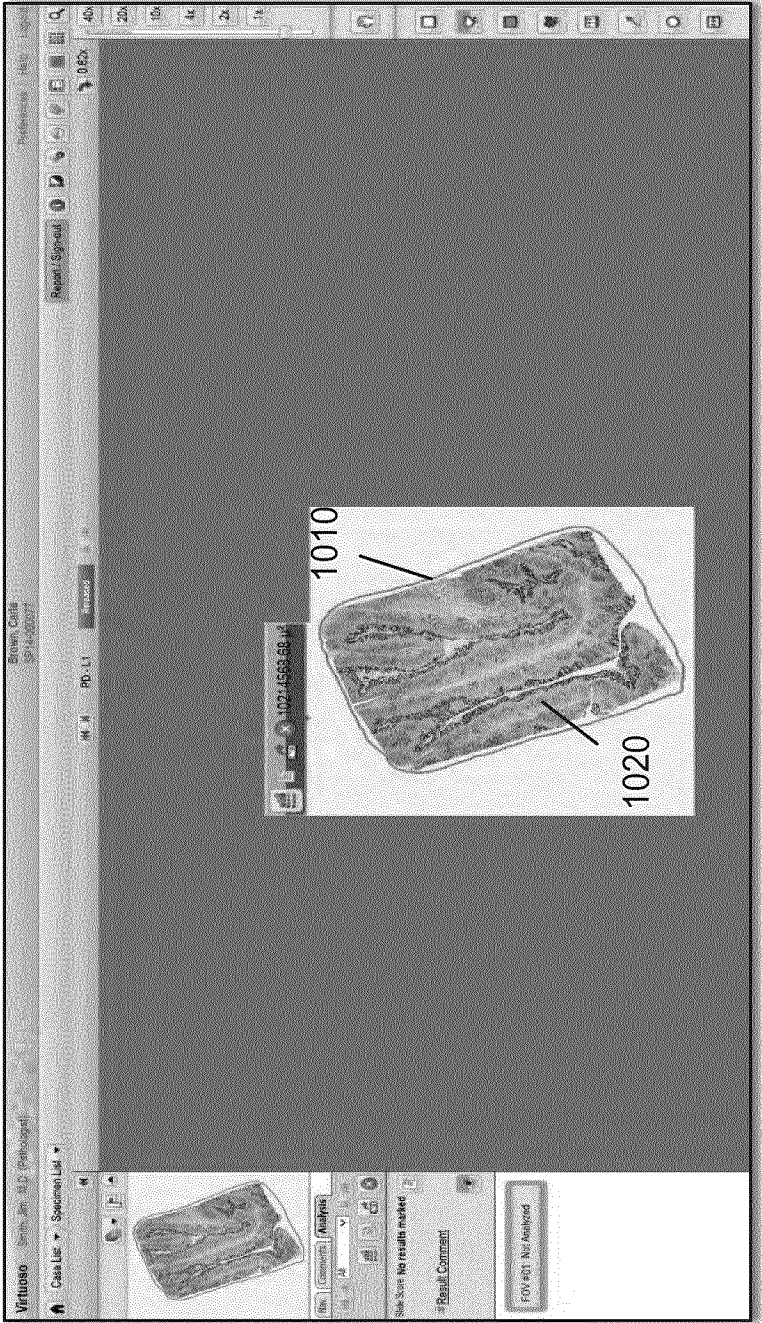


FIG. 10

1100

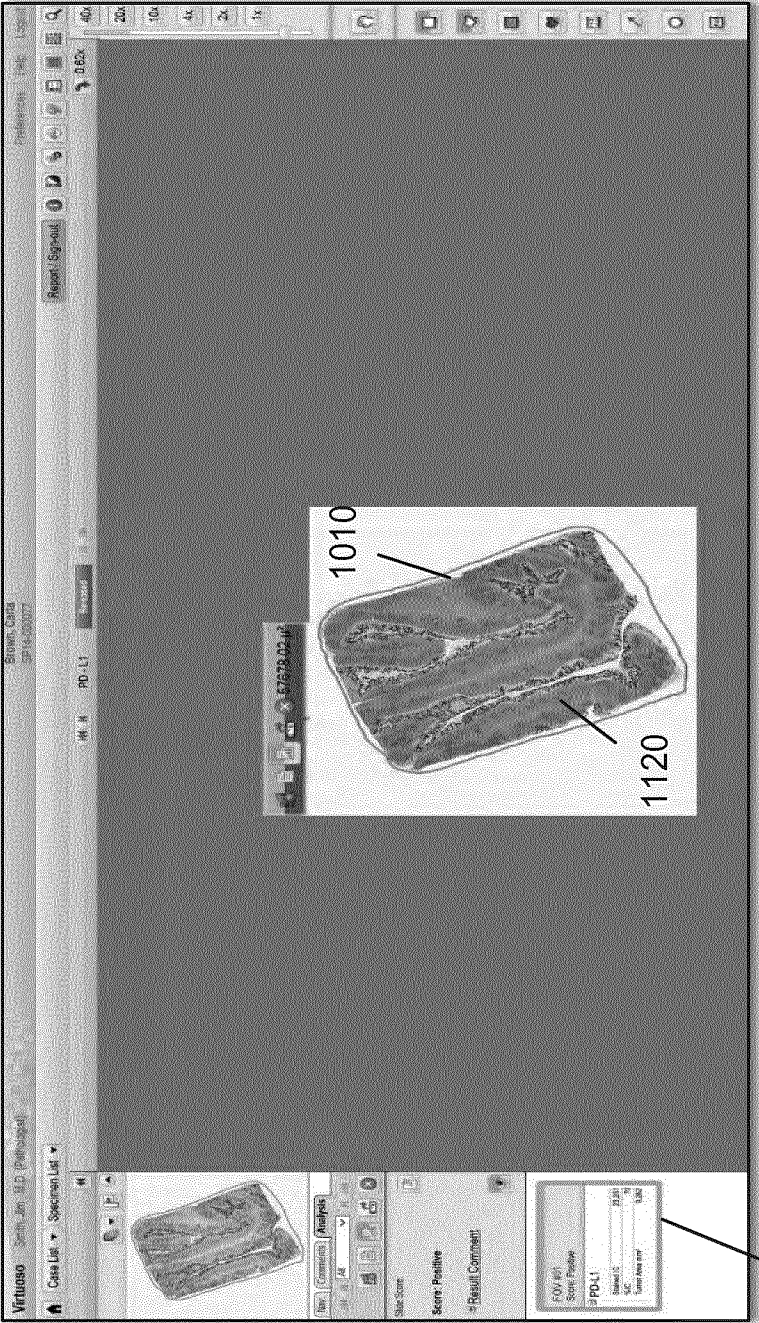


FIG. 11

1130

1200

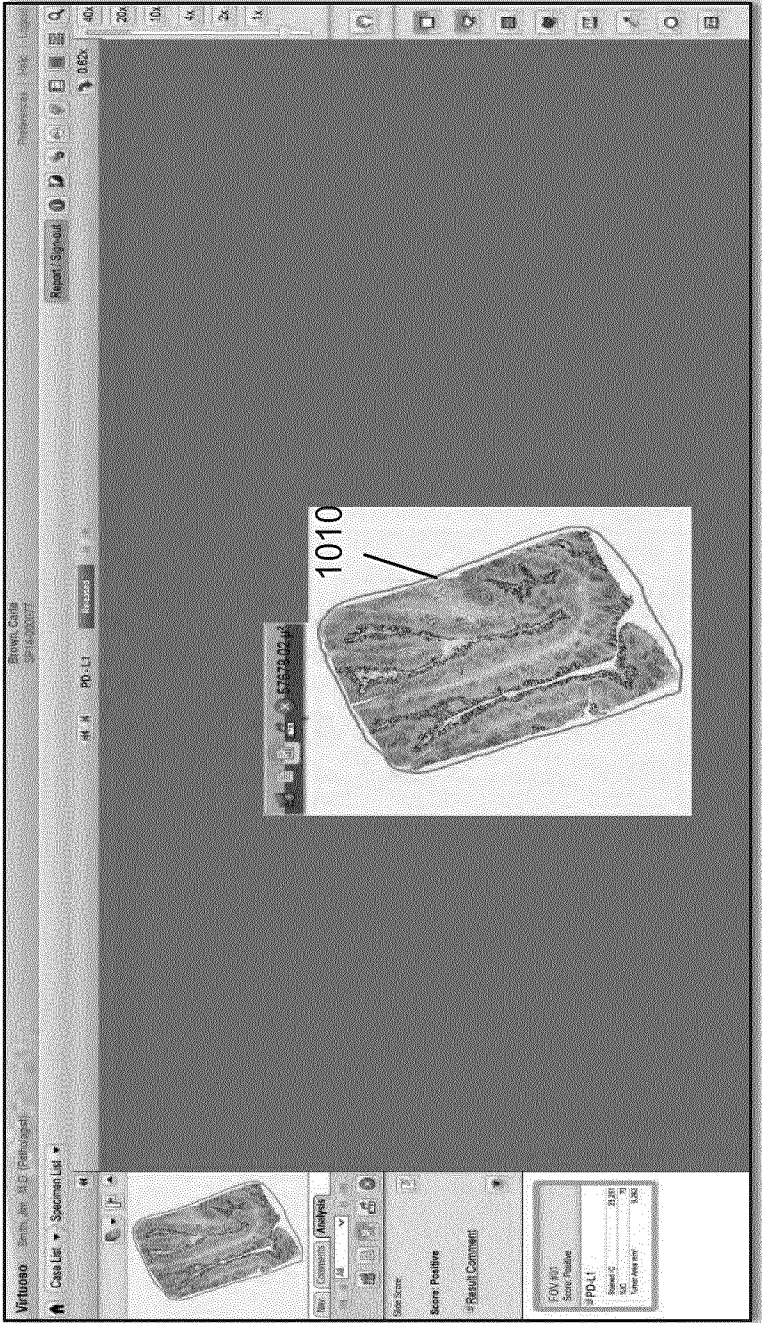


FIG. 12

1300

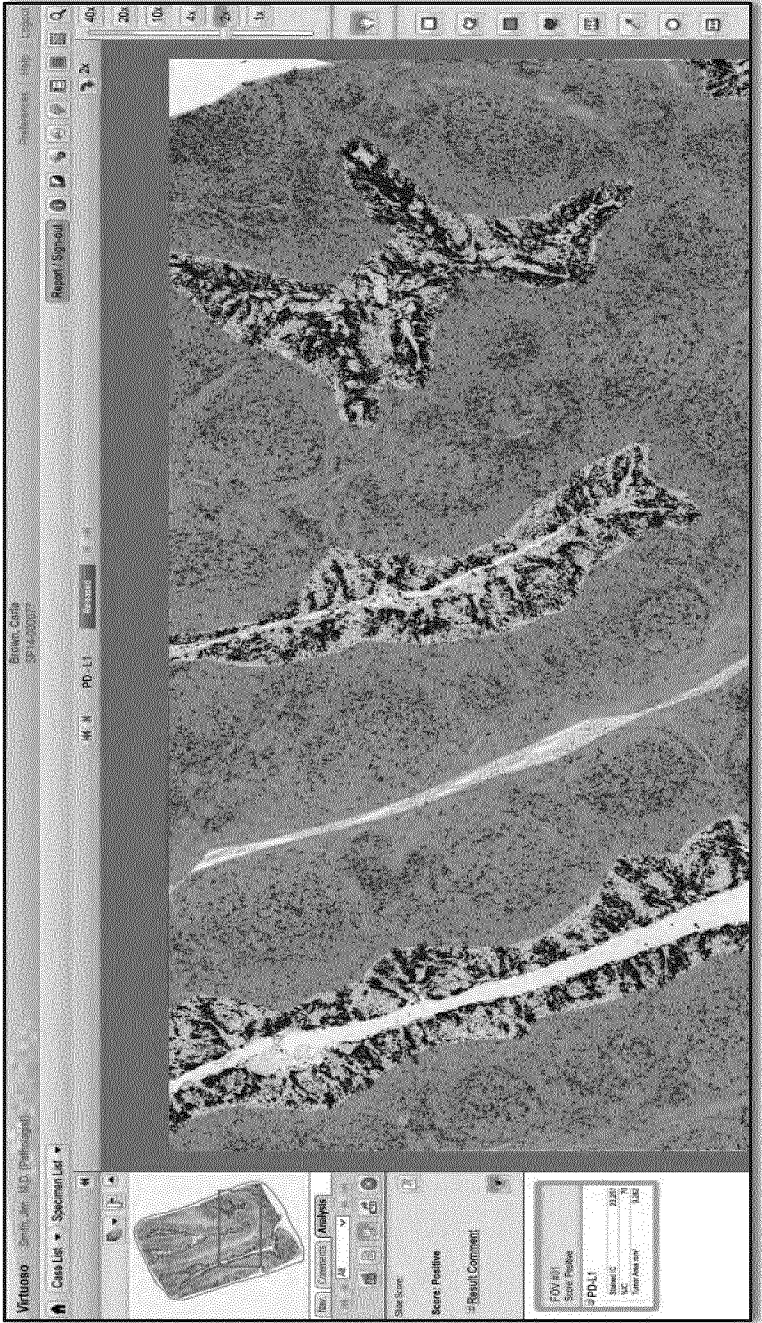


FIG. 13

1400

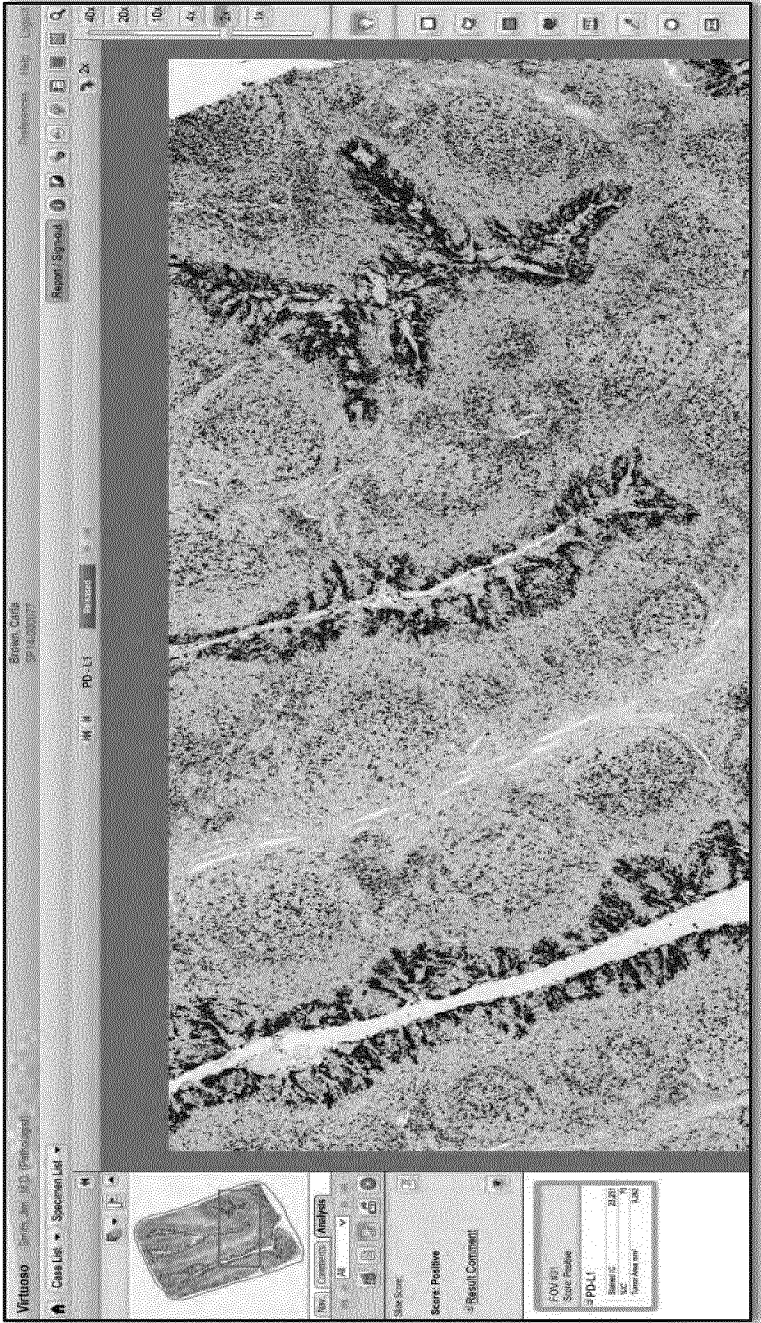


FIG. 14

1500
1555

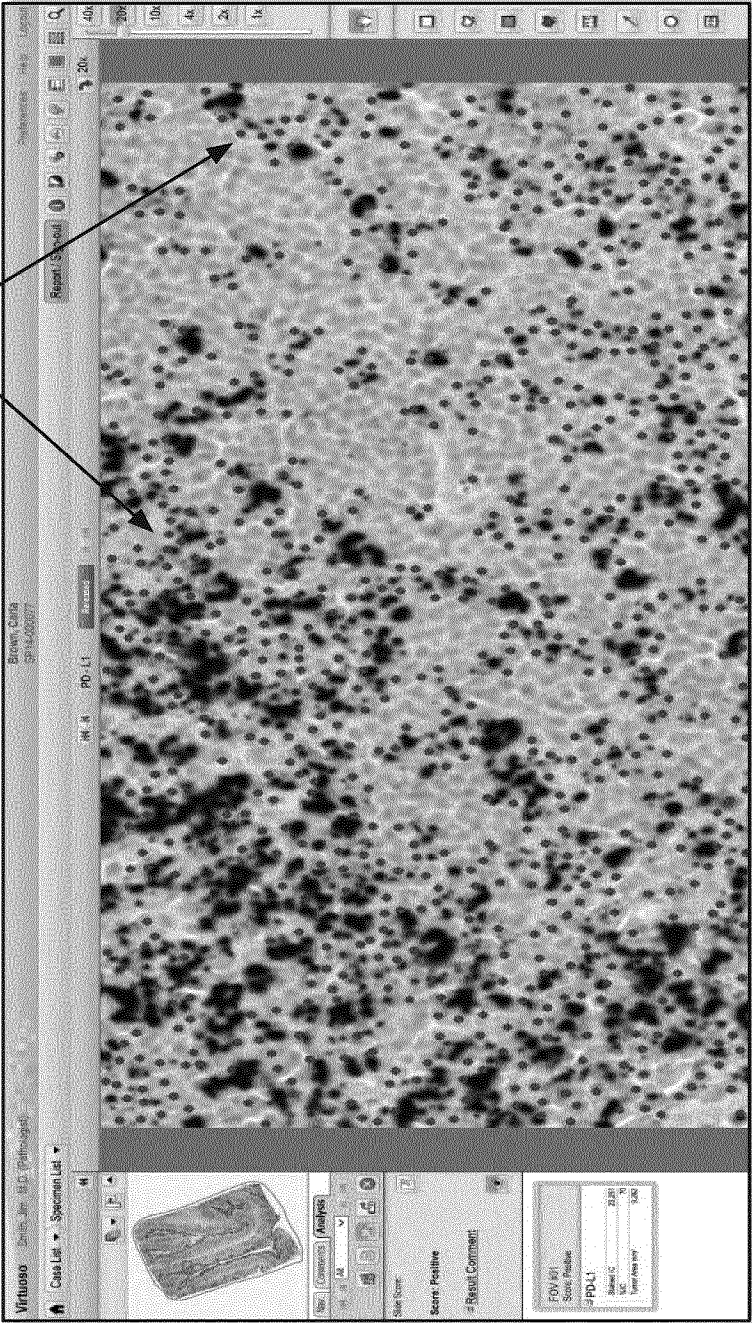


FIG. 15

1600

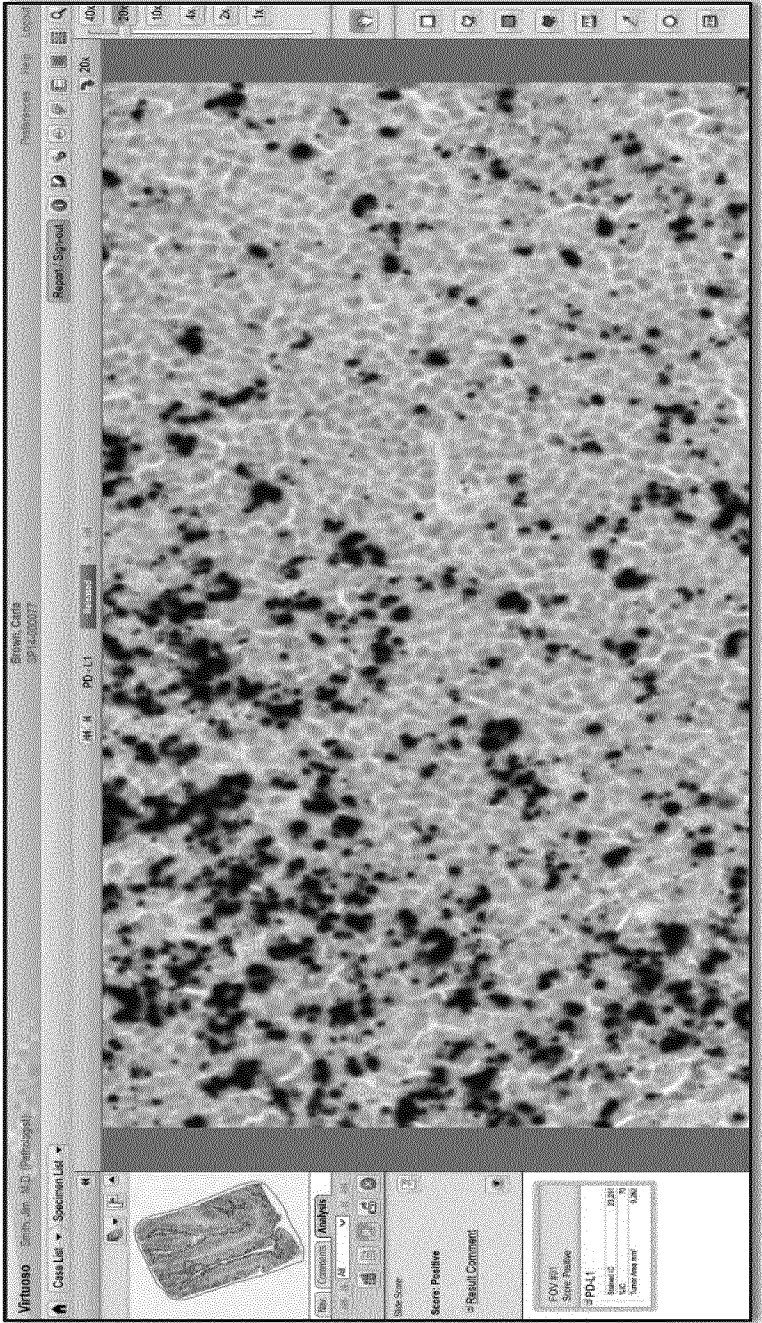
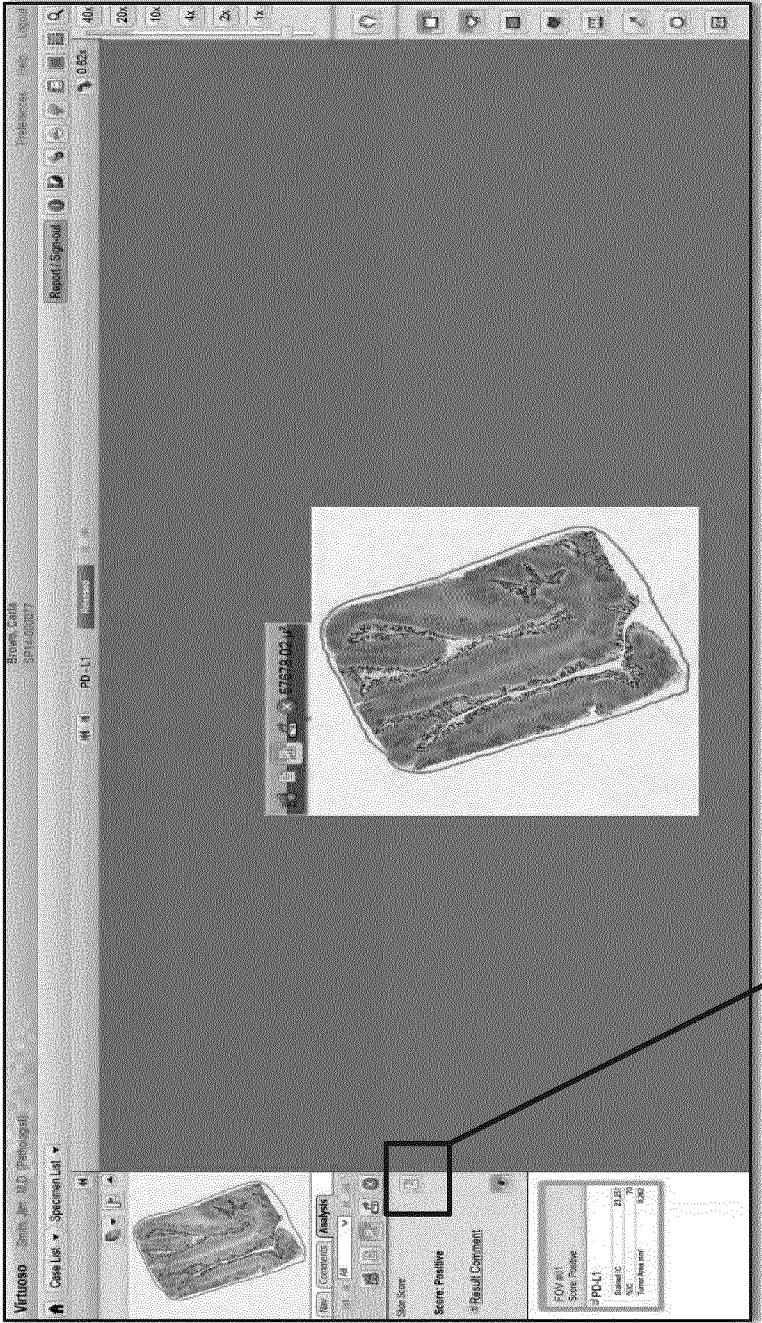


FIG. 16

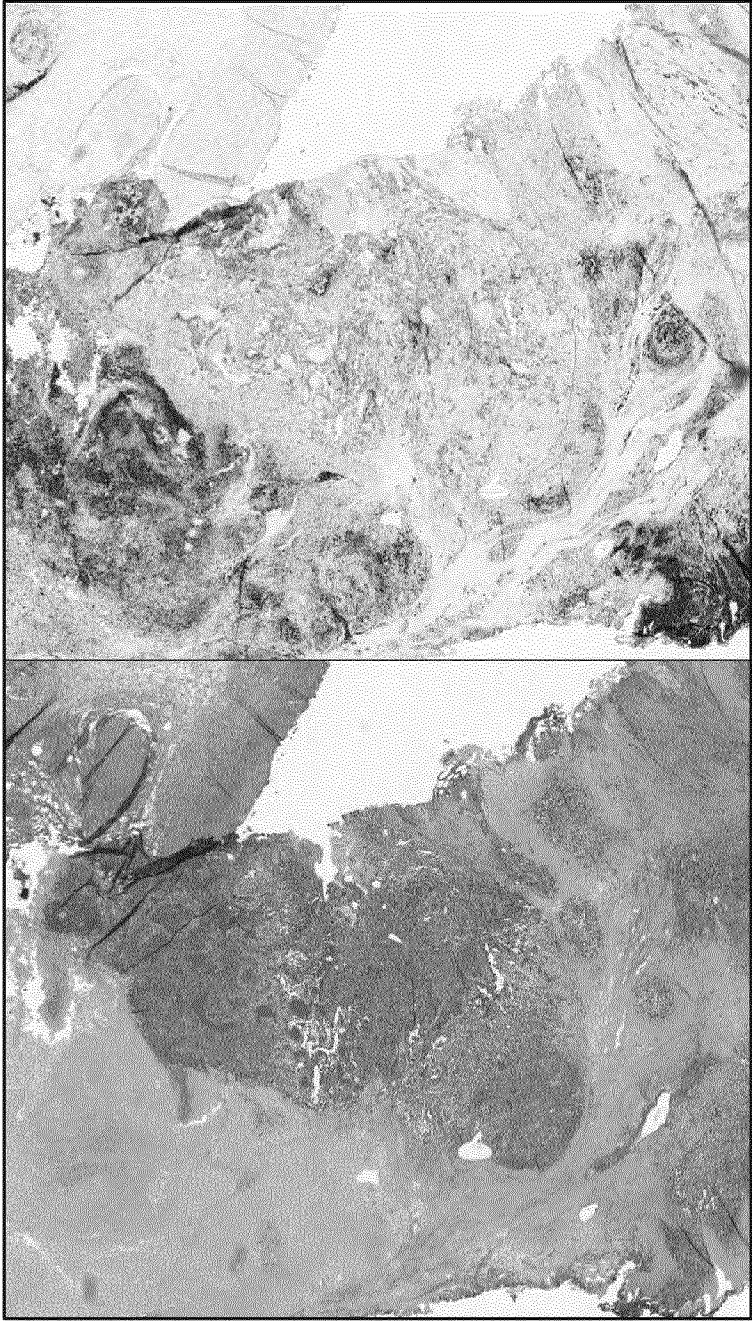
1700



Override invoked if pathologist disagrees with IA score

FIG. 17

1800



B

A

FIG. 18

1900

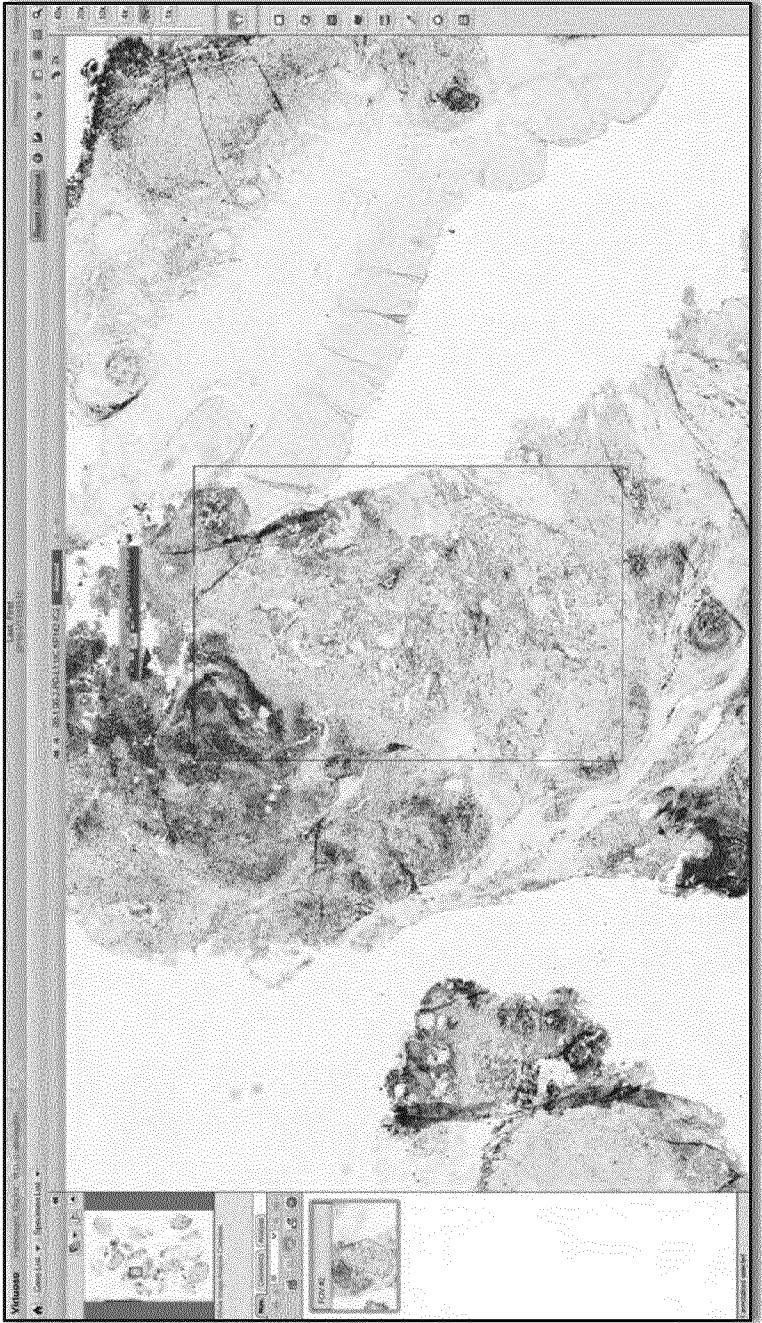


FIG. 19

2000



FIG. 20

2100

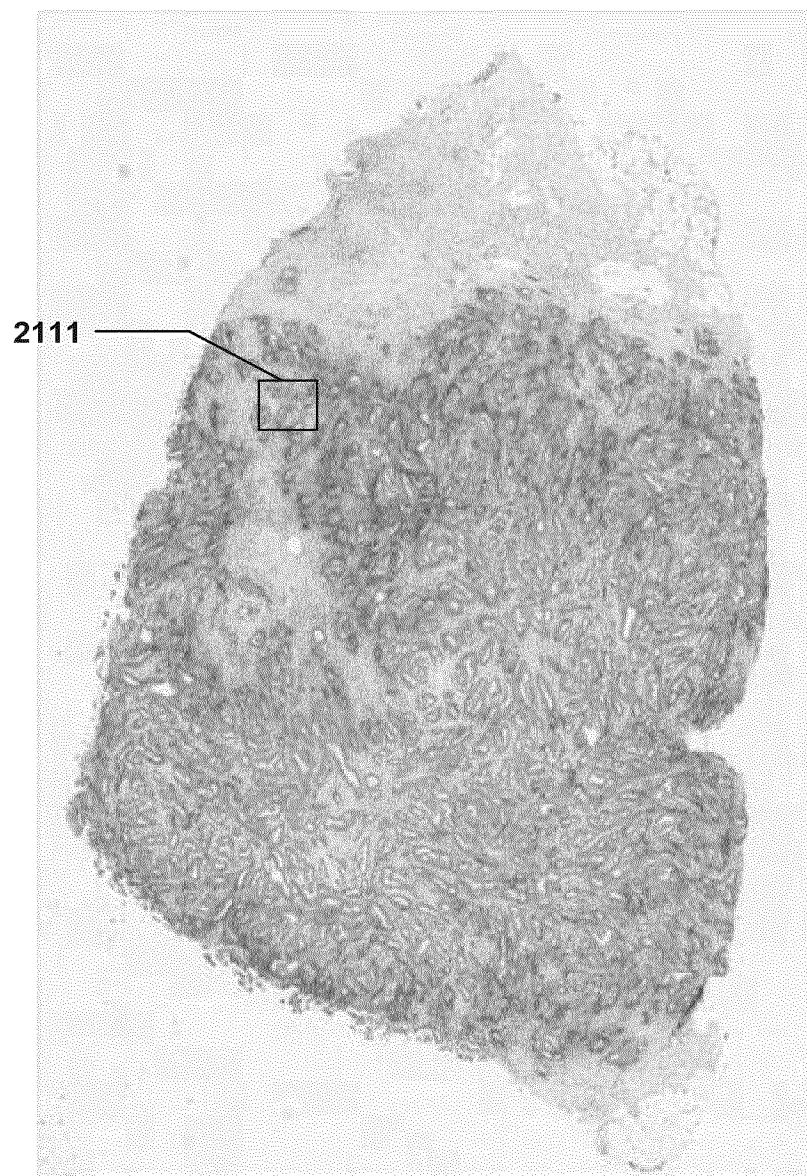


FIG. 21

2111

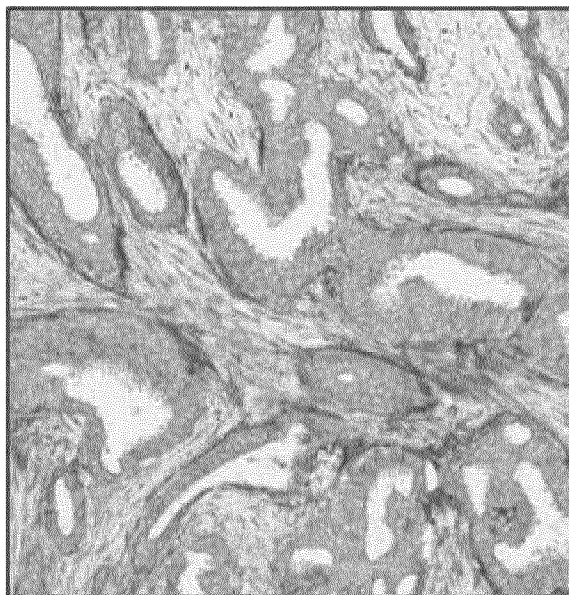


FIG. 22

2300

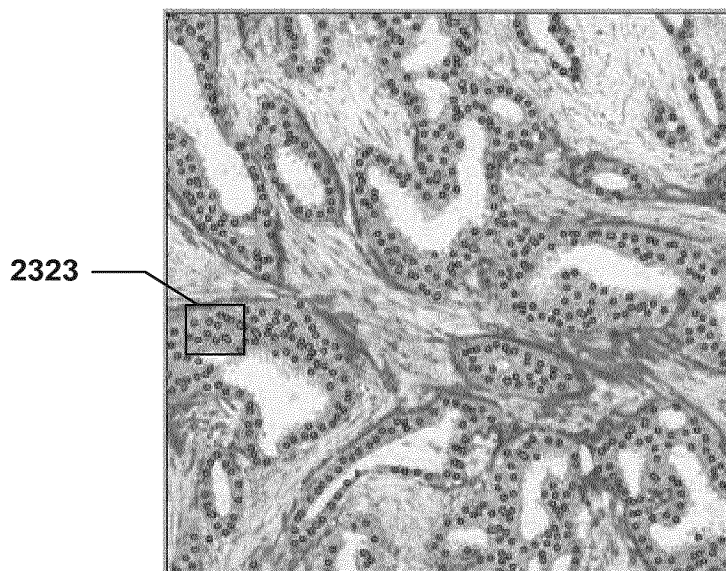


FIG. 23

2400

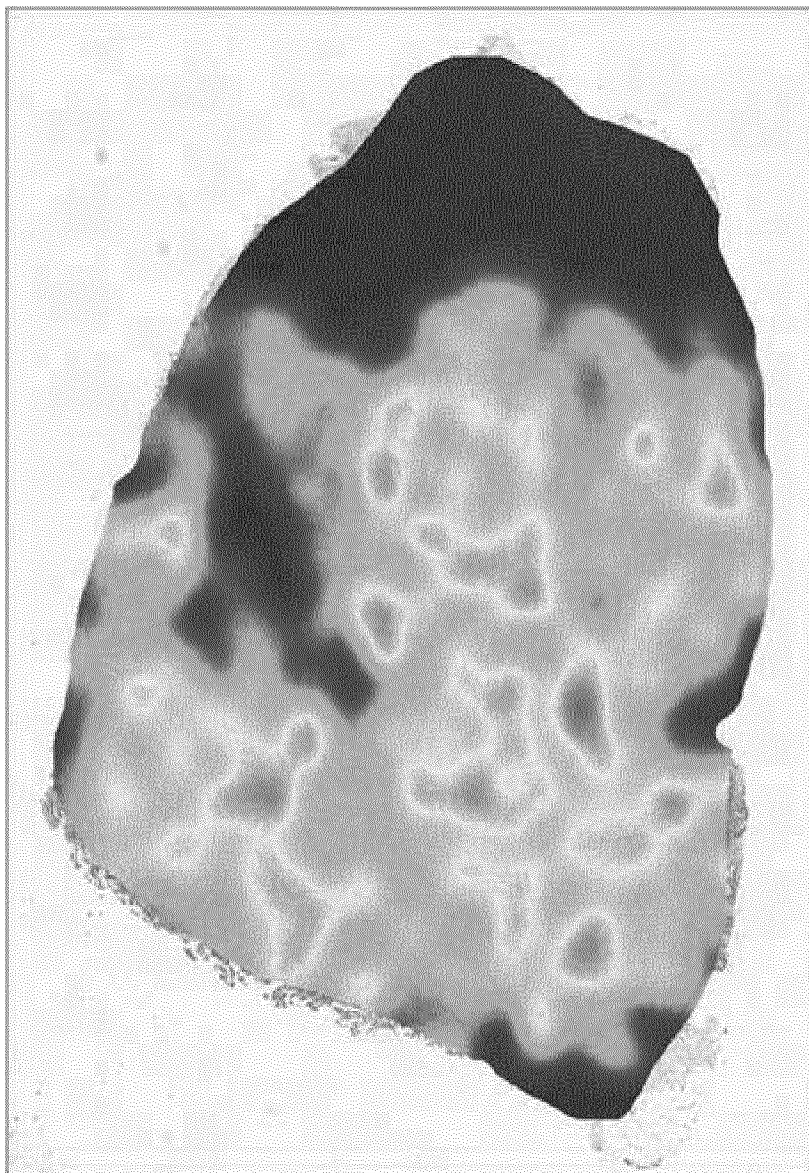


FIG. 24

2500



FIG. 25

2600

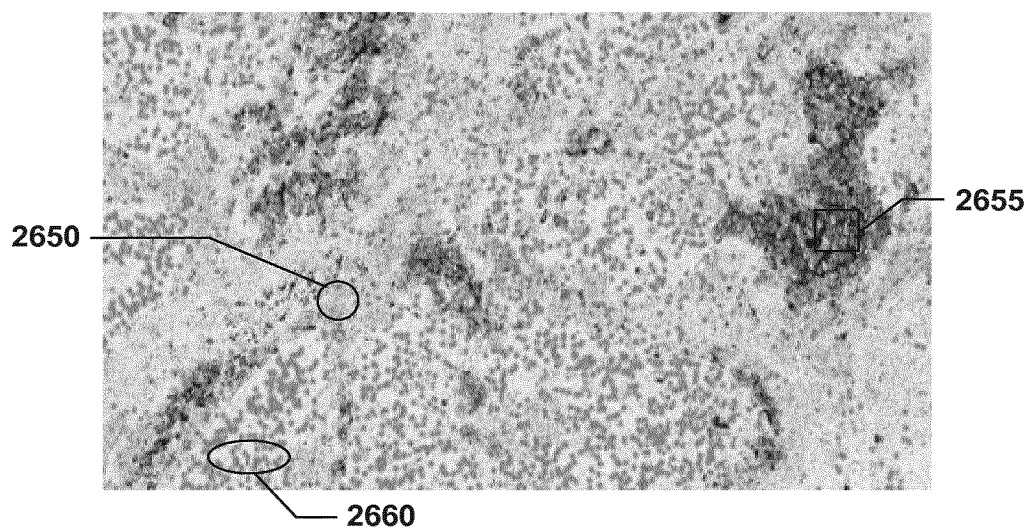


FIG. 26

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2017/075113

A. CLASSIFICATION OF SUBJECT MATTER

INV. G06T7/00
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06T

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2016/120418 A1 (VENTANA MED SYST INC [US]; HOFFMANN LA ROCHE [CH]; HOFFMANN LA ROCHE []) 4 August 2016 (2016-08-04) abstract paragraph [0041] - paragraph [0043]; figures 2,7	1-22
A	----- WO 2016/087592 A1 (VENTANA MED SYST INC [US]; HOFFMANN LA ROCHE [CH]) 9 June 2016 (2016-06-09) the whole document	1-22
A	----- WO 2013/113707 A1 (VENTANA MED SYST INC [US]; HOFFMANN LA ROCHE [CH]; BAMFORD PASCAL [US]) 8 August 2013 (2013-08-08) the whole document ----- -/-	1-22

☒ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

24 January 2018

Date of mailing of the international search report

30/01/2018

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Błaszczuk, Marek

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2017/075113

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	<p>WO 2014/140070 A2 (VENTANA MED SYST INC [US]; HOFFMANN LA ROCHE [CH]) 18 September 2014 (2014-09-18) the whole document -----</p>	1-22

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2017/075113

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2016120418 A1	04-08-2016	AU 2016211981 A1 CA 2973418 A1 CN 107209111 A EP 3251052 A1 US 2018012360 A1 WO 2016120418 A1	27-07-2017 04-08-2016 26-09-2017 06-12-2017 11-01-2018 04-08-2016
WO 2016087592 A1	09-06-2016	AU 2015357091 A1 CA 2965431 A1 EP 3227833 A1 US 2017270666 A1 WO 2016087592 A1	27-04-2017 09-06-2016 11-10-2017 21-09-2017 09-06-2016
WO 2013113707 A1	08-08-2013	AU 2013214345 A1 CA 2859269 A1 CN 104081412 A EP 2810213 A1 JP 6140192 B2 JP 2015508639 A JP 2017163990 A KR 20140120321 A KR 20170010088 A US 2014377753 A1 WO 2013113707 A1	26-06-2014 08-08-2013 01-10-2014 10-12-2014 31-05-2017 23-03-2015 21-09-2017 13-10-2014 25-01-2017 25-12-2014 08-08-2013
WO 2014140070 A2	18-09-2014	AU 2014230809 A1 CA 2897614 A1 CN 105027164 A EP 2973396 A2 JP 2016517559 A US 2016019695 A1 WO 2014140070 A2	23-07-2015 18-09-2014 04-11-2015 20-01-2016 16-06-2016 21-01-2016 18-09-2014

Publication Number: WO 2018/067937 A1

Publication Date: Apr 12, 2018

Abstract

A method for generating an immune score, the method comprising the steps of: (i) determining a qualitative and/or quantitative assessment of tumor infiltrating lymphocytes in a sample; (ii) determining a qualitative and/or quantitative assessment of T-cell receptor signaling in the sample; (iii) determining a qualitative and/or quantitative assessment of mutation burden in the sample; (iv) generating, using a predictive algorithm, an immune score based on the determined qualitative and/or quantitative assessment of tumor infiltrating lymphocytes, the determined qualitative and/or quantitative assessment of T-cell receptor signaling, and the determined qualitative and/or quantitative assessment of mutation burden.



(51) International Patent Classification:

C12Q 1/68 (2018.01) *A61P 35/00* (2006.01)
G01N 33/574 (2006.01)

(21) International Application Number:

PCT/US2017/055545

(22) International Filing Date:

06 October 2017 (06.10.2017)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/405,609 07 October 2016 (07.10.2016) US
62/456,072 07 February 2017 (07.02.2017) US

(71) Applicant: OMNISEQ, INC. [US/US]; Hauptman Woodward Medical Research Institute, 700 Ellicott Street, Buffalo, NY 14203 (US).

(72) Inventors: MORRISON, Carl; 9055 Fredonia Stockton Road, Ferdonia, NY 14063 (US). PABLA, Sarabjot; 141 Hampton Parkway, Buffalo, NY 14217 (US). CONROY, Jeffrey; 4162 Vinewood Drive, Williamsville, NY 14221 (US). NESLINE, Mary; 57 Dana Road, Buffalo, NY 14216 (US). GARDNER, Mark; 9685 Rocky Point, Clarence, NY 14031 (US). HE, Ji; 11716 Owens Glen Way, North Po-

tomac, MD 20878 (US). GLENN, Sean; 6605 Yorktown Circle, East Amherst, NY 14051 (US).

(74) Agent: BETTINGER, Blaine #66,254 et al.; Bond, Schoeneck & King, Plc, One Lincoln Center, Syracuse, NY 13202 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,

(54) Title: METHODS AND SYSTEMS FOR DETERMINING PERSONALIZED THERAPIES

Bayesian Model Averaging for Final
Prediction

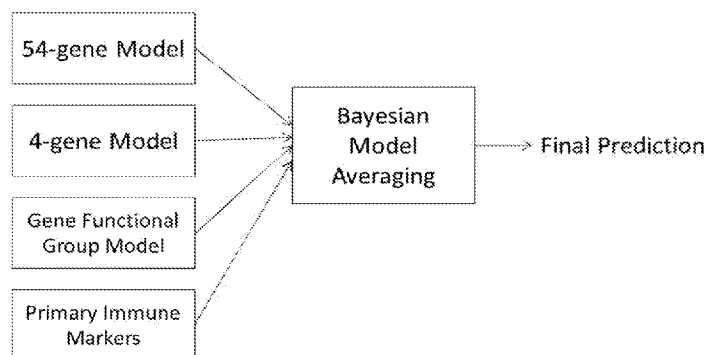


FIG. 7

(57) Abstract: A method for generating an immune score, the method comprising the steps of: (i) determining a qualitative and/or quantitative assessment of tumor infiltrating lymphocytes in a sample; (ii) determining a qualitative and/or quantitative assessment of T-cell receptor signaling in the sample; (iii) determining a qualitative and/or quantitative assessment of mutation burden in the sample; (iv) generating, using a predictive algorithm, an immune score based on the determined qualitative and/or quantitative assessment of tumor infiltrating lymphocytes, the determined qualitative and/or quantitative assessment of T-cell receptor signaling, and the determined qualitative and/or quantitative assessment of mutation burden.

TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

METHODS AND SYSTEMS FOR DETERMINING PERSONALIZED THERAPIES**Cross-reference to Related Applications**

[0001] This application claims priority to pending U.S. Provisional Patent Application Nos. 62/405,609, filed October 7, 2016 and entitled “Methods and Systems for Determining Personalized Therapies,” and 62/456,072, filed February 7, 2017 and entitled “Methods and Systems for Determining Personalized Therapies,” the entireties of which are hereby incorporated by reference.

Field of the Invention

[0002] The present disclosure is directed generally to methods and systems for generating tumor treatment recommendations.

Background

[0003] In melanoma and non-small cell lung cancer (NSCLC) patients, a high proportion of somatic mutations – the so-called “mutational burden” (MuB) - as well as elevated intratumoral expression of checkpoint blockades, including the immunosuppressive molecule CD274 (best known as programmed death-ligand 1 protein (PD-L1)), that have been shown to correlate with improved clinical responses to immune checkpoint blocker (ICB)-based immunotherapies. However, predicting the response of patients with other malignancies to immunotherapies – and perhaps in combination of immunotherapy with other targets for immune checkpoint blockade – requires a more profound deconvolution of the immunological tumor microenvironment. Similarly, an in-depth characterization of the immunological configuration of malignancies may be necessary to assist clinical decision making, especially for patients that fail standard ICB-based immunotherapy.

[0004] For example, therapeutic antibodies targeting immune checkpoint molecules have been approved by the FDA for the treatment of several types of cancer. However, evaluation of the tumor checkpoint blockade is limited to FDA-approved IHC assays measuring programmed death-ligand 1 (PD-L1) protein status, which is subjective and not analytically robust. As the number of antibodies targeting immune checkpoints expands, assays that can evaluate additional biomarkers in tumor specimens are needed to accurately predict patient response to these drugs.

[0005] Accordingly, there is a need for assays capable of characterizing an immunological tumor microenvironment as a guide for therapeutic decisions.

Summary of the Invention

[0006] The present disclosure is directed to inventive methods for determining the presence or absence of cancer cell sensitivity to one or more personalized oncology therapies.

[0007] According to one aspect of the invention is a method for generating an immune score, the method comprising the steps of: (i) determining a qualitative and/or quantitative assessment of tumor infiltrating lymphocytes in a sample; (ii) determining a qualitative and/or quantitative assessment of T-cell receptor signaling in the sample; (iii) determining a qualitative and/or quantitative assessment of mutational burden in the sample; and (iv) generating, using a predictive algorithm, an immune score based on the determined qualitative and/or quantitative assessment of tumor infiltrating lymphocytes, the determined qualitative and/or quantitative assessment of T-cell receptor signaling, and the determined qualitative and/or quantitative assessment of mutation burden.

[0008] According to an embodiment, the method further includes the step of determining, based on the immune score, one or more possible treatment therapies.

[0009] According to an embodiment, the predictive algorithm is trained and updated using machine learning.

[0010] According to one aspect of the invention is a method for analyzing a tumor, the method comprising the steps of: (i) determining a qualitative and/or quantitative assessment of tumor infiltrating lymphocytes in a sample; (ii) determining a qualitative and/or quantitative assessment of T-cell receptor signaling in the sample; (iii) determining a qualitative and/or quantitative assessment of mutational burden in the sample; and (iv) classifying, based on the determined qualitative and/or quantitative assessment of tumor infiltrating lymphocytes, the determined qualitative and/or quantitative assessment of T-cell receptor signaling, and the determined qualitative and/or quantitative assessment of mutational burden, the sample as a responder to a therapy or a non-responder to a therapy.

[0011] According to an embodiment, the method further includes the step of determining a response of the tumor, based on the determined classification, to one or more possible treatment therapies.

[0012] According to an embodiment, the sample is classified as an indeterminate responder to therapy. According to an embodiment, a sample classified as a non-responder to therapy and non-responder to therapy may be further classified as being at risk for hyper-progression of cancer.

[0013] According to an embodiment, the method further includes the step of correlating the determined classification of the sample with a second classification.

[0014] According to an embodiment, the method further includes the step of determining a response of the tumor, based on the correlation, to one or more possible treatment therapies.

[0015] According to an embodiment, the method further includes the step of determining immunohistochemistry data for the sample, wherein said classifying step is further based on the determined immunohistochemistry data.

[0016] According to another aspect is a method for analyzing a tumor, the method comprising the steps of: (i) determining a quantitative assessment of an expression level of at least four genes in a sample; (ii) classifying, if the determined quantitative assessment of an expression level of the first of the four genes exceeds a predetermined threshold, the sample as a responder to a therapy, or proceeding to the next step if the determined quantitative assessment of the expression level of the first of the four genes does not exceed the predetermined threshold; (iii) classifying, if the determined quantitative assessment of an expression level of the second of the four genes is greater than a predetermined threshold, the sample as a non-responder to a therapy, or proceeding to the next step if the determined quantitative assessment of the expression level of the second of the four genes does exceed the predetermined threshold; (iv) classifying, if the determined quantitative assessment of an expression level of the third of the four genes exceeds a predetermined threshold, the sample as a responder to a therapy, or proceeding to the next step if the determined quantitative assessment of the expression level of the third of the four genes does exceed the predetermined threshold; and (v) classifying, if the determined quantitative assessment of an expression level of the fourth of the four genes exceeds a predetermined threshold, the sample as a non-responder to a therapy, or classifying, if the determined quantitative assessment of the expression level of the fourth of the four genes does exceed the predetermined threshold, the sample as a responder to a therapy.

[0017] According to another aspect is a method for analyzing a tumor, the method comprising the steps of: (i) determining a qualitative and/or quantitative assessment of a plurality of genes related to immune cell infiltration in a sample; (ii) determining a qualitative and/or quantitative assessment of a plurality of genes related to T-cell activation in the sample; (iii) determining a qualitative and/or quantitative assessment of a plurality of genes related to cytokine signaling in the sample; (iv) determining a qualitative and/or quantitative assessment of a plurality of genes related to immune response regulation in the sample; (v) normalizing each of the determined qualitative and/or quantitative assessments; (vi) if the normalized qualitative and/or quantitative assessment of immune cell infiltration exceeds a predetermined threshold, proceeding to an analysis of the normalized qualitative and/or

quantitative assessment of T-cell activation in the sample, wherein the sample is classified as a responder if the normalized qualitative and/or quantitative assessment of T-cell activation exceeds a predetermined threshold, and wherein the sample is classified as a non-responder if the normalized qualitative and/or quantitative assessment of T-cell activation does not exceed the predetermined threshold; (vii) proceeding, if the normalized qualitative and/or quantitative assessment of immune cell infiltration falls below the predetermined threshold, to the next step; (viii) classifying, if the normalized qualitative and/or quantitative assessment of immune regulation response exceeds a predetermined threshold, the sample as an indeterminate responder, or proceeding, if the normalized qualitative and/or quantitative assessment of immune regulation response does not exceed the predetermined threshold, to the next step; and (ix) classifying, if the normalized qualitative and/or quantitative assessment of cytokine signaling exceeds a predetermined threshold, the sample as an indeterminate responder, or classifying, if the normalized qualitative and/or quantitative assessment of cytokine signaling does not exceed the predetermined threshold, the sample as a non-responder.

[0018] According to another aspect is a method for providing a comprehensive immune profiling clinical report to a patient's clinician, the method comprising the steps of: (i) obtaining one or more samples from a tumor of the patient; (ii) generating, from the one or more samples, RNA sequencing data comprising information about expression of a plurality of tumor-infiltrating lymphocyte proteins and a plurality of T-cell receptor signaling proteins; (iii) generating, from the one or more samples, DNA sequencing data comprising mutational burden information about a plurality of genes; (iv) generating, from the one or more samples, immunohistochemistry data and fluorescent in situ hybridization (FISH) data to measure, among other proteins, PD-L1 protein expression and copy number gain and patterns of tumor infiltrating lymphocyte protein expression for CD3 and CD8; (v) calculating, from the RNA sequencing data, the DNA sequencing data, and the immunohistochemistry and FISH data, a likelihood of the patient's tumor to respond to a plurality of possible treatments; and (vi) providing a comprehensive immune profiling clinical report to the patient's clinician, the report comprising the calculated likelihoods of the patient's tumor to respond to the plurality of possible treatments.

[0019] These and other aspects of the invention will be apparent from the embodiments described below.

Brief Description of the Drawings

[0020] In the drawings, like reference characters generally refer to the same parts throughout the different views. Also, the drawings are not necessarily to scale, emphasis instead generally being placed upon illustrating the principles of the invention.

[0021] FIG. 1 is a graph of results for a 54 gene model using a panel of retrospective samples for training, in accordance with an embodiment.

[0022] FIG. 2 is a schematic representation of feature space for a model utilizing the 54 genes and the mutation burden (MuB), in accordance with an embodiment.

[0023] FIG. 3 is a schematic representation of a decision tree for a 4 gene model, in accordance with an embodiment.

[0024] FIG. 4 is a graph of results for the 4 gene model using a panel of retrospective samples for training, in accordance with an embodiment.

[0025] FIG. 5 is a schematic representation of a decision tree for an immune function model, in accordance with an embodiment.

[0026] FIG. 6 is graph of results for an immune function model using a panel of retrospective samples for training, in accordance with an embodiment.

[0027] FIG. 7 is a schematic representation of a Bayesian Model Averaging (BMA) for final prediction using output from the 54 gene model, the 4 gene model, and the immune function model, in accordance with an embodiment.

[0028] FIG. 8 is a graph of results for Bayesian Model Averaging using a panel of 87 retrospective samples for training, and results from the 54 gene model, the 4 gene model, and the immune function model, in accordance with an embodiment.

[0029] FIG. 9 is a table of overall results from four different models for determining cancer cell sensitivity to one or more personalized oncology therapies, in accordance with an embodiment.

[0030] FIG. 10 is a flowchart of a method for providing a report to a patient's clinician, in accordance with an embodiment.

Detailed Description

[0031] The present disclosure is directed to embodiments of a method and system for determining cancer cell sensitivity to one or more personalized oncology therapies.

[0032] According to one embodiment, the disclosure is directed to methods for applying a Multi-Analyte Assay with Algorithm Analyses (MAAA) approach to predict therapeutic efficacy. Immune Advance (IA) is a next generation sequencing (NGS) assay designed to

provide information about tumor infiltrating lymphocytes (TILs), mutation burden (MuB), T-cell receptor signaling (TCRS), immune-related drug target scores (IRDTs), and overall immune activation status (i.e., Immune Score). The evaluation of TILs focuses on CD8+ T-cells, but does provide additional information about other subsets of T-cells and related immune effector cells such as B-cells and macrophages. The evaluation of MuB is designed to provide information about the overall number of somatic mutations in a tumor (high versus low), and with no intent to report on specific mutations at the gene level. TCRS, or the signaling of T-cells with both neoplastic cells and other immune-related cells, utilizes expression information from genes representing both receptors and ligands in the interaction of these various cell types. IRDTs are a set of genes that are the direct targets of one or more immunomodulatory agents, such as CTLA4 and ipilimumab. The Immune Score (IS) utilizes information from TILs, MuB, and TCRS in the context of a reference database of prior clinical outcomes for patients treated with one or more checkpoint inhibitors (CPIs) to provide an assessment of activation of the immune state of the tumor tested.

[0033] Immune Advance (IA) is a next generation sequencing assay that uses RNA-Seq to evaluate the mRNA expression of numerous immune related (IR) genes and multiple expression control genes, and DNA-Seq to evaluate mutational burden in a 1.5 Mb targeted capture. The RNA-Seq component validates the expression of the most clinically important genes, and the DNA-Seq component reports a count on non-synonymous mutations with no reference to the specific genes involved. A representative table of 767 genes that, for example, may be utilized in one or more of the RNA-Seq, DNA-Seq, and mutation burden analyses is provided herein as Table 2.

[0034] The RNA-Seq component of IA interrogates numerous genes that at the highest level represent more than 40 unique gene functions. IA is focused on the most clinically important genes, of which many are directly related to T-cell receptor signaling (TCRS) and several others for classifying tumor infiltrating lymphocytes (TILs). In the clinical validation portion of IA for all identified clinically important genes, the RNA-Seq data for multiple formalin-fixed, paraffin-embedded (FFPE) specimens is compared to results for custom TaqMan assays, as well as publicly available information for the same set of samples for whole transcriptome analysis via the Cancer Genome Atlas project (TCGA).

[0035] Genes related to TCRS may be expressed on immune infiltrating cells or neoplastic cells and are typically classified as pairs of a receptor and associated ligand. For all genes classified in IA as related to TCRS, either the ligand or the receptor is expressed by one or more subsets of T-cells. The current approach to this list of TCRS genes is to divide them

into either co-stimulatory or co-inhibitory T-cell function, or checkpoint pathway. Genes related to TCRS can be further divided into those that are the direct target of one or more checkpoint inhibitor drugs versus those that are not. Direct targets of checkpoint inhibitor drugs may be either the receptor or ligand, but not both simultaneously. An example of a checkpoint inhibitor drug and target is ipilimumab and the receptor CTLA-4 that is expressed on activated T-cells.

[0036] Genes related to TILs encompass a wide variety of infiltrating immune cells and IA has adopted an existing classification and associated gene expression markers. Classically, an “immunoscore” has been reported as a prognostic marker in multiple tumor types using three (3) or fewer markers of TILs. IA utilizes many more than three genes related to TILs to stratify patients for response to CPIs.

[0037] The DNA-Seq component of IA is, for example, a 1.5 Mb AmpliSeq capture of many different cancer-related genes. For IA specific mutations may not be reported, but rather the number of non-synonymous mutations may be reported. The output of the DNA-Seq component of IA, therefore, is an assessment of mutational burden.

[0038] Drug-target gene expression analysis is focused on each gene targeted by one or more checkpoint inhibitor drugs, and categorizes each result based upon pre-specified thresholds as High, Intermediate or Moderate, or Low association for the specified drug target. The high, intermediate/moderate, or low association for each drug target is more than a simple evaluation of reads per million for a target gene and encompasses upstream and downstream effectors for that target gene. For example, the primary function of CTLA-4 signaling is to down-regulate T-cell activation by countering the co-stimulatory signal delivered by a second receptor, CD28. Both CTLA-4 and CD28 share the same ligands, CD80 (also known as B7.1) and CD86 (also known as B7.2). CTLA-4 has a higher affinity for both of these ligands than CD28 for ligand binding resulting in an overall co-inhibition signal when expressed at equivalent levels. Clinically, this is countered by administering the CTLA-4 checkpoint inhibitor ipilimumab.

[0039] TIL analysis is focused on a semi-quantitative assessment and/or quantitative assessment of TIL(s) and a qualitative and/or quantitative assessment of additional subsets of immune effector cells. In the first instance, tumor infiltrating lymphocytes (TIL) are reported as number of CD3+ and CD8+ transcripts expressed (reads). To aid in the interpretation of the expression level of TIL(s), a direct immunohistochemical comparison is performed for CD3 and CD8 using an Aperio image analysis platform. In the second instance, additional qualitative and/or quantitative markers of TIL(s) such as FOXP3 for T-regulatory cells,

CD163 and CD68 for macrophages, ICOS and CD28 for T-helper cells, and several other markers of subsets of immune cells are categorized for each as a relative expression result based upon pre-specified transcript read thresholds as High, Intermediate, or Low. In contrast to drug-target evaluation the analysis for additional various markers of Ttic(s) is simpler and reflects a ranking to prior observed values.

[0040] Mutational burden analysis is focused on number of non-synonymous mutations equivalent to exome sequencing. Results of the 1.5 Mb targeted capture sequence are provided in the context of exome sequencing in the validation of the DNA-Seq component of IA. These results are provided as a High, Intermediate, or Low mutational burden. A mutation burden greater than the equivalent of 200 exomic non-synonymous mutations is reported as high, less than 200 and 150 or greater as intermediate, and less than 150 as low. The cut-offs are arbitrary and reflect a summary of what has been reported in the literature in regard to clinical response to one or more checkpoint inhibitors. Accordingly, many other thresholds are possible.

[0041] To summarize the results of drug-target gene expression, immune cell infiltration, and mutation burden, IA provides a single immune score ("Immune Score") on a scale of 0-100 for the overall assessment of immune activation. The IS represents a single patient result that was tested for clinical utility in this validation.

[0042] According to another embodiment, the disclosure is directed to methods and systems for generating tumor treatment recommendations using at least three unique models or approaches, independently developed, that can be utilized for comparative purposes and optionally for Bayesian modeling. The multiple models address both a machine learning approach as well as a biological approach to tumor treatment. According to an embodiment, the at least three models will provide a similar response to checkpoint inhibitors, and a Bayesian average model can be utilized to represent the best fit.

[0043] The first model, referred to as the 54 gene model and discussed in greater detail below, is a polynomial machine learning regression model. According to an embodiment, the 54 gene model or approach uses 11 genes representing TILs and 43 genes for TCRS combined with MuB for prediction, although other genes and combinations are possible.

[0044] The second model, referred to as the 4 gene model and discussed in greater detail below, represents a biological approach at the gene level. According to an embodiment, the 4 gene model or approach utilizes a decision tree model to select the best minimal set of TILs or T-cell activation genes for prediction.

[0045] The third model, referred to as the immune function model or the gene functional group model and discussed in greater detail below, represents a biological approach at the functional level. According to one embodiment, this model or approach utilizes 13 genes representing immune cell infiltration, 23 genes for T-cell activation, 10 genes for cytokine signaling, and 8 genes for immune response regulation, although other embodiments are possible.

[0046] The fourth model, which is optional, is referred to as the primary immune marker model or approach. The primary immune marker model or approach analyzes one or more immune markers. According to an embodiment, the primary immune marker model or approach analyzes PD-L1 protein expression and/or tumor infiltrating lymphocyte (TIL) expression (including but not limited to CD3 and/or CD8) using immunohistochemistry. According to another embodiment, the primary immune marker model or approach analyzes PD-L1 and/or PD-L2 copy number gain utilizing fluorescent in situ hybridization (FISH) methodology.

[0047] According to another embodiment, the method may also optionally utilize PCR analysis or any other methodology to analyze microsatellite instability (MSI), among other possible factors.

[0048] Each of the two approaches is described in greater detail below.

[0049] Approach #1 - The IA Multi-Analyte Assay with Algorithmic Analysis (MAAA)

[0050] According to a first approach to deriving a personalized oncology therapy, a multi-factor analysis referred to as Immune Advance (IA) provides a qualitative and/or quantitative assessment of tumor infiltrating lymphocytes (TILs), mutation burden (MuB), and T-cell receptor signaling (TCRS). The three values are used to derive an overall score, referred to as an Immune Score, which is a predictor of the tumor's response to one or more checkpoint inhibitors (CPIs). As just one example, the Immune Score can be reported on a scale of 1-100 and can be grouped as three clinically relevant groups such as high, indeterminate, or low overall response rate to CPIs, among many other possible reporting mechanisms.

[0051] IA utilizes a unique algorithmic analysis to provide a qualitative and/or quantitative assessment of tumor infiltrating lymphocytes (TILs), mutation burden (MuB), and T-cell receptor signaling (TCRS). These three values are then used to derive an overall Immune Score. All of these analyses utilize a reference database of prior IA results developed through this validation for future evaluations. This reference database will be updated, for

example, on a quarterly basis as prior clinical results are added to the existing results. This update to the reference database will not impact any clinical results prior to the time of the update. Evaluation of TILs, MuB, and TCRS are independent of any known clinical outcomes, but are dependent on the observed values for these parameters in the reference database of prior IA results. In comparison, the Immune Score requires utilization of clinical outcomes in the reference database regarding response to current FDA-approved checkpoint inhibitors (CPIs).

[0052] The approach for the IA analysis was developed by first analyzing a group of co-expressed TCRS genes discovered using RNA-seq data from a cohort of specimens. The concept of using TILs and MuB was added to the assessment based upon information provided in the peer-reviewed literature pertaining to response to CPIs as well as prognostic outcomes in various tumor types. Based upon the peer-reviewed literature and analysis of the RNA-seq data from the RPCI TCGA cohort, several genes may be utilized for TIL evaluation. This approach shows that TILs and TCRS are highly associated and co-expressed, while MuB has some but a lesser concordance with these two parameters.

[0053] The final approach for the IA analysis was to develop mathematical formulas (i.e., algorithms) that could be used for evaluation of TILs, MuB, TCRS, and IAS in the context of patient stratification of response to CPIs. The endpoint in this analysis was to normalize each of these parameters to a score of 100 to provide a numerical reference of results. The algorithm utilizes 3 different analyses, or steps, in a unique order to derive the final patient stratification for response to CPIs.

[0054] **Step 1** in the algorithmic analysis is based upon the ranking, or Score, of the observed test result to the reference database for TILs, MuB, and TCRS. For purposes of reporting patient results the TCRS Score is an intermediate value that is not reported while the TILs Score and MuB Score are reported values. The mathematical equation used for this ranking, or Score, is as listed below:

[0055] TIL Score means the number of reference samples that have less than or equal to the normalized log2 reads per million of the average of the TIL-identified genes to the test sample / total number of reference samples * 100 and rounded to the closest integer.

[0056] TCRS Score means the number of reference samples that have less than or equal to the normalized log2 reads per million of the average of the TCRS-identified genes to the test sample / total number of reference samples * 100 and rounded to the closest integer.

[0057] MuB Score means the number of reference samples that have less than or equal to the number of somatic mutations to the test sample / total number of reference samples * 100 and rounded to the closest integer.

[0058] **Step 2** in the algorithmic analysis, the Scores for TILs, MuB, and TCRS are combined into a single value, i.e. Immune Activated Weighted Score (IAWS), using a weighted value for each of these parameters. The mathematical equation used for this Immune Weighted Score (IWS) is as follows:

[0059]
$$\text{IWS Score} = (\text{TILs Score} \times \text{weighted value for TILs}) + (\text{MuB Score} \times \text{weighted value for MuB}) + (\text{TCRS Score} \times \text{weighted value for TCRS})$$
 and rounded to the closest integer.

[0060] The weighted values for TILs, MuB and TCRS may be based, for example, on a machine-learning approach that trains a classifier to produce the optimal IWS (see Step 3) for known therapeutic responders versus non-responders in the reference dataset. This classifier functions such that the IWS of the two classes will be best distinguished, and that weighted value of TILs + weighted value of MuB + weighted value of TCR = 1.0. Within the ranked IWS of the reference dataset with responder / nonresponder indications, two thresholds are further determined for calling High / Low scores, so that all reference samples above the high threshold will have $\geq 95\%$ PPV in predicting responders and below the low threshold will have $\geq 95\%$ NPV in predicting nonresponders.

[0061] Optional **Step 3** in the algorithmic analysis involves transforming the IWS into an Immune Score (IS) based upon the ranking, or Score, of the observed test IWS result to the reference database. For purposes of reporting patient results, the IWS is an intermediate value that is not reported while the IS is a reported value. The mathematical equation used for this ranking, or Score, is as listed below:

[0062]
$$\text{IS} = (\text{Number of samples in the reference dataset that have less than or equal to the IWS score of the test sample}) / \text{total number of reference samples} \times 100$$
 and rounded to the closest integer

[0063] While the IA algorithm does not change, the reference database of patient responses to CPIs will continue to expand with future clinical testing and follow-up of patient responses. Once a score is assigned to one of the qualitative and/or quantitative assessments for TILs, MuB, TCRS, IWS, or IS that value will not change with future additions to the reference database.

[0064] **Approach #2 – The Integration of Three or More Models**

[0065] According to a second approach to deriving a personalized oncology therapy, three independent models – the 54 gene model, the 4 gene model, and the immune function model – are utilized individually or in one or more possible combinations. For example, two or more of the three models may be utilized for comparative purposes, and/or may be used for Bayesian or other types of modeling. Each of the three models provide a prediction of the tumor's response to one or more checkpoint inhibitors, and advanced comparative and modeling techniques can provide an average or overall prediction using output from two or more of the three models.

[0066] According to the second approach to deriving a personalized oncology therapy, three independent models – the 54 gene model, the 4 gene model, and the immune function model – are utilized individually or in one or more possible combinations. The first model, referred to as the 54 gene model, is a polynomial machine learning regression model that uses 11 genes representing TILs and 43 genes for TCRS combined with MuB for prediction. The second model, referred to as the 4 gene model, represents a biological approach at the gene level and utilizes a decision tree model to select the best minimal set of TILs or T-cell activation genes for prediction. The third model, referred to as the immune function model, represents a biological approach at the functional level and utilizes 13 genes representing immune cell infiltration, 23 genes for T-cell activation, 10 genes for cytokine signaling, and 8 genes for immune response regulation. As described below, the three models can be utilized independently and/or in collaboration to generate or derive a personalized oncology therapy.

[0067] According to a further embodiment of the second approach, a fourth model called the primary immune marker model or approach is also utilized. The fourth model analyzes one or more immune markers using immunohistochemistry as described or otherwise envisioned herein.

[0068] Model #1 – The 54 Gene Model

[0069] According to an embodiment, the 54 gene model is a polynomial machine learning regression model that uses 11 genes representing TILs and 43 genes for TCRS combined with MuB for prediction, although other gene combinations are possible.

[0070] The 54 gene model was derived by benchmarking different combinations of training and testing data sizes based on a plurality of retrospective samples where the treatment method and tumor response were known. For a training size of N samples, 20,000 iterations of trainings were carried out, each using randomly-drawn N samples and evaluating the performance of the classifier using the rest ((Total # of retrospective samples) – N) testing samples.

[0071] The overall performance of the classifier for N-training-sample was calculated from the mean of the 20,000 benchmarks. From an internal benchmark, it was observed that the classifier's ROC / AUC performance converges once N reaches ~ 50% of the total initial retrospective population. The published performance metrics (ROC plot, AUC score, PPV and NPV etc.) is calculated out of a more explicit leave-one-out test, i.e. with 87 iterations of unique N = 86 tests, each iteration using 86 samples for training while leaving one unique sample for testing purpose. The final prediction model for future testing purpose however used all retrospective samples for training, which in one experimental study was a total of 87 samples.

[0072] Referring to FIG. 1, in one embodiment, is a graph of results for the 54 gene model using a panel of retrospective samples for training. As shown in the graph, the results have a positive predictive value (PPV) of 96% for 26% of the population, a negative predictive value (NPV) of 90% for 49% of the population, and an indeterminate group representing 25% of the population.

[0073] According to one embodiment, not shown in FIG. 1, the 54 gene model can be determined or designed to classify tumors as being a responder, indeterminate responder, or non-responder to a therapy, and can also be determined or designed to predict the risk of hyper-progression of cancer. For example, the 54 gene model can be determined or designed to identify a subset of non-responders that have a risk of hyper-progressive disease. The risk of hyper-progressive disease may be qualitative and/or quantitative.

[0074] According to an embodiment, the 54 gene model utilizes expression information for 54 different genes, and/or mutational burden as the sum of mutation counts in 409 genes. Table 1 includes a representative list of genes for which expression information can be utilized in the 54 gene model. However, many other genes are possible for expression analysis in this model, including but not limited to the genes identified in Table 2.

[0075] TABLE 1. Genes Utilized for an Embodiment of the 54 Gene Model

TILs Genes (11)	TCRS Genes (43)			
CD163	ADORA2A	CD40LG	TIM3	PD-L2
CD2	BTLA	CD80 (B7-1)	ICOS	STAT1
CD3D	VISTA (B7-H5)	CD86 (B7-2)	ICOSLG	TBX21
CD3E	CCL2	CSF1R	IDO1	TGFB1
CD3G	CCR2	CTLA4	IFNG	TNF

CD4	SLAMF4	CXCL10	IL10	TNFRSF14
CD68	CD27 (TNFRSF27)	CXCR6	IL1B	GITR
CD8A	PD-L1	DDX58	KLRD1	OX40
CD8B	CD28	ENTPD1	LAG3	CD137
FOXP3	CD38	GATA3	MX1	OX-40L
CD20	CD40	GZMB	PD-1	

[0076] Referring to FIG. 2, in one embodiment, is a schematic representation of feature space for a model utilizing the 54 genes, or 54 features, combined with mutational burden (MuB) as an additional feature. According to an embodiment, MuB, as an individual feature, is treated equally to the 54 genes, or 54 features.

[0077] In addition to listing genes that may be suitable for the 54 gene model, Table 2 also provides genes for which expression information can be utilized in the immune function model described herein. Accordingly, the genes listed in Table 2 may be utilized in one or more of the RNA-Seq, DNA-Seq, and mutation burden analyses.

[0078] **TABLE 2. Genes Utilized for an Embodiment of the 54 Gene Model**

ABCF1	CIITA	HLA-B	MADCAM1	PTPRC
ABL1	CKS1B	HLA-C	MAF	PTPRCAP
ABL2	CLEC4C	HLA-DMA	MAFB	PTPRD
ACVR2A	CMKLR1	HLA-DMB	MAGEA1	PTPRT
ADAMTS20	CMPK1	HLA-DOA	MAGEA10	PVR
ADGRE5	COL1A1	HLA-DOB	MAGEA12	PYGL
ADORA2A	CORO1A	HLA-DPA1	MAGEA3	RAD50
AFF1	CRBN	HLA-DPB1	MAGEA4	RAF1
AFF3	CREB1	HLA-DQA1	MAGEC2	RALGDS
AIF1	CREBBP	HLA-DQA2	MAGI1	RARA
AKAP9	CRKL	HLA-DQB2	MALT1	RB1
AKT1	CRTAM	HLA-DRA	MAML2	RECQL4
AKT2	CRTC1	HLA-DRB1	MAP2K1	REL
AKT3	CSF1R	HLA-E	MAP2K2	RET
ALK	CSF2RB	HLA-F	MAP2K4	RHOH

ALOX15B	CSMD3	HLA-F-AS1	MAP3K7	RNASEL
APC	CTAG1B	HLA-G	MAPK1	RNF2
AR	CTAG2	HLF	MAPK14	RNF213
ARG1	CTLA4	HMBS	MAPK8	RORC
ARID1A	CTNNA1	HNF1A	MARK1	ROS1
ARID2	CTNNB1	HOOK3	MARK4	RPS6
ARNT	CTSS	HRAS	MBD1	RPS6KA2
ASXL1	CX3CL1	HSP90AA1	MCL1	RRM1
ATF1	CX3CR1	HSP90AB1	MDM2	RUNX1
ATM	CXCL1	ICAM1	MDM4	RUNX1T1
ATR	CXCL10	ICK	MELK	S100A8
ATRX	CXCL11	ICOS	MEN1	S100A9
AURKA	CXCL13	ICOSLG	MET	SAMD9
AURKB	CXCL8	ID2	MIF	SAMHD1
AURKC	CXCL9	ID3	MITF	SBDS
AXL	CXCR2	IDH1	MKI67	SDHA
B3GAT1	CXCR3	IDH2	MLANA	SDHB
BAGE	CXCR4	IDO1	MLH1	SDHC
BAI3	CXCR5	IDO2	MLL	SDHD
BAP1	CXCR6	IFI27	MLL2	SELL
BATF	CYBB	IFI35	MLL3	9-Sep
BCL10	CYLD	IFI44L	MLLT10	SETD2
BCL11A	CYP2C19	IFI6	MMP2	SF3B1
BCL11B	CYP2D6	IFIH1	MMP9	SGK1
BCL2	DAXX	IFIT1	MN1	SH2D1A
BCL2L1	DCC	IFIT2	MPL	SH2D1B
BCL2L11	DDB2	IFIT3	MPO	SIT1
BCL2L2	DDIT3	IFITM1	MRC1	SKAP2
BCL3	DDR2	IFITM2	MRE11A	SLAMF7
BCL6	DDX58	IFNA17	MS4A1	SLAMF8
BCL9	DEK	IFNB1	MSH2	SMAD2
BCR	DGAT2	IFNG	MSH6	SMAD4

BIRC2	DICER1	IGF1R	MTOR	SMARCA4
BIRC3	DMBT1	IGF2	MTR	SMARCB1
BIRC5	DNMT3A	IGF2R	MTRR	SMO
BLM	DPYD	IGSF6	MUC1	SMUG1
BLNK	DST	IKBKB	MUTYH	SNAI1
BMPR1A	EBI3	IKBKE	MX1	SNAI2
BRAF	EFNA4	IKZF1	MYB	SOCS1
BRCA1	EGFR	IKZF2	MYC	SOX11
BRCA2	EGR2	IKZF3	MYCL1	SOX2
BRD3	EGR3	IKZF4	MYCN	SRC
BST2	EIF2AK2	IL10	MYD88	SRGN
BTK	EML4	IL10RA	MYH11	SSX1
BTLA	ENTPD1	IL12A	MYH9	SSX2
BUB1	EOMES	IL12B	NBN	STAT1
BUB1B	EP300	IL13	NCAM1	STAT3
C10orf54	EP400	IL15	NCF1	STAT4
C1QA	EPHA3	IL17A	NCOA1	STAT5A
C1QB	EPHA7	IL17F	NCOA2	STAT6
CA4	EPHB1	IL18	NCOA4	STK11
CARD11	EPHB4	IL1A	NCR1	STK36
CASC5	EPHB6	IL1B	NCR3	SUFU
CBL	ERBB2	IL2	NECTIN2	SYK
CBLB	ERBB3	IL21	NF1	SYNE1
CCL17	ERBB4	IL21R	NF2	TAF1
CCL18	ERCC1	IL22	NFATC1	TAF1L
CCL2	ERCC2	IL23A	NFE2L2	TAGAP
CCL20	ERCC3	IL2RA	NFKB1	TAL1
CCL21	ERCC4	IL2RB	NFKB2	TAP1
CCL22	ERCC5	IL2RG	NFKBIA	TARP
CCL3	ERG	IL3RA	NIN	TBP
CCL4	ESR1	IL4	NKG7	TBX21
CCL5	ETS1	IL6	NKX2-1	TBX22

CCNB2	ETV1	IL6ST	NLRP1	TCF12
CCND1	ETV4	IL7	NOS2	TCF3
CCND2	EXT1	IL7R	NOTCH1	TCF7
CCNE1	EXT2	ING4	NOTCH2	TCF7L1
CCR1	EZH2	IRF1	NOTCH3	TCF7L2
CCR2	FAM123B	IRF4	NOTCH4	TCL1A
CCR4	FANCA	IRF9	NPM1	TDO2
CCR5	FANCC	IRS1	NRAS	TET1
CCR6	FANCD2	IRS2	NRP1	TET2
CCR7	FANCF	ISG15	NSD1	TFE3
CD14	FANCG	ISG20	NT5E	TFRC
CD160	FANCI	ITGA1	NTN3	TGFB1
CD163	FAS	ITGA10	NTRK1	TGFBR2
CD19	FASLG	ITGA9	NTRK3	TGM7
CD1C	FBXW7	ITGAE	NUMA1	THBS1
CD1D	FCER1G	ITGAL	NUP214	TIGIT
CD2	FCGR1A	ITGAM	NUP98	TIMP3
CD209	FCGR2B	ITGAX	OAS1	TLR3
CD22	FCGR3A	ITGB1	OAS2	TLR4
CD226	FCGR3B	ITGB2	OAS3	TLR7
CD244	FCRLA	ITGB3	PAK3	TLR8
CD247	FGFR1	ITGB7	PALB2	TLR9
CD27	FGFR2	ITK	PARP1	TLX1
CD274	FGFR3	JAK1	PAX3	TNF
CD276	FGFR4	JAK2	PAX5	TNFAIP3
CD28	FH	JAK3	PAX7	TNFAIP8
CD33	FLCN	JAML	PAX8	TNFRSF14
CD37	FLI1	JCHAIN	PBRM1	TNFRSF17
CD38	FLT1	JUN	PBX1	TNFRSF18
CD3D	FLT3	KAT6A	PDCD1	TNFRSF4
CD3E	FLT4	KAT6B	PDCD1LG2	TNFRSF9
CD3G	FN1	KDM5C	PDE4DIP	TNFSF10

CD4	FOXL2	KDM6A	PDGFB	TNFSF13B
CD40	FOXM1	KDR	PDGFRA	TNFSF14
CD40LG	FOXO1	KEAP1	PDGFRB	TNFSF18
CD44	FOXO3	KIAA0101	PECAM1	TNFSF4
CD47	FOXP1	KIR2DL1	PER1	TNFSF9
CD48	FOXP3	KIR2DL2	PGAP3	TNK2
CD52	FOXP4	KIR2DL3	PGF	TOP1
CD53	FUT4	KIT	PHOX2B	TOP2A
CD6	FYB	KLF2	PIK3C2B	TP53
CD63	FZR1	KLF6	PIK3CA	TP63
CD68	G6PD	KLRB1	PIK3CB	TPR
CD69	GADD45GIP1	KLRD1	PIK3CD	TRIM24
CD70	GAGE1,GAGE12I,GAGE12F	KLRF1	PIK3CG	TRIM29
CD74	GAGE10	KLRG1	PIK3R1	TRIM33
CD79A	GAGE12J	KLRK1	PIK3R2	TRIP11
CD79B	GAGE13	KRAS	PIM1	TRRAP
CD80	GAGE2C,GAGE2A,GAGE2E	KREMEN1	PKHD1	TSC1
CD83	GATA1	KRT5	PLAG1	TSC2
CD86	GATA2	KRT7	PLCG1	TSHR
CD8A	GATA3	LAG3	PLEKHG5	TUBB
CD8B	GBP1	LAMP1	PMEL	TWIST1
CDC73	GDNF	LAMP3	PML	TYROBP
CDH1	GNA11	LAPTM5	PMS1	UBR5
CDH11	GNAQ	LCK	PMS2	UGT1A1
CDH2	GNAS	LCN2	POLR2A	USP9X
CDH20	GNLY	LEXM	POT1	VCAM1
CDH5	GPR124	LIFR	POU2AF1	VEGFA
CDK1	GPR18	LILRB1	POU5F1	VHL
CDK12	GRAP2	LILRB2	PPARG	VTCN1
CDK4	GRM8	LMNA	PPP2R1A	WAS
CDK6	GUCY1A2	LPHN3	PRDM1	WHSC1
CDK8	GUSB	LPP	PRF1	WRN

CDKN2A	GZMA	LRG1	PRKAR1A	WT1
CDKN2B	GZMB	LRP1	PRKDC	XAGE1B
CDKN2C	GZMH	LRP1B	PSIP1	XPA
CDKN3	GZMK	LST1	PSMB9	XPC
CEACAM1	HAVCR2	LTF	PTCH1	XPO1
CEACAM8	HCAR1	LTK	PTEN	XRCC2
CEBPA	HERC6	LY9	PTGS2	ZAP70
CHEK1	HGF	LYZ	PTK7	ZBTB46
CHEK2	HIF1A	M6PR	PTPN11	ZEB1
CIC	HLA-A	MAD2L1	PTPN6	ZNF384
			PTPN7	ZNF521

[0079] Model #2 – The 4 Gene Model

[0080] According to an embodiment, the 4 gene model utilizes a decision tree model to select the best minimal set of TILs or T-cell activation genes for prediction. The 4-gene model was derived, for example, with an independent decision tree machine learning approach from the initial selection of 54 genes, although other genes are possible. The machine learning algorithm selected a subset of genes and built a human-interpretable decision tree that best distinguishes the responders and nonresponders from the whole population of 87 training samples. The automatically selected four genes includes two genes related to T-cell activation, a gene related to immune response regulation and a gene related to cytokine signaling.

[0081] Referring to FIG. 3, in one embodiment, is a schematic representation of a decision tree for the 4 gene model. The identity of the four genes utilized, and the specific cut-off values, in the 4 gene model can vary. According to one embodiment, the four genes are PD-L1, TGFB1, TBX21, and BTLA, and the cut-off values for each of these genes are 171, 2043, 70.56, and 26.38, respectively. According to a further embodiment, the first gene is PD-L1, the second gene is TGFB1, the third gene is TBX21, and the fourth gene is BTLA. The decision tree provides a series of YES or NO decisions utilizing the expression level (in normalized reads per million (nRPM)) of each gene. For example, if the nRPM of gene 1 is less than a certain threshold or cutoff, the tumor is determined to be a responder. If the nRPM of gene 1 is more than the threshold or cutoff, the decision tree proceeds to the next gene. The

results of the decision tree classify the tumor as either a responder or a non-responder, as shown in FIG. 3.

[0082] According to one embodiment, not shown in FIG. 3, the 4 gene model can be determined or designed to classify tumors as being a responder or non-responder to a therapy, and can also be determined or designed to predict the risk of hyper-progression of cancer. For example, the 4 gene model can be determined or designed to identify a subset of non-responders that have a risk of hyper-progressive disease. The risk of hyper-progressive disease may be qualitative and/or quantitative.

[0083] Referring to FIG. 4, in one embodiment, is a graph of results for the 4 gene model using a panel of 87 retrospective samples for training. As shown in the graph, the results have a PPV of 72% for 43% of the population, a NPV of 92% for 49% of the population, and no indeterminate group.

[0084] Model #3 – The Immune Function Model

[0085] According to an embodiment, the immune function model utilizes 13 genes representing immune cell infiltration, 23 genes for T-cell activation, 10 genes for cytokine signaling, and 8 genes for immune response regulation, although other genes are possible. The immune function model utilizes a decision tree learning method, similar to the 4-gene model. However, instead of evaluating each individual gene's predictive importance, the immune function model takes the weighted average relative rank of multiple genes in a given immune functional group that include immune cell infiltration, immune response regulation, T-cell activation, and cytokine signaling. The relative rank is established by ranking a gene's normalized expression value (nRPM) against those of a reference population, and further normalizing the rank into a same range from 0 to 100. The immune function relative rank collectively reflects the degree of expressions of multiple genes with same function in comparison to a reference population.

[0086] Referring to FIG. 5, in one embodiment, is a schematic representation of a decision tree for the immune function model. The identity of the gene sets can vary. The decision tree provides a series of YES or NO decisions utilizing a relative rank cutoff for each of the four different immune functional groups (immune cell infiltration, immune response regulation, T-cell activation, and cytokine signaling), and can classify tumors as being a responder to therapy, a non-responder to therapy, or an indeterminate responder to therapy.

[0087] According to one embodiment, not shown in FIG. 5, the immune function model can be determined or designed to classify tumors as being a responder, indeterminate

responder, or non-responder to a therapy, and can also be determined or designed to predict the risk of hyper-progression of cancer. According to one embodiment, the four functions are immune cell infiltration, immune response regulation, T-cell activation, and cytokine signaling, and the cut-off values for each of these functions are 58.9, 60.25, 42.98 and 69.78, respectively. For example, the immune function model can be determined or designed to identify a subset of non-responders that have a risk of hyper-progressive disease. The risk of hyper-progressive disease may be qualitative and/or quantitative.

[0088] Referring to FIG. 6, in one embodiment, is a graph of results for the immune function model using a panel of 87 retrospective samples for training. As shown in the graph, the results have a PPV of 72% for 21% of the population, a NPV of 85% for 55% of the population, and an indeterminate group representing 24% of the population.

[0089] According to an embodiment, the immune function model utilizes expression information for 54 genes (13 genes representing immune cell infiltration, 23 genes for T-cell activation, 10 genes for cytokine signaling, and 8 genes for immune response regulation), although many other gene combinations are possible, including but not limited to the genes identified in Table 2. Table 3 includes a list of genes for which expression information can be utilized in the immune function model.

[0090] TABLE 3. Genes Utilized for an Embodiment of the Immune Function Model

Immune cell infiltration genes (13)	T-cell activation genes (23)		Cytokine signaling genes (10)	Immune response regulation genes (8)
CD8A	CD27	CD274 (PD-L1)	IL10	ADORA2A
CD8B	CD28	PDCD1LG2 (PD-L2)	IL1B	GATA3
CD3D	CD40	CTLA4	TNF	CD38
CD3E	CD40LG	OX40	TGFB1	ENTPD1 (CD39)
CD3G	CD80	OX40LG	CCR2	IDO1
CD2	CD86	GZMB	MX1	KLRD1
CD4	TNFRSF9 (CD137)	IFNG	CXCR6	STAT1
FOXP3	TNFRSF18 (GITR)	TNFRSF14	CXCL10	BTLA

CD68	ICOS	TBX21	DDX58	
CD163	ICOSLG	VISTA	CCL2	
MS4A1 (CD20)	LAG3			
CSF1R	HAVCR2 (TIM3)			
SLAMF4	PDCD1 (PD-1)			

[0091] Model #4 – The Primary Immune Marker Model

[0092] The primary immune marker model or approach is optional, and analyzes one or more immune markers utilizing immunohistochemistry. According to an embodiment, the primary immune marker model or approach analyzes PD-L1 protein expression and/or tumor infiltrating lymphocyte (TIL) expression (including but not limited to CD3 and/or CD8) using immunohistochemistry. According to another embodiment, the primary immune marker model or approach analyzes PD-L1 and/or PD-L2 copy number gain utilizing fluorescent in situ hybridization (FISH) methodology.

[0093] Multiple Model Correlation

[0094] According to an embodiment, the output from two or more of the 54 gene model, the 4 gene model, and the immune function model are combined to provide a final recommendation for personalized oncology therapies.

[0095] For example, referring to FIG. 7, in one embodiment, output from the 54 gene model, the 4 gene model, and the immune function model were combined using a Bayesian Model Averaging (BMA) for final prediction. According to an embodiment, the BMA algorithm is similar to the concept of majority voting, however, the algorithm also takes advantage of each individual model's performance prior probability distribution to optimize the final prediction.

[0096] According to a further embodiment as shown in FIG. 7, output from the 54 gene model, the 4 gene model, the immune function model, and the primary immune marker model were combined using a BMA for a final prediction.

[0097] Referring to FIG. 8, in one embodiment, is a graph of results for the BMA using a panel of 87 retrospective samples for training, and results from the 54 gene model, the 4 gene model, and the immune function model. As shown in the graph, the results have a PPV of 96% for 30% of the population, a NPV of 90% for 70% of the population, and no indeterminate group.

[0098] Referring to FIG. 9, in one embodiment, is a table of overall results from each of the 54 gene model, the 4 gene model, the immune function model, and the Bayesian Model Averaging (BMA).

[0099] The various approaches and models described or otherwise envisioned herein utilize a retrospective training panel and machine learning to derive a 54 gene model, a 4 gene model, and an immune function model. As the various approaches and models utilize larger and/or different retrospective training panels, the various genes and models may vary.

[00100] The Immune Report Card

[00101] Referring to FIG. 10 is a method 100 for generating a report comprising a likelihood of a patient's tumor microenvironment to respond to immunotherapy. According to an embodiment, the report also comprises one or more personalized treatment options based on the patient's comprehensive immune profile. According to an embodiment, the method utilizes three or more of five data inputs to generate the data necessary to determine the likelihood(s) and the personalized treatment options. According to an embodiment, those five data inputs can comprise at least:

1. RNA-seq to measure relative transcript levels of genes associated with tumor infiltrating lymphocytes (TILs) and T-cell receptor signaling (TCRS) genes associated with anti-cancer immune response and immunotherapeutic targets;
2. DNA-seq to estimate mutational burden (MUB);
3. Immunohistochemistry (IHC) to measure PD-L1 protein expression and pattern of tumor infiltrating lymphocytes (TILS) expression (CD3 and CD8);
4. PCR to assess microsatellite instability (MSI); and/or
5. Fluorescent in situ hybridization (FISH) to detect PD-L1/L2 copy number gain

[00102] Accordingly, at step 110 of the method, one or more tumor samples or specimens are collected. The samples may be collected using any method now known or developed in the future. The sample may be obtained directly from the patient and/or tumor, or may be obtained from a sample previously obtained from the patient and/or tumor. The sample may be a tumor sample, or may be a non-tumor sample obtained from an individual. The sample may be analyzed immediately, and/or may be stored for future analysis. Accordingly, the sample may be processed for shipping, storage, and/or for any other current or future use.

[00103] At step 120 of the method, RNA-seq data input is obtained. According to an embodiment, the RNA-Seq data input comprises a next generation sequencing (NGS) assay that uses amplicon-based targeted NGS for digital gene expression detection to interrogate 395 genes representing immune-related gene functions. One embodiment may be focused on

54 of these 395 genes, with an additional 10 genes used as controls. In this or other embodiments, one or more additional genes – such as the genes listed in one or more of the tables herein – may be analyzed and provided in the method results.

[00104] According to an embodiment, the TILs gene component of the RNA-seq data input comprises genes identified to classify subsets of infiltrating immune cells. The genes may be the 11 genes identified herein (CD163, CD2, CD3D, CD3E, CD3G, CD4, CD68, CD8A, CD8B, FOXP3, and CD20), or may comprise different and/or additional genes. The application of TILs as predictive immune biomarkers or prognostic markers of survival has been studied in a wide variety of tumor types. Evidence has shown that CD3+ CD8+ cytotoxic lymphocytes are positively associated with response to CPIs, while FOXP3+ Tregs have a negative association. The pattern of TILs in a tumor has also been shown to have importance.

[00105] According to an embodiment, the T-cell receptor signaling (TCRS) component of the RNA-seq data input comprises genes expressed on immune infiltrating cells, neoplastic cells, or other cells of the tumor microenvironment, and are classified by Immune Phenotypes as either directly involved in checkpoint blockade or other functions related to the adaptive immune response. The genes may be the 43 genes identified herein (see, e.g., Table 1), or may comprise different and/or additional genes.

[00106] According to an embodiment, Immune Phenotypes associated with checkpoint blockade are typically classified as a receptor and associated ligand. According to the subset of TCRS associated genes, either the ligand or the receptor is expressed by one or more subsets of T-cells. Genes related to checkpoint blockade can be further divided into those that are the direct target of one or more checkpoint inhibitor drugs versus those that are not. Direct targets of checkpoint inhibitor drugs may be either the receptor or ligand, but not both simultaneously. An example of a checkpoint inhibitor drug and target is ipilimumab and the receptor CTLA-4 that is expressed on activated T-cells. According to an embodiment, the report of likelihood(s) and the personalized treatment options also reports genes related to checkpoint blockade, as Checkpoint Blockade (PD-1, CTLA-4), Checkpoint Blockade (Other), or T-cell Primed, as shown in Table 3. Checkpoint Blockade and Checkpoint Blockade (Other) are related to co-inhibitory signaling for effector T- cells, while T-cell Primed is co-stimulatory. Other immune phenotypes that are associated with the adaptive immune response that indirectly impact TCRS include Myeloid Suppression, Pro-inflammatory Response, Anti-inflammatory Response, and Metabolic Immune Escape.

[00107] TABLE 3. Analyzed immune phenotypes.

Immune Phenotype	Key Markers	Action
Checkpoint Blockade (PD-1/CTLA4)	PD-L1, PD-1, CTLA4, PD-L2	T-cell Inhibition
Checkpoint Blockade (Other)	BTLA, LAB3, VISTA (B7-H5), TIM3, TNFRSF14 (HVEM; CD270)	T-cell Inhibition
T-cell Primed	CD27, CD28, CD40, CD40LG, TNFRSF9 (CD137), TNFRSF18 (GITR), ICOS, ICOSLG, OX40, OX40LG, INFG, GZMB, TBX21 (T-bet)	T-cell Activation
Myeloid Suppression	CSF1R, CD68, CD163, CCR2, CCL2	Promote M2 TAMs
Anti-Inflammatory Response	IL10, TGFB1	Promote MDSCs
Metabolic Immune Escape	IDO1, ADORA2A, CD39	Self-amplifying T-reg loop
Pro-Inflammatory Response	CXCL10, CXCR6, IL1B, STAT1, TNF, DDX58, MX1	Promote NK T-cell functions

[00108] The expression of each gene is compared to a reference population, normalized to a value between 1 and 100, and referred to as the relative rank. According to one embodiment, the baseline reference population for this method consisted of RNA-seq results derived from 167 unique tumors. The top 95th percentile of scores, or relative rank, which are values equal to or greater than 95 are interpreted as very high expression, while the 85th to 94th percentile are interpreted as high expression. The bottom 50th percentile of scores, or relative rank, which are values less than 49 are considered low or very low expression. Scores between 50 and 85 are considered moderate expression. The interpretation of immune phenotypes is derived from the mean of all genes for that class and are ranked as normalized values in the same manner as expression of individual genes.

[00109] At step 130 of the method, of the method, DNA-seq data input is obtained. According to an embodiment, the DNA-Seq data input comprises a 1.75 Mb AmpliSeq capture of 409 oncogenes with full exon coverage that evaluates a total of 6,602 exons

covering 1,165,294 base pairs of unique exon DNA in an all-exon mutational profiling assay. Mutational Burden (MUB) is reported as the number of mutations per megabase (Mb) of exonic DNA. MuB can be calibrated against a subset of samples with whole exome sequencing for development of a variant calling pipeline that provides 20x coverage at $\geq 90\%$ of the unique exon DNA in the IRC panel. According to an embodiment, the MuB was calibrated against four clinically relevant peer-reviewed publications reporting a correlation of high mutational burden with response to checkpoint inhibitors in melanoma. Using an equivalent whole exome and these four references as a calibrator the cut-off values for MuB as measured by number of mutations per Mb DNA was established using an internal reference population of 167 patients. In this regard MuB is classified as “very high”, “high”, “intermediate”, “low”, and “very low”. While classifications of high and very high MuB are more likely to be responders, as a single biomarker, this measurement lacks sensitivity and specificity, and should not be used independently of other assay results in IRC.

[00110] At step 140 of the method, immunohistochemistry data input is obtained. According to an embodiment, the immunohistochemistry data input comprises a measurement of PD-L1 protein expression and tumor infiltrating lymphocytes (TILs) expression (CD3 and CD8). According to an embodiment, the immunohistochemistry data input is obtained utilizing an automated DAKO platform and commercially available antibodies to provide expression data for PD-L1, CD3 and CD8. The method can report protein expression patterns for all three analytes to better define the multi-dimensional interactions that occur in the tumor microenvironment, as well as a semi-quantitative measurement of expression for PD-L1.

[00111] According to an embodiment, for melanoma, PD-L1 is performed using the PD-L1 IHC 28-8 FDA approved assay, and follows scoring guidelines for reporting the percentage of neoplastic cells displaying membranous staining of any intensity. The PD-L1 22C3 FDA approved assay is used to test non-small cell lung cancer and other tumor types, with PD-L1 protein expression determined by using Tumor Proportion Score (TPS), which is the percentage of viable tumor cells showing partial or complete membrane staining at any intensity.

[00112] According to an embodiment, the TILs expression pattern (CD3 and CD8 as measured by IHC) is reported as “infiltrating”, “non-infiltrating”, or “minimal to absent”. An “infiltrating” pattern refers to staining of TILs within groups of neoplastic cells in the majority of the tumor examined. “Non-infiltrating” refers to TILs present, but inconsistent pattern of infiltrating groups of neoplastic cells in the majority of the tumor examined. The

non-infiltrating pattern of staining for CD3 and CD8 includes cases with an abundance of TILs at the advancing edge of the tumor, but without infiltration of the neoplastic cells. The minimal to absent pattern is essentially minimal to no TILs present within any part of the tumor.

[00113] According to an embodiment, CD3 highlights T-cells (referred to as TILs in this assay) and is useful to identify the T-cell population associated with the neoplasm. CD8 highlights cytotoxic T-cells which (when found in the midst of neoplastic cells) tend to be reflective of response to checkpoint inhibitors (CPIs). The information provided in this report may be used by physicians to decide if immunotherapy with one or more FDA-approved checkpoint inhibitors may benefit this patient.

[00114] According to one embodiment, the PCR component of the method uses five markers including two mononucleotide repeat markers (BAT-25, BAT-26) and three dinucleotide repeat markers (D2S123, D5S346 and D17S250) to detect microsatellite instability (MSI). According to another embodiment, the NGS component of the method uses up to 100 homopolymer, dinucleotide, trinucleotide and/or tetranucleotide markers to detect microsatellite instability (MSI). In either case, the results are reported as “MSI-high”, “MSI-low”, or “MSS” (microsatellite stable). MSI typically used as a prognostic marker of survival in the setting of Lynch syndrome, also called hereditary nonpolyposis colorectal cancer (HNPCC), is an FDA marker of response to checkpoint inhibitors in colorectal cancer and second line treatment in solid tumors with no other therapeutic options.

[00115] Colorectal carcinoma, endometrial carcinomas, and other types of neoplasms with MSI-H may be sporadic (ie, microsatellite instability is found only in the neoplasm and is therefore not part of a hereditary condition) or secondary to Lynch syndrome, a hereditary condition (ie, related to a familial genetic mutation in a DNA repair gene, typically MLH1, PMS2, MSH2, or MSH6). Although an MSI test can determine the presence of microsatellite instability it cannot determine which specific DNA-repair gene is affected. Immunohistochemistry against MLH1, PMS2, MSH2, or MSH6 can be used to find the specific protein that is affected. Depending on the expression loss pattern and the carcinoma tissue of origin, slightly different strategies are indicated. For specifics, a review of the National Comprehensive Cancer Network (NCCN) guidelines is recommended. The presence of Lynch syndrome is definitely determined when a pathogenic DNA-repair gene mutation is found in non-neoplastic tissue from the patient. In such cases, genetic counseling is recommended with involvement of family members who are at risk of also having Lynch syndrome. MSI-H colorectal cancers (both sporadic and Lynch syndrome related) are known

to have a different clinical profile (prognosis and chemotherapy responsiveness) from MSS (microsatellite stable) colorectal cancers.

[00116] According to an embodiment, the FISH component of the method measures copy number of PD-L1 (CD274) and PD-L2 (PDCD1LG2), two PD-1 ligands that when amplified are associated with PD-L1 expression. The two genes are located 40kb apart on 9p24.1 and are detected using a pool of fluorescently labeled BAC clones that map to the gene region (RP11-635N21, RP11-812M23 and RP11-485M14). The results are reported using ASCO–CAP HER2 Test Guideline Recommendations for determining copy number as amplified, equivocal, or not amplified. An amplified result is when the ratio of PD-L1/2 to CEP9 is equal to or greater than 2.0 for any copy number value of PD-L1/2, or when the ratio of PD-L1/2 to CEP9 is less than 2.0 and the copy number value of PD-L1/2 is equal to or greater than 6.0. An equivocal result occurs when the ratio of PD-L1/2 to CEP9 is less than 2.0 and the copy number value of PD-L1/2 is equal to or greater than 4.0 and less than 6.0. A not amplified result occurs when the ratio of PD-L1/2 to CEP9 is less than 2.0 and the copy number value of PD-L1/2 is equal to or greater than 4.0. IRC uses the prior NYS-CLEP approved OmniSeq “PD-L1 and PD-L2 amplification and Validation Data SOP” for PD-L1/2 copy number testing.

[00117] At step 150 of the method, the data input from one or more of steps 120, 130, and 140 is collated and analyzed. For example, according to an embodiment, the output are combined to provide a final recommendation for personalized oncology therapies. For example, according to an embodiment, the output are combined using a Bayesian Model Averaging (BMA) for final prediction. According to an embodiment, the BMA algorithm is similar to the concept of majority voting, however, the algorithm also takes advantage of each individual model’s performance prior probability distribution to optimize the final prediction. However, many other methods for collating the data from the two or more data inputs are possible.

[00118] At step 160 of the method, a clinical report is provided to the patient’s clinician. The report comprises at least a likelihood of the patient’s tumor microenvironment to respond to particular immunotherapies, and/or combination immunotherapies, and/or immunotherapy clinical trials. According to an embodiment, at step 170 of the method the data inputs, collated data, and/or final analysis is utilized to generate one or more personalized treatment options based on the patient’s comprehensive immune profile. Thus, the report provided to the patient may also comprise one or more personalized treatment options.

[00119] According to an embodiment, the report provided to the patient's clinician can comprise one or more of the following:

[00120] • Priority Immune Markers - At-a-glance information for clinically relevant immune markers for therapeutic management including but not limited PD-L1 expression, pattern and expression of tumor infiltrating lymphocytes (TILS), microsatellite instability (MSI) mutational burden (MUB), and PD-L1/L2 copy number gain. Evidence for therapeutic associations for each marker can be provided in the context of the tumor type tested.

[00121] • Summary Interpretation – Pathologist interpretation and characterization of the overall immune status of this tumor based on both priority immune markers and TCRS gene expression. The Summary Interpretation can provide an assertion of likelihood of response to FDA approved checkpoint inhibitors and an overview of clinical trial opportunities trials based on immune phenotype assessment.

[00122] • Immune Phenotypes Summary – Overview of gene expression for several immune phenotypes, showing highly expressed genes within each phenotype as applicable. The genes associated with each immune phenotype are not intended to be all inclusive of genes with a similar or identical action, but rather are genes for which expression was rigorously validated. Some genes within a given immune phenotype are associated immunotherapeutic agents available as either FDA-approved therapies or clinical trials.

[00123] • Immune Phenotype Details – Gene level expression rank and interpretation of TCRS genes by immune phenotype. The expression of each gene is provided as a rank and interpretations as previously described in the RNA-Seq component. Immunotherapies associated with genes are listed without regard to tumor type tested. With the exception of PD-L1, evidence supporting over-expression of any immune phenotype gene and response to any associated immunotherapy, may be limited.

[00124] • Tumor Infiltrating Lymphocytes – Gene level expression rank and interpretation of TILS genes associated with differentiation of various immune-related cells.

[00125] • Clinical Trials – Clinical trials are displayed for the tumor type tested, for over-expressed markers with therapies in clinical development that are ranked high or very high. When there are no markers ranked high or very high, clinical trials are displayed for markers ranked moderately high. Trials related to the recruitment of lymphocytes into the tumor are always displayed when a tumor is considered non- inflamed.

[00126] Example – Analyzing Immune Response in Solid Tumors

[00127] According to an embodiment, the methodologies described or otherwise envisioned herein were utilized to analyze immune response in formalin-fixed paraffin-

embedded (FFPE) tumor specimens, in order to provide a characterization of the immunological tumor microenvironment as a guide for therapeutic decisions on patients with solid tumors.

[00128] As described herein, the analysis utilized RNA-seq data to semi-quantitatively measure the levels of transcripts related to anticancer immune responses and transcripts reflecting the relative abundance of tumor-infiltrating lymphocytes (TILs), as well as on DNA-seq data to estimate mutational burden. Although not described in this example, the analysis could comprise primary marker immune data as described or otherwise envisioned herein.

[00129] An embodiment of a methodology for obtaining and/or analyzing RNA-seq data and DNA-seq data is described below. However, it is understood that this is just one possible embodiment, and is therefore not limiting in any way. Other methodologies for obtaining and/or analyzing RNA-seq data and DNA-seq data are possible.

[00130] This assay, unlike existing mutational profiling assays, accurately matches patients to immunotherapeutic treatments based on the immunological configuration of their tumor, using a wide range of biomarkers.

[00131] Methods

[00132] In order to evaluate the analytical performance of the assay, 167 FFPE specimens and a subset of matched fresh frozen (FF) tissues from NSCLC, melanoma, renal cell carcinoma, head and neck squamous cell carcinoma (HNSCC), and bladder cancer patients were obtained. Specimens were collected under an institutional banking policy with informed patient consent, the study was approved by an internal review board review (IRB Protocol # BDR 073116) as per institutional policy for non-human subjects research.

[00133] The specimen included fine-needle aspiration biopsies, punch biopsies, needle core biopsies, incisional biopsies, excisional biopsies, and resection specimens from 2002-2016. For a subset of specimens, whole-exome sequencing and whole-transcriptome RNA-seq data were available as part of The Cancer Genome Atlas (TCGA) project for comparative purposes. Four human cell lines, lymphoblastoid GM12878 cells (ATCC, Manassas, VA), colorectal cancer KM-12 cells (NCI-Frederick Cancer DCTD, Bethesda, MD), NSCLC HCC-78 cells (DSMZ, Braunschweig, Germany), and large cell lymphoma SU-DHL-1 cells (ATCC) processed as FFPE blocks were also used for development and as internal run controls.

[00134] A board-certified anatomical pathologist reviewed a hematoxylin and eosin (H&E)-stained tumor section to identify the region(s) to be tested. Tumor surface area on the

H&E-stained section was ≥ 2 mm² per slide, with tumor cellularity $\geq 50\%$ and necrosis $\leq 50\%$. For the examination of potential pre-analytic interferences, non-malignant and necrotic tissues were also macro-dissected from corresponding unstained slides for nucleic acid extraction. Regions identified by the pathologist were used as guides to scrape tissue from 3-5 unstained slides. Genomic DNA and total RNA were simultaneously extracted from this material by means of the truXTRAC FFPE extraction kit (Covaris, Inc., Woburn, MA), following the manufacturer's instructions with some modifications. Following purification, RNA and DNA were eluted in 30 and 50 μ L water, respectively, and yield was determined by the Quant-iT RNA HS Assay and Quantifiler Human DNA Quantification Kits (both from Thermo Fisher Scientific, Waltham, MA), as per manufacturer's recommendation. A predefined yield of 10 ng RNA and 30 ng DNA was used as acceptance criteria to ensure adequate library preparation.

[00135] Run controls were established and used for library preparation, enrichment and NGS. They included both positive (MuB-DNA, GEX-RNA) and negative (MuB-DNA Negative, GEX-RNA Negative) controls, as well as a no template control (NTC, water). Positive controls provide templates for all targets for qualification and downstream normalization purposes, while negative controls monitor assay specificity. For RNA-seq, the NTC is used to identify the limit of detection at the individual sample level. For DNA-seq, the NTC is used to identify the threshold for false positivity at the run level. The performance characteristics of these five run controls were assessed across multiple weeks to develop thresholds and filters that serve as daily QC parameters on runs and samples.

[00136] The assay utilized the Oncomine Immune Response Research Assay (OIRRA) for GEX and the Comprehensive Cancer Panel (CCP) for MuB (Thermo Fisher Scientific). Both these panels use multiplexed gene-specific primer pairs and NGS to amplify nucleic acids extracted from FFPE slides. The OIRRA was adapted to quantify expression of 54 target genes using 10 constitutively expressed housekeeping (HK) genes as normalizers. All 409 cancer-related genes included in the CCP panel were used to estimate MuB from genomic DNA.

[00137] OIRRA libraries were prepared using the Ion AmpliSeq targeted sequencing technology (Thermo Fisher Scientific). Briefly, 10 ng RNA was reverse transcribed into cDNA and targets were amplified with a multiplex primer pool. For DNA-seq, CCP libraries were prepared using 30 ng DNA. Barcode adapters were ligated to partially digested amplicons, purified and normalized to 50 pM. Up to 16 equimolar RNA and DNA libraries were pooled prior to enrichment and template preparation using the Ion Chef system (Thermo

Fisher Scientific). 200-bp sequencing was performed on the Ion S5XL 540 chip to obtain 1.5-2.5M RNA-seq mapped reads and 100-150X DNA-seq mean depth per sample.

[00138] RNA-seq accuracy was evaluated by comparison with qRT-PCR for all reported genes for all samples, and with IHC for CD8A (based on automated image analysis for a subset of 57 samples with adequate material). For orthogonal qRT-PCR analysis, 100 ng RNA was reverse-transcribed, amplified and measured in triplicate by means of the TaqMan Gene Expression Assay monitoring the 54 target and 10 HK genes using the QuantStudio 7 Real-Time PCR System (Applied BioSystems, Foster City, CA). For IHC, 5 μ m-thick sections from tissue microarrays (TMAs) with three 0.6-mm tissue cores arrayed per tumor were stained with antibodies specific for CD8A (C8/144B, Dako, Agilent Technologies, Santa Clara, CA) as per standard procedures, and PD-L1 is run specific to tumor histology (22C3, Dako, or 28-8 Dako, or SP142 Ventana), as per FDA guidelines. Quantitative IHC data on CD8A+ T-cell counts were obtained using the Aperio Scanscope (Aperio Technologies, Inc., Vista, CA), based on 20X bright-field optical microscopy. Images were analyzed using eSlide Manager v12.2.1 (Aperio Technologies, Inc.) and the number of positive cells per sqmm for each TMA core was counted. Quantitative IHC expression data for PD-L1 was obtained by trained pathologists interpreting the tumor proportion score (TPS), H-score (HS), and modified H-score (MHS). A minimum of two evaluable cores out of three was required for inclusion in final analysis. Average number of CD8A+ cells per sqmm, TPS, HS, and MHS were derived from each sample upon individual analysis of at least two cores.

[00139] Sequencing data were first processed using the Torrent Suite software (v5.2.0) for reference mapping and base calling, during which validation-defined QC specifications for mapped reads, on-target reads, mean read length, mean depth, uniformity, and percent valid reads were used as acceptance criteria. To ensure high quality results, a QC system was developed based on NGS data generated at validation. The QC criteria were established for several metrics at the run, sample, amplicon and base-pair level for each nucleic acid type and run control thresholds, with defined values to accept or reject one or more aspects of sequencing. Likewise, specific QC metrics are monitored over time to detect any potential long-term assay drift. Quality filters are used at the amplicon level to remove counts below the threshold for detection, and at the base-pair level for low-quality variant calls.

[00140] RNA-seq absolute reads were generated using Torrent Suite's plugin immuneResponseRNA (v5.2.0.0). For each transcript, absolute read counts from the NTC were considered as the library preparation background and hence were subtracted from

absolute read counts of the same transcript of all other samples in the same preparation batch. To allow NGS measurements across runs to be comparable for evaluation and interpretation, background-subtracted read counts were subsequently normalized into normalized reads per million (nRPM) values as follows. Each HK gene background-subtracted reads was compared against a pre-determined HK reads per million (RPM) profile. The HK RPM profile was established based on the average RPM of multiple replicates of the GM12878 sample across different sequencing runs of the validation. This produced a fold-change ratio for each HK gene:

$$\text{Ratio of HK} = \frac{\text{Background Subtracted Read Count of HK}}{\text{RPM Profile of HK}}$$

[00141] After this, the median value of all HK ratios was used as the normalization ratio for the particular sample:

$$\text{Normalization Ratio} = \text{Median}(\text{all HK ratios})$$

[00142] The nRPM of all genes (G) of the particular sample (S) was then calculated as:

$$\text{nRPM}_{(S,G)} = \frac{\text{Background subtracted Read Count}_{(S,G)}}{\text{Normalization Ratio}_{(S)}}$$

[00143] DNA-seq variant calling was conducted using Ion Torrent Suite software's (v5.2.0) variantCaller (v5.2.0.34) plugin, which requires a minimal minor allele frequency (MAF) of 0.1 and a minimal coverage of 20X. A series of filters were applied to variants that generated the MuB-qualified variant subset meeting following criteria: at least one minor allele reads on both strands; < 0.2% MAF in 1,000 genomes, Exome Aggregation Consortium (ExAC), and Exome Sequencing Project (ESP) databases; missense or nonsense; has co-located somatic variants as in Ensembl team's curated database. The count of MuB-qualified variants was further normalized over the number of exonic bases with $\geq 20X$ coverage, as reflected in the input BAM file, to generate the normalized MuB score, interpreted as the number of mutations per million exonic bases. High MuB was defined as $> 2X$ standard deviation of the mean number of mutations per megabase DNA in a tumor reference population of varying histology.

[00144] For RNA-seq, the suitability of FFPE specimens was established by comparing RNA-seq results obtained from FFPE versus matched FF samples. Principal component analysis (PCA) was used to show that different sections obtained from a given FFPE sample do not lead to different RNA-seq results, that potential confounders such as the amount or quality of stroma are well tolerated by the assay, and that testing either multiple foci of a metastatic tumor or the primary lesion has a minimal impact on results (data not shown). For

MuB, sensitivity of variant detection was used to evaluate minimal threshold for percent neoplastic nuclei.

[00145] To compare RNA-seq with gold-standard Taqman qRT-PCR, nRPM values of target genes were log2 transformed, which allowed for appropriate comparison with ΔC_t values from qRT-PCR measurements. The Pearson product-moment correlation coefficient (R) was then calculated for each gene on log-transformed GEX measurements and ΔC_t values. Any gene with poor correlation values (< 0.7) was excluded from final reporting. To assess GEX correlation for pre-analytical, analytical and reproducibility studies, the R value for nRPM of 54 target genes was used as part of the variables. R was calculated for FFPE versus unstained sections, FF versus FFPE sections, varying percentage of non-neoplastic tissue content, necrotic tissue content, amount of input RNA (ng), genomic DNA (gDNA) contamination, batch size, linearity of expression, inter-run reproducibility, intra-run reproducibility, inter-operator reproducibility and inter-day reproducibility. Additionally, to demonstrate the linearity in absolute reads for different library dilutions, coefficient of determination (R^2) was used. Also, average coefficient of variation (CV) for the nRPM of the 54 target genes was calculated as a measure of dispersion of GEX measurements for various batch sizes.

[00146] To assess the correlation of MuB counts to gold standard TCGA whole-exome data, TCGA whole-exome counts were filtered to select genomic regions from the MuB panel. The TCGA variant count mapped to the panel's Browser Extensible Data (BED) file was then correlated with MuB counts using Pearson product-moment correlation. DNA stability was tested using two-tailed Student's t test between the mean MuB values of FFPE versus unstained sections. Average CV was calculated as a measure of variability in MuB measurements for FFPE versus unstained sections. CV was also used to demonstrate the effect of DNA input amount (ng) on MuB measurements.

[00147] While various embodiments have been described and illustrated herein, those of ordinary skill in the art will readily envision a variety of other means and/or structures for performing the function and/or obtaining the results and/or one or more of the advantages described herein, and each of such variations and/or modifications is deemed to be within the scope of the embodiments described herein. More generally, those skilled in the art will readily appreciate that all parameters, dimensions, materials, and configurations described herein are meant to be exemplary and that the actual parameters, dimensions, materials, and/or configurations will depend upon the specific application or applications for which the teachings is/are used. Those skilled in the art will recognize, or be able to ascertain using no

more than routine experimentation, many equivalents to the specific embodiments described herein. It is, therefore, to be understood that the foregoing embodiments are presented by way of example only and that, within the scope of the appended claims and equivalents thereto, embodiments may be practiced otherwise than as specifically described and claimed. Embodiments of the present disclosure are directed to each individual feature, system, article, material, and/or method described herein. In addition, any combination of two or more such features, systems, articles, materials, and/or methods, if such features, systems, articles, materials, and/or methods are not mutually inconsistent, is included within the scope of the present disclosure.

[00148] The claims should not be read as limited to the described order or elements unless stated to that effect. It should be understood that various changes in form and detail may be made by one of ordinary skill in the art without departing from the spirit and scope of the appended claims. All embodiments that come within the spirit and scope of the following claims and equivalents thereto are claimed.

Claims

What is claimed is:

1. A method for generating an immune score, the method comprising the steps of:
 - determining a qualitative and/or quantitative assessment of tumor infiltrating lymphocytes in a sample;
 - determining a qualitative and/or quantitative assessment of T-cell receptor signaling in the sample;
 - determining a qualitative and/or quantitative assessment of mutational burden in the sample; and
 - generating, using a predictive algorithm, an immune score based on the determined qualitative and/or quantitative assessment of tumor infiltrating lymphocytes, the determined qualitative and/or quantitative assessment of T-cell receptor signaling, and the determined qualitative and/or quantitative assessment of mutation burden.
2. The method of claim 1, further comprising the step of determining, based on the immune score, one or more possible treatment therapies.
3. The method of claim 1, wherein the predictive algorithm is trained and updated using machine learning.
4. A method for analyzing a tumor, the method comprising the steps of:
 - determining a qualitative and/or quantitative assessment of tumor infiltrating lymphocytes in a sample;
 - determining a qualitative and/or quantitative assessment of T-cell receptor signaling in the sample;
 - determining a qualitative and/or quantitative assessment of mutational burden in the sample; and
 - classifying, based on the determined qualitative and/or quantitative assessment of tumor infiltrating lymphocytes, the determined qualitative and/or quantitative assessment of T-cell receptor signaling, and the determined qualitative and/or quantitative assessment of mutational burden, the sample as a responder to a therapy or a non-responder to a therapy.
5. The method of claim 4, further comprising the step of determining a response of the tumor, based on the determined classification, to one or more possible treatment therapies.

6. The method of claim 4, wherein the sample is classified as an indeterminate responder to therapy.

7. The method of claim 4, wherein a sample classified as a non-responder to therapy and non-responder to therapy may be further classified as being at risk for hyper-progression of cancer.

8. The method of claim 4, further comprising the step of correlating the determined classification of the sample with a second classification.

9. The method of claim 8, further comprising the step of determining a response of the tumor, based on the correlation, to one or more possible treatment therapies.

10. The method of claim 4, further comprising the step of determining immunohistochemistry data for the sample, wherein said classifying step is further based on the determined immunohistochemistry data.

11. A method for analyzing a tumor, the method comprising the steps of:
determining a quantitative assessment of an expression level of at least four genes in a sample;

classifying, if the determined quantitative assessment of an expression level of the first of the four genes exceeds a predetermined threshold, the sample as a responder to a therapy, or proceeding to the next step if the determined quantitative assessment of the expression level of the first of the four genes does not exceed the predetermined threshold;

classifying, if the determined quantitative assessment of an expression level of the second of the four genes is greater than a predetermined threshold, the sample as a non-responder to a therapy, or proceeding to the next step if the determined quantitative assessment of the expression level of the second of the four genes does exceed the predetermined threshold;

classifying, if the determined quantitative assessment of an expression level of the third of the four genes exceeds a predetermined threshold, the sample as a responder to a therapy, or proceeding to the next step if the determined quantitative assessment of the expression level of the third of the four genes does exceed the predetermined threshold; and

classifying, if the determined quantitative assessment of an expression level of the fourth of the four genes exceeds a predetermined threshold, the sample as a non-responder to a therapy, or classifying, if the determined quantitative assessment of the expression level of the fourth of the four genes does exceed the predetermined threshold, the sample as a responder to a therapy.

12. The method of claim 11, further comprising the step of determining response, based on one of the determined classification, to one or more possible treatment therapies.

13. The method of claim 11, wherein the sample is classified as a non-responder to therapy.

14. The method of claim 11, wherein a sample classified as a non-responder to therapy is further classified as being at risk for hyper-progression of cancer.

15. The method of claim 11, further comprising the step of correlating the determined classification of the sample with a second classification.

16. The method of claim 15, further comprising the step of determining, based on the correlation, one or more possible treatment therapies.

17. A method for analyzing a tumor, the method comprising the steps of:

- determining a qualitative and/or quantitative assessment of a plurality of genes related to immune cell infiltration in a sample;
- determining a qualitative and/or quantitative assessment of a plurality of genes related to T-cell activation in the sample;
- determining a qualitative and/or quantitative assessment of a plurality of genes related to cytokine signaling in the sample;
- determining a qualitative and/or quantitative assessment of a plurality of genes related to immune response regulation in the sample;
- normalizing each of the determined qualitative and/or quantitative assessments;
- if the normalized qualitative and/or quantitative assessment of immune cell infiltration exceeds a predetermined threshold, proceeding to an analysis of the normalized qualitative and/or quantitative assessment of T-cell activation in the sample, wherein the sample is classified as a responder if the normalized qualitative and/or quantitative assessment of T-cell activation exceeds a predetermined threshold, and wherein the sample is classified as a non-responder if the normalized qualitative and/or quantitative assessment of T-cell activation does not exceed the predetermined threshold;
- proceeding, if the normalized qualitative and/or quantitative assessment of immune cell infiltration falls below the predetermined threshold, to the next step;
- classifying, if the normalized qualitative and/or quantitative assessment of immune regulation response exceeds a predetermined threshold, the sample as an indeterminate responder, or proceeding, if the normalized qualitative and/or quantitative

assessment of immune regulation response does not exceed the predetermined threshold, to the next step; and

classifying, if the normalized qualitative and/or quantitative assessment of cytokine signaling exceeds a predetermined threshold, the sample as an indeterminate responder, or classifying, if the normalized qualitative and/or quantitative assessment of cytokine signaling does not exceed the predetermined threshold, the sample as a non-responder.

18. The method of claim 17, further comprising the step of determining a response of the tumor, based on one of the determined classification, to one or more possible treatment therapies.

19. The method of claim 17, wherein a sample classified as a non-responder to therapy is further classified as being at risk for hyper-progression of cancer.

20. The method of claim 17, further comprising the step of correlating the determined classification of the sample with a second classification.

21. The method of claim 20, further comprising the step of determining, based on the correlation, one or more possible treatment therapies.

22. The method of claim 17, further comprising the step of determining immunohistochemistry data for the sample, wherein one or both of said classifying steps is further based on the determined immunohistochemistry data.

23. A method for providing an immune profile tumor report, the method comprising the steps of:

obtaining one or more samples from a tumor of the patient;

generating, from the one or more samples, RNA sequencing data comprising information about expression of a plurality of tumor-infiltrating lymphocyte proteins and a plurality of T-cell receptor signaling proteins;

generating, from the one or more samples, DNA sequencing data comprising mutational burden information about a plurality of genes;

generating, from the one or more samples, immunohistochemistry data;

calculating, from the RNA sequencing data, the DNA sequencing data, and the immunohistochemistry data, a likelihood of the patient's tumor to respond to a plurality of possible treatments; and

providing an immune profile report, the immune profile report comprising the calculated likelihoods of the patient's tumor to respond to the plurality of possible treatments.

54-Gene + MuB Model Predictions

RESPONDERS AND NON-RESPONDERS

Percentage of Samples, n = 87

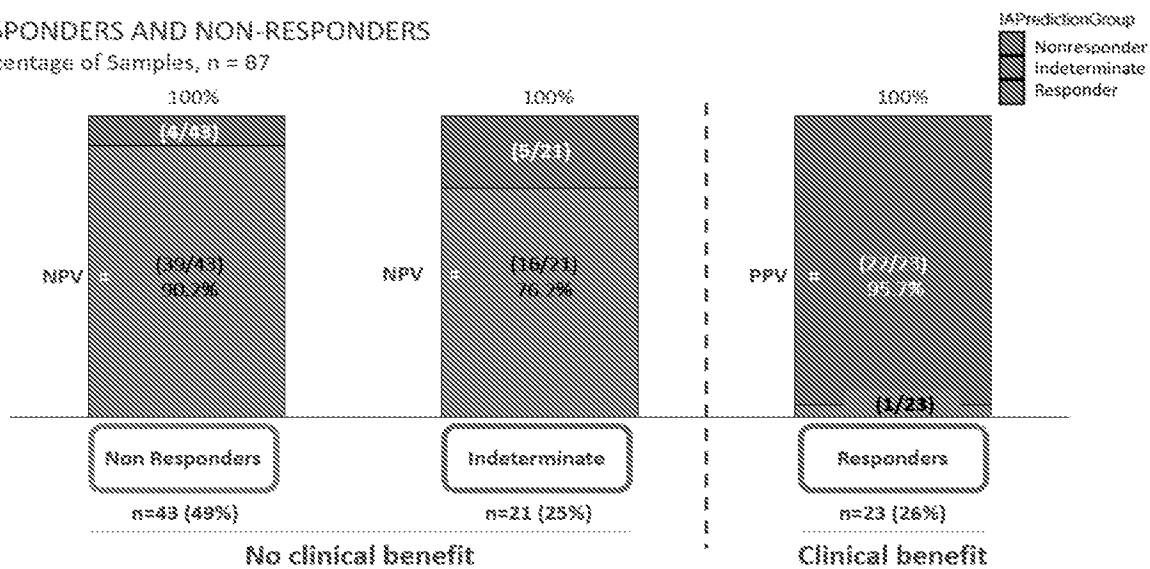
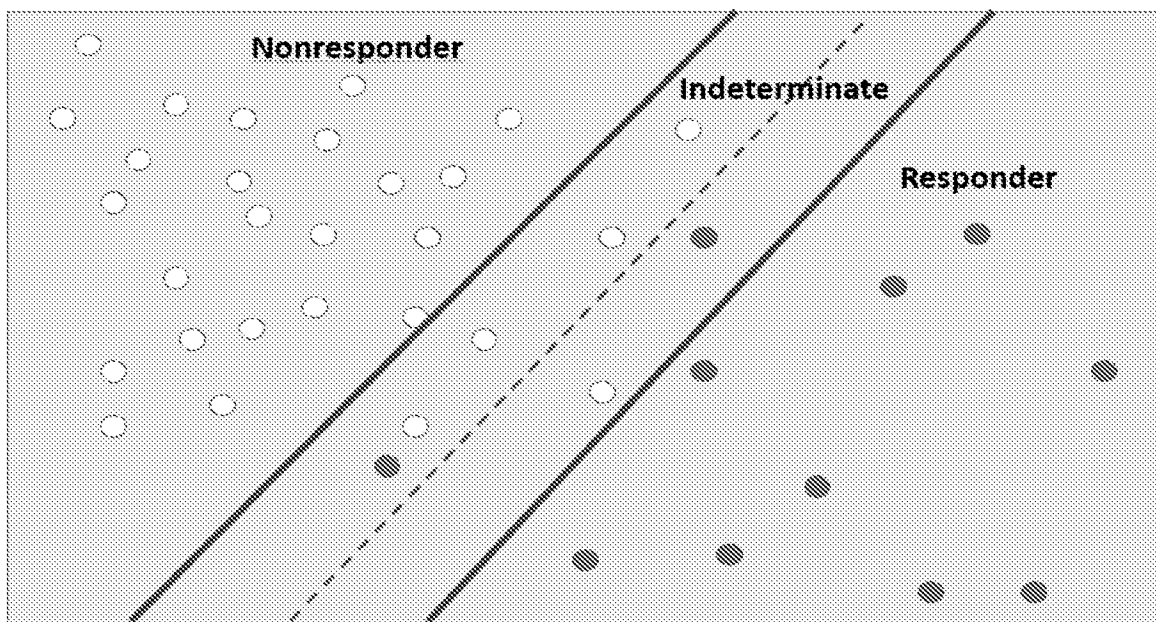


FIG. 1



54-gene GeX + MuB (55 Dimensional) Feature Space

FIG. 2

54-Gene + MuB Model Predictions

RESPONDERS AND NON-RESPONDERS

Percentage of Samples, n = 87

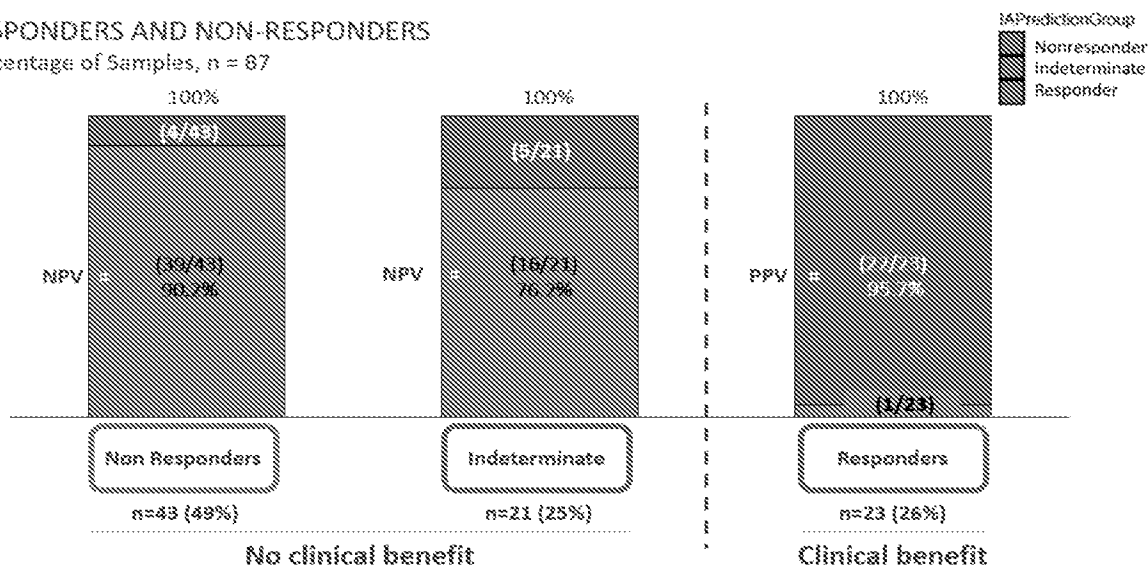
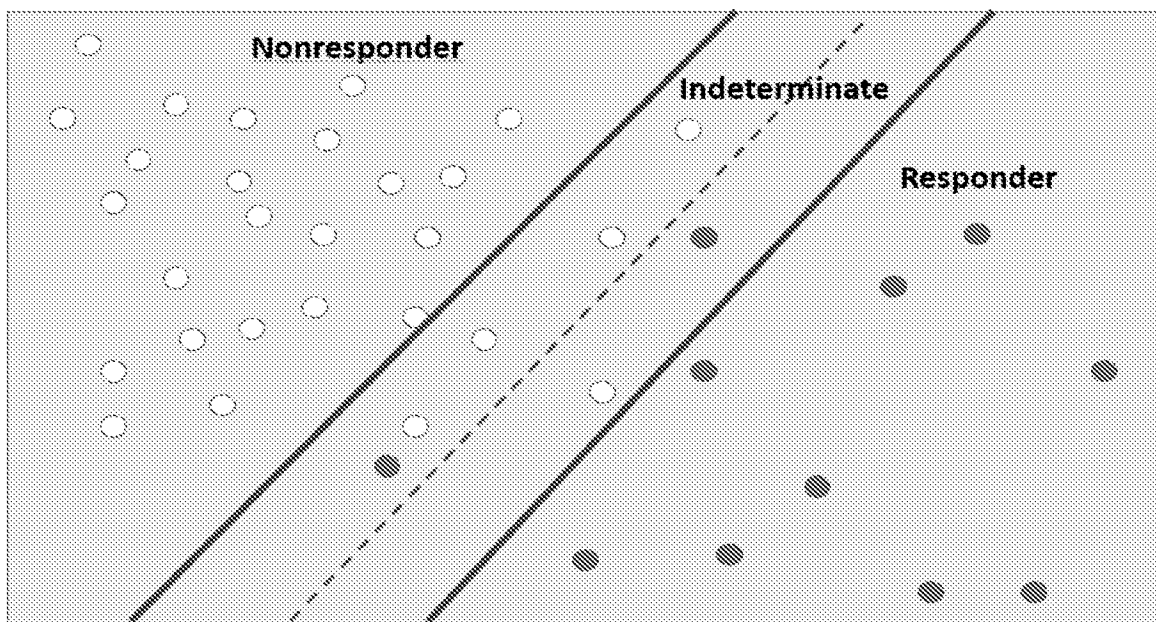


FIG. 1



54-gene GeX + MuB (55 Dimensional) Feature Space

FIG. 2

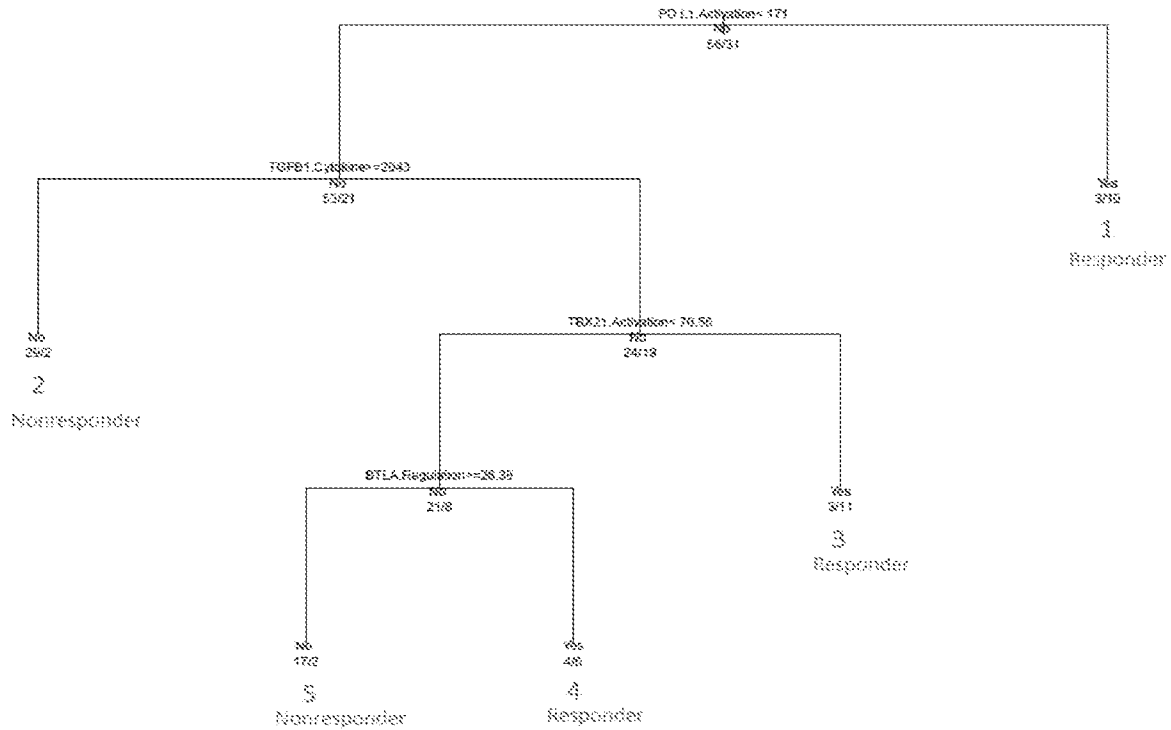


FIG. 3

4 Gene Model Predictions

RESPONDERS AND NON-RESPONDERS
Percentage of Samples, n = 87

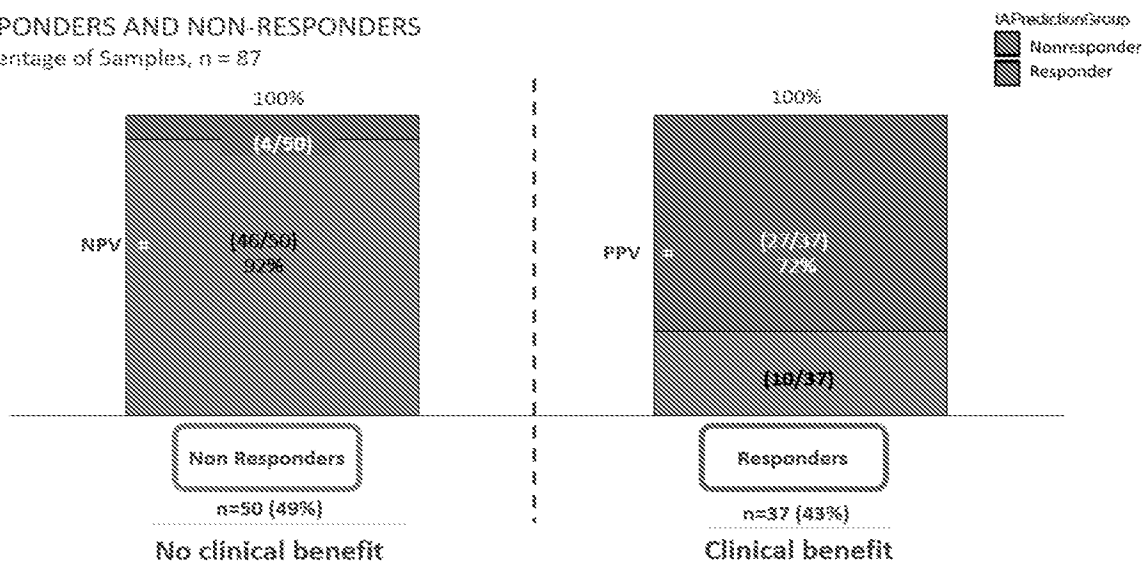


FIG. 4

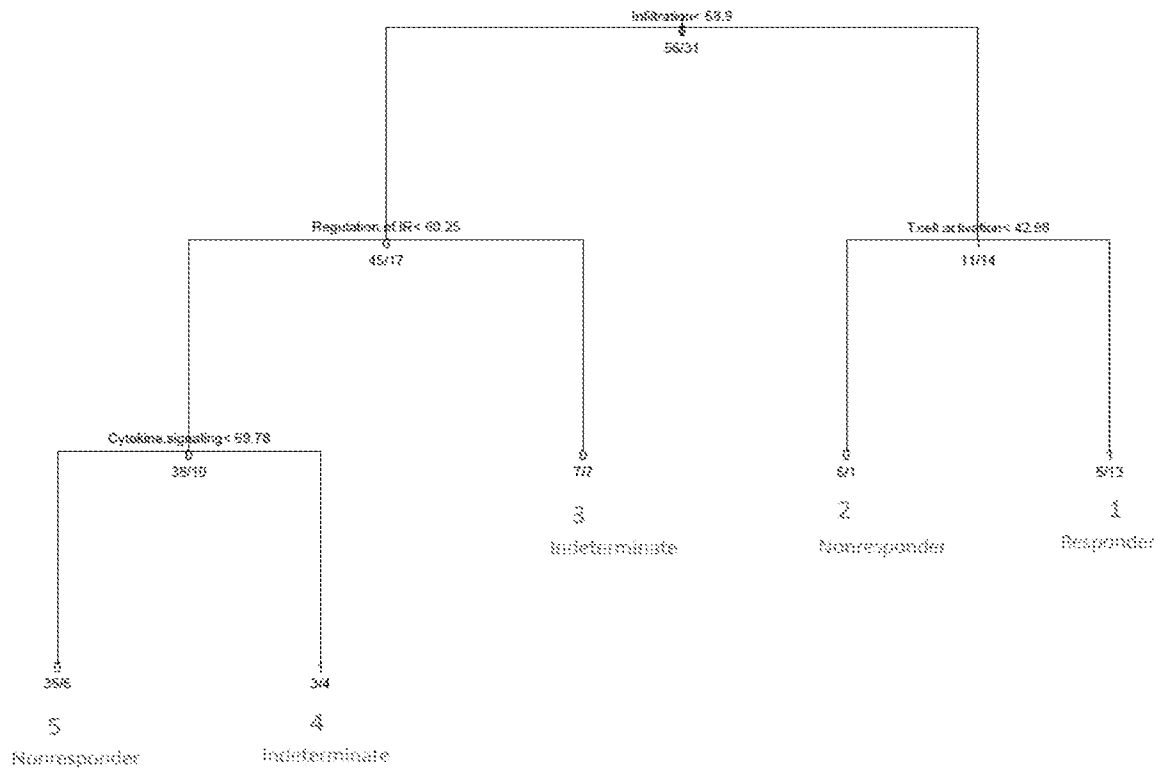


FIG. 5

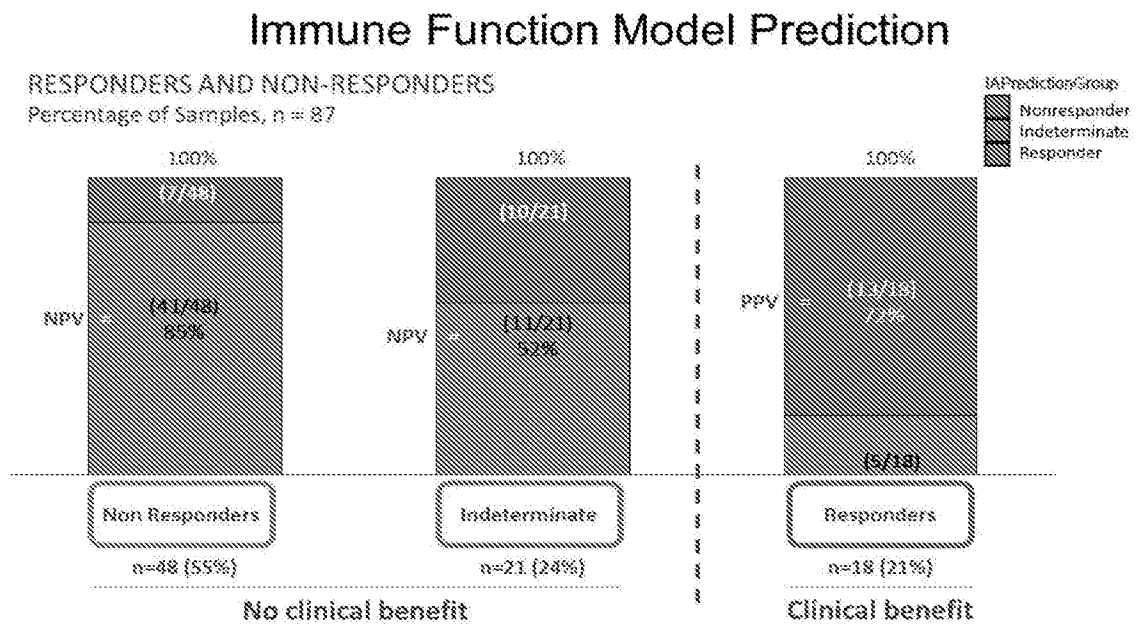


FIG. 6

Bayesian Model Averaging for Final Prediction

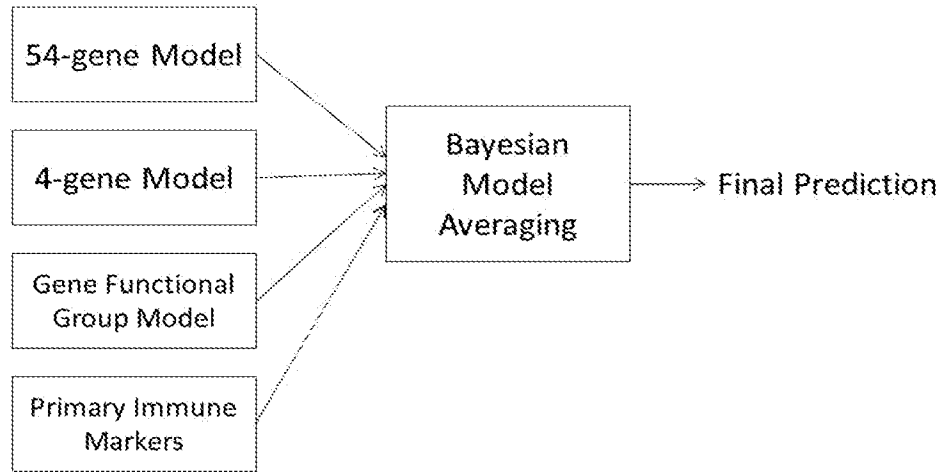


FIG. 7

Bayesian Model Averaging Predictions

RESPONDERS AND NON-RESPONDERS
Percentage of Samples, n = 87

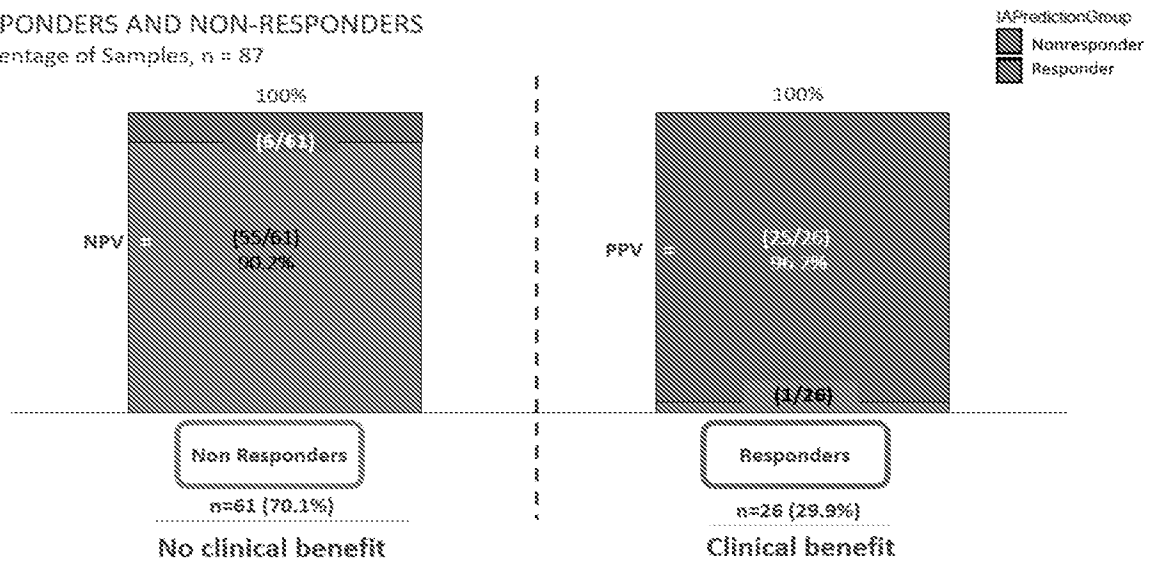


FIG. 8

Model	54-gene	4-gene	Immune Function	BMA Final Prediction
Predicted Responder	23 (26.4%)	37 (42.5%)	18 (20.7%)	26 (29.3%)
PPV (Response Rate in Predicted Responders)	22/23 (95.7%)	27/37 (73.0%)	13/18 (72.2%)	25/26 (96.2%)
Predicted Indeterminate	21 (24.1%)	-	21 (24.1%)	-
Response Rate in Predicted Indeterminate	5/21 (23.8%)	-	11/21 (52.4%)	-
Predicted Nonresponder	43 (49.4%)	50 (57.5%)	48 (55.2%)	61 (70.1%)
NPV (Nonresponse Rate in Predicted Nonresponders)	39/43 (90.7%)	46/50 (92.0%)	41/48 (85.4%)	55/61 (90.2%)

FIG. 9

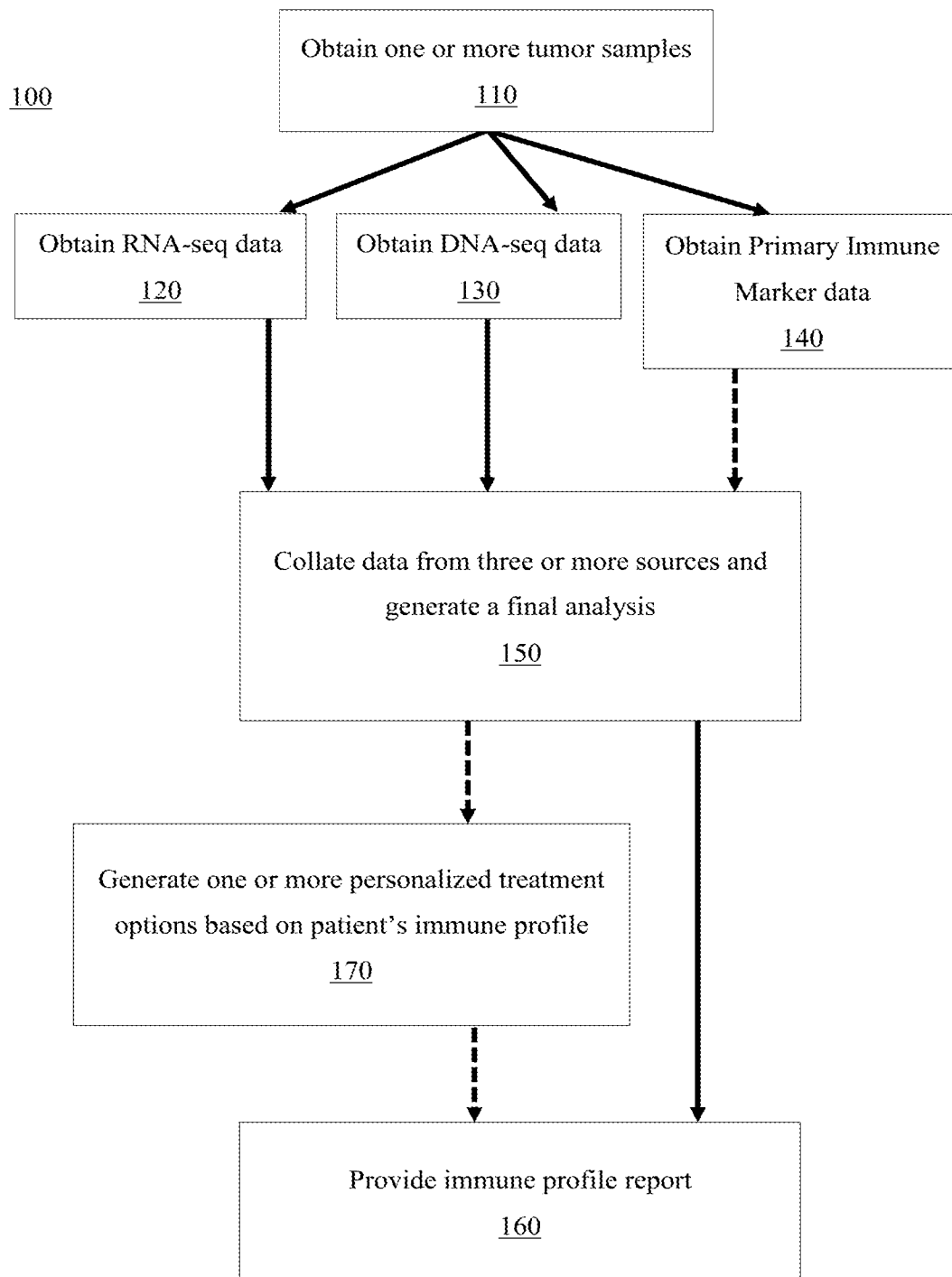


FIG. 10

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2017/055545

A. CLASSIFICATION OF SUBJECT MATTER INV. C12Q1/68 G01N33/574 ADD. A61P35/00		
According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) C12Q G01N C07K A61K		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) EP0-Internal, BIOSIS		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	ROSENBERG JONATHAN E ET AL: "Atezolizumab in patients with locally advanced and metastatic urothelial carcinoma who have progressed following treatment with platinum-based chemotherapy: a single-arm, multicentre, phase 2 trial.", LANCET (LONDON, ENGLAND) 07 MAY 2016, vol. 387, no. 10031, 7 May 2016 (2016-05-07), pages 1909-1920, XP029530539, ISSN: 1474-547X the whole document <div style="text-align: center; margin-top: 10px;"> ----- -/-- </div>	1-10,23
<div style="display: flex; justify-content: space-between;"> ☒ Further documents are listed in the continuation of Box C. ☒ See patent family annex. </div>		
* Special categories of cited documents : <div style="display: flex;"> <div style="flex: 1;"> <p>"A" document defining the general state of the art which is not considered to be of particular relevance</p> <p>"E" earlier application or patent but published on or after the international filing date</p> <p>"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)</p> <p>"O" document referring to an oral disclosure, use, exhibition or other means</p> <p>"P" document published prior to the international filing date but later than the priority date claimed</p> </div> <div style="flex: 1;"> <p>"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention</p> <p>"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone</p> <p>"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art</p> <p>"&" document member of the same patent family</p> </div> </div>		
Date of the actual completion of the international search <div style="text-align: center; font-size: 1.2em;">8 December 2017</div>		Date of mailing of the international search report <div style="text-align: center; font-size: 1.2em;">12/02/2018</div>
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016		Authorized officer <div style="text-align: center; font-size: 1.2em;">Bumb, Peter</div>

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2017/055545

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-10, 23

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2017/055545

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	<p>Rosenberg et al: "Supplement to: Rosenberg JE, Hoffman-Censits J, Powles T, et al. Atezolizumabin patients with locally advanced and metastatic urothelial carcinoma who have progressed following treatment with platinum-based chemotherapy: a single-arm, multicentre, phase 2 trial.", THE LANCET</p> <p>4 March 2016 (2016-03-04), XP002776516, Retrieved from the Internet: URL: https://ars.els-cdn.com/content/image/1-s2.0-S0140673616005614-mmcl.pdf [retrieved on 2017-12-08] pages 2, 23 figure 5A</p>	1-10,23
X	<p>-----</p> <p>TRAN ERIC ET AL: "Immunogenicity of somatic mutations in human gastrointestinal cancers.", SCIENCE (NEW YORK, N.Y.) 11 DEC 2015, vol. 350, no. 6266, 11 December 2015 (2015-12-11), pages 1387-1390, XP002776517, ISSN: 1095-9203 page 1387, left-hand column page 1388, left-hand column figure 2F page 1389, right-hand column</p>	1,2,4-10
X	<p>-----</p> <p>VAN ALLEN ELIEZER M ET AL: "Genomic correlates of response to CTLA-4 blockade in metastatic melanoma.", SCIENCE (NEW YORK, N.Y.) 09 OCT 2015, vol. 350, no. 6257, 9 October 2015 (2015-10-09), pages 207-211, XP002776518, ISSN: 1095-9203 figure 1A page 209, right-hand column table 1 abstract page 210, left-hand column</p>	1,2,4-9
Y	<p>-----</p> <p>WO 2016/100975 A1 (MASSACHUSETTS INST OF TECHNOLOGY [US]; GEN HOSPITAL CORP [US]) 23 June 2016 (2016-06-23) paragraph [0451]; claims 1-3,25-26 claims 1-3,25-26</p> <p>-----</p> <p style="text-align: center;">-/--</p>	1-10,23

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2017/055545

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	N. A. RIZVI ET AL: "Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer", SCIENCE, vol. 348, no. 6230, 3 April 2015 (2015-04-03), pages 124-128, XP055322846, ISSN: 0036-8075, DOI: 10.1126/science.aaa1348 the whole document	1-10,23
Y	----- Rizvi et al: "Supplementary Materials for Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer", Science 3 April 2015 (2015-04-03), XP002776519, Retrieved from the Internet: URL:http://science.sciencemag.org/content/sci/suppl/2015/03/11/science.aaa1348.DC1/Rizvi-SM.pdf [retrieved on 2017-12-08] the whole document	1-10,23
Y	----- WO 2016/094377 A1 (MERCK SHARP & DOHME [US]; AYERS MARK D [US]; LOBODA ANDREY [US]; LUNCE) 16 June 2016 (2016-06-16) table 1 page 56, lines 23-26	1-10,23
A	----- SHIEN KAZUHIKO ET AL: "Predictive biomarkers of response to PD-1/PD-L1 immune checkpoint inhibitors in non-small cell lung cancer.", LUNG CANCER (AMSTERDAM, NETHERLANDS) SEP 2016, vol. 99, September 2016 (2016-09), pages 79-87, XP002776520, ISSN: 1872-8332 the whole document -----	1-10,23

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2017/055545

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2016100975 A1	23-06-2016	EP 3234193 A1 WO 2016100975 A1	25-10-2017 23-06-2016

WO 2016094377 A1	16-06-2016	AU 2015360736 A1 CA 2968406 A1 CN 107109700 A EP 3230498 A1 KR 20170086661 A WO 2016094377 A1	01-06-2017 16-06-2016 29-08-2017 18-10-2017 26-07-2017 16-06-2016

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-10, 23

A method of generating an immune score, the method comprising (1) assessing tumour infiltrating lymphocytes in a sample, (2) assessing T-cell receptor signaling in the sample and (3) determining mutational burden in the sample, and optionally either (A) classifying a sample as a responder to therapy or (B) as a non-responder, i.e. wherein the special technical feature is TIL+TCRS+MuB.

2. claims: 11-16

A method of analyzing a tumour, comprising determining expression level of at least four genes (1), (2), (3), (4) and either (A) classifying a sample as responder to a therapy if (A1) expression of the first gene exceeds a threshold, or (B) classifying a sample as a non-responder to a therapy if (B1) expression of the first gene falls below a threshold and (B2) expression of the second gene exceeds a threshold, or (C) classifying a sample as a responder to a therapy if (C1) expression of the first gene falls below a threshold and (C2) expression of the second gene falls below a threshold and (C3) expression of the third gene exceeds a threshold, or (D) classifying a sample as a non-responder to a therapy if (D1) expression of the first gene falls below a threshold and (D2) expression of the second gene falls below a threshold and (D3) expression of the third gene exceeds a threshold and (D4) expression of the fourth gene exceeds a threshold, or (E) classifying a sample as a responder to a therapy if (E1) expression of the first gene falls below a threshold and (E2) expression of the second gene falls below a threshold and (E3) expression of the third gene exceeds a threshold and (E4) expression of the fourth gene exceeds a threshold, i.e. wherein the special technical features are four (unspecified) genes and a particular decision tree.

3. claims: 17-22

A method of analyzing a tumor, comprising (1) assessing a plurality of genes related to immune cell infiltration in a sample, (2) assessing a plurality of genes related to T-cell activation in the sample, (3) assessing a plurality of genes related to cytokine signaling in the sample and (4) assessing a plurality of genes related to immune response regulation in the sample, then either

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

(A) classifying the sample as a responder if (A1) immune cell infiltration exceeds a threshold and (A2) the T-cell activation value exceeds a threshold, or
(B) classifying the sample as a non-responder if (A1) immune cell infiltration exceeds a threshold and (A2) the T-cell activation value falls below a threshold,
or (B) classifying the sample as an intermediate responder if (B1) immune cell infiltration falls below a threshold and (B4) immune response regulation exceeds a threshold,
or (C) classifying the sample as intermediate responder if (C1) immune cell infiltration falls below a threshold, (C4) immune response regulation falls below a threshold and (C3) cytokine signaling response exceeds a threshold,
or (D) classifying the sample as a non-responder if (D1) immune cell infiltration falls below a threshold, (D4) immune response regulation falls below a threshold and (D3) cytokine signaling falls below a threshold,
i.e. wherein the special technical feature is
TIL+TCRS+cytokine+immune response regulation.

Abstract

Systems and methods for determining the likely responsiveness of a human cancer subject to a checkpoint blockade immunotherapy regimen are provided. Sequencing reads are obtained from samples from the subject representative of the cancer. A human leukocyte antigen type and a plurality of clones is determined from the sequencing reads. For each clone, an initial frequency X in the one or more samples is determined and a corresponding clone fitness score of the clone is computed, thereby computing clone fitness scores. Each such fitness score is computed by identifying neoantigens in the respective clone, computing a recognition potential for each neoantigen, and determining the corresponding clone fitness score of the respective clone as an aggregate of these recognition potentials. A total fitness, quantifying the likely responsiveness of the subject to the regimen, is computed by summing the clone fitness scores across the plurality of clones.



(51) International Patent Classification:

A61K 39/00 (2006.01) *A61K 35/15* (2015.01)
G01N 33/574 (2006.01) *A61K 35/17* (2015.01)
C12Q 1/68 (2018.01) *G01N 33/68* (2006.01)
G06F 19/22 (2011.01) *A61K 39/395* (2006.01)
A61P 35/00 (2006.01)

(21) International Application Number:

PCT/US2018/014282

(22) International Filing Date:

18 January 2018 (18.01.2018)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/447,852	18 January 2017 (18.01.2017)	US
62/448,291	19 January 2017 (19.01.2017)	US
62/448,247	19 January 2017 (19.01.2017)	US
62/554,232	05 September 2017 (05.09.2017)	US
62/582,851	07 November 2017 (07.11.2017)	US
62/618,540	17 January 2018 (17.01.2018)	US

(71) Applicants: **ICHAN SCHOOL OF MEDICINE AT MOUNT SINAI** [US/US]; One Gustave L. Levy Place, Box 1675, New York, NY 10029 (US). **THE SIMONS CENTER FOR SYSTEMS BIOLOGY AT THE INSTITUTE FOR ADVANCED STUDY** [US/US]; Einstein Drive, Princeton, NJ 08540 (US). **MEMORIAL SLOAN KETTERING CANCER CENTER** [US/AE]; 1275 York Avenue, New York, NY 10065 (US). **UNIVERSITÄT ZU KÖLN | UNIVERSITY OF COLOGNE** [DE/DE]; D7 Forschungsmanagement, Research Management Transfer Office, Besucheradresse Berrenrather Straße 154-156, EG, 0.25, 50923 Köln (DE).

(72) Inventors; and

(71) Applicants: **LUKSZA, Marta** [PL/PL]; 150 East 42nd Street, Second Floor, New York, NY 10017 (US). **BALACHANDRAN, Vinod, P.** [US/US]; 200 East 94th Street Apt. 3012, New York, NY 10128 (US). **LEVINE, Arnold, J.** [US/US]; 150 East 42nd Street, Second Floor, New York, NY 10017 (US). **WOLCHOK, Jedd, D.** [US/US]; 333 East 30th Street Apt. 17C, New York, NY 10016 (US). **MERGHOU, Taha** [US/US]; 10 Huron Avenue Apt 14K, Jersey City, NJ 07306 (US). **LEACH, Steven, D.** [US/US];

515 E 72nd Street, New York, NY 10021 (US). **CHAN, Timothy, A.** [US/US]; 1275 York Avenue, New York, NY 10065 (US). **GREENBAUM, Benjamin, D.** [US/US]; 150 East 42nd Street, Second Floor, New York, NY 10017 (US). **LAESSIG, Michael** [DE/DE]; 150 East 42nd Street, Second Floor, New York, NY 10017 (US).

(74) Agent: **LOVEJOY, Brett, A.** et al.; Morgan Lewis & Bockius LLP, One Market, Spear Street Tower, San Francisco, CA 94105 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

- with international search report (Art. 21(3))
- with sequence listing part of description (Rule 5.2(a))

(54) Title: NEOANTIGENS AND USES THEREOF FOR TREATING CANCER

(57) Abstract: Systems and methods for determining the likely responsiveness of a human cancer subject to a checkpoint blockade immunotherapy regimen are provided. Sequencing reads are obtained from samples from the subject representative of the cancer. A human leukocyte antigen type and a plurality of clones is determined from the sequencing reads. For each clone, an initial frequency X_0 in the one or more samples is determined and a corresponding clone fitness score of the clone is computed, thereby computing clone fitness scores. Each such fitness score is computed by identifying neoantigens in the respective clone, computing a recognition potential for each neoantigen, and determining the corresponding clone fitness score of the respective clone as an aggregate of these recognition potentials. A total fitness, quantifying the likely responsiveness of the subject to the regimen, is computed by summing the clone fitness scores across the plurality of clones.



NEOANTIGENS AND USES THEREOF FOR TREATING CANCER**CROSS REFERENCE TO RELATED APPLICATIONS**

[0001] This application claims priority to United States Provisional Patent Application No. 62/582,851 entitled “Systems and Methods for Predicting Tumor Response to Checkpoint Blockade Immunotherapy,” filed November 7, 2017, which is hereby incorporated by reference.

[0002] This application also claims priority to United States Provisional Patent Application No. 62/554,232 entitled “Neoantigens and Uses Thereof for Treating Cancer,” filed September 5, 2017, which is hereby incorporated by reference.

[0003] This application also claims priority to United States Provisional Patent Application No. 62/448,291 entitled “Neoantigens and Uses Thereof for Treating Cancer,” filed January 19, 2017, which is hereby incorporated by reference.

[0004] This application also claims priority to United States Provisional Patent Application No. 62/448,247 entitled “Neoantigens and Uses Thereof for Treating Cancer,” filed January 19, 2017, which is hereby incorporated by reference.

[0005] This application also claims priority to United States Provisional Patent Application No. 62/447,852 entitled “Neoantigen Fitness Model Predicts Tumor Response to CheckPoint Blockade Immunotherapy,” filed January 18, 2017, which is hereby incorporated by reference.

[0006] This application also claims priority to United States Provisional Patent Application No. 62/618,540 entitled “Neoantigens and Uses Thereof for Treating Cancer,” filed January 17, 2018, which is hereby incorporated by reference.

STATEMENT REGARDING FEDERALLY FUNDED RESEARCH

[0007] This invention was made with government support under R01DK097087-01 awarded by National Institute of Health, K12CA184746-01A1 awarded by National Cancer Institute, 1545935 awarded by National Science Foundation, and P30 CA08748 awarded by National Cancer Institute Cancer Center. The government has certain rights in this invention reference.

SEQUENCE LISTING

[0008] The instant application contains a Sequence Listing that has been submitted electronically in ASCII format and is hereby incorporated by reference in its entirety. The ASCII copy, created on January 18, 2018, is named 104593-5011-WO_ST25.txt and is 30 kilobytes in size.

TECHNICAL FIELD

[0009] The present disclosure relates generally to systems and methods for determining a likelihood that a human subject afflicted with a cancer will be responsive to a treatment regimen, where the treatment regimen that comprises administering a checkpoint blockade immunotherapy directed to the cancer to the subject.

BACKGROUND

[0010] Pancreatic ductal adenocarcinoma (PDAC) will be diagnosed in approximately 53,000 patients in the United States in 2016, and an estimated 41,000 will die from its effects¹. PDAC is one of the most lethal forms of cancer, less than 7% of PDAC patients survive 5 years after diagnosis². A large proportion of patients present with advanced and metastatic disease at initial diagnosis and offered treatment options restricted to cytotoxic chemotherapies which extend life by no more than a few months³. Research aims to identify new therapeutic agents and effective combinations of existing therapies for PDAC patients have not yet significantly improved patient survival⁴.

[0011] The immune system plays an important role in controlling and eradicating cancer. Nevertheless, in the setting of malignancy, multiple mechanisms of immune suppression can exist that prevent effective antitumor immunity. Antibody therapy directed against several negative immunologic regulators (checkpoints) is demonstrating significant success and is likely to be a major component of treatment for patients with a variety of malignancies. Immunologic checkpoint blockade with antibodies that target cytotoxic T lymphocyte-associated antigen 4 (CTLA-4) and the programmed cell death protein 1 pathway (PD-1/PD-L1) have demonstrated promise in a variety of malignancies. However, these immune checkpoint inhibitors have demonstrated limited impact on overall patient survival in PDAC⁵. A possible reason for this can be the relatively low mutational load observed in PDAC⁶.

[0012] Checkpoint blockade immunotherapies enable the host immune system to recognize and destroy tumor cells. Recent clinical trials using immune checkpoint blocking antibodies, such as anti-cytotoxic T-lymphocyte-associated protein 4 (anti-CTLA4), or anti-programmed cell death protein-1 (anti-PD-1), have improved overall survival in many malignancies by disinhibiting the immune system. *See Topalian et al.*, 2015, “Immune checkpoint blockade: a common denominator approach to cancer therapy,” 2015, *Cancer Cell* 27, pp. 450-61. Their clinical activity depends on activated T-cell recognition of neoantigens, which are tumor-specific, mutated peptides presented on the surface of cancer cells. *See Schumacher and Schreiber*, 2015, “Neoantigens in cancer immunotherapy,” *Science* 348, pp. 69-74; and *Gubin et al.*, 2015, “Tumor neoantigens: building a framework for personalized cancer immunotherapy,” *J. Clin. Invest.* pp. 3413-3421. How these underlying processes determine the success of immunotherapies has remained unclear. Furthermore, only a minority of patients achieves a durable clinical benefit, suggesting there may be genetic determinants of response.

[0013] De novo somatic mutations within coding regions can create *neoantigens* – novel protein epitopes specific to tumors, which MHC molecules present to the immune system and which may be recognized by T-cells as non-self. An elevated number of mutations or neoantigens has been linked to improved response to checkpoint blockade therapy in multiple malignancies. (*See Snyder et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” *N. Engl. J. Med.* 371, pp. 2189-2199; *Van Allen et al.*, 2015, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma,” *Science* 350, pp. 207-211, and *Rizvi et al.*, 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer,” *Science* 348, pp. 124-128). Hence, inferred neoantigen burden is a coarse-grained proxy for whether a tumor is likely to respond to therapy.

[0014] Other implicated biomarkers of response include T-cell receptor (TCR) repertoire profiles (leading to hypothesized roles for the microbiome (*Vétizou et al.*, 2015, “Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota,” *Science* 350, pp. 1079-1084; and *Zitvogel et al.*, 2016, “Microbiome and Anticancer Immunosurveillance,” *Cell* 165, pp. 276-287), immune based microenvironment signatures (*Snyder et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” *N. Engl. J. Med.* 371, pp. 2189-2199; *de Henau et al.*, 2016 “Overcoming resistance to checkpoint blockade therapy by targeting PI3K γ in myeloid cells,” *Nature* 539, pp. 443-447) and tumor

heterogeneity (McGranahan *et al.*, 2016, “Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade,” *Science* 351, pp. 1463-1469.

[0015] Despite high overall mutational load, a heterogeneous tumor may have immunogenic neoantigens present only in certain subclones. As a result, therapies targeting only a fraction of the tumor could disrupt clonal competitive balance and inadvertently stimulate growth of untargeted clones (Fisher *et al.*, 2015, “The value of monitoring to control evolving populations,” *Proc. Natl. Acad. Sci.* 112(4), pp. 1007-1012; and Anagnostu *et al.*, 2016, “Evolution of neoantigen landscape during immune checkpoint blockade in non-small cell lung cancer,” *Cancer Discov.*, 7(3), pp. 264-276). Moreover, mass spectrometry based validation of neoantigens, already limited by sensitivity, does not sample all of the many relevant clones in heterogeneous tumors nor account for clonal variations across metastases (Purcell *et al.*, “More than one reason to rethink the use of peptides in vaccine design,” *Nature Rev. Drug Discov.* 6, pp. 404-414). Worldwide efforts are being undertaken to model neoantigens and quantify neoantigen features from genomic data, and a predictive neoantigen-based model for immunotherapy response is a highly sought-after goal.

[0016] Given the above background, what is needed in the art are systems and methods for determining the likely responsiveness of a human cancer subject to a checkpoint blockade immunotherapy regimen. Further, given the above background, there remains a need for developing novel therapies that improve the survival of PDAC patients.

SUMMARY

The present disclosure addresses the need in the art for systems and methods for determining the likely responsiveness of a human cancer subject to a checkpoint blockade immunotherapy regimen. Such a mathematical model using genomic data has the advantage of broad consideration of neoantigen space. The disclosed recognition potential fitness model of immune interactions is used to describe the evolutionary dynamics of cancer cell populations under checkpoint-blockade immunotherapy.

To calculate the recognition potential fitness model, sequencing reads (*e.g.*, whole genome sequencing reads, exome sequencing reads, targeted sequencing reads, *etc.*) are obtained from samples from the subject representative of the cancer. A human leukocyte antigen (HLA) type and a plurality of clones is determined (*e.g.*, from the sequencing reads). For each clone, an initial frequency X_{α} in the one or more samples is determined and a corresponding clone fitness score of the clone is computed, thereby computing clone fitness scores. Each such fitness score is computed by identifying neoantigens in the respective

clone, computing a recognition potential for each neoantigen, and determining the corresponding clone fitness score of the respective clone as an aggregate of these recognition potentials. A total fitness, quantifying the likely responsiveness of the subject to the regimen, is then computed by summing the clone fitness scores across the plurality of clones.

[0017] As such, one aspect of the present disclosure provides a method for determining a likelihood that a human subject afflicted with a cancer will be responsive to a treatment regimen, where the treatment regimen comprises administering a checkpoint blockade immunotherapy directed to the cancer to the subject. In some embodiments, the checkpoint blockade immunotherapy comprises administering an anti-CTLA-4, anti-PD1, anti-PD-L1, anti-LAG3, anti-TIM-3, anti-GITR, anti-OX40, anti-CD40, anti-TIGIT, anti-4-1BB, anti-B7-H3, anti-B7-H4, or anti-BTLA compound to the cancer subject. In some embodiments, the cancer is a carcinoma, a melanoma, a lymphoma/leukemia, a sarcoma, or a neuro-glial tumor. In some embodiments, the cancer is lung cancer, pancreatic cancer, colon cancer, stomach or esophagus cancer, breast cancer, ovary cancer, prostate cancer, or liver cancer.

[0018] In the present disclosure, a plurality of sequencing reads (*e.g.*, whole genome sequencing reads, exome sequencing reads, targeted sequencing reads, *etc.*) is obtained from one or more samples from the human cancer subject that is representative of the cancer. In some embodiments, the plurality of sequencing reads exhibits an average read depth of less than 40. In some embodiments, the plurality of sequencing reads exhibits an average read depth of between 25 and 60.

[0019] In the method in accordance with the present disclosure, a human leukocyte antigen (HLA) type of the human cancer subject is determined. In some embodiments, the HLA type of the human cancer subject is determined from the plurality of sequencing reads. In some embodiments, the determining the HLA type of the human cancer subject is determined using a polymerase chain reaction using a biological sample from the cancer subject.

[0020] In the method in accordance with the present disclosure, a plurality of clones is determined from the plurality of sequencing reads. For each respective clone α in the plurality of clones, an initial frequency X_{α} of the respective clone α in the one or more samples is determined.

[0021] In some embodiments, each clone α in the plurality of clones is uniquely defined by a unique set of somatic mutations (*e.g.*, single nucleotide variant or an indel). In some embodiments, the plurality of clones is determined by a variant allele frequency of each respective somatic mutation in a plurality of somatic mutations determined from the whole-genome sequencing data.

[0022] In some embodiments, the plurality of clones is determined by identifying a plurality of inferred copy number variations using the whole-genome sequencing data.

[0023] In some embodiments, each clone α in the plurality of clones is uniquely defined by a unique set of somatic mutations. In such embodiments, the plurality of clones is determined by a combination of (i) a variant allele frequency of each respective somatic mutation in the plurality of somatic mutations determined from the whole-genome sequencing data and (ii) an identification of a plurality of inferred copy number variations using the whole-genome sequencing data (330).

[0024] In some embodiments, the plurality of clones consists of two clones (332). In some embodiments, the plurality of clones consists of between two clones and ten clones (334). In some embodiments, the initial frequency X_α of the respective clone α in the one or more samples is determined using the plurality of sequencing reads from the one or more samples from the human cancer subject (336).

[0025] In the method in accordance with the present disclosure, for each respective clone α in the plurality of clones, a corresponding clone fitness score of the respective clone is computed, thereby computing a plurality of clone fitness scores, each corresponding clone fitness score computed for a respective clone α by a first procedure.

[0026] In the first procedure, a plurality of neoantigens in the respective clone α is identified. In some embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a nonamer peptide. In some embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide that is eight, nine, ten, or eleven residues in length. In certain embodiments each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide that is 3-30 amino acids, *e.g.*, about 3-5, about 5-15 (*e.g.*, about 8-11, about 5-10, or about 10-15), about 15-20, about 20-25, or about 20-30 amino acids, in length. In certain embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide that is about 8-11 amino acids in length. In certain embodiments, each neoantigen in the plurality of

neoantigens of a clone in the plurality of clones is a peptide that is about 3, about 4, about 5, about 6, about 7, about 8, about 9, about 10, about 11, about 12, about 13, about 14, about 15, about 16, about 17, about 18, about 19, about 20, about 21, about 22, about 23, about 24, about 25, about 26, about 27, about 28, about 29, or about 30 amino acids in length. In certain embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide that is at least about 3, at least about 5, or at least about 8 amino acids in length. In certain embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide is less than about 30, less than about 20, less than about 15, or less than about 10 amino acids in length.

[0027] In some embodiments, the method further comprises identifying a population of neoantigens present in the one or more samples by a procedure comprising: determining a plurality of somatic single nucleotide polymorphisms (SNPs) in the plurality of sequencing reads by comparison of the plurality of sequencing reads to a reference human genome, evaluating each respective somatic SNP in the plurality of SNPs as a neoantigen candidate by evaluation of a peptide encoded by a portion of one or more sequencing reads in the sequencing reads that includes the respective somatic SNP against a classifier that has been trained to predict peptide binding to class I MHC of the HLA type of the cancer subject, where a neoantigen candidate having a binding score below a threshold value is deemed to be a neoantigen in the population of neoantigens. Further, the identifying the plurality of neoantigens in the respective clone α comprises matching the SNPs in the respective clone α to respective neoantigens in the population of neoantigens. In some such embodiments, the threshold value is 500 nM.

[0028] Continuing with the method in accordance with the present disclosure, a recognition potential of each respective neoantigen in the plurality of neoantigens in the respective clone α is computed by a second procedure. In the second procedure, an amplitude A of the respective neoantigen is computed as a function of the relative major histocompatibility complex (MHC) affinity of the respective neoantigen and the wildtype counterpart of the respective neoantigen given the HLA type of the subject. In some such embodiments, the function of the relative class I MHC affinity of the respective neoantigen and the wildtype counterpart of the respective neoantigen given the HLA type of the human cancer subject is a ratio of the relative class I MHC affinity of the respective neoantigen and the wildtype counterpart of the respective neoantigen given the HLA type of the subject.

[0029] In some embodiments, the function of the relative class I MHC affinity of the respective neoantigen and the wildtype counterpart of the respective neoantigen given the HLA type of the subject is a ratio of: (1) a dissociation constant between the respective neoantigen and the class I MHC presented by the cancer subject given the HLA type of the cancer subject, and (2) a dissociation constant between the wildtype counterpart of the respective neoantigen and the class I MHC presented by the cancer subject given the HLA type of the cancer subject. In some such embodiments, the dissociation constant between the respective neoantigen and the class I MHC presented by the cancer subject is obtained as output from a first classifier upon inputting into the first classifier the amino acid sequence of the neoantigen. The dissociation constant between the wildtype counterpart of the respective neoantigen and the class I MHC presented by the cancer subject of the HLA type of the subject is obtained as output from the first classifier upon inputting into the first classifier the amino acid sequence of the respective wildtype counterpart of the neoantigen (e.g., the first classifier is specific to the HLA type of the cancer subject and has been trained with the respective class I MHC binding coefficient and sequence data of each peptide epitope in a plurality of epitopes presented by class I MHC in a training population having the HLA type of the subject).

[0030] Further, in the second procedure, a probability of T-cell receptor recognition R of the respective neoantigen is computed as a probability that the respective neoantigen binds one or more epitopes that are positively recognized by T-cells after class I MHC presentation.

[0031] In some such embodiments, the probability that the respective neoantigen binds one or more epitopes that are positively recognized by T-cells after class I MHC presentation is determined by a third procedure that comprises (a) selecting a respective epitope $\mathbf{e}$ from an epitope database IEDB, where the respective epitope $\mathbf{e}$ is positively recognized by T-cells after class I MHC presentation, (b) computing, for the respective epitope $\mathbf{e}$, the probability

$$\text{Pr}_{\text{binding}}(\mathbf{s}, \mathbf{e}) = \frac{1}{1 + e^{-k(|\mathbf{s}, \mathbf{e}| - \alpha)}}$$

where, $|\mathbf{s}, \mathbf{e}|$ is a sequence alignment score between the sequence of the respective neoantigen and the sequence of the respective epitope, and k and α are constants, (c) performing the selecting (a) and the computing (b) for each respective epitope $\mathbf{e}$ in a plurality of epitopes in the epitope database IEDB, thereby computing a plurality of probabilities $\text{Pr}_{\text{binding}}(\mathbf{s}, \mathbf{e})$; and (d) computing the probability of T-cell receptor recognition R of the respective neoantigen as:

$$R = 1 - \prod_{e \in \text{IEDB}} [1 - \text{Pr}_{\text{binding}}(\mathbf{s}, \mathbf{e})],$$

where IEDB is the plurality of epitopes. In some such embodiments, $|\mathbf{s}, \mathbf{e}|$ is computed as an alignment (*e.g.*, gapless, or an alignment that allows gaps with suitable gap introduction and extension penalties) between the sequence of the respective neoantigen and the sequence of the respective epitope using an amino-acid similarity matrix. In some such embodiments, the amino-acid similarity matrix is a BLOSUM62 matrix. In some embodiments, for each patient (*e.g.*, in a cohort), sequence alignments of IEDB sequences and the patient's neoantigen sequences is performed. In some embodiments, BLAST and a blastp program with BLOSUM62 matrix and a strong gap penalty -11 is used to prevent gapped alignments. In some embodiments, the gap extension cost is set to the default value -1. In some embodiments, a threshold on alignment E-values is not imposed and all alignments are considered. In some embodiments, alignment scores for these identified alignments are then computed with Biopython Bio.pairwise2 package.

[0032] In the second procedure, the recognition potential of the respective neoantigen is computed as a function (*e.g.*, product) of the amplitude A of the respective neoantigen and the probability of T-cell receptor recognition R of the respective neoantigen.

[0033] In accordance with the present disclosure, the method further comprises determining the corresponding clone fitness score of the respective clone α as an aggregate of the neoantigen recognition potential across the plurality of neoantigens in the respective clone α . In some such embodiments, the aggregate of the neoantigen recognition potentials across the plurality of neoantigens in the respective clone α is computed as:

$$F_{\alpha} = - \max_{i \in \text{Clone } \alpha} (A_i \times R_i)$$

where i is an index iterating over each neoantigen in the plurality of neoantigens in the respective clone α .

[0034] In some such embodiments, the aggregate of the neoantigen recognition potentials across the plurality of neoantigens in the respective clone α is computed as a summation of the recognition potential of each respective neoantigen in the plurality of neoantigens. In some embodiments, the aggregate of the neoantigen recognition potentials across the plurality of neoantigens in the respective clone α is computed as a summation of the recognition potentials of a subset of the neoantigens in the plurality of neoantigens. In some embodiments, the subset of the neoantigens in the plurality of neoantigens constitutes a predetermined number of neoantigens in the plurality of neoantigens that have the top recognition potential for the respective clone α .

[0035] In some embodiments, the aggregate of the neoantigen recognition potentials across the plurality of neoantigens in the respective clone α is computed as a nonlinear combination of the recognition potential of all or a subset of the neoantigens in the plurality of neoantigens. In some such embodiments, the computing a total fitness for the one or more samples as a sum of the clone fitness scores across the plurality of clones is computed as:

$$n(\tau) = \sum_{\alpha} X_{\alpha} \exp(F_{\alpha} \tau),$$

where τ is a characteristic evolutionary time scale.

[0036] In accordance with the present disclosure, the method further comprises computing a total fitness for the one or more samples as a sum of the clone fitness scores across the plurality of clones, where each clone fitness score is weighted by the initial frequency X_{α} of the corresponding clone α , and the total fitness quantifies the likelihood that the human subject afflicted with the cancer will be responsive to the treatment regimen. In some such embodiments, the computing a total fitness for the one or more samples as a sum of the clone fitness scores across the plurality of clones is computed as:

$$n(\tau) = \sum_{\alpha} X_{\alpha} \exp(F_{\alpha} \tau),$$

where τ is a characteristic evolutionary time scale (374). In some such embodiments, τ is between 0.0 and 0.5. In some such embodiments, τ is 0.06. In some such embodiments, τ is 0.09. In some such embodiments, τ is between 0.01 and 1.0. In some such embodiments, a lower total fitness score is associated with (a) a higher likelihood that the cancer subject will be responsive to the immunotherapy and (b) a longer term survival of the cancer patient.

[0037] Another aspect of the present disclosure provides a method for identifying an immunotherapy for a cancer. In some embodiments of the method, a plurality of sequencing reads is obtained from one or more samples from a human cancer subject that is representative of the cancer. A human leukocyte antigen (HLA) type of the human cancer subject is determined from the plurality of sequencing reads. A plurality of clones is determined. For each respective clone α in the plurality of clones, an initial frequency X_{α} of the respective clone α in the one or more samples is determined from the plurality of sequencing reads. Further, for each respective clone α in the plurality of clones, a corresponding clone fitness score of the respective clone is computed, thereby computing a plurality of clone fitness scores, each corresponding clone fitness score computed for a respective clone α by a first procedure.

[0038] In the first procedure, a plurality of neoantigens in the respective clone α are identified.

[0039] A recognition potential of each respective neoantigen in the plurality of neoantigens in the respective clone α is then computed by a second procedure. In the second procedure, an amplitude A of the respective neoantigen as a function of the relative major histocompatibility complex (MHC) affinity of the respective neoantigen and the wildtype counterpart of the respective neoantigen given the HLA type of the subject is computed. Further, a probability of T-cell receptor recognition R of the respective neoantigen as a probability that the respective neoantigen binds one or more epitopes that are positively recognized by T-cells after class I MHC presentation is computed. The recognition potential of the respective neoantigen is computed as a function of (*e.g.*, the product of) the amplitude A of the respective neoantigen and the probability of T-cell receptor recognition R of the respective neoantigen.

[0040] In the first procedure, the corresponding clone fitness score of the respective clone α is determined as an aggregate of the neoantigen recognition potentials across the plurality of neoantigens in the respective clone α .

[0041] In the disclosed first procedure, a first neoantigen is selected from a plurality of neoantigens for a respective clone α in the plurality of respective clones based upon the recognition potential of the first neoantigen as the immunotherapy for the cancer.

[0042] In some embodiments, the first procedure is repeated for a plurality of human cancer subjects across a plurality of HLA types and the first neoantigen is selected on the basis of the recognition potential of the first neoantigen across the plurality of HLA types.

[0043] In some embodiments, the first procedure is repeated for a plurality of human cancer subjects and the first neoantigen is selected on the basis of the recognition potential of the first neoantigen across the plurality of human cancer subject.

[0044] In some embodiments, the cancer is a carcinoma, a melanoma, a lymphoma/leukemia, a sarcoma, or a neuro-glial tumor. In some embodiments, the cancer is lung cancer, pancreatic cancer, colon cancer, stomach or esophagus cancer, breast cancer, ovary cancer, prostate cancer, or liver cancer.

[0045] In some embodiments, each clone α in the plurality of clones is uniquely defined by a unique set of somatic mutations, and the plurality of clones is determined by a variant allele frequency of each respective somatic mutation in a plurality of somatic mutations determined from the whole-genome sequencing data. For instance, in some embodiments, the somatic mutation is a single nucleotide variant or an indel.

[0046] In some embodiments, the plurality of clones is determined by identifying a plurality of inferred copy number variations using the whole-genome sequencing data.

[0047] In some embodiments, each clone α in the plurality of clones is uniquely defined by a unique set of somatic mutations, and the plurality of clones is determined by a combination of (i) a variant allele frequency of each respective somatic mutation in the plurality of somatic mutations determined from the whole-genome sequencing data and (ii) an identification of a plurality of inferred copy number variations using the whole-genome sequencing data.

[0048] In some embodiments, the plurality of sequencing reads exhibits an average read depth of less than 40. In some embodiments, the plurality of sequencing reads exhibits an average read depth of between 25 and 60.

[0049] In some embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a nonamer peptide. In some embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide that is eight, nine, ten, or eleven residues in length. In certain embodiments each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide that is 3-30 amino acids, *e.g.*, about 3-5, about 5-15 (*e.g.*, about 8-11, about 5-10, or about 10-15), about 15-20, about 20-25, or about 20-30 amino acids, in length. In certain embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide that is about 8-11 amino acids in length. In certain embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide that is about 3, about 4, about 5, about 6, about 7, about 8, about 9, about 10, about 11, about 12, about 13, about 14, about 15, about 16, about 17, about 18, about 19, about 20, about 21, about 22, about 23, about 24, about 25, about 26, about 27, about 28, about 29, or about 30 amino acids in length. In certain embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide that is at least about 3, at least about 5, or at least about 8 amino acids in length. In certain embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide is less than about 30, less than about 20, less than about 15, or less than about 10 amino acids in length.

[0050] In some embodiments, the method further comprises identifying a population of neoantigens present in the one or more samples by a third procedure in which a plurality of somatic single nucleotide polymorphisms (SNPs) in the plurality of sequencing reads is

determined by comparison of the plurality of sequencing reads to a reference human genome. In the third procedure, each respective somatic SNP in the plurality of SNPs is evaluated as a neoantigen candidate by evaluation of a peptide encoded by a portion of one or more sequencing reads in the sequencing reads that includes the respective somatic SNP against a classifier that has been trained to predict peptide binding to class I MHC of the HLA type of the cancer subject, where a neoantigen candidate having a binding score below a threshold value (e.g., 500 nM) is deemed to be a neoantigen in the population of neoantigens. Further, when this third procedure is used, the identifying the plurality of neoantigens in the respective clone α comprises matching the SNPs in the respective clone α to respective neoantigens in the population of neoantigens.

[0051] In some embodiments, the HLA type determination of the human cancer subject is made from the plurality of sequencing reads. In some embodiments, the HLA type determination of the human cancer subject is made using a polymerase chain reaction using a biological sample from the cancer subject.

[0052] In some embodiments, the plurality of clones consists of two clones. In some embodiments, the plurality of clones consists of between two clones and ten clones. In some embodiments, the initial frequency X_α of the respective clone α in the one or more samples is determined using the plurality of sequencing reads from the one or more samples from the human cancer subject.

[0053] In some embodiments, the function of the relative class I MHC affinity of the respective neoantigen and the wildtype counterpart of the respective neoantigen given the HLA type of the subject is a ratio of the relative class I MHC affinity of the respective neoantigen and the wildtype counterpart of the respective neoantigen given the HLA type of the subject.

[0054] In some embodiments, the function of the relative class I MHC affinity of the respective neoantigen and the wildtype counterpart of the respective neoantigen given the HLA type of the human cancer subject is a ratio of: (1) a dissociation constant between the respective neoantigen and the class I MHC presented by the cancer subject given the HLA type of the cancer subject, and (2) a dissociation constant between the wildtype counterpart of the respective neoantigen and the class I MHC presented by the cancer subject given the HLA type of the cancer subject. In some such embodiments, the dissociation constant between the respective neoantigen and the class I MHC presented by the cancer subject is

obtained as output from a first classifier upon inputting into the first classifier the amino acid sequence of the neoantigen, and the dissociation constant between the wildtype counterpart of the respective neoantigen and the class I MHC presented by the cancer subject of the HLA type of the subject is obtained as output from the first classifier upon inputting into the first classifier the amino acid sequence of the respective wildtype counterpart of the neoantigen. In some such embodiments, the first classifier is specific to the HLA type of the cancer subject and has been trained with the respective class I MHC binding coefficient and sequence data of each peptide epitope in a plurality of epitopes presented by class I MHC in a training population having the HLA type of the subject.

[0055] In some embodiments, the probability that the respective neoantigen binds one or more epitopes that are positively recognized by T-cells after class I MHC presentation is determined by a third procedure that comprises (a) selecting a respective epitope $\mathbf{e}$ from an epitope database IEDB, where the respective epitope $\mathbf{e}$ is positively recognized by T-cells after class I MHC presentation, (b) computing, for the respective epitope $\mathbf{e}$, the probability

$$\text{Pr}_{\text{binding}}(\mathbf{s}, \mathbf{e}) = \frac{1}{1 + e^{-k(|\mathbf{s}, \mathbf{e}| - \alpha)}}$$

where $|\mathbf{s}, \mathbf{e}|$ is a sequence alignment score between the sequence of the respective neoantigen and the sequence of the respective epitope, and k and α are constants, and (c) performing the selecting (a) and the computing (b) for each respective epitope $\mathbf{e}$ in a plurality of epitopes in the epitope database IEDB, thereby computing a plurality of probabilities $\text{Pr}_{\text{binding}}(\mathbf{s}, \mathbf{e})$, and (d) computing the probability of T-cell receptor recognition R of the respective neoantigen as:

$$R = 1 - \prod_{\mathbf{e} \in \text{IEDB}} [1 - \text{Pr}_{\text{binding}}(\mathbf{s}, \mathbf{e})],$$

where IEDB the plurality of epitopes. In some such embodiments, α is set to 23 and k is set to 1. In some such embodiments, $|\mathbf{s}, \mathbf{e}|$ is computed as an alignment (*e.g.*, gapless, or an alignment that allows gaps with suitable gap introduction and extension penalties) between the sequence of the respective neoantigen and the sequence of the respective epitope using an amino-acid similarity matrix (*e.g.*, a BLOSUM62 matrix).

[0056] In some embodiments, the first neoantigen from a plurality of neoantigens for a respective clone α in the plurality of respective clones is selected when it has a recognition potential that is lower than the recognition potential of other neoantigens in each plurality of neoantigens for each respective clone α in the plurality of respective clones of the subject.

[0057] The presently disclosed subject matter further provides neoantigens, methods for detecting neoantigens, uses of the neoantigens for identifying cancer subjects as candidates for an immunotherapy, and uses of the neoantigens for predicting responsiveness of cancer subjects to an immunotherapy. In addition, the presently disclosed subject matter provides a population of T cells that target one or more of the presently disclosed neoantigens, and vaccines comprising one or more of the presently disclosed neoantigens.

[0058] In one aspect, the presently disclosed subject matter provides a method for identifying a subject having a cancer as a candidate for treatment with an immunotherapy. In certain non-limiting embodiments, the method comprises (a) obtaining a biological sample from the subject; (b) measuring the number of neoantigens (neoantigen number) in the biological sample; and (c) measuring the homology between each of the neoantigen and a microbial epitope (neoantigen-microbial homology); wherein a neoantigen number higher than the median neoantigen number obtained from a population of subjects having the cancer and a neoantigen-microbial homology higher than the median neoantigen-microbial homology obtained from subjects having the cancer indicate that the subject is a candidate for an immunotherapy. In certain embodiments, the neoantigen-microbial homology is measured by calculating a recognition potential score between each neoantigen and microbial epitope.

[0059] In certain embodiments, the method for identifying a subject having a cancer as a candidate for treatment with an immunotherapy comprises: (a) measuring the number of neoantigens (neoantigen number) in a biological sample of the subject; and (b) measuring the number of activated T cells (activated T cell number) in a biological sample of the subject; wherein a neoantigen number higher than the median neoantigen number obtained from a population of subjects having the cancer and an activated T cell number higher than the median activated T cell number obtained from subjects having the cancer indicate that the subject is a candidate for an immunotherapy.

[0060] The presently disclosed subject matter further provides a method of predicting the responsiveness of a subject having a cancer to an immunotherapy. In certain embodiments, the method comprises: (a) obtaining a biological sample from the subject; (b) measuring the number of neoantigens (neoantigen number) in the biological sample; and (c) measuring the homology between each of the neoantigen and a microbial epitope (neoantigen-microbial homology); wherein a neoantigen number higher than the median neoantigen number obtained from a population of subjects having the cancer and a neoantigen-microbial homology higher than the median neoantigen-microbial homology

obtained from subjects having the cancer indicate that the subject is likely to be responsive to an immunotherapy. In certain embodiments, the neoantigen-microbial homology is measured by calculating a recognition potential score between each neoantigen and microbial epitope.

[0061] In certain embodiments, the method of predicting the responsiveness of a subject having a cancer to an immunotherapy comprises: (a) measuring the number of neoantigens (neoantigen number) in a biological sample of the subject; and (b) measuring the number of activated T cells (activated T cell number) in a biological sample of the subject; wherein a neoantigen number higher than the median neoantigen number obtained from a population of subjects having the cancer and an activated T cell number higher than the median activated T cell number obtained from subjects having the cancer indicate that the subject is likely to be responsive to an immunotherapy.

[0062] In certain non-limiting embodiments, the subject's neoantigen number, the neoantigen-microbial homology, and activated T cell numbers are at least about 1%, at least about 2%, at least about 3%, at least about 4%, at least about 5%, at least about 6%, at least about 7%, at least about 8%, at least about 9%, at least about 10%, or at least about 20%, or at least about 30%, higher than the median values.

[0063] In certain embodiments, the activated T cells are T cells expressing one or more T cell activation marker. In certain embodiments, the one or more T cell activation marker is selected from the group consisting of CD3, CD8, PD-1, 4-1BB, CD69, CD107a, Granzyme B, and combinations thereof. In certain embodiments, the activated T cells are selected from the group consisting of CD3⁺CD8⁺ T cells, CD3⁺CD8⁺Granzyme-B⁺ T cells, and polyclonal activated T cells.

[0064] In certain embodiments, the cancer is a solid tumor. In certain embodiments, the cancer is pancreatic cancer. In certain embodiments, the pancreatic cancer is pancreatic ductal adenocarcinoma (PDAC). In certain embodiments, the one or more neoantigen is selected from the neoantigenic peptides listed in Table 2.

[0065] The presently disclosed subject matter also provides a method for identifying a subject having pancreatic cancer as a candidate for treatment with an immunotherapy. In certain embodiments, the method comprises: (a) obtaining a biological sample from the subject; and (b) detecting one or more neoantigen of MUC16 in the biological sample; wherein the presence of one or more neoantigen of MUC16 in the biological sample indicates that the subject is a candidate for an immunotherapy. In certain embodiments, the method

comprises detecting a plurality of neoantigens of MUC16. In certain embodiments, the one or more neoantigen is selected from the neoantigenic peptides listed in Table 1.

[0066] Additionally, the presently disclosed subject matter provides a method for predicting the responsiveness of a subject having pancreatic cancer to an immunotherapy. In certain embodiments, the method comprises: (a) obtaining a biological sample from the subject; and (b) detecting one or more neoantigen in MUC16 in the biological sample; wherein the presence of one or more neoantigen, and preferably a plurality of neoantigens, in MUC16 in the biological sample indicates that the subject is likely to be responsive to an immunotherapy. In certain embodiments, the method comprises detecting a plurality of neoantigens of MUC16. In certain embodiments, the one or more neoantigen is selected from the neoantigenic peptides listed in Table 1.

[0067] In certain embodiments, the neoantigen is a peptide, for example a peptide incorporated into a larger protein. In certain embodiments, the neoantigen is about 8 to 11 amino acids in length. In certain embodiments each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide that is 3-30 amino acids, *e.g.*, about 3-5, about 5-15 (*e.g.*, about 8-11, about 5-10, or about 10-15), about 15-20, about 20-25, or about 20-30 amino acids, in length. In certain embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide that is about 8-11 amino acids in length. In certain embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide that is about 3, about 4, about 5, about 6, about 7, about 8, about 9, about 10, about 11, about 12, about 13, about 14, about 15, about 16, about 17, about 18, about 19, about 20, about 21, about 22, about 23, about 24, about 25, about 26, about 27, about 28, about 29, or about 30 amino acids in length. In certain embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide that is at least about 3, at least about 5, or at least about 8 amino acids in length. In certain embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide is less than about 30, less than about 20, less than about 15, or less than about 10 amino acids in length.

[0068] In certain embodiments, the neoantigen is identified by exome sequencing. In certain embodiments, the immunotherapy is selected from the group consisting of therapies comprising one or more immune checkpoint-blocking antibody, adoptive T cell therapies, and combinations thereof. In certain embodiments, the one or more immune checkpoint-blocking antibody is selected from the group consisting of anti-CTLA-4 antibodies, anti-PD-1

antibodies, anti-PD-L1 antibodies, anti-TIM3 antibodies, anti-LAG3 antibodies, anti-GITR antibodies, anti-OX40 antibodies, anti-CD40 antibodies, anti-TIGIT antibodies, and anti-4-1BB antibodies, anti-B7-H3 antibodies, anti-B7-H4 antibodies, anti-BTLA antibodies.

[0069] Furthermore, the presently disclosed subject matter provides a population of neoantigen-specific T cells. In certain embodiments, the T cells target one or more neoantigen in MUC16. In certain embodiments, the one or more neoantigen in MUC16 is selected from the neoantigenic peptides listed in Table 1. In certain embodiments, the T cells target one or more neoantigen associated with a cancer, said one or more neoantigen correlating with a neoantigen-microbial homology that is higher than the median neoantigen-microbial homology occurring in subjects with the cancer. In certain embodiments, the T cells target one or more neoantigen associated with a cancer, said neoantigen correlating with an activated T cell number that is higher than the median activated T cell number occurring in subjects with the cancer.

[0070] In certain embodiments, the T cells are selectively expanded to target the one or more neoantigen in MUC16. In certain embodiments, the one or more neoantigen in MUC16 is selected from the neoantigenic peptides listed in Table 1. In certain embodiments, the neoantigen to be targeted is selected based, at least in part, on predicted immunogenicity. In certain embodiments, a neoantigen may be selectively targeted based on one or more of the following: (i) homology to an epitope of a known pathogen or microbe; and/or (ii) ability to activate T cells, e.g. in an *in vitro* assay. In certain embodiments, the T cells are selectively expanded to target the one or more neoantigen associated with a cancer, said one or more neoantigen correlating with a neoantigen-microbial homology that is higher than the median neoantigen-microbial homology occurring in subjects with the cancer, where said neoantigen less frequently occurs in subjects with the cancer and having neoantigen-microbial homology at or less than the median value. In certain embodiments, the T cells are selectively expanded to target a neoantigen occurring in a subject with a cancer where said subject has an activated T cell number that is higher than the median activated T cell number of subjects with the cancer, where said neoantigen less frequently occurs in subjects with the cancer and having activated T cell numbers at or less than the median value.

[0071] In certain embodiments, the T cells comprise a recombinant antigen receptor that specifically binds to one or more neoantigen of MUC16. In certain embodiments, the one or more neoantigen of MUC16 is selected from the neoantigenic peptides listed in Table 1. In certain embodiments, the T cells comprise a recombinant antigen receptor that

specifically binds to the one or more neoantigen associated with a cancer, said one or more neoantigen correlating with a neoantigen-microbial homology that is higher than the median neoantigen-microbial homology occurring in subjects with the cancer. In certain embodiments, the T cells comprise a recombinant antigen receptor that specifically binds to a neoantigen associated with a cancer, said neoantigen correlating with an activated T cell number that is higher than the median activated T cell number occurring in subjects with the cancer. In certain embodiments, the one or more neoantigen is selected from the neoantigenic peptides listed in Table 2.

[0072] In certain embodiments, the recombinant antigen receptor is a T cell receptor (TCR). In certain embodiments, the recombinant antigen receptor is a chimeric antigen receptor (CAR).

[0073] The presently disclosed subject further provides a vaccine comprising one or more neoantigen described herein or a polynucleotide encoding said neoantigen or a protein or peptide comprising said neoantigen. In certain embodiments, the vaccine comprises one or more neoantigen of MUC16, or a polynucleotide encoding said neoantigen or a protein or peptide comprising said neoantigen. In certain embodiments, the one or more neoantigen of MUC16 is selected from the neoantigenic peptides listed in Table 1. In certain embodiments, the vaccine is comprised in a vector. In certain embodiments, the vector is a viral vector. In certain embodiments, the polynucleotide is an RNA or a DNA.

[0074] The presently disclosed subject matter provides compositions comprising the T cell population described herein. The presently disclosed subject also provides compositions comprising the vaccine described herein. In certain embodiments, the composition is a pharmaceutical composition that comprises a pharmaceutically acceptable carrier.

[0075] The presently disclosed subject provides methods of treating pancreatic cancer. In certain embodiments, the method comprises administering to the subject the T cell population or the composition comprising the T cell population as described herein. In certain embodiments, the method comprises administering to the subject the vaccine or the composition comprising the vaccine as described herein.

BRIEF DESCRIPTION OF THE DRAWINGS

[0076] Figure 1 illustrates an exemplary system topology for determining a likelihood that a human subject afflicted with a cancer will be responsive to a treatment regimen that comprises administering a checkpoint blockade immunotherapy directed to the cancer to the subject, in accordance with an embodiment of the present disclosure.

[0077] Figure 2 illustrates a device for determining a likelihood that a human subject afflicted with a cancer will be responsive to a treatment regimen that comprises administering a checkpoint blockade immunotherapy directed to the cancer to the subject in accordance with an embodiment of the present disclosure.

[0078] Figures 3A, 3B, 3C, 3D, 3E, and 3F collectively provide a flow chart of processes and features for determining a likelihood that a human subject afflicted with a cancer will be responsive to a treatment regimen that comprises administering a checkpoint blockade immunotherapy directed to the cancer to the subject, in accordance with some embodiments of the present disclosure.

[0079] Figure 4 illustrates evolutionary tumor dynamics under strong immune selection and a neoantigen recognition potential fitness model based on immune interactions. In the left panel, clones are inferred from a tumor's genealogical tree. The value $n(\tau)$, the future effective size of the cancer cell population, relative to its size at the start of therapy is predicted using the equation $n(\tau) = \sum_{\alpha} X_{\alpha} \exp(F_{\alpha} \tau)$, by evolving clones under the model over a fixed time-scale, τ . Application of therapy can decrease fitness of clones depending on their neoantigens. Clones with strongly negative fitness have greater loss of population size than more fit ones. In the right panel, the disclosed model accounts for the presence of dominant neoantigens within a clone, α , by modeling presentation and recognition of inferred neoantigens, assigning fitness to a clone, F_{α} .

[0080] Figure 5 illustrates a neoantigen recognition potential fitness model based on immune interactions that accounts for the presence of dominant neoantigens within a clone, α , by modeling the presentation and recognition of inferred neoantigens and assigning a fitness to a clone, F_{α} in accordance with some embodiments of the present disclosure.

[0081] Figure 6 illustrates (A) how positions 2 and 9 in neoantigens are of less predictive value in some embodiments of the present disclosure, with neoantigens with mutations at anchor residues at position 2 and 9 have highly diverging amplitude values and are of less overall predictive value than neoantigens at other positions; and (B) how patients classified in studies as responders are marked with solid circles and non-responders are

marked with hollow circles. Positions 2 and 9 are highly constrained by a bias to be hydrophobic. Their Shannon entropy is lower than that of other residues, across all three datasets regardless of classification of their neoantigens in those datasets. Other residue sites have the same entropy as the overall proteome (Lehmann *et al.*, 2016, “Fundamental amino acid mass distributions and entropy costs in proteomes,” J. Theor. Biol. 410, pp. 119-124, which is hereby incorporated by reference) and are therefore unconstrained.

[0082] Figure 7 illustrates survival landscape as a function of TCR binding model by illustrating (A) the landscape is a contour plot of log-rank test scores in survival analysis with patient data split by median relative population size ($n(\tau) = \sum_{\alpha} X_{\alpha} \exp(F_{\alpha}\tau)$). The locally smoothed landscape is plotted for the Van Allen *et al.*, 2015, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma,” Science 350, pp. 207-211 dataset as a function of the model parameters for the logistic curve midpoint (a) and steepness (k) and (B) a logistic binding curve at inferred midpoint and steepness parameters used across all three datasets from parameters in Van Allen *et al.*, 2015, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma,” Science 350, pp. 207-211. The curve represents the binding probability of a neoantigen to a T-cell receptor associated with an IEDB epitope as a function of its alignment score to that epitope, in accordance with an embodiment of the present disclosure.

[0083] Figure 8 illustrates distribution of predicted relative population size $n(\tau)$ for responders and non-responders at consistent parameters across the Van Allen *et al.*, 2015, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma,” Science 350, pp. 207-211 cohort. Responders and non-responders are as defined in Van Allen *et al.* Error bars are 95% confidence intervals around the population average. The dashed line indicates the consistent time scale $\tau=0.09$ used for patient survival predictions. The significance of the separation of the two groups was computed with Kolmogorov-Smirnov test, the p-value at $\tau=0.09$ is 0.0016. Background shading represents significance of separation of the two groups as a continuous function of τ (** $p<0.01$, *** $p<0.001$) in accordance with an embodiment of the present disclosure.

[0084] Figure 9 illustrates distribution of predicted relative population size $n(\tau)$ for responders and non-responders at consistent parameters across the Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” N. Engl. J. Med. 371, pp. 2189-2199 cohort. Responders and non-responders are as defined in Snyder *et al.* Error bars are 95% confidence intervals around the population average. The dashed line

indicates the consistent choice of $\tau=0.09$ used for patient survival predictions. The significance of the separation of the two groups was computed with Kolmogorov-Smirnov test, the p-value at $\tau=0.09$ is 0.00084. Background shading represents significance of separation of the two groups as a continuous function of τ (** $p<0.01$, *** $p<0.001$) in accordance with an embodiment of the present disclosure.

[0085] Figure 10 illustrates distribution of predicted relative population size $n(\tau)$ for responders and non-responders at consistent parameters across the Rizvi *et al.*, 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer,” Science 348, pp. 124-128 cohort. Responders and non-responders are as defined in Snyder *et al.* Error bars are 95% confidence intervals around the population average. The dashed line indicates the consistent choice of $\tau=0.09$ used for patient survival predictions. The significance of the separation of the two groups was computed with Kolmogorov-Smirnov test, the p-value at $\tau=0.09$ is 0.00071. Background shading represents significance of separation of the two groups as a continuous function of τ (** $p<0.01$, *** $p<0.001$) in accordance with an embodiment of the present disclosure.

[0086] Figure 11 illustrates how a disclosed neoantigen recognition potential fitness model is predictive of patient survival after checkpoint blockade immunotherapy. Kaplan-Meier survival curves are calculated across a melanoma patient dataset treated with anti-CTLA4 antibodies (Van Allen *et al.*, 2015, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma,” Science 350, pp. 207-211). The samples are split in an unsupervised manner by the median value of their tumor’s relative population size $n(\tau) = \sum_{\alpha} X_{\alpha} \exp(F_{\alpha}\tau)$, where $F_{\alpha} = - \max_{i \in \text{Clone } \alpha} (A_i \times R_i)$. Curve 1102 represents low fitness tumors while curve 1104 represents high fitness tumors. Error bars represent standard error due to sample size and were calculated in GraphPad Prism 7 and are defined by the standard error of the Kaplan-Meier estimator using Greenwood’s formula (Greenwood, 1926, “The natural duration of cancer. Rep Public Health and Related Subjects. 33). The p-values from log-rank test comparing the two KM curves are shown above each plot.

[0087] Figure 12 illustrates how a disclosed neoantigen recognition potential fitness model is predictive of patient survival after checkpoint blockade immunotherapy. Kaplan-Meier survival curves are calculated across a melanoma patient dataset treated with anti-CTLA4 antibodies (See, Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” N. Engl. J. Med. 371, pp. 2189-2199). The samples are split in an

unsupervised manner by the median value of their tumor's relative population size $n(\tau) = \sum_{\alpha} X_{\alpha} \exp(F_{\alpha}\tau)$, where $F_{\alpha} = - \max_{i \in \text{Clone } \alpha} (A_i \times R_i)$. Curve 1202 represents low fitness tumors while curve 1204 represents high fitness tumors. Error bars represent standard error due to sample size and were calculated in GraphPad Prism 7 and are defined by the standard error of the Kaplan-Meier estimator using Greenwood's formula (Greenwood, 1926, "The natural duration of cancer. Rep Public Health and Related Subjects. 33). The p-values from log-rank test comparing the two KM curves are shown above each plot.

[0088] Figure 13 illustrates how a disclosed neoantigen recognition potential fitness model is predictive of patient survival after checkpoint blockade immunotherapy. Kaplan-Meier survival curves are calculated across a melanoma patient dataset treated with anti-CTLA4 antibodies (Rizvi *et al.*, 2015, "Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer," Science 348, pp. 124-128). The samples are split in an unsupervised manner by the median value of their tumor's relative population size $n(\tau) = \sum_{\alpha} X_{\alpha} \exp(F_{\alpha}\tau)$, where $F_{\alpha} = - \max_{i \in \text{Clone } \alpha} (A_i \times R_i)$. Curve 1302 represents low fitness tumors while curve 1304 represents high fitness tumors. Error bars represent standard error due to sample size and were calculated in GraphPad Prism 7 and are defined by the standard error of the Kaplan-Meier estimator using Greenwood's formula (Greenwood, 1926, "The natural duration of cancer. Rep Public Health and Related Subjects. 33). The p-values from log-rank test comparing the two KM curves are shown above each plot.

[0089] Figure 14 illustrates the log-rank test score (higher is better) for the model of the Van Allen *et al.*, 2015, "Genomic correlates of response to CTLA-4 blockade in metastatic melanoma," Science 350, pp. 207-211 cohort of Figure 11, which accounts for removal of one feature of the model: full model (1401), an MHC-presentability only model in which the recognition factor is ignored and fitness is assumed to be determined only by MHC-amplitude of neoantigens in accordance with the equation $n(\tau) = \sum_{\alpha} X_{\alpha} \exp[- \max_{i \in \text{Clone } \alpha} A_i \tau]$ (1402) and a TCR-recognition only model in which the MHC-presentation factor is ignored and fitness is assumed to be determined only by TCR-recognition of neoantigens as given by $n(\tau) = \sum_{\alpha} X_{\alpha} \exp[- \max_{i \in \text{Clone } \alpha} R_i \tau]$ (1404). These values are compared with a tumors neoantigen burden, computed as a uniform fitness cost to each neoantigen, where for L_{α} the number of neoantigens in clone, this model is defined by $n(\tau) = \sum_{\alpha} X_{\alpha} \exp[- L_{\alpha} \tau]$ (1406), and the model is computed both over a tumor's clonal structure (heterogeneous, left 1408) and without taking heterogeneity into account, where for

each of the MHC-presentability only model, the TCR-recognition only model, and the tumors neoantigen burden model, the homogenous structure equivalent is computed by assuming the tumor is strictly clonal with all neoantigens in the same clone at frequency 1 (homogenous, right 1410). The dashed line 1412 marks the score value corresponding to the significance threshold of 5%. The error bars are the standard deviation of log-rank test score acquired from the survival analysis with one sample removed from the cohort at a time (n=64).

[0090] Figure 15 illustrates the log-rank test score (higher is better) for the model of the Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” N. Engl. J. Med. 371, pp. 2189-2199 cohort of Figure 12, which accounts for removal of one feature of the model: full model (1501), an MHC-presentability only model (1502) and a TCR-recognition only model (1504). These values are compared with a tumors neoantigen burden (1506), and the model is computed both over a tumor’s clonal structure (clonal, left 1508) and without taking heterogeneity into account (homogenous, right 1510). The dashed line 1512 marks the score value corresponding to the significance threshold of 5%. The error bars are the standard deviation of log-rank test score acquired from the survival analysis with one sample removed from the cohort at a time (n=103).

[0091] Figure 16 illustrates the log-rank test score (higher is better) for the model of the Rizvi *et al.*, 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer,” Science 348, pp. 124-128 cohort of Figure 13, which accounts for removal of one feature of the model: full model (1601), an MHC-presentability only model (1602) and a TCR-recognition only model (1604). These values are compared with a tumors neoantigen burden (1606), and the model is computed both over a tumor’s clonal structure (clonal, left 1608) and without taking heterogeneity into account (homogenous, right 1610). The dashed line 1612 marks the score value corresponding to the significance threshold of 5%. The error bars are the standard deviation of log-rank test score acquired from the survival analysis with one sample removed from the cohort at a time (n=34).

[0092] Figure 17 illustrates a survival landscape for Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” N. Engl. J. Med. 371, pp. 2189-2199 (A) and Rizvi *et al.*, 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer, Science 348, pp. 124-128 (B) cohorts in accordance with an embodiment of the present disclosure.

[0093] Figure 18 illustrates the distribution of characteristic times scales of samples with clonal fitness heterogeneity for the three patient cohorts (Van Allen *et al.*, 2015, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma,” Science 350, pp. 207-211, Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” N. Engl. J. Med. 371, pp. 2189-2199, and Rizvi *et al.*, 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer, Science 348, pp. 124-128). These distributions consistently define the interval for relevant time scales of τ , in all datasets investigated $\tau \in [0,0.5]$, in accordance with an embodiment of the present disclosure.

[0094] Figure 19 illustrates significance of survival analysis reported as the result of the log-rank test on the Van Allen *et al.*, 2015, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma,” Science 350, pp. 207-211 dataset with sample split at a median value $n(\tau)$ plotted as a function of τ , in accordance with an embodiment of the present disclosure. The chosen value of parameter $\tau=0.06$ and a broad surrounding interval gives highly significant sample segregation in the Van Allen *et al.*, 2015, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma,” Science 350, pp. 207-211 dataset.

[0095] Figure 20 illustrates significance of survival analysis reported as the result of the log-rank test on the Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” N. Engl. J. Med. 371, pp. 2189-2199 dataset with sample split at a median value $n(\tau)$ plotted as a function of τ , in accordance with an embodiment of the present disclosure. The chosen value of parameter $\tau=0.06$ and a broad surrounding interval gives highly significant sample segregation in the Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” N. Engl. J. Med. 371, pp. 2189-2199 dataset.

[0096] Figure 21 illustrates significance of survival analysis reported as the result of the log-rank test on the Rizvi *et al.*, 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer, Science 348, pp. 124-128 dataset with sample split at a median value $n(\tau)$ plotted as a function of τ , in accordance with an embodiment of the present disclosure. The chosen value of parameter $\tau=0.06$ and a broad surrounding interval gives highly significant sample segregation in the Rizvi *et al.*, 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer, Science 348, pp. 124-128 dataset.

[0097] Figure 22 illustrates how word usage in a proteome is exhausted between 5 and 6 letter words. Given the entropy of the genome, see Lehmann et al., 2016, “Fundamental amino acid mass distributions and entropy costs in proteomes,” *J. Theor. Biol.* 410, pp. 119-124 (2016), hereby incorporated by reference, the expected number of words of a given length in the proteome is calculated as a function of word length. This is compared that to the number of unique words in the proteome of a given length. Between 5 and 6 letters the two curves diverge due to the finite size of the genome. By the time one reaches 9 letter nonamers (the length of a neoantigen) this divergence is of several orders of magnitude.

[0098] Figure 23 illustrates the ranking of fitness models when accounting for tumor subclonal composition in which the ranking of fitness models evaluated for the Van Allen *et al.*, 2015, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma,” *Science* 350, pp. 207-211 (n=103) cohort in accordance with an embodiment of the present disclosure.

[0099] Figure 24 illustrates the ranking of fitness models when accounting for tumor subclonal composition in which the ranking of fitness models evaluated for the Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” *N. Engl. J. Med.* 371, pp. 2189-2199 (n=64) cohort in accordance with an embodiment of the present disclosure.

[00100] Figure 25 illustrates the ranking of fitness models when accounting for tumor subclonal composition in which the ranking of fitness models evaluated for the Rizvi *et al.*, 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer,” *Science* 348, pp. 124-128 (n=34) cohort in accordance with an embodiment of the present disclosure.

[00101] Figure 26 illustrates inferred MHC binding affinities of mutant versus wildtype peptides in the Van Allen *et al.*, 2015, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma,” *Science* 350, pp. 207-211 dataset in accordance with an embodiment of the present disclosure.

[00102] Figure 27 illustrates inferred MHC binding affinities of mutant versus wildtype peptides in the Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” *N. Engl. J. Med.* 371, pp. 2189-2199 dataset in accordance with an embodiment of the present disclosure.

[00103] Figure 28 illustrates inferred MHC binding affinities of mutant versus wildtype peptides in the Rizvi *et al.*, 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer, Science 348, pp. 124-128 dataset in accordance with an embodiment of the present disclosure.

[00104] Figure 29 illustrates alignments to IEDB epitopes in which the TCR recognition probability for a neoantigen is a sigmoidal function of the neoantigen’s alignment scores with IEDB epitopes, here shown as evaluated for the set of neoantigens from Van Allen *et al.* cohort patients, using a consistent set of parameters in accordance with an embodiment of the present disclosure.

[00105] Figures 30A, 30B, 30C, 30D, 30E, and 30F illustrate the effect of IEDB sequence content on predictive power of neoantigen recognition potential fitness model. Predictions were performed using subsampled IEDB epitope sequences, with subsampling rate varying between 0.1 and 0.9. For each rate, 10,000 iterations were performed to obtain a distribution of log-rank test scores. The violin plots represent data density at a given value on a vertical axis (n=10,000). Solid black lines mark the log-rank test score of the prediction on the full set of epitope sequences and gray thick lines mark the median scores on subsampled data. Figures 30A, 30B, 30C, subsampling of the original set of IEDB sequences, supported by positive T-cell assays, shows that quality of predictions decreases with subsampling rate. Prediction quality is more robust in the Snyder *et al.* and Rizvi *et al.* datasets. Figures 30D, 30E, and 30F, analogous subsampling procedure was repeated on IEDB sequences not supported by positive T-cell assays. For Van Allen *et al.* and Snyder *et al.*, model performance is substantially lowered.

[00106] Figures 31A, 31B, and 31C illustrate how neoantigen residue positions 2 and 9 provide less predictive value. The violin plots represent data density at a given value on a vertical axis. In Figure 31A, neoantigens coming from mutations at position 2 or 9 tend to have wildtype peptides with larger predicted affinities. In particular, this is magnified if the corresponding wildtype residue is non-hydrophobic. In Figure 31B, those biases are reflected in a wider distribution of amplitudes

$$n(\tau, \beta) = \sum_{\alpha} X_{\alpha} \exp \left[\sum_{i \in \text{Clone } \alpha}^{\max} \frac{\exp(-\beta f_i)}{Z(\beta)} f_i \tau \right]$$

for wildtype peptides with non-hydrophobic residues at positions 2 and 9. Figure 31C illustrates the Shannon entropy of amino acid diversity by position in neoantigens, shown for

all distinct HLA-types and computed based on neoantigens across all datasets. Positions 2 and 9 have lower entropy than other residues. Other sites have the same entropy as the overall proteome (Lehmann *et al.*, 2016, “Fundamental amino acid mass distributions and entropy costs in proteomes,” J. Theor. Biol. 410, pp. 119-124) and are therefore unconstrained. Five HLA with non-canonical entropy profiles are singled out in the plot. These HLA types contributed only five informative neoantigens across all datasets and therefore are not treated differentially in the model.

[00107] Figures 32A, 32B, and 32C illustrate ranking of fitness models disregarding subclonal composition of tumors. Fitness models described in conjunction with Figures 23, 24, and 25 are evaluated without accounting for subclonal composition of tumors on the same three cohorts, Van Allen *et al.* (n=103), Snyder *et al.* (n=64) and Rizvi *et al.* (n=34). As in Figures 23, 24, and 24, parameters used for predictions and error bars for these parameters are reported. Parameter τ is not a free parameter when disregarding subclonal composition of tumors and is not reported. Log-rank test scores are reported for all models and the log-rank test p-value for models with significant patient segregation ($p < 0.05$). The significant models are highlighted: models significant in a single cohort are shown with a solid line box, and models significant across two cohorts are shown with a dashed line box.

[00108] Figures 33A, 33B, 33C, 33D, 33E, and 33F illustrate survival analysis score landscape as a function of model parameters. In Figures 33A, 33B, and 33C, the landscape of log-rank test scores as the function of the parameters of the TCR binding model (a and $1/k$), shown for the consistent choice of $\tau = 0.09$, different shadings represent the significance level of the long-rank test. The regions of high scores are similar across all three datasets. The point corresponding to consistent parameters ($a = 26$ and $k = 4.87$) is marked by a black dot in each plot. In Figures 33D, 33E, and 33F, log-rank score for fitness model at consistent binding function parameters, plotted as a function of τ . Dashed vertical lines are at $\tau = 0.09$, thin solid lines mark the score values corresponding to significance of 0.05, 0.01 and 0.005. (n=103 (Figures 33A and 33D), n=64 (Figures 33B and 33E), n=34 (Figures 33C and 33F)).

[00109] Figures 34A, 34B, and 34C illustrate how cytolytic score improves prediction quality. Figure 34A illustrates a Kaplan-Meier curve of overall survival shown for the disclosed model applied to Van Allen *et al.* for n=40 patient subset with transcriptional data. Samples are split by the median value of their tumor’s relative population size $n(\tau)$ from the equation:

$$n(\tau) = \sum_{\alpha} X_{\alpha} \exp(F_{\alpha} \tau),$$

Error bars represent standard error due to sample size. Figure 34B illustrates that the model optimized for cytolytic score significantly separates patients as described in further detail in Example IV. Figure 34C illustrates how inclusion of cytolytic score in the model improves prediction on 40 patient subset. The p-values from log-rank tests comparing the two KM curves are shown above each plot. In Figures 34A and 34C, consistent parameters are used to trained on the three cohorts (Figure 31). In Figure 34B, parameter τ is optimized.

Figure 35 illustrates how the reshuffling of patient HLA-types reduces predictive power of the disclosed neoantigen recognition potential fitness model. In each cohort, ten iterations of reshuffling patient HLA-types, followed by computational neoantigen prediction, recognition potential fitness model calculation and survival analysis is performed. The distribution of log-rank test scores is reported over these iterations: boxes mark 75% confidence intervals and whiskers mark the range of scores (n=10). The score values for the model on original data are marked with squares.

[00110] Figure 36 illustrates a multivariate analysis with a Cox proportional hazards model that was performed to adjust for clinical covariates, while assessing for the predictive value of $n(\tau)$ values. In melanoma cohorts (Van Allen *et al.*, n=103 and Snyder *et al.*, n=64), stage, gender, and age were controlled for. Stage IIIC and IVa are combined together, as both of these stages had limited number of patients in either cohort. Stage IIIC/IVa serve as the reference in the table. In both the Van Allen *et al.* and Snyder *et al.* cohorts, $n(\tau)$ predictions are independently associated with overall survival after anti-CTLA4 therapy. In the lung cancer cohort (Rizvi *et al.*, n=34), all patients are Stage IV, so we correct for age, gender, and number of pack years smoked, and continued to find that $n(\tau)$ predictions are independently associated with overall survival after anti-PD1 therapy.

[00111] Figure 37 illustrates a script for computing neoantigen fitness scores (recognition potentials) for the Van Allen *et al.*, Snyder *et al.*, and Rizvi *et al.*, datasets in accordance with an embodiment of the present disclosure.

[00112] Figures 38A, 38 B and 38C collectively illustrate the “main.py” python script that is called by the script of Figure 37 in order to compute neoantigen fitness scores (recognition potentials) for the Van Allen *et al.*, Snyder *et al.*, and Rizvi *et al.*, datasets in accordance with an embodiment of the present disclosure.

[00113] Figures 39A and 39B collectively illustrate the “aligner.py” python script that is called by the “main.py” script of Figure 38 in order to compute neoantigen fitness scores

(recognition potentials) for the Van Allen *et al.*, Snyder *et al.*, and Rizvi *et al.*, datasets in accordance with an embodiment of the present disclosure.

[00114] Figures 40A, 40B, and 40C collectively illustrate the “neoantigen.py” python script that is called by the “main.py” script of Figure 38 in order to compute neoantigen fitness scores (recognition potentials) for the Van Allen *et al.*, Snyder *et al.*, and Rizvi *et al.*, datasets in accordance with an embodiment of the present disclosure.

[00115] Figures 41A and 41B collectively depict overall Survival and patient overlap of short and long term survivors in tissue microarray, whole exome sequencing, TCR V β sequencing, and bulk tumor transcriptomic profiling cohorts.

[00116] Figures 42A, 42B, 42C, 42D, 42E, 42F, 42G, 42H, 42K, 42L, and 42M collectively present data showing that long term pancreatic cancer survivors displayed enhanced intratumoral T cell immunity. (A) Overall survival of short (>3m, <1yr) and long term (>3yr) survivor cohorts. Composite images (B) and quantification (C, D, E) of multiplexed immunophenotyping. In (B), red rectangular sections are enlarged to 50X. CK19 stains tumor cells. Arrows indicate CD3⁺CD8⁺Granzyme-B⁺ T cells. Data in (E) indicates log₂(cells/mm²). Highlighted boxes indicated significantly increased parameters in long term survivors. (F) Bulk tumor transcriptomic immune profiling. Ligands of PD-1 (PDL-1) and TIGIT (CD226) are also shown. (G) T cell frequency (top) and repertoire clonality (1=complete oligoclonality, 0=complete polyclonality, bottom) in tumor and matched adjacent non-tumor pancreatic tissue by quantitative TCR V β sequencing. (H) Percentage of tumor T cell clones unique to tumors, or shared with matched adjacent non-tumor pancreatic tissue by quantitative TCR V β sequencing. (I) Frequency of PD-1, CD69, CD45RA, and CD45RO expressing CD8⁺ T cells in blood, tumor draining lymph nodes (DLN), and tumor in unselected patients by flow cytometry. (J) Clonality of tumor T cell repertoire in short and long term survivors by quantitative TCR V β sequencing. All data represent values in individual patients. C, D, E represent median values of 3 cores per patient. Horizontal bars represent median values and are indicated in (C). In (I, bottom), data are mean $\pm$ SEM, * and ** indicate comparisons between blood, DLN, and tumor. *P<0. 05; ** P<0. 01; *** P<0. 001; **** P<0. 0001. (K-M) Depict overall survival of patients who did or did not receive adjuvant chemotherapy (adjuvant chemotherapy +/- respectively, top left), and of patients with tumors harboring greater or less than the median number of CD3-CD8Granzyme B triple positive cells (CD3-CD8-GranzymeB^{Hi/Low} respectively, top right) .

Overall survival of all four groups shown in bottom. Table shows univariate and multivariate Cox regression analysis of clinicopathologic features, adjuvant chemotherapy, and CD3-CD8-GranzymeB density association with overall survival.

[00117] Figures 43A and 43B collectively depict representative sequential immunohistochemical staining of a single short term and a single long term core tumor section. Sections bounded by black rectangles (100X) are magnified to 275X (right) for each core section. Merged images are shown in Figure 42B.

[00118] Figure 44 depicts immunofluorescent quantification of CD8⁺ and CD4⁺ cells in tumor tissue microarray of short and long term survivors. Slides used above were cut from separate sections of the block as those used for sequential immunohistochemistry (Figure 42B-42E). * P<0.05.

[00119] Figure 45 depicts the number of T cell clones in matched tumor and adjacent non-tumor pancreatic tissue. **** P<0.0001.

[00120] Figures 46A-46D collectively depict a representative flow cytometric gating strategy to phenotype human T cells. Single cell suspensions from peripheral blood, tumor draining lymph nodes, and tumor tissue were subjected to flow cytometric analysis. First plot is pre-gated on live cells, followed by CD45⁺, and CD3⁺CD56⁻ cells. Values indicate percentage of cells within the red boxes, and are gated based on isotype controls.

[00121] Figure 47 depicts an oncoprint demonstrating the frequency of oncogenic driver gene mutations in the MSKCC PDAC cohort.

[00122] Figures 48A, 48B1, 48B2, and 48C collectively depict data showing that neoantigen quantity and cytotoxic CD8⁺ T cell infiltrate identified long term pancreatic cancer survivors. (A) Number of nonsynonymous, missense, and neoantigenic mutations per tumor. (B1-B2) Overall survival of patients with tumors harboring greater than the median number of neoantigens (NeoantigenHi), CD3-CD8 double positive cells (CD3-CD8Hi), CD3-CD8-Granzyme-B triple positive cells (CD3-CD8-Granzyme-BHi), or median polyclonality (PolyclonalHi), compared to all other patients (Rest). Neoantigens were determined using the MSKCC (left) and the pVAC-Seq (right) neoantigen prediction pipelines. (C) Unsupervised hierarchical clusters (cluster 1-4) of bulk tumor whole transcriptomic profiling in short and long term survivors (left). Each column represents a patient, each row a gene. Gene list available in methods. Overlap of NeoantigenHiCD8Hi, NeoantigenHiCD3-CD8-Granzyme-BHi, or all other tumors (Rest) with transcriptionally defined clusters (right). Each bar in (A)

represents data from one patient's tumor. Mutations in (A), (B1-B2), and (C) were determined by whole exome sequencing. Polyclonality in (B1-B2) was calculated by quantitative TCR V β sequencing. Number of immune cells in (B1-B2) and (C) were quantified by immunophenotyping as the median of three cores per tumor in short and long term survivor tumor tissue microarrays. *P<0.05.

[00123] Figures 49A and 49B collectively depict number of neoantigens per tumor (n=58) as determined by the MSKCC and pVAC-Seq neoantigen calling pipelines (A). Tick marks on the x-axis correspond to individual tumors. Correlation matrix of neoantigens as determined by the MSKCC and pVAC-Seq neoantigen calling pipelines (B). Solid red line indicates line of best fit, dotted lines indicate 95% confidence intervals.

[00124] Figures 50A, 50B, and 50C collectively depict survival of patients relative to number of neoantigens. (A) depicts overall survival of patients with tumors harbouring more or fewer than the median number of neoantigens (neoantigen^{hi/low}) and CD4 single positive cells (CD4^{hi/low}), compared to all other patients (Rest). (B) depicts overall survival of patients with tumors harbouring CD3-CD8 double-positive cells (CD3-CD8^{hi/low}), compared to all other patients (Rest). (C) depicts overall survival of patients with tumors harbouring polyclonality (polyclonal^{hi/low}) and mutations (mutation^{hi/low}), compared to all other patients (Rest).

[00125] Figure 51 depicts an oncoprint demonstrating the frequency of oncogenic driver mutations in short and long term tumors.

[00126] Figure 52 depicts the number of nonsynonymous, missense, and immunogenic mutations (neoantigens) in short and long term PDAC tumors.

[00127] Figure 53 depicts overall survival stratified by mutations in ARID1A, KRASQ61H, RBM10, and MLL related genes (MLL, MLL2, MLL3, MLL5) in accordance with an embodiment of the present disclosure.

[00128] Figures 54A, 54B, 54C, 54D, 54E1, 54E2, 54E3, 54F, 54G1, 54G2, 54G3, 54G4, 54H1, and 54H2 collectively depict neoantigens with microbial homology stratify long term pancreatic cancer survivors. (A) Schematic of neoantigen immune fitness models. Each circle represents a tumor clone in an evolutionary tree. Clones in both models are identical with respect to the number of mutations and neoantigens. Numbers represent hypothetical neoantigens gained in a successive tumor clone. Shades of red indicate immunogenicity of each clone, as ascribed by the two models, namely neoantigen quality (recognition potential

model or quality model) or neoantigen quantity (neoantigen load model or quantity model).

(B) Parameters defining the recognition potential score in the recognition potential model. In (a), amino acid sequences of a hypothetical wild type (WT) epitope, tumor neoepitope, and a homologous microbial epitope are shown. Yellow highlights the changing amino acid between the WT and tumor sequence as a consequence of a tumor specific mutation. The amino acids in red indicate homology between the tumor neoepitope and the microbial epitope. (C) Overall survival of patients in the MSKCC cohort whose tumors displayed high neoantigen recognition potential (recognition potential^{Hi}) compared to low neoantigen recognition potential (recognition potential^{Low}) to pathogenic epitopes (top). Overall survival of patients whose tumors displayed high neoantigen load (Neoantigen^{Hi}) compared to low neoantigen load (Neoantigen^{Low}) (bottom). (D) Distribution of high and low recognition potential neoantigens in Neoantigen^{Hi} CD3-CD8^{Hi} and Neoantigen^{Hi} CD3-CD8-Granzyme-B^{Hi} long term pancreatic cancer survivors compared to all other patients (Rest). ** P<0. 01. (E) Neoantigen quality is independently associated with long term survival. Overall survival of patients whose tumors displayed high compared to low neoantigen recognition potential (Neoantigen Quality^{Hi/Low}) to microbial epitopes (E1 top). Overall survival of patient who did or did not receive adjuvant chemotherapy (E2 top). Overall survival of all four groups is shown at the bottom. Neoantigen quality defined by neoantigen fitness modeling of sequence homology to microbial epitopes. Table shows univariate and multivariate Cox regression analysis of the associations of clinicopathologic features, adjuvant chemotherapy, and neoantigen quality with overall survival. Data include all patients in the whole exome sequencing MSKCC cohort. (F) Overall survival of patients in the ICGC cohort whose tumors displayed high neoantigen recognition potential (Neoantigen Quality^{Hi}) compared to low neoantigen recognition potential (Neoantigen Quality^{Low}) to pathogenic epitopes (top). Overall survival of patients whose tumors displayed high neoantigen load (Neoantigen Quantity^{Hi}) compared to low neoantigen load (Neoantigen Quantity^{Low}) (bottom). (G1) Number of nonsynonymous, missense, and neoantigenic mutations per patient in the ICGC cohort (n=166). (G2) Overall survival of patients whose tumors displayed high compared to low neoantigen recognition potential (Neoantigen Quality^{Hi/Low}) to microbial epitopes in the ICGC cohort (top). Overall survival of patients in the ICGC cohort stratified by adjuvant chemotherapy administration (bottom). (G3) Overall survival of all four groups. Neoantigen quality defined by neoantigen fitness modeling of sequence homology to microbial epitopes. Table shows univariate and multivariate Cox regression analysis of the associations of clinicopathologic features, adjuvant chemotherapy, and neoantigen quality with overall

survival in the ICGC cohort. (H) depicts parameters of the neoantigen recognition potential fitness (quality model) for a, the MSKCC cohort b, and the ICGC cohort. (top) Log-rank test score landscape as a function of the model parameters, the horizontal alignment score displacement a, and the characteristic time T, the significance of the score is denoted in the legend. (bottom) Two dimensional histograms showing distributions of optimal parameters obtained on subsampled datasets with 50, 70, and 80% of patients left, over 500 iterations of subsampling at each frequency.

[00129] Figures 55A, 55B, 55C, 55D, 55E, 55F, 55G, 55H, 55I, 55J, 55K, and 55L collectively depict MUC16 to be a neoantigen enriched immunogenic hotspot in long term pancreatic cancer survivors. (A) Frequency (left) and distribution (right) of patients with neoantigens. Genes harboring neoantigens in >15% of patients are indicated. (B) Frequency of patients with MUC16 neoantigens and (C) the number of MUC16 neoantigens per tumor. (D) Bulk tumor MUC16 mRNA and (E) quantified immunohistochemical protein expression. (F) MUC16 mutant allele frequency in non-hypermethylated tumors with MUC16 mutations. (G) Frequency of patients with neoantigens in genes recurrently harboring neoantigens in > 5% of patients in both MSKCC and ICGC cohorts. (H) Frequency of MUC16 neoantigens in MSKCC and ICGC cohorts. (I, J, K) Unique wild type (WT peptide 1,2) and mutant (Mutant peptide 1,2) MUC16 nonamers were generated for two long term PDAC survivors (Patients 1, 2) respectively based on in silico neoantigen predictions. Peripheral blood mononuclear cells were pulsed in vitro for 3 weeks and CD8⁺ T cell expansion (I), CD107a expression (J), and TCR V β clonotyping (K) were determined. Data in (I, J) indicate n=2/group and are representative of two experiments with similar results. Venn diagrams in (K) indicate clonal overlap. (L) Peripheral blood mononuclear cells of healthy donors (neopeptide 1, 2 – n=5; neopeptide 3 – n=6) were stained with MUC16 neopeptide-MHC multimers based on in silico somatic neoantigen predictions in tumors of long term survivors. For each neopeptide, healthy donors were HLA-matched to alleles predicted to bind the neopeptide. HLA-matched irrelevant multimers (control multimer) were available for neopeptides 1, and 2 and were used as controls. Data represent mean $\pm$ SEM. *P<0. 05, ** P<0. 01, *** P<0. 001.

[00130] Figure 56 depicts the recognition potential load fitness model parameter stability analysis. The boxplots (top) and the histograms (bottom) illustrate distributions of the difference between the optimal shift parameter obtained on the full dataset and the subsampled datasets, with frequencies 0.8, 0.7, 0.6, 0.5, and 0.4.

- [00131] Figure 57 depicts the frequency of MUC16 mutations in short and long term PDAC tumors.
- [00132] Figure 58 depicts a lollipop plot showing location of *MUC16* mutations in short and long term pancreatic cancer survivors.
- [00133] Figure 59 depicts a lollipop plot showing location of *MUC16* neoantigens in short and long term pancreatic cancer survivors.
- [00134] Figure 60 depicts genes most frequently generating neoantigens as determined by pVACSeq. Frequency of patients and raw numbers are indicated.
- [00135] Figure 61 depicts MUC16 mRNA and protein expression in MUC16 non-mutated (WT) and mutated (mutant) tumors.
- [00136] Figure 62 depicts mRNA expression of transcriptional activators of MUC16.
- [00137] Figure 63 depicts mRNA expression of mediators implicated in MUC16 dependent tumor progression in short and long term tumors.
- [00138] Figure 64 depicts mRNA expression of tissue expression antigens MUC1, MUC4, WT1, mesothelin, and Annexin A2 in short and long term tumors.
- [00139] Figure 65 depicts quantified protein expression of tissue expression antigens MUC1, mesothelin, and Annexin A2 in short and long term tumors. WT1 protein was undetectable in tumors of both short and long term survivors.
- [00140] Figure 66 depicts number of nonsynonymous, missense, and neoantigen mutations per patient in the ICGC cohort (n=169).
- [00141] Figures 67A and 67B collectively depict representative gating strategy to identify CD8⁺ T cells in peripheral blood of healthy donors (67A). Identification of CD8⁺ T cells in healthy donors reactive to 2 unique MUC16 neoepitopes (NE#1, NE#2; NE=Neoepitope) predicted to bind to the B*0801 HLA-allele, using MUC16-neoepitope-HLA multimers (67B).
- [00142] Figure 68 depicts identification of CD8⁺ T cells in healthy donors reactive to MUC16 neoepitopes (NE#3, NE=Neoepitope) predicted to bind to the A*2402 HLA-allele, using MUC16-neoepitope-HLA multimers.
- [00143] Figures 69A, 69B, 69C1, 69C2, 69D1, 69D2, 69D3, and 69D4 collectively depict neoantigen and cross-reactive microbial peptide T cells detected in blood and tumours.

(A) Left, gene expression in the presence (red) or absence (grey) of high-quality neoantigenic mutations. x axis, genes; shaded circles, biologically independent samples in individual patients (n = 30). Right, median non-neoantigenic and neoantigenic expression. All high-quality neoantigenic genes with available mRNA expression are shown. (B) Metastatic propagation of all clones in the primary tumour stratified by neoantigen quality. Mutant allele frequencies in matched primary–metastatic tumours (left) and metastatic tumours alone (right) are shown in biologically independent samples in one patient. MAF, mutant allele frequency; M, metastasis; P, primary tumour. (C and D) Peripheral blood mononuclear cells (PBMCs) pulsed with no (N), wild-type control (WT), cross-reactive (CR), and high-quality neo (M) peptides (n = 7). c, CD8+ T-cell expansion and degranulation. (D) Clonal overlap of expanded T-cell clones in (C) and archival tumours by TCR V β sequencing. Arrows indicate clones in archival primary tumours with rank frequencies. Venn diagrams show the number of T-cell clones expanding with mutant peptides, with cross-reactive peptides, their respective clonal overlap, and clonal overlap with archival primary tumours. Note the presence of clones recognizing both neopeptides and cross-reactive peptides in archival tumours. Years surviving after surgery are shown for each individual patient. AWD, alive with disease; NED, no evidence of disease. Horizontal bars indicate median values, error bars represent the s.e.m. n is the number of biologically independent samples in individual patients in (A) and (C). P values were determined using a two-tailed Student's t-test (A), a two-tailed Mann–Whitney U-test (B), a one-way analysis of variance (ANOVA) with Tukey's multiple comparison test (C) and as described in Example X (D).

[00144] Figures 70A1, 70A2, 70B1, 70B2, and 70C collectively show decreased MUC16 protein expression in long term survivors is not a consequence of differential antibody binding to neoantigenic MUC16 mutations. (A) representative immunohistochemical staining and quantification of MUC16 expression in tissue microarrays of short and long term pancreatic cancer survivors as assessed using three independent anti MUC16 antibodies Ab#1-clone EPSISR23, purchased from Abcam; Ab#2-polyclonal, purchased from Abcam ab1333419; Ab#3- clone 4H11. (B) Western blot (B1) and immunocytochemistry (B2) of untransfected (-), empty vector (vector), MUC16 wild type (MUC16WT), and MUC16 mutant (MUC16 R15C) HEK293T cells. The top left blot was probed with anti MUC16 specific antibody (clone 4H11) and the right blot with anti β -actin. Red rectangle indicates MUC16 specific band. All bottom cells were probed with anti MUC16 antibody (clone 4H11). The inserted mutation was identical to a neoantigenic

MUC16 mutation. (C) MUC16 immunohistochemistry on two long term pancreatic cancer survivors with MUC16 neoepitopes in primary resected tumors. Areas in rectangular low power fields are magnified on the right.

[00145] Figure 71 depicts a comprehensive flowchart of neoantigen fitness modelling pipeline. Software programs utilized for each step are indicated in bold, colored text. Mathematical formulae for calculation of individual components of neoantigen quality are shown above and further defined in Methods. All software components of the pipeline are published and/or publically available.

[00146] Like reference numerals refer to corresponding parts throughout the several views of the drawings.

DETAILED DESCRIPTION

[00147] The presently disclosed subject matter provides strategies for improving cancer therapy (*e.g.*, immunotherapy) by selection of particular patients to receive therapy based on the neoantigen portfolio of the patients (*e.g.*, immunogenic hotspot (*e.g.*, MUC16 for pancreatic cancer), neoantigen quantity and neoantigen immunogenicity). The presently disclosed subject matter also provides methods for predicting responsiveness of cancer patients to immunotherapy based on the patients' neoantigen portfolio. Furthermore, based on the fact that, at least in tumors with high mutational loads, the amount of DNA damage is sufficient for the immune system to perceive one or more neoantigens as foreign, it becomes of interest to stimulate neoantigen-specific T cell responses in cancer patients. The presently disclosed subject matter further provides vaccines comprising one or more neoantigen, and/or therapeutic T cell preparations comprising neoantigen-specific T cells.

[00148] The present disclosure further provides a recognition potential fitness model for tumors based on the immune interactions of dominant neoantigens predicts response to immunotherapy. The disclosed recognition potential fitness model associates each neoantigen with a fitness cost, which is termed the recognition potential of a neoantigen. The recognition potential of a neoantigen is the likelihood it is productively recognized by the TCR repertoire. It is defined by two components. The first is the amplitude A , which is given by the relative probability that a neoantigen will be presented by on class I major histocompatibility complex (MHC) and the relative probability that its wildtype counterpart will not be presented. The second component is the probability R that a presented neoantigen will be recognized by the TCR repertoire, which is inferred from its sequence homology to

known antigens. For a given neoantigen their product defines its recognition potential, $A \times R$. In some embodiments, this recognition potential is computed using the scripts illustrated in Figures 37 through 40 in accordance with the present disclosure.

[00149] Reference will now be made in detail to embodiments, examples of which are illustrated in the accompanying drawings. In the following detailed description, numerous specific details are set forth in order to provide a thorough understanding of the present disclosure. However, it will be apparent to one of ordinary skill in the art that the present disclosure may be practiced without these specific details. In other instances, well-known methods, procedures, components, circuits, and networks have not been described in detail so as not to unnecessarily obscure aspects of the embodiments.

[00150] 1. *Definitions*

[00151] Unless defined otherwise, all technical and scientific terms used herein have the meaning commonly understood by a person skilled in the art to which this invention belongs. The following references provide one of ordinary skill in the art with a general definition of many of the terms used herein: Singleton *et al.*, Dictionary of Microbiology and Molecular Biology (2nd ed. 1994); The Cambridge Dictionary of Science and Technology (Walker *ed.*, 1988); The Glossary of Genetics, 5th Ed., R. Rieger *et al.* (eds.), Springer Verlag (1991); and Hale & Marham, The Harper Collins Dictionary of Biology (1991); Molecular Cloning: a Laboratory Manual 3rd edition, J. F. Sambrook and D. W. Russell, ed. Cold Spring Harbor Laboratory Press 2001; Recombinant Antibodies for Immunotherapy, Melvyn Little, ed. Cambridge University Press 2009; "Oligonucleotide Synthesis" (M. J. Gait, ed., 1984); "Animal Cell Culture" (R. I. Freshney, ed., 1987); "Methods in Enzymology" (Academic Press, Inc.); "Current Protocols in Molecular Biology" (F. M. Ausubel *et al.*, eds., 1987, and periodic updates); "PCR: The Polymerase Chain Reaction", (Mullis *et al.*, ed., 1994); "A Practical Guide to Molecular Cloning" (Perbal Bernard V., 1988); "Phage Display: A Laboratory Manual" (Barbas *et al.*, 2001). The contents of these references and other references containing standard protocols, widely known to and relied upon by those of skill in the art, including manufacturers' instructions are hereby incorporated by reference as part of the presently disclosed subject matter. As used herein, the following terms have the meanings ascribed to them below, unless specified otherwise.

[00152] As used herein, the term "about" or "approximately" means within an acceptable error range for the particular value as determined by one of ordinary skill in the

art, which will depend in part on how the value is measured or determined, i.e., the limitations of the measurement system. For example, “about” can mean within 3 or more than 3 standard deviations, per the practice in the art. Alternatively, “about” can mean a range of up to 20%, *e.g.*, up to 10%, up to 5%, or up to 1% of a given value. Alternatively, particularly with respect to biological systems or processes, the term can mean within an order of magnitude, *e.g.*, within 5-fold, or within 2-fold, of a value.

[00153] As used herein, the term “neoantigen”, “neoepitope” or “neopeptide” refers to a tumor-specific antigen that arises from one or more tumor-specific mutation, which alters the amino acid sequence of genome encoded proteins.

[00154] As used herein, the term “neoantigen number” or “neoantigen burden” refers to the number of neoantigen(s) measured, detected or predicted in a sample (*e.g.*, a biological sample from a subject). In certain embodiments, the neoantigen number is measured by using whole exome sequencing and *in silico* prediction.

[00155] As used herein, the term “mutation” refers to permanent change in the DNA sequence that makes up a gene. In certain embodiments, mutations range in size from a single DNA building block (DNA base) to a large segment of a chromosome. In certain embodiments, mutations can include missense mutations, frameshift mutations, duplications, insertions, nonsense mutation, deletions and repeat expansions. In certain embodiments, a missense mutation is a change in one DNA base pair that results in the substitution of one amino acid for another in the protein made by a gene. In certain embodiments, a nonsense mutation is also a change in one DNA base pair. Instead of substituting one amino acid for another, however, the altered DNA sequence prematurely signals the cell to stop building a protein. In certain embodiments, an insertion changes the number of DNA bases in a gene by adding a piece of DNA. In certain embodiments, a deletion changes the number of DNA bases by removing a piece of DNA. In certain embodiments, small deletions can remove one or a few base pairs within a gene, while larger deletions can remove an entire gene or several neighboring genes. In certain embodiments, a duplication consists of a piece of DNA that is abnormally copied one or more times. In certain embodiments, frameshift mutations occur when the addition or loss of DNA bases changes a gene's reading frame. A reading frame consists of groups of 3 bases that each code for one amino acid. In certain embodiments, a frameshift mutation shifts the grouping of these bases and changes the code for amino acids. In certain embodiments, insertions, deletions, and duplications can all be frameshift mutations. In certain embodiments, a repeat expansion is another type of mutation. In

certain embodiments, nucleotide repeats are short DNA sequences that are repeated a number of times in a row. For example, a trinucleotide repeat is made up of 3-base-pair sequences, and a tetranucleotide repeat is made up of 4-base-pair sequences. In certain embodiments, a repeat expansion is a mutation that increases the number of times that the short DNA sequence is repeated.

[00156] As used herein, the “median” value (*e.g.*, median neoantigen number, median neoantigen-microbial homology - *e.g.*, median cross-reactivity score, median recognition potential score), or median activated T cell number - refers to the median value obtained from a population of subjects having a cancer (*e.g.*, pancreatic cancer, *e.g.*, PDAC). The median values may be previously determined reference values, or may be contemporaneously determined values.

[00157] As used herein, the term “cell population” refers to a group of at least two cells expressing similar or different phenotypes. In non-limiting examples, a cell population can include at least about 10, at least about 100, at least about 200, at least about 300, at least about 400, at least about 500, at least about 600, at least about 700, at least about 800, at least about 900, or at least about 1000 cells expressing similar or different phenotypes.

[00158] As used herein, the terms “antibody” and “antibodies” refer to antigen-binding proteins of the immune system. As used herein, the term “antibody” includes whole, full length antibodies having an antigen-binding region, and any fragment thereof in which the “antigen-binding portion” or “antigen-binding region” is retained, or single chains, for example, single chain variable fragment (scFv), thereof. The term “antibody” means not only intact antibody molecules, but also fragments of antibody molecules that retain immunogen-binding ability. Such fragments are also well known in the art and are regularly employed both *in vitro* and *in vivo*. Accordingly, as used herein, the term “antibody” means not only intact immunoglobulin molecules but also the well-known active fragments F(ab')₂, and Fab. F(ab')₂, and Fab fragments that lack the Fc fragment of intact antibody, clear more rapidly from the circulation, and can have less non-specific tissue binding of an intact antibody (Wahl et al., J. Nucl. Med. 24:316-325 (1983)). In certain embodiments, an antibody is a glycoprotein comprising at least two heavy (H) chains and two light (L) chains interconnected by disulfide bonds. Each heavy chain is comprised of a heavy chain variable region (abbreviated herein as V_H) and a heavy chain constant (C_H) region. The heavy chain constant region is comprised of three domains, C_H 1, C_H 2 and C_H 3. Each light chain is comprised of a light chain variable region (abbreviated herein as V_L) and a light chain

constant C_L region. The light chain constant region is comprised of one domain, C_L. The V_H and V_L regions can be further sub-divided into regions of hypervariability, termed complementarity determining regions (CDR), interspersed with regions that are more conserved, termed framework regions (FR). Each V_H and V_L is composed of three CDRs and four FRs arranged from amino-terminus to carboxy-terminus in the following order: FR1, CDR1, FR2, CDR2, FR3, CDR3, FR4. The variable regions of the heavy and light chains contain a binding domain that interacts with an antigen. The constant regions of the antibodies can mediate the binding of the immunoglobulin to host tissues or factors, including various cells of the immune system (*e.g.*, effector cells) and the first component (C1 q) of the classical complement system.

[00159] The term “antigen-binding portion”, “antigen-binding fragment”, or “antigen-binding region” of an antibody, as used herein, refers to that region or portion of an antibody that binds to the antigen and which confers antigen specificity to the antibody; fragments of antigen-binding proteins. It has been shown that the antigen-binding function of an antibody can be performed by fragments of a full-length antibody. Examples of antigen-binding portions encompassed within the term “antibody fragments” of an antibody include a Fab fragment, a monovalent fragment consisting of the V_L, V_H, C_L and C_{H1} domains; a F(ab)₂ fragment, a bivalent fragment comprising two Fab fragments linked by a disulfide bridge at the hinge region; a Fd fragment consisting of the V_H and C_{H1} domains; a Fv fragment consisting of the V_L and V_H domains of a single arm of an antibody; a dAb fragment (Ward *et al.*, 1989 Nature 341:544-546), which consists of a V_H domain; and an isolated complementarity determining region (CDR).

[00160] As used herein, the term “single-chain variable fragment” or “scFv” is a fusion protein of the variable regions of the heavy (V_H) and light chains (V_L) of an immunoglobulin (*e.g.*, mouse or human) covalently linked to form a V_H:V_L heterodimer. The heavy (V_H) and light chains (V_L) are either joined directly or joined by a peptide-encoding linker (*e.g.*, 10, 15, 20, 25 amino acids), which connects the N-terminus of the V_H with the C-terminus of the V_L, or the C-terminus of the V_H with the N-terminus of the V_L. The linker is usually rich in glycine for flexibility, as well as serine or threonine for solubility. Despite removal of the constant regions and the introduction of a linker, scFv proteins retain the specificity of the original immunoglobulin. Single chain Fv polypeptide antibodies can be expressed from a nucleic acid comprising V_H - and V_L -encoding sequences as described by Huston *et al.* (Proc. Nat. Acad. Sci. USA, 85:5879-5883, 1988). See, also, U.S. Patent Nos. 5,091,513,

5,132,405 and 4,956,778; and U.S. Patent Publication Nos. 20050196754 and 20050196754. Antagonistic scFvs having inhibitory activity have been described (*see, e.g., Zhao et al. Hybridoma (Larchmt) 2008 27(6):455-51; Peter et al., J Cachexia Sarcopenia Muscle 2012 August 12; Shieh et al., J Immunol 2009 183(4):2277-85; Giomarelli et al., Thromb Haemost 2007 97(6):955-63; Fife et al., J Clin Invest 2006 116(8):2252-61; Brocks et al., Immunotechnology 1997 3(3):173-84; Mooscaner et al., Ther Immunol 1995 2(10):31-40*). Agonistic scFvs having stimulatory activity have been described (*see, e.g., Peter et al., J Bio Chem 2003 25278(38):36740-7; Xie et al., Nat Biotech 1997 15(8):768-71; Ledbetter et al., Crit Rev Immunol 1997 17(5-6):427-55; Ho et al., Biochim Biophys Acta 2003 1638(3):257-66*).

[00161] As used herein, “F(ab)” refers to a fragment of an antibody structure that binds to an antigen but is monovalent and does not have a Fc portion, for example, an antibody digested by the enzyme papain yields two F(ab) fragments and an Fc fragment (e.g., a heavy (H) chain constant region; Fc region that does not bind to an antigen).

[00162] As used herein, “F(ab')₂” refers to an antibody fragment generated by pepsin digestion of whole IgG antibodies, wherein this fragment has two antigen binding (ab') (bivalent) regions, wherein each (ab') region comprises two separate amino acid chains, a part of a H chain and a light (L) chain linked by an S-S bond for binding an antigen and where the remaining H chain portions are linked together. A “F(ab')₂” fragment can be split into two individual Fab' fragments.

[00163] As used herein, the term “antigen-binding protein” refers to a protein or polypeptide that comprises an antigen-binding region or antigen-binding portion, that is, has a strong affinity to another molecule to which it binds. Antigen-binding proteins encompass antibodies, chimeric antigen receptors (CARs) and fusion proteins.

[00164] As used herein, the term “treating” or “treatment” refers to clinical intervention in an attempt to alter the disease course of the individual or cell being treated, and can be performed either for prophylaxis or during the course of clinical pathology. Therapeutic effects of treatment include, without limitation, preventing occurrence or recurrence of disease, alleviation of symptoms, diminishment of any direct or indirect pathological consequences of the disease, preventing metastases, decreasing the rate of disease progression, amelioration or palliation of the disease state, and remission or improved prognosis. By preventing progression of a disease or disorder, a treatment can prevent

deterioration due to a disorder in an affected or diagnosed subject or a subject suspected of having the disorder, but also a treatment may prevent the onset of the disorder or a symptom of the disorder in a subject at risk for the disorder or suspected of having the disorder.

[00165] As used herein, the term “subject” refers to any animal (*e.g.*, a mammal), including, but not limited to, humans, and non-human animals (including, but not limited to, non-human primates, dogs, cats, rodents, horses, cows, pigs, mice, rats, hamsters, rabbits, and the like (*e.g.*, which is to be the recipient of a particular treatment, or from whom cells are harvested). In certain embodiments, the subject is a human.

[00166] As used herein, an “effective amount” or “therapeutically effective amount” is an amount sufficient to affect a beneficial or desired clinical result upon treatment. An effective amount can be administered to a subject in one or more doses. In terms of treatment, an effective amount is an amount that is sufficient to palliate, ameliorate, stabilize, reverse or slow the progression of the disease, or otherwise reduce the pathological consequences of the disease. The effective amount is generally determined by the physician on a case-by-case basis and is within the skill of one in the art. Several factors are typically taken into account when determining an appropriate dosage to achieve an effective amount. These factors include age, sex and weight of the subject, the condition being treated, the severity of the condition and the form and effective concentration of the immunoresponsive cells administered.

[00167] As used herein, the term “response” or “responsiveness” refers to an alteration in a subject's condition that occurs as a result of or correlates with treatment. In certain embodiments, a response is a beneficial response. In certain embodiments, a beneficial response can include stabilization of the condition (*e.g.*, prevention or delay of deterioration expected or typically observed to occur absent the treatment), amelioration (*e.g.*, reduction in frequency and/or intensity) of one or more symptoms of the condition, and/or improvement in the prospects for cure of the condition, etc. In certain embodiments, “response” can refer to response of an organism, an organ, a tissue, a cell, or a cell component or *in vitro* system. In certain embodiments, a response is a clinical response. In certain embodiments, presence, extent, and/or nature of response can be measured and/or characterized according to particular criteria. In certain embodiments, such criteria can include clinical criteria and/or objective criteria. In certain embodiments, techniques for assessing response can include, but are not limited to, clinical examination, positron emission tomography, chest X-ray CT scan, MRI, ultrasound, endoscopy, laparoscopy, presence or level of a particular marker in a

sample, cytology, and/or histology. Where a response of interest is a response of a tumor to a therapy, ones skilled in the art will be aware of a variety of established techniques for assessing such response, including, for example, for determining tumor burden, tumor size, tumor stage, *etc.* The likelihood of a subject having predictive features identified herein (*e.g.*, one or more MUC16 neoantigen, number of neoantigens, number of activated T cells) exhibiting a particular response is relative to a similarly situated second subject (*e.g.*, a second subject having the same type cancer, optionally the same type of cancer with similar characteristics (*e.g.* stage and/or location/distribution)) or group of subjects that lack one or more predictive feature.

[00168] As used herein, the term “sample” refers to a biological sample obtained or derived from a source of interest, as described herein. In certain embodiments, a source of interest comprises an organism, such as an animal or human. In certain embodiments, a biological sample is a biological tissue or fluid. Non-limiting biological samples include bone marrow, blood, blood cells, ascites, (tissue or fine needle) biopsy samples, cell-containing body fluids, free floating nucleic acids, sputum, saliva, urine, cerebrospinal fluid, peritoneal fluid, pleural fluid, feces, lymph, gynecological fluids, swabs (*e.g.*, skin swabs, vaginal swabs, oral swabs, and nasal swabs), washings or lavages such as a ductal lavages or bronchoalveolar lavages, aspirates, scrapings, specimens (*e.g.*, bone marrow specimens, tissue biopsy specimens, and surgical specimens), feces, other body fluids, secretions, and/or excretions, and cells therefrom, *etc.*

[00169] As used herein, the term “substantially” refers to the qualitative condition of exhibiting total or near-total extent or degree of a characteristic or property of interest. One of ordinary skill in the art will understand that biological and chemical phenomena rarely, if ever, go to completion and/or proceed to completeness or achieve or avoid an absolute result. The term “substantially” is therefore used herein to capture the potential lack of completeness inherent in many biological and chemical phenomena.

[00170] As used herein, the term “vaccine” refers to a composition for generating immunity for the prophylaxis and/or treatment of diseases (*e.g.*, neoplasia/tumor). In certain embodiments, vaccines are medicaments that comprise antigens and are intended to be used in humans or animals for generating specific defense and protective substance by vaccination.

[00171] It will also be understood that, although the terms first, second, *etc.* may be used herein to describe various elements, these elements should not be limited by these terms.

These terms are only used to distinguish one element from another. For example, a first subject could be termed a second subject, and, similarly, a second subject could be termed a first subject, without departing from the scope of the present disclosure. The first subject and the second subject are both subjects, but they are not the same subject. Furthermore, the terms “subject,” “user,” and “patient” are used interchangeably herein.

[00172] The terminology used in the present disclosure is for the purpose of describing particular embodiments only and is not intended to be limiting of the invention. As used in the description of the invention and the appended claims, the singular forms “a,” “an” and “the” are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will also be understood that the term “and/or” as used herein refers to and encompasses any and all possible combinations of one or more of the associated listed items. It will be further understood that the terms “comprises” and/or “comprising,” when used in this specification, specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof.

[00173] As used herein, the term “if” may be construed to mean “when” or “upon” or “in response to determining” or “in response to detecting,” depending on the context. Similarly, the phrase “if it is determined” or “if [a stated condition or event] is detected” may be construed to mean “upon determining” or “in response to determining” or “upon detecting [the stated condition or event]” or “in response to detecting [the stated condition or event],” depending on the context.

[00174] 2. *Neoantigens*

[00175] The presently disclosed subject matter provides identification of neoantigens in subjects having cancer, e.g., pancreatic cancer (e.g., pancreatic ductal adenocarcinoma (PDAC)).

[00176] 2.1 *Neoantigens*

[00177] Neoantigen is a tumor-specific antigen that arises from one or more tumor-specific mutation. In certain embodiments, the neoantigen is a tumor-specific antigen that arises from a tumor-specific mutation. In certain embodiments, a neoantigen is not expressed by healthy cells (e.g., non-tumor cells or non-cancer cells) in a subject.

[00178] Many antigens expressed by cancer cells are self-antigens which are selectively expressed or overexpressed on the cancer cells. These self-antigens are difficult

to target with immunotherapy because they require overcoming both central tolerance (whereby autoreactive T cells are deleted in the thymus during development) and peripheral tolerance (whereby mature T cells are suppressed by regulatory mechanisms). Targeting neoantigens can abrogate these tolerance mechanisms. In certain embodiments, a neoantigen is recognized by cells of the immune system of a subject (*e.g.*, T cells) as “non-self”. Neoantigens are not recognized as “self-antigens” by immune system, T cells that are capable of targeting neoantigens are not subject to central and peripheral tolerance mechanisms to the same extent as T cells which recognize self-antigens.

[00179] In certain embodiments, the tumor-specific mutation that results in a neoantigen is a somatic mutation. Somatic mutations comprise DNA alterations in non-germline cells and commonly occur in cancer cells. Certain somatic mutations in cancer cells result in the expression of neoantigens, that in certain embodiments, transform a stretch of amino acids from being recognized as “self” to “non-self”. Human tumors without a viral etiology can accumulate tens to hundreds of fold somatic mutations in tumor genes during neoplastic transformation, and some of these somatic mutations can occur in protein-coding regions and result in the formation of neoantigens. The exome is the protein-encoding part of the genome. Based on the mutations present within the protein-encoding part of the genome (*i.e.*, the exome) of an individual tumor, potential neoantigens can be predicted for the individual tumor. In certain embodiments, whole exome sequencing is performed in a biological sample (*e.g.*, a tumor sample) obtained from a subject having cancer.

[00180] In certain embodiments, a neoantigen is a neoantigenic peptide comprising a tumor specific mutation. The neoantigenic peptide can be a peptide that is incorporated into a larger protein. In certain embodiments, a neoantigenic peptide is a series of residues, typically L-amino acids, connected one to the other, typically by peptide bonds between the α -amino and carboxyl groups of adjacent amino acids. A neoantigenic peptide can be a variety of lengths, either in their neutral (uncharged) forms or in forms which are salts, and either free of modifications such as glycosylation, side chain oxidation, or phosphorylation.

[00181] In certain embodiments, the size of the neoantigenic peptide is about 3-30 amino acids, *e.g.*, about 3-5, about 5-15 (*e.g.*, about 8-11, about 5-10, or about 10-15), about 15-20, about 20-25, or about 20-30 amino acids, in length. In certain embodiments, the neoantigenic peptide is about 8-11 amino acids in length. In certain embodiments, the neoantigenic peptide is about 3, about 4, about 5, about 6, about 7, about 8, about 9, about 10, about 11, about 12, about 13, about 14, about 15, about 16, about 17, about 18, about 19,

about 20, about 21, about 22, about 23, about 24, about 25, about 26, about 27, about 28, about 29, or about 30 amino acids in length. In certain embodiments, the neoantigenic peptide is at least about 3, at least about 5, or at least about 8 amino acids in length. In certain embodiments, the neoantigenic peptide is less than about 30, less than about 20, less than about 15, or less than about 10 amino acids in length.

[00182] In certain embodiments, a neoantigenic peptide is selected from the neoantigenic peptides listed in Table 1.

TABLE 1		
NEOANTIGENIC PEPTIDE^A	GENE THE NEOANTIGENIC PEPTIDE IS LOCATED	MUTANT ALLELE^B IN MUC16
vmKhllspl (SEQ ID NO: 1)	MUC16	B0801
mHhpgrskf (SEQ ID NO: 2)	MUC16	C0701
nstLiptli (SEQ ID NO: 3)	MUC16	A6802
sssgvnstI (SEQ ID NO: 4)	MUC16	A6802
stLiptliI (SEQ ID NO: 5)	MUC16	A6802
aeassAvpt (SEQ ID NO: 6)	MUC16	B4501
eassAvptv (SEQ ID NO: 7)	MUC16	C1203
sravtsttI (SEQ ID NO: 8)	MUC16	C0602
sttipiltL (SEQ ID NO: 9)	MUC16	C1402
tipiltLsl (SEQ ID NO: 10)	MUC16	C1402
gLrktnmsl (SEQ ID NO: 11)	MUC16	B0702
smpanfetI (SEQ ID NO: 12)	MUC16	A2402
stvrKspwk (SEQ ID NO: 13)	MUC16	A0301
stvrKspwk (SEQ ID NO: 14)	MUC16	A1101
mgKsthtsm (SEQ ID NO: 15)	MUC16	B0801
smkAerppa (SEQ ID NO: 16)	MUC16	B0801
ttspSNtlv (SEQ ID NO: 17)	MUC16	C0701
sssptCslm (SEQ ID NO: 18)	MUC16	C1502

^a the capitalized letter represents the changed amino acid between the WT and mutated sequences

^b “MT allele” represents the HLA allele predicted to bind the mutated neopeptide

[00183] In certain embodiments, a neoantigenic peptide is selected from the neoantigenic peptides listed in Table 2.

TABLE 2

NEOANTIGENIC PEPTIDE	GENE THE NEOANTIGENIC PEPTIDE IS LOCATED	ALIGNMENT SCORE	S-TAU ^c
APLGAPPPL (SEQ ID NO: 19)	FBR3	29	0.690625
ASLHHHHHR (SEQ ID NO: 20)	CACNA1B	24	0.987118
ATYHFHFNL (SEQ ID NO: 21)	COL6A5	25	0.955964
DWPVFPGFL (SEQ ID NO: 22)	HDAC3	27	0.772377
EAFTLKATV (SEQ ID NO: 23)	PZP	26	0.920519
EAHHHFPSL (SEQ ID NO: 24)	TTLL3	29	0.843503
FLNRWMANT (SEQ ID NO: 25)	AADACL2	28	0.920174
GIVSWDTFL (SEQ ID NO: 26)	ELMO1	28	0.646096
GTPRAATMK (SEQ ID NO: 27)	PRKAR1B	18	0.999997
HWPEKEWPI (SEQ ID NO: 28)	N/A	22	0.982173
ILFDEAVKL (SEQ ID NO: 29)	IQCA1	20	0.973993
ILIAICRLNK (SEQ ID NO: 30)	LRPPRC	21	0.679058
ILPTCSPLV (SEQ ID NO: 31)	LRP4	27	0.419822
KPRFLVGLW (SEQ ID NO: 32)	IGFN1	27	0.872717
KYIAFCINI (SEQ ID NO: 33)	S1PR3	26	0.691118
LFHQCLSIY (SEQ ID NO: 34)	C5orf42	31	0.006421
LLLMSTLGI (SEQ ID NO: 35)	KLRD1	32	0.565677
LLPPQDPHL (SEQ ID NO: 36)	PCDHB15	24	0.679058
LQDFYLGTY (SEQ ID NO: 37)	SLC15A5	29	0.652087
LTPPQAQEL (SEQ ID NO: 38)	#N/A	25	0.811812
QTYQHMYNY (SEQ ID NO: 39)	GRIK5	34	0.973993
REFKFRVSA (SEQ ID NO: 40)	PCDHB16	18	0.99305
STGFPHMLF (SEQ ID NO: 41)	CSNK2B- LY6G5B-1181	22	0.953153
TLVGHQGPV (SEQ ID NO: 42)	TRAF7	21	0.459383
VDWFLDWLR (SEQ ID NO: 43)	SLC1A6	28	0.811812
GGAPHFGHF (SEQ ID NO: 44)	ZNRF1	17	0.904809
CYYELNQCL (SEQ ID NO: 45)	ZNF430	26	0.48456
AYPQYVIEY (SEQ ID NO: 46)	ZC3HAV1	26	0.994797
HLETHNTDK (SEQ ID NO: 47)	ZBTB17	24	0.964887
AEVEEEVV (SEQ ID NO: 48)	WWC1	26	0.806188
RGMQCAICK (SEQ ID NO: 49)	WDR59	26	0.833236
MPEDYMYVY (SEQ ID NO: 50)	VTN	17	0.946362

TABLE 2			
NEOANTIGENIC PEPTIDE	GENE THE NEOANTIGENIC PEPTIDE IS LOCATED	ALIGNMENT SCORE	S-TAU^C
SSYGRNHYY (SEQ ID NO: 51)	VCPIPI	25	0.976041
AMDDLDTDM (SEQ ID NO: 52)	UTRN	27	0.45028
PTDPMGLA (SEQ ID NO: 53)	TSPAN10	24	0.935737
FSSNLPTY (SEQ ID NO: 54)	TPTE	27	0.571614
FSSNLPTY (SEQ ID NO: 55)	TPTE	27	0.571614
FRHSMVVPY (SEQ ID NO: 56)	TP53	23	0.920174
NLLGRNSFK (SEQ ID NO: 57)	TP53	36	0.771043
GLGFYNDVV (SEQ ID NO: 58)	TNFRSF4	23	0.45028
AQTHEPRQW (SEQ ID NO: 59)	TMEM132C	26	0.887438
STHPSLSQW (SEQ ID NO: 60)	TGOLN2	27	0.677777
VYMPPRLL (SEQ ID NO: 61)	TENM2	23	0.902688
NHDDDDVEI (SEQ ID NO: 62)	TADA2B	32	0.18872
YVKIYLLPY (SEQ ID NO: 63)	SYT9	28	0.904809
SYQSTGDPK (SEQ ID NO: 64)	SVIL	24	0.732722
RPRKAWAWC (SEQ ID NO: 65)	SUOX	26	0.679058
WVLHHMGGM (SEQ ID NO: 66)	SRP54	30	0.18872
LAGEWRL (SEQ ID NO: 67)	SPTBN2	16	0.646096
TVWPSLAPL (SEQ ID NO: 68)	SPEG	23	0.955964
QEASNKHAE (SEQ ID NO: 69)	SMC2	27	0.970774
RRRLCILRM (SEQ ID NO: 70)	SMAD4	22	0.872717
VRLGPVKSI (SEQ ID NO: 71)	SETX	28	0.964887
FFVEKRHAF (SEQ ID NO: 72)	SCAND3	24	0.994797
HTSLRGFLY (SEQ ID NO: 73)	SBK2	26	0.419822
SLAETKTLY (SEQ ID NO: 74)	SARDH	17	0.99983
TYAPLFIWV (SEQ ID NO: 75)	RXFP1	27	0.691118
SPAPERCMV (SEQ ID NO: 76)	RTL1	26	0.872717
EQLKLGAIF (SEQ ID NO: 77)	RNGTT	23	0.907671
RPQGQRPAL (SEQ ID NO: 78)	RNF5	26	0.771676
APRGVCYGA (SEQ ID NO: 79)	RETSAT	24	0.973993
APRGVCYGA (SEQ ID NO: 80)	RETSAT	24	0.772377
LAAPRGVCY (SEQ ID NO: 81)	RETSAT	24	0.955964
PPRYIGIPI (SEQ ID NO: 82)	RAPGEF2	25	0.48456

TABLE 2			
NEOANTIGENIC PEPTIDE	GENE THE NEOANTIGENIC PEPTIDE IS LOCATED	ALIGNMENT SCORE	S-TAU^C
HVWLCDLPV (SEQ ID NO: 83)	RABGGTA	30	0.772377
QLYMNPKTW (SEQ ID NO: 84)	PYGL	31	0.946362
KYSNYVWPI (SEQ ID NO: 85)	PRSS50	26	0.419822
LPRQYWEAL (SEQ ID NO: 86)	POLRMT	27	0.771676
VLNGWLRVS (SEQ ID NO: 87)	PNPLA6	28	0.780199
YLALAAQCL (SEQ ID NO: 88)	PLEKHH3	28	0.48456
CPLPRPPPI (SEQ ID NO: 89)	PCDH15	28	0.970774
GIICLDYKL (SEQ ID NO: 90)	OR2A14	24	0.998936
TLGVFCLGL (SEQ ID NO: 91)	OLR1	32	0.909408
ERPCHREPL (SEQ ID NO: 92)	OBSCN	24	0.971063
RLALSTFEW (SEQ ID NO: 93)	NLRP12	26	0.403676
VVWATKYFL (SEQ ID NO: 94)	NCEH1	25	0.920174
NPEAMCSDL (SEQ ID NO: 95)	MYRIP	25	0.679058
I PLEVMEPF (SEQ ID NO: 96)	MORC1	26	0.728968
EEAFVPILY (SEQ ID NO: 97)	MLK4	24	0.887438
HHHHHHQAW (SEQ ID NO: 98)	MAFB	28	0.459383
LVWSLPCGF (SEQ ID NO: 99)	LBR	29	0.45028
MPDVVHQSL (SEQ ID NO: 100)	L3MBTL1	30	0.971063
CRPQCCQSV (SEQ ID NO: 101)	KRTAP4-7	28	0.677777
MTPSVYGGA (SEQ ID NO: 102)	KRT20	25	0.982173
YKLVVVGAV (SEQ ID NO: 103)	KRAS	24	0.690625
YKLVVVGAV (SEQ ID NO: 104)	KRAS	24	0.891553
YKLVVVGAV (SEQ ID NO: 105)	KRAS	24	0.811812
APAQPPMLA (SEQ ID NO: 106)	KMT2D	24	0.732722
EPPPPPSPL (SEQ ID NO: 107)	KMT2D	22	0.946362
YLWEDPVCV (SEQ ID NO: 108)	KIF26A	25	0.780199
SFADFVWHF (SEQ ID NO: 109)	KIF22	26	0.909408
YQQSNTWSL (SEQ ID NO: 110)	KIAA1407	30	0.891553
SPGGWRSGW (SEQ ID NO: 111)	ITPKB	32	0.403676
RVWDIVPTL (SEQ ID NO: 112)	ILVBL	37	0.945017
MLAIGCALL (SEQ ID NO: 113)	IL6R	29	0.652087
WEEEYTVWI (SEQ ID NO: 114)	IFFO1	29	0.893521

TABLE 2			
NEOANTIGENIC PEPTIDE	GENE THE NEOANTIGENIC PEPTIDE IS LOCATED	ALIGNMENT SCORE	S-TAU^c
VWPKKINNI (SEQ ID NO: 115)	HYDIN	24	0.987118
WPQCHPEEI (SEQ ID NO: 116)	HN1	36	0.971063
QEFENIKSY (SEQ ID NO: 117)	HIVEP1	23	0.708595
TMDVATPSV (SEQ ID NO: 118)	HERC2	26	0.853911
LPLHLYDTL (SEQ ID NO: 119)	HECTD1	23	0.893521
RPGQSPGQL (SEQ ID NO: 120)	HAND1	29	0.780199
ELLDYIRAV (SEQ ID NO: 121)	GRM8	25	0.833236
LPPSLQGAV (SEQ ID NO: 122)	GPR179	26	0.970774
FLTQPVAPK (SEQ ID NO: 123)	FGD3	24	0.806188
LASSCGCTF (SEQ ID NO: 124)	FCGBP	24	0.843503
RPRGDNGYT (SEQ ID NO: 125)	FBN1	19	0.987118
QAMYDVLTF (SEQ ID NO: 126)	FAT3	29	0.45028
LTISGECPK (SEQ ID NO: 127)	FAM83H	21	0.45028
SWKSPGWSF (SEQ ID NO: 128)	FAM124B	26	0.902688
RKREEEERW (SEQ ID NO: 129)	EFHD1	27	0.691118
NHLCFGHCF (SEQ ID NO: 130)	DAND5	28	0.907671
YVYSLYWSI (SEQ ID NO: 131)	CNGA1	29	0.000624
MVLWHLPAV (SEQ ID NO: 132)	CLCF1	29	0.18872
FSVSPEWAV (SEQ ID NO: 133)	CDHR2	22	0.891553
GYTLLNVF (SEQ ID NO: 134)	CACNA1A	26	0.883775
QALIRPTTF (SEQ ID NO: 135)	C18orf8	27	0.976041
KQLPRILEA (SEQ ID NO: 136)	C17orf100	18	0.677777
YQQALGKRF (SEQ ID NO: 137)	C16orf78	21	0.920519
QLAWVPSPY (SEQ ID NO: 138)	AUTS2	27	0.953153
HIQDLYTVL (SEQ ID NO: 139)	ATP6V0A2	28	0.883775
MNRGRRSSL (SEQ ID NO: 140)	ARID4A	25	0.982173
YAYTFWTYI (SEQ ID NO: 141)	APEX1	28	0.888813
SEVLGYWAF (SEQ ID NO: 142)	ADRA1A	27	0.99305
KPLLSGPWA (SEQ ID NO: 143)	ADAMTS5	27	0.780199
FNGNFLLSM (SEQ ID NO: 144)	ADAMTS20	22	0.907671
AHPDGSWTF (SEQ ID NO: 145)	ABT1	30	0.972244

^c this parameter is a metric of the whole tumor, not just the neoantigen

[00184] Neoantigens may vary in different subjects, *e.g.*, different subjects may have different combinations of neoantigens, also referred to as “neoantigen signatures”. For example, each subject may have a unique neoantigen signature. In certain embodiments, a neoantigen signature is a combination of one or more neoantigen listed in Tables 1 and 2.

[00185] The presently disclosed subject matter also provides compositions comprising one or more presently disclosed neoantigen. In certain embodiments, the compositions are pharmaceutical compositions comprising pharmaceutically acceptable carriers. In certain embodiments, the composition comprises two or more neoantigens. The two or more can be linked, *e.g.*, by any biochemical strategy to link two proteins.

[00186] 2.2 *Detection of Neoantigens*

[00187] Cancers can be screened to detect neoantigens using any of a variety of known technologies. In certain embodiments, neoantigens or expression thereof is detected at the nucleic acid level (*e.g.*, in DNA or RNA). In certain embodiments, neoantigens or expression thereof is detected at the protein level (*e.g.*, in a sample comprising polypeptides from cancer cells, which sample can be or comprise polypeptide complexes or other higher order structures including but not limited to cells, tissues, or organs).

[00188] In certain embodiments, a neoantigen is detected by the method selected from the group consisting of whole exome sequencing, immunoassay, microarray, genome sequencing, RNA sequencing, ELISA, Western Transfer, DNA or RNA sequencing, mass spectrometry, and combinations thereof. In certain embodiments, one or more neoantigen is detected by whole exome sequencing. In certain embodiments, one or more neoantigen is detected by the immunogenicity analysis method of somatic mutations described in Snyder et al. *Engl J Med* 371, 2189–2199 (2014). In certain embodiments, one or more neoantigen is detected by the pVAC-Seq method described in Hndal *et al.*, *Genome Med* (2016);8:11. In certain embodiments, one or more neoantigen is detected by the *in silico* neoantigen prediction pipeline method described in Rizvi et al., *Science* 348, 124–128 (2015). In certain embodiments, one or more neoantigen is detected by any of the methods described in WO2015/103037 and WO2016/081947.

[00189] 3. *Uses of Neoantigens for Patient Selection for and Responsiveness Prediction to Immunotherapies*

[00190] The neoantigens of the presently disclosed subject matter can be used to identify cancer subjects as candidates for immunotherapies, and to predict responsiveness of

cancer subjects to immunotherapies. In certain embodiments, the presently disclosed subject matter provides methods of identifying subjects (e.g., cancer subjects) as candidates for treatment with an immunotherapy (hereinafter “the patient selection method”). In certain embodiments, the presently disclosed subject matter provides methods of predicting the responsiveness of subjects (e.g., cancer subjects) to an immunotherapy (hereinafter “the responsiveness prediction method”).

[00191] 3.1 *MUC16 as Immunogenic Hotspot*

[00192] As used herein, the term “immunogenic hotspot” refers to a genetic locus that is enriched with neoantigens (e.g., a genetic locus that frequently generates neoantigens). The immunogenic hotspot can vary depending on the type of disease (e.g., tumor). Detecting one or more neoantigen of an immunogenic hotspot can be used to identify cancer subjects as candidates for immunotherapies, and to predict responsiveness of cancer subjects to immunotherapies. Furthermore, immunotherapies (e.g., vaccines, T cells (including modified T cells, e.g., T cells comprising a T cell receptor (TCR) or a chimeric antigen receptor (CAR)) targeting one or more neoantigen of an immunogenic hotspot can be used for neoantigen-directed therapies, e.g., cancer therapies. In certain embodiments, MUC16 is an immunogenic hotspot for pancreatic cancer, more specifically, for pancreatic ductal adenocarcinoma (PDAC). In certain embodiments, MUC16 is an immunogenic hotspot for checkpoint blockade refractory tumor (e.g., PDAC).

[00193] *MUC16*-neoantigen specific T cell immunity induces immunoediting of *MUC16*-expressing clones in primary tumors, curtails the development of metastases, and prolongs survival, given the cell-autonomous roles of MUC16 in promoting metastases.

[00194] In certain non-limiting embodiments, the patient selection method comprises obtaining a biological sample from the subject, and detecting one or more neoantigen of MUC16 in the biological sample, wherein the presence of one or more neoantigen of MUC16 in the biological sample indicates that the subject is a candidate for an immunotherapy.

[00195] In certain non-limiting embodiments, the responsiveness prediction method comprises obtaining a biological sample from the subject, and detecting one or more neoantigen of MUC16 in the biological sample, wherein the presence of one or more neoantigen of MUC16 in the biological sample indicates that the subject is likely to be responsive to an immunotherapy.

[00196] In certain embodiments, the patient selection method and responsiveness prediction method comprise detecting a plurality of neoantigens of MUC16. In certain embodiments, the methods comprise detecting one or more of the neoantigenic peptides listed in Table 1. In certain embodiments, the methods comprise detecting one or more of the neoantigenic peptides having the amino acid sequences set forth in SEQ ID NOS: 1-18.

[00197] In certain embodiments, the subject has pancreatic cancer. In certain embodiments, the pancreatic cancer is pancreatic ductal adenocarcinoma (PDAC).

[00198] In some embodiments, neoantigens of MUC16 are detected by any of the methods for detecting neoantigens, e.g., those described in Section 2.2. In certain embodiments, the neoantigens of MUC16 are detected by whole exome sequencing the biological sample. In certain embodiments, the neoantigens of MUC16 are detected by the immunogenicity analysis method of somatic mutations described in Snyder et al. *Engl J Med* 371, 2189–2199 (2014). In certain embodiments, the neoantigens of MUC16 are detected by the pVAC-Seq method described in Hndal et al., *Genome Med* (2016);8:11. In certain embodiments, the neoantigens of MUC16 are detected by the *in silico* neoantigen prediction pipeline described in Rizvi et al., 2015, *Science* 348, 124–128. In certain embodiments, the neoantigens of MUC16 are detected by any of the methods described in WO2015/103037 and WO2016/081947.

[00199] 3.2 *Neoantigen Quantity and Neoantigen Immunogenicity*

[00200] In certain embodiments, the patient selection method and responsiveness prediction method relate to the quantity of the neoantigens (e.g., the number of neoantigens (hereinafter “neoantigen number”) in the subject. In certain embodiments, the neoantigen number includes but is not limited to MUC16 neoantigens.

[00201] In some embodiments, neoantigen number is measured by any methods for detecting neoantigens, e.g., those described in Section 2.2. In certain embodiments, the neoantigen number is measured by whole exome sequencing the biological sample. In certain embodiments, the neoantigen number is measured by the immunogenicity analysis method of somatic mutations described in Snyder et al., 2014, *Engl J Med* 371, 2189–2199. In certain embodiments, the neoantigen number is measured by the pVAC-Seq method described in Hndal et al., 2016, *Genome Med* 8, 11. In certain embodiments, the neoantigen number is measured by the *in silico* neoantigen prediction pipeline described in Rizvi et al.,

2015 *Science* 348, 124–128. In certain embodiments, the neoantigen number is measured by any of the methods described in WO2015/103037 and WO2016/081947.

[00202] The neoantigen number obtained from a population of subjects having the cancer may vary depending on the type of the cancer, therapy (*e.g.*, immunotherapy, chemotherapy), the cancer histology, and/or the site of disease (tumor). In certain embodiments, the median neoantigen number obtained from a population of subjects having cancer is between 10 and 10,000, between 10 and 50, between 30 and 50, between 30 and 40, between 50 and 100, between 100 and 200, between 200 and 300, between 300 and 400, between 400 and 500, between 500 and 600, between 600 and 700, between 700 and 800, between 800 and 900, between 900 and 1,000, between 1,000 and 2,000, between 2,000 and 3,000, between 3,000 and 4,000, between 4,000 and 5,000, between 5,000 and 6,000, between 6,000 and 7,000, between 7,000 and 8,000, between 8,000 and 9,000, or between 9,000 and 10,000. In certain embodiments, the median neoantigen number obtained from a population of subjects having cancer is about 40. In certain embodiments, the median neoantigen number obtained from a population of subjects having cancer is about 38. In certain embodiments, the cancer is pancreatic cancer, *e.g.*, PDAC.

[00203] The neoantigens can be located in any genes in a subject. In certain embodiments, the neoantigen is located in MUC16. In certain embodiments, the neoantigen is located in any of the genes listed in Table 2.

[00204] In certain embodiments, the patient selection method and responsiveness prediction method relate to the quantity of the neoantigens and the immunogenicity of the neoantigens (hereinafter “neoantigen immunogenicity”) in the subject. In certain embodiments, the patient selection method and responsiveness prediction method comprise assessing the neoantigen immunogenicity. Immunogenicity is the ability of a particular substance, such as an antigen (*e.g.*, a neoantigen), to induce or stimulate an immune response in the cells expressing such antigen. In certain embodiments, assessing the neoantigen immunogenicity comprises measuring one or more surrogate for the neoantigen immunogenicity.

[00205] In certain embodiments, the surrogate is the homology between a neoantigen and a microbial epitope (hereinafter “neoantigen-microbial homology”).

[00206] In certain non-limiting embodiments, the patient selection method comprises obtaining a biological sample from the subject, measuring the neoantigen number in the

biological sample; and measuring the homology between each of the neoantigen and a microbial epitope (neoantigen-microbial homology), where a neoantigen number higher than the median neoantigen number obtained from a population of subjects having the cancer (*e.g.*, referred to as “neoantigen number^{high}”) and a neoantigen-microbial homology higher than the median neoantigen-microbial homology obtained from subjects having the cancer (*e.g.*, referred to as “neoantigen-microbial homology^{high}”) indicate that the subject is a candidate for an immunotherapy.

[00207] In certain non-limiting embodiments, the responsiveness prediction method comprises obtaining a biological sample from the subject, measuring the neoantigen number in the biological sample, and measuring the neoantigen-microbial homology, where a neoantigen number higher than the median neoantigen number obtained from a population of subjects having the cancer (*e.g.*, referred to as “neoantigen number^{high}”) and a neoantigen-microbial homology higher than the median neoantigen-microbial homology obtained from subjects having the cancer (*e.g.*, referred to as “neoantigen-microbial homology^{high}”) indicate that the subject is likely to be responsive to an immunotherapy.

[00208] In certain embodiments, the surrogate is the number of activated T cells (hereinafter “activated T cell number”). The immune system plays an important role in controlling and eradicating cancer. The number and frequency of the activated T cells present in the tumor can affect the fraction of tumor cells eventually killed *in vitro*. Therefore, the baseline infiltration by activated T cells in the tumor can have important prognostic implications for solid tumors.

[00209] In certain non-limiting embodiments, the subject’s neoantigen number, the neoantigen-microbial homology, and activated T cell numbers are at least about 1%, at least about 2%, at least about 3%, at least about 4%, at least about 5%, at least about 6%, at least about 7%, at least about 8%, at least about 9%, at least about 10%, or at least about 20%, or at least about 30%, higher than the median values.

[00210] In certain embodiments, the one or more neoantigen is selected from the neoantigenic peptides listed in Tables 1 and 2.

[00211] In certain embodiments, activated T cells are T cells expressing one or more T cell activation marker. Non-limiting examples of T cell activation markers include CD3, CD8, PD-1, 4-1BB, CD69, CD107a, and Granzyme B. In certain embodiments, the activated

T cells are selected from the group consisting of CD3⁺CD8⁺ T cells, CD3⁺CD8⁺Granzyme-B⁺ T cells, polyclonal activated T cells, and combinations thereof.

[00212] In certain non-limiting embodiments, the patient selection method comprises measuring the neoantigen number in a biological sample of the subject, and measuring the number of activated T cells (hereinafter “activated T cell number”) in a biological sample of the subject. In such embodiments, a neoantigen number higher than the median neoantigen number obtained from a population of subjects having the cancer (*e.g.*, referred to as “neoantigen number^{high}”) and an activated T cell number higher than the median activated T cell number obtained from subjects having the cancer (*e.g.*, referred to as “activated T cell number^{high}”) indicate that the subject is a candidate for an immunotherapy.

[00213] In certain non-limiting embodiments, the responsiveness prediction method comprises measuring the neoantigen number in a biological sample of the subject; and measuring the activated T cell number a biological sample of the subject. In such embodiments, a neoantigen number higher than the median neoantigen number obtained from a population of subjects having the cancer (*e.g.*, referred to as “neoantigen number^{high}”) and an activated T cell number higher than the median activated T cell number obtained from subjects having the cancer (*e.g.*, referred to as “activated T cell number^{high}”) indicate that the subject is likely to be responsive to an immunotherapy.

[00214] The patient selection method and responsiveness prediction method further comprise obtaining one or more biological sample from the subject. The biological sample used to measure the neoantigen number, the biological sample used to measure the neoantigen-microbial homology, and the biological sample used to measure the activated T cell numbers can be the same or different. In certain embodiments, the biological sample can be a tumor tissue. In certain embodiments, the biological sample is a blood sample.

[00215] In certain embodiments, the neoantigen-microbial homology is measured by calculating a recognition potential (cross reactivity) score between each neoantigen and microbial epitope. As used herein, the terms “cross reactivity” and “recognition potential” are used interchangeably. In certain embodiments, calculating a recognition potential score between each neoantigen and microbial epitope is performed by a recognition potential model. As used herein the terms “recognition potential model,” “neoantigen recognition potential model,” and “cross reactivity model” are used interchangeably. In certain embodiments, the neoantigen recognition potential model comprises measuring sequence

alignment scores for each neoantigen to a microbial epitope (*e.g.*, with positive immune assays from the Immune Epitope Database). The neoantigen recognition potential model can also comprise scaling the alignment scores to binding probability of a T cell to cross reacting microbial epitope by fitting a sigmoid function. In some embodiments, the binding probabilities are amplified by relative wild type (“WT”) and neoantigenic peptide (“Mutant”). A neoantigen recognition potential score for a given neoantigen is a function of the alignment score and the amplitude, *i.e.*, K_d^{WT}/K_d^{Mutant} . In some embodiments, the calculation of a recognition potential score for a neoantigen is done using any of the fitness models disclosed in Łuksza *et al.*, 2017, “A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy,” *Nature* 551, 517-520. In some embodiments, the calculation of a recognition potential score for a neoantigen is done using any of the fitness models disclosed in Łuksza *et al.*, 2017, “A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy,” *Nature* 551, 517-520. In some embodiments, the calculation of a recognition potential score for a neoantigen is done in accordance with any of the models disclosed below in conjunction with blocks 346 through 358 of Figure 3.

[00216] In certain embodiments, the neoantigen-microbial homology^{high} neoantigen is a cross-reactivity score^{high} neoantigen, *e.g.*, a neoantigen having a recognition potential score (*e.g.*, calculated by i) a recognition potential model described above, ii) any of the fitness models disclosed in Łuksza *et al.*, 2017, “A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy,” *Nature* 551, 517-520, and/or iii) any of the models disclosed below in conjunction with blocks 346 through 358 of Figure 3) higher than the median recognition potential obtained from a population of subjects with the cancer. The median recognition potential score obtained from a population of subjects having the cancer may vary depending on the type of the cancer, the type of therapy (*e.g.*, immunotherapy, chemotherapy), the cancer histology, and/or presence/absence of neoantigens. In certain embodiments, the median recognition potential score (*e.g.*, calculated by i) a neoantigen recognition potential model described above, ii) any of the fitness models disclosed in Łuksza *et al.*, 2017, “A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy,” *Nature* 551, 517-520, and/or iii) any of the models disclosed below in conjunction with blocks 346 through 358 of Figure 3) obtained from a population of subjects having a cancer is between about 0 and 1, between about 0 and 0.5, between about 0 and 0.1, between about 0.1 and 0.2, between about 0.2 and 0.3, between about 0.3 and 0.4, between about 0.4 and 0.5, between about 0.5 and 1, between about 0.5 and 0.6, between about 0.6

and 0.7, between about 0.7 and 0.8, between about 0.8 and 0.9, or between about 0.9 and 0.1. In certain embodiments, the median neoantigen number (*e.g.*, calculated by i) a recognition potential model described above, ii) any of the fitness models disclosed in Łuksza *et al.*, 2017, “A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy,” Nature 551, 517-520, and/or iii) any of the models disclosed below in conjunction with blocks 346 through 358 of Figure 3) obtained from a population of subjects having cancer is about 0.8. In certain embodiments, the median neoantigen number (*e.g.*, calculated by i) a recognition potential model described above, ii) any of the fitness models disclosed in Łuksza *et al.*, 2017, “A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy,” Nature 551, 517-520, and/or iii) any of the models disclosed below in conjunction with blocks 346 through 358 of Figure 3) obtained from a population of subjects having cancer is about 0.9. In certain embodiments, the cancer is pancreatic cancer, *e.g.*, PDAC.

[00217] In some embodiments, the median activated T cell number obtained from subjects having the cancer varies depending on the type of the cancer, the type of therapy (*e.g.*, immunotherapy, chemotherapy), the cancer histology, presence/absence of ongoing immune response, and/or presence/absence of neoantigens. In certain embodiments, the median activated T cell number obtained from a population of subjects having a cancer is between 0 and 2,000 cells/mm², between 0 and 10 cells/mm², between 10 and 20 cells/mm², between 20 and 30 cells/mm², between 30 and 40 cells/mm², between 40 and 50 cells/mm², between 10 and 50 cells/mm², between 50 and 100 cells/mm², between 50 and 60 cells/mm², between 60 and 70 cells/mm², between 70 and 80 cells/mm², between 80 and 90 cells/mm², between 90 and 100 cells/mm², between 100 and 200 cells/mm², between 200 and 500 cells/mm², between 500 and 1,000 cells/mm², between 1,000 and 1,500 cells/mm², or between 1,500 and 2,000 cells/mm². In certain embodiments, the median activated T cell number obtained from a population of subjects having a cancer is about 60 cells/mm². In certain embodiments, the median activated T cell number obtained from a population of subjects having a cancer is about 58 cells/mm². In certain embodiments, the cancer is pancreatic cancer, *e.g.*, PDAC.

[00218] In some embodiments, the cancer is a solid tumor. In some embodiments, the cancer is a liquid tumor. Non-limiting examples of solid tumor include pancreatic cancer, gastric cancer, bile duct cancer (*e.g.*, cholangiocarcinoma), liver cancer, colorectal cancer, melanoma, lung cancer, and breast cancer. Non-limiting examples of liquid tumor include

acute leukemia and chronic leukemia. In certain embodiments, the cancer is pancreatic cancer. In certain embodiments, the pancreatic cancer is pancreatic ductal adenocarcinoma (PDAC).

[00219] 3.3. *Immunotherapy*

[00220] Immunotherapies that boost the ability of endogenous T cells to destroy cancer cells have demonstrated therapeutic efficacy in a variety of human malignancies. However, some cancer patients have resistance to certain immunotherapies. The presently disclosed subject matter provides methods for identifying cancer patients who would be candidates and/or who would likely to respond to an immunotherapy.

[00221] Non-limiting examples of immunotherapies include therapies comprising one or more immune checkpoint blocking antibody, adoptive T cell therapies, non-checkpoint blocking antibody based immunotherapies, small molecule inhibitors, cancer vaccines, and combinations thereof.

[00222] Non-limiting examples of immune checkpoint blocking antibodies include antibodies against CTLA cytotoxic T-lymphocyte antigen 4 (anti-CTLA4 antibodies), antibodies against programmed death 1 (anti-PD-1 antibodies), antibodies against Programmed death-ligand 1 (anti-PD-L1 antibodies), antibodies against lymphocyte activation gene-3 (anti-LAG3 antibodies), antibodies against T cell immunoglobulin and mucin domain-containing protein 3 (anti-TIM-3 antibodies), antibodies against glucocorticoid-induced TNFR-related protein (GITR), antibodies against OX40, antibodies against CD40, antibodies against T cell immunoreceptor with Ig and ITIM domains (TIGIT), antibodies against 4-1BB, antibodies against B7 homolog 3 (anti-B7-H3 antibodies), antibodies against B7 homolog 4 (anti-B7-H4 antibodies), and antibodies against B- and T-lymphocyte attenuator (anti-BTLA antibodies).

[00223] Adoptive T cell therapy involves the isolation and *ex vivo* expansion of tumor specific T cells to achieve greater number of T cells. The tumor specific T cells are infused into cancer patients to give their immune system the ability to overwhelm remaining tumor via T cells which can attack and kill cancer. Non-limiting examples of adoptive T cell therapy include tumor-infiltrating lymphocyte (TIL) cell therapies, therapies comprising engineered or modified T cells, *e.g.*, T cells engineered or modified with T cell receptor (TCR-transduced T cells), or T cells engineered or modified with chimeric antigen receptor

(CAR-transduced T cells). These engineered or modified T cells recognize specific antigens associated with the cancers and attack cancers.

[00224] 4. *Therapeutic Uses of the Neoantigens*

[00225] In some embodiments, the neoantigens of the present disclosure, or identified using the methods of the present disclosure, are used for vaccine therapy and/or adoptive T cell therapies.

[00226] 4.1 *Vaccines*

[00227] Neoantigens can be an attractive source of targets for vaccine therapy. Neoantigen-based cancer vaccine can induce more robust and specific anti-tumor T-cell responses compared with conventional shared-antigen-targeted vaccine. Certain genetic loci, also known as immunogenic hotspots, can be preferentially enriched for neoantigens in specific tumors that display great T cell infiltration and adaptive immune activation. Vaccine targeting tumor-specific immunogenic hotspot generated neoantigen can induce robust and specific anti-tumor T cell response against the tumor cell. The vaccine can be used along or in combination with other cancer treatment, *e.g.*, immunotherapy.

[00228] In one aspect, the present disclosure provides a vaccine comprising one or more of the presently disclosed neoantigens, or a polynucleotide encoding the neoantigen, or a protein or peptide comprising the neoantigen. In some embodiments, the neoantigen is selected based, at least in part, on predicted immunogenicity, for example, *in silico*. In some embodiments, the predicted immunogenicity is analyzed using computational algorithms for MHC class I and class II binding as well as use of tandem minigene libraries for class II epitope screening. In addition, in some embodiments, neoantigen specific T cell assays are used to differentiate true immunogenic neoepitopes from putative ones (*see* Kvistborg *et. al.*, 2016, J. ImmunoTherapy of Cancer 4:22 for detailed review). In some embodiments, any methods and tools known in the art are used to predict immunogenicity of a neoantigen. For example, in some embodiments, the Immune Epitope Database (IEDB) T Cell Epitope-MHC Binding Prediction Tool disclosed in Brown *et. al.* 2010, Nucleic Acids Res. Jan;38(Database issue):D854-62) is used to predict the binding of neoantigen to autologous HLA-A encoded MHC proteins. In certain embodiments, immunogenicity analysis strategy and tools disclosed in WO2015/103037 are used in accordance with the disclosed subject matter, and the content of the forgoing patent is incorporated herein by reference in its entirety. In certain embodiments, a neoantigen is selectively targeted based on one or more of the

following: (i) homology to an epitope of a known pathogen or microbe and/or (ii) ability to activate T cells, *e.g.* in an *in vitro* assay.

[00229] In another aspect, the present disclosure provides a vaccine comprising one or more neoantigens identified using any of the neoantigen identification methods disclosed herein.

[00230] In certain embodiments, the vaccine comprises one or more neoantigen of MUC16, or a polynucleotide encoding said neoantigen or a protein or peptide comprising said neoantigen. In certain embodiments, the one or more neoantigen of MUC16 is selected from the neoantigenic peptides listed in Table 1. In certain embodiments, the vaccine comprises one or more neoantigen associated with a cancer, the one or more neoantigen correlating with a neoantigen-microbial homology that is higher than the median neoantigen-microbial homology occurring in subjects with the cancer, or a polynucleotide encoding the neoantigen or a protein or peptide comprising the neoantigen. In certain embodiments, the vaccine comprises one or more neoantigen associated with a cancer, the neoantigen correlating with an activated T cell number that is higher than the median activated T cell number occurring in subjects with a cancer, or a polynucleotide encoding the neoantigen or a protein or peptide comprising the neoantigen.

[00231] In certain embodiments, the vaccine comprises a neoantigen that occurs in a subject with a cancer, where this subject has an activated T cell number that is higher than the median activated T cell number of a population of subjects with the cancer. In certain embodiments, the neoantigen less frequently occurs in subjects with the cancer and having activated T cell numbers at or less than the median activated T cell number.

[00232] In some embodiments, the number of different neoantigens in the vaccine varies, for example, in some embodiments the vaccine comprises 2, 3, 4, 5, 6, 7, 8, 9, 10 or more different neoantigens. In some embodiments, the neoantigens of a given vaccine are linked, *e.g.*, by any biochemical strategy to link two proteins. In some embodiments, the neoantigens of a given vaccine are not linked.

[00233] In certain embodiments, the vaccine comprises one or more polynucleotides. In some embodiments, the one or more polynucleotides are RNA, DNA, or a mixture thereof. In certain embodiments, the vaccine is in the form of DNA or RNA vaccines relating to neoantigens. In some embodiments, the one or more neoantigen are delivered via a bacterial or viral vector containing DNA or RNA sequences that encode one or more neoantigen.

[00234] Non-limiting examples of vaccines of the present disclosure include tumor cell vaccines, antigen vaccines, and dendritic cell vaccines, RNA vaccines, and DNA vaccines. Tumor cell vaccines are made from cancer cells removed from the patient. In some such embodiments, the cells are altered (and killed) to make them more likely to be attacked by the immune system and then injected back into the patient. In some embodiments, the tumor cell vaccines are autologous, *e.g.*, the vaccine is made from killed tumor cells taken from the same person who receives the vaccine. In some embodiments, the tumor cell vaccines are allogeneic, *e.g.*, the cells for the vaccine come from someone other than the patient being treated.

[00235] In some embodiments, the vaccine is an antigen presenting cell vaccine, *e.g.*, a dendritic cell vaccine. Dendritic cells are special immune cells in the body that help the immune system recognize cancer cells. They break down cancer cells into smaller pieces (including antigens), and then hold out these antigens so T cells can see them. The T cells then start an immune reaction against any cells in the body that contain these antigens.

[00236] In some embodiments, the antigen presenting cell such as a dendritic cell is pulsed or loaded with the neoantigen, or genetically modified (via DNA or RNA transfer) to express one or more neoantigen (*see, e.g.*, Butterfield, 2015, *BMJ*. 22, 350; and Palucka, 2013, *Immunity* 39, 38-48). In certain embodiments, the dendritic cell is genetically modified to express one or more neoantigen peptide. In some embodiments, any suitable method known in the art is used for preparing dendritic cell vaccines of the present disclosure. For example, in some embodiments, immune cells are removed from the patient's blood and exposed to cancer cells or cancer antigens, as well as to other chemicals that turn the immune cells into dendritic cells and help them grow. The dendritic cells are then injected back into the patient, where they can cause an immune response to cancer cells in the body.

[00237] Furthermore, the presently disclosed subject matter provides methods for treating pancreatic cancer in a subject. In certain embodiments, the method comprises administering to the subject a presently disclosed vaccine.

[00238] In certain embodiments, the vaccination is therapeutic vaccination, administered to a subject who has pancreatic cancer. In certain embodiments, the vaccination is prophylactic vaccination, administered to a subject who can be at risk of developing

pancreatic cancer. In certain embodiments, the vaccine is administered to a subject who has previously had cancer and in whom there is a risk of the cancer recurring.

[00239] Vaccines can be administered in any suitable way as known in the art. In certain embodiments, the vaccine is delivered using a vector delivery system. In some embodiments, the vector delivery system is viral, bacterial or makes use of liposomes. In certain embodiments, a listeria vaccine or electroporation is used to deliver the vaccine.

[00240] The presently disclosed subject matter further provides a composition comprising a presently disclosed vaccine. In certain embodiments, the composition is a pharmaceutical composition. In some embodiments, the pharmaceutical composition further comprises a pharmaceutically acceptable carrier, diluent or excipient.

[00241] In certain embodiments, the vaccine leads to generation of an immune response in a subject. In some embodiments, the immune response is humoral and/or cell-mediated immunity, for example the stimulation of antibody production, or the stimulation of cytotoxic or killer cells, which can recognize and destroy (or otherwise eliminate) cells (*e. g.*, tumor cells) expressing antigens (*e.g.*, neoantigen) corresponding to the antigens in the vaccine on their surface. In some embodiments, inducing or stimulating an immune response includes all types of immune responses and mechanisms for stimulating them. In certain embodiments, the induced immune response comprises expansion and/or activation of Cytotoxic T Lymphocytes (CTLs). In certain embodiments, the induced immune response comprises expansion and/or activation of CD8⁺ T cells. In certain embodiments, the induced immune response comprises expansion and/or activation of helper CD4⁺T Cells. In some embodiments, the extent of an immune response is assessed by production of cytokines, including, but not limited to, IL-2, IFN- γ , and/or TNF α .

[00242] 4.2 *Adoptive T-cell Therapy*

[00243] In some embodiments of the present disclosure, the neoantigens of the present disclosure are used in adoptive T cell therapy. The presently disclosed subject matter provides a population of T cells that target one or more of the presently disclosed neoantigens. In some embodiments, the neoantigen are selected based, at least in part, on predicted immunogenicity, for example, *in silico*. In some embodiments, the predicted immunogenicity is analyzed using computational algorithms for MHC class I and class II binding as well as use of tandem minigene libraries for class II epitope screening. In addition, or alternatively, in some embodiments neoantigen specific T cell assays are used to

differentiate true immunogenic neoepitopes from putative ones (*see*, Kvistborg *et. al.*, 2016, J. ImmunoTherapy of Cancer 4:22 for a detailed review). In some embodiments, any methods and tools known in the art are used to predict immunogenicity of a neoantigen. For example, in some embodiments the Immune Epitope Database (IEDB) T Cell Epitope-MHC Binding Prediction Tool disclosed in Brown *et. al.*, 2010, Nucleic Acids Res. 38 (Database issue): D854-62 is used to predict the binding of neoantigen to autologous HLA-A encoded MHC proteins. In certain embodiments, immunogenicity analysis strategy and tools disclosed in WO2015/103037 are used in accordance with the disclosed subject matter, and the content of the forgoing patent is incorporated herein by reference in its entirety. In certain embodiments, a neoantigen is selectively targeted based on one or more of the following: (i) homology to an epitope of a known pathogen or microbe; and/or (ii) ability to activate T cells, *e.g.* in an *in vitro* assay.

[00244] In certain embodiments, the population of T cells target one or more neoantigen of MUC16. In certain embodiments, the one or more neoantigen of MUC16 is selected from the neoantigenic peptides listed in Table 1. In certain embodiments, the population of T cells target one or more neoantigen associated with a cancer, the one or more neoantigen correlating with a neoantigen-microbial homology that is higher than the median neoantigen-microbial homology occurring in subjects with the cancer. In certain embodiments, the population of T cells target one or more neoantigen associated with a cancer, the neoantigen correlating with an activated T cell number that is higher than the median activated T cell number occurring in subjects with the cancer. In certain embodiments, the neoantigen comprised in the vaccine occurs in a subject with a cancer, where the subject has an activated T cell number that is higher than the median activated T cell number of a population of subjects with the cancer. In certain embodiments, the neoantigen occurs less frequently in subjects with the cancer and having activated T cell numbers at or less than the median activated T cell number.

[00245] In certain embodiments, the T cells are selectively expanded to target the one or more neoantigen. T cells are lymphocytes that mature in the thymus and are chiefly responsible for cell-mediated immunity. T cells are involved in the adaptive immune system. In some embodiments, the T cells of the presently disclosed subject matter are any type of T cells, including, but not limited to, helper T cells, cytotoxic T cells, memory T cells (including central memory T cells, stem-cell-like memory T cells (or stem-like memory T cells), and two types of effector memory T cells: *e.g.*, T_{EM} cells and T_{EMRA} cells, regulatory T

cells (also known as suppressor T cells), natural killer T cells, mucosal associated invariant T cells, and $\gamma\delta$ T cells. Cytotoxic T cells (CTL or killer T cells) are a subset of T lymphocytes capable of inducing the death of infected somatic or tumor cells.

[00246] In certain embodiments, the T cells that specifically target one or more neoantigen are engineered or modified T cells. In certain embodiments, the engineered T cells comprise a recombinant antigen receptor that specifically targets or binds to one or more of the presently disclose neoantigens. In certain embodiments, the recombinant antigen receptor specifically targets one or more neoantigen of MUC16. In certain embodiments, the recombinant antigen receptor specifically targets one or more neoantigen associated with a cancer, the one or more neoantigen correlating with a neoantigen-microbial homology that is higher than the median neoantigen-microbial homology occurring in subjects with the cancer. In certain embodiments, the recombinant antigen receptor specifically targets one or more neoantigen associated with a cancer, the neoantigen correlating with an activated T cell number that is higher than the median activated T cell number occurring in subjects with the cancer. In certain embodiments, the recombinant antigen receptor is a chimeric antigen receptor (CAR). In certain embodiments, the recombinant antigen receptor is a T cell receptor (TCR). In certain embodiments, the CAR comprises an extracellular antigen-binding domain that specifically binds to one or more neoantigen, a transmembrane domain, and an intracellular signaling domain. CARs can activate the T-cell in response to recognition by the extracellular antigen-binding domain of its target. When T cells express such a CAR, they recognize and kill cells that express one or more neoantigen.

[00247] Affinity-enhanced TCRs are generated by identifying a T cell clone from which the TCR α and β chains with the desired target specificity are cloned. The candidate TCR then undergoes PCR directed mutagenesis at the complimentary determining regions (“CDR”) of the α and β chains. The mutations in each CDR region are screened to select for mutants with enhanced affinity over the native TCR. Once completed, lead candidates are cloned into vectors to allow functional testing in T cells expressing the affinity-enhanced TCR.

[00248] In certain embodiments, the T cell population is enriched with T cells that are specific to one or more neoantigen, *e.g.*, having an increased number of T cells that target one or more neoantigen. Therefore, the T cell population differs from a naturally occurring T cell population, in that the percentage or proportion of T cells that target a neoantigen is increased.

[00249] In certain embodiments, the T cell population comprises at least about 10%, about 20%, about 30%, about 40%, about 50%, about 60%, about 70%, about 80%, about 90%, about 95%, or about 100% T cells that target one or more neoantigen. In certain embodiments, the T cell population comprises no more than about 5%, about 10%, about 15%, about 20%, about 25%, about 30%, about 35%, about 40%, about 45%, or about 50% T cells that do not target one or more neoantigen.

[00250] In certain embodiments, the T cell population is generated from T cells isolated from a subject with cancer. In one non-limiting example, the T cell population is generated from T cells in a biological sample isolated from a subject with cancer. In some embodiments, the biological sample is a tumor sample, a peripheral blood sample, or a sample from a tissue of the subject. In certain embodiments, the T cell population is generated from a biological sample in which the one or more neoantigen is identified or detected.

[00251] The presently disclosed subject matter further provides a composition comprising such T cell populations as described herein. In certain embodiments, the composition is a pharmaceutical composition that comprises a pharmaceutically acceptable carrier. Furthermore, the presently disclosed subject matter provides a method of treating cancer in a subject, comprising administering to the subject a composition comprising such T cell population as described herein. In some embodiments, the cancer is any of the cancers enumerated in the present disclosure. In some embodiments the cancer is pancreatic cancer. In certain embodiments, the composition comprises a population of T cells that target one or more neoantigen of MUC16. In certain embodiments, the one or more neoantigen of MUC16 is selected from the neoantigenic peptides listed in Table 1.

[00252] In some embodiments, the methods are used *in vitro*, *ex vivo* or *in vivo*, for example, either for *in situ* treatment or for *ex vivo* treatment followed by the administration of the treated cells to the subject. In certain embodiments, the T cell population or composition is reinfused into the subject, for example following T cell isolation and expansion to target the one or more MUC16 neoantigen.

[00253] *5. Further Uses of Neoantigen Fitness Models for Determining A Likelihood That A Human Subject Afflicted With A Cancer Will Be Responsive to a Treatment Regimen that Comprises Administering a Checkpoint Blockade Immunotherapy and/or for Identifying an Immunotherapy for a Cancer*

[00254] 5.1 *Heterogeneous Tumor Evolution*

[00255] To describe the evolution of a heterogeneous tumor, its fitness is evaluated as a weighted average over dominant neoantigens in the tumor's subclones. The disclosed neoantigen recognition potential model predicts survival in anti-CTLA4 treated melanoma patients (Snyder *et al.*, 2014, "Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma," N. Engl. J. Med. 371, pp. 2189-2199; and Van Allen *et al.*, 2015, "Genomic correlates of response to CTLA-4 blockade in metastatic melanoma," Science 350, pp. 207-211, each of which is hereby incorporated by reference) and anti-PD1 treated lung cancer patients (Rizvi *et al.*, 2015, "Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer, Science 348, pp. 124-128, which is hereby incorporated by reference). Importantly, in some embodiments, low-fitness neoantigens identified by the systems and methods of the present disclosure are leveraged for developing novel immunotherapies as disclosed below and as discussed in the sections above. In a broader context, the present disclosure reveals evolutionary similarities between cancers and fast-evolving pathogens (Łuksza and Lässig, 2014, "Predictive fitness model for influenza, Nature 507, pp. 57-61; Wang *et al.*, 2015, "Manipulating the selection forces during affinity maturation to generate cross-reactive HIV antibodies," Cell 160, pp. 785-797; and Nourmohammad *et al.*, 2016, "Host-pathogen coevolution and the emergence of broadly neutralizing antibodies in chronic infections," PLoS Genet 12, e1006171, each of which is hereby incorporated by reference).

[00256] 5.2 *Example Systems for Determining a Likelihood that a Human Subject Afflicted with a Cancer will be Responsive to a Treatment Regimen and/or for Identifying an Immunotherapy for a Cancer Using a Neoantigen Fitness Model*

[00257] One aspect of the present disclosure relies upon the acquisition of a data set comprising a plurality of sequence reads (*e.g.*, whole genome sequencing reads, exome sequencing reads, targeted sequencing reads, *etc.*) of a subject. Figure 1 illustrates an example of an integrated system 502 for the acquisition of such data, and Figure 2 provides more details of such a system 502. The integrated system 502 obtains sequencing data (*e.g.*, whole genome sequencing reads, exome sequencing reads, targeted sequencing reads, *etc.*) from one or more samples 102 from a human cancer subject that is representative of the cancer, one or more glucose monitors 102, and a data collection device 250.

[00258] A detailed description of a data collection device 250 for determining a likelihood that a human subject afflicted with a cancer will be responsive to a treatment regimen (e.g., a treatment regimen comprising administering a checkpoint blockade immunotherapy directed to the cancer to the subject in accordance with the present disclosure) and/or for identifying an immunotherapy for a cancer using a neoantigen fitness model is described in conjunction with Figures 1 and 2. As such, Figures 1 and 2 collectively illustrate the topology of the system in accordance with the present disclosure. In the topology, there is a data collection device 250 for determining a likelihood that a human subject afflicted with a cancer will be responsive to a treatment regimen and/or for identifying an immunotherapy for a cancer using a neoantigen fitness model.

[00259] Referring to Figure 2, the data collection device 250 determines a likelihood that a human subject afflicted with a cancer will be responsive to a treatment regimen and/or for identifying an immunotherapy for a cancer using a neoantigen fitness model. To do this, the data collection device 250, receives sequencing reads (e.g., whole genome sequencing reads, exome sequencing reads, targeted sequencing reads, *etc.*) originating from one or more biological samples (e.g., biopsies) 102 of a subject. In some embodiments, the data collection device 250 receives such data directly from nucleic acid sequencers. For instance, in some embodiments the data collection device 250 receives this data wirelessly through radio-frequency signals. In some embodiments such signals are in accordance with an 802.11 (WiFi), Bluetooth, or ZigBee standard. In some embodiments, the data collection device 250 receives such data directly. In some embodiments the data collection device 250 receives this data across a communications networks.

[00260] Examples of networks 106 include, but are not limited to, the World Wide Web (WWW), an intranet and/or a wireless network, such as a cellular telephone network, a wireless local area network (LAN) and/or a metropolitan area network (MAN), and other devices by wireless communication. The wireless communication optionally uses any of a plurality of communications standards, protocols and technologies, including but not limited to Global System for Mobile Communications (GSM), Enhanced Data GSM Environment (EDGE), high-speed downlink packet access (HSDPA), high-speed uplink packet access (HSUPA), Evolution, Data-Only (EV-DO), HSPA, HSPA+, Dual-Cell HSPA (DC-HSPDA), long term evolution (LTE), near field communication (NFC), wideband code division multiple access (W-CDMA), code division multiple access (CDMA), time division multiple access (TDMA), Bluetooth, Wireless Fidelity (Wi-Fi) (e.g., IEEE 802.11a, IEEE 802.11ac,

IEEE 802.11ax, IEEE 802.11b, IEEE 802.11g and/or IEEE 802.11n), voice over Internet Protocol (VoIP), Wi-MAX, a protocol for e-mail (e.g., Internet message access protocol (IMAP) and/or post office protocol (POP)), instant messaging (e.g., extensible messaging and presence protocol (XMPP), Session Initiation Protocol for Instant Messaging and Presence Leveraging Extensions (SIMPLE), Instant Messaging and Presence Service (IMPS)), and/or Short Message Service (SMS), or any other suitable communication protocol, including communication protocols not yet developed as of the filing date of the present disclosure.

[00261] Of course, other topologies of the system 48 of Figure 1 are possible. For instance, rather than relying on a communications network 106, information may be sent directly to the data collection device 250. Further, the data collection device 250 may constitute a portable electronic device, a server computer, or in fact constitute several computers that are linked together in a network or be a virtual machine in a cloud computing context. As such, the exemplary topology shown in Figure 1 merely serves to describe the features of an embodiment of the present disclosure in a manner that will be readily understood to one of skill in the art.

[00262] Referring to Figure 2, in typical embodiments, the data collection device 250 comprises one or more computers. For purposes of illustration in Figure 2, the data collection device 250 is represented as a single computer that includes all of the functionality for determining a likelihood that a human subject afflicted with a cancer will be responsive to a treatment regimen and/or for identifying an immunotherapy for a cancer using a neoantigen fitness model. However, the disclosure is not so limited. In some embodiments, the functionality for determining a likelihood that a human subject afflicted with a cancer will be responsive to a treatment regimen and/or for identifying an immunotherapy for a cancer using a neoantigen fitness model is spread across any number of networked computers and/or resides on each of several networked computers and/or is hosted on one or more virtual machines at a remote location accessible across the communications network 106. One of skill in the art will appreciate that any of a wide array of different computer topologies are used for the application and all such topologies are within the scope of the present disclosure.

[00263] Turning to Figure 2 with the foregoing in mind, an exemplary data collection device 250 for determining a likelihood that a human subject afflicted with a cancer will be responsive to a treatment regimen and/or for identifying an immunotherapy for a cancer using a neoantigen fitness model comprises one or more processing units (CPU's) 274, a network or other communications interface 284, a memory 192 (e.g., random access memory), one or

more magnetic disk storage and/or persistent devices 290 optionally accessed by one or more controllers 288, one or more communication busses 213 for interconnecting the aforementioned components, a user interface 278, the user interface 278 including a display 282 and input 280 (e.g., keyboard, keypad, touch screen), and a power supply 276 for powering the aforementioned components. In some embodiments, the input 280 is a touch-sensitive display, such as a touch-sensitive surface. In some embodiments, the user interface 278 includes one or more soft keyboard embodiments. The soft keyboard embodiments may include standard (QWERTY) and/or non-standard configurations of symbols on the displayed icons. In some embodiments, data in memory 192 is seamlessly shared with non-volatile memory 290 using known computing techniques such as caching. In some embodiments, memory 192 and/or memory 290 includes mass storage that is remotely located with respect to the central processing unit(s) 274. In other words, some data stored in memory 192 and/or memory 290 may in fact be hosted on computers that are external to the data collection device 250 but that can be electronically accessed by the data collection device 250 over an Internet, intranet, or other form of network or electronic cable (illustrated as element 106 in Figure 2) using network interface 284.

[00264] In some embodiments, the memory 192 of the data collection device 250 for determining a likelihood that a human subject afflicted with a cancer will be responsive to a treatment regimen and/or for identifying an immunotherapy for a cancer using a neoantigen fitness model stores:

- an operating system 202 that includes procedures for handling various basic system services;
- a subject assessment module 204;
- whole genome sequencing reads 206 for the subject, the whole genome sequencing reads comprising, for each biological sample 208 from the subject, a plurality of sequence reads 210;
- a human leukocyte antigen (HLA) type of the subject 212;
- a plurality of clones 214 associated with a subject, and, for each respective clone 216 in the plurality of clones 214, an initial frequency X_{α} 218 of the clone 216, a clone fitness score 220, and a plurality of neoantigens 222 of the clone 216, each such neoantigen 222 including a neoantigen recognition potential 224, an amplitude A 226, an optional MHC affinity 228, and optional MHC affinity of the wildtype sequence corresponding to the neoantigen 230, and a probability of T-cell receptor recognition

232; and

- a total fitness 234 of the cancer.

[00265] In some embodiments, the subject assessment module 204 is accessible within any browser (phone, tablet, laptop/desktop). In some embodiments the subject assessment module 204 runs on native device frameworks, and is available for download onto the data collection device 250 running an operating system 202 such as Android or iOS.

[00266] In some implementations, one or more of the above identified data elements or modules of the data collection device 250 for determining a likelihood that a human subject afflicted with a cancer will be responsive to a treatment regimen and/or for identifying an immunotherapy for a cancer using a neoantigen fitness model are stored in one or more of the previously described memory devices, and correspond to a set of instructions for performing a function described above. The above-identified data, modules or programs (*e.g.*, sets of instructions) need not be implemented as separate software programs, procedures or modules, and thus various subsets of these modules may be combined or otherwise re-arranged in various implementations. In some implementations, the memory 192 and/or 290 optionally stores a subset of the modules and data structures identified above. Furthermore, in some embodiments the memory 192 and/or 290 stores additional modules and data structures not described above. Further still, in some embodiments, the data collection device 250 stores data for determining a likelihood that a human subject afflicted with a cancer will be responsive to a treatment regimen and/or for identifying an immunotherapy for a cancer using a neoantigen fitness model for two or more subjects, five or more subjects, one hundred or more subjects, or 1000 or more subjects.

[00267] In some embodiments, a data collection device 250 for determining a likelihood that a human subject afflicted with a cancer will be responsive to a treatment regimen and/or for identifying an immunotherapy for a cancer using a neoantigen fitness model is a smart phone (*e.g.*, an iPhone), laptop, tablet computer, desktop computer, or other form of electronic device (*e.g.*, a gaming console). In some embodiments, the data collection device 250 is not mobile. In some embodiments, the data collection device 250 is mobile.

[00268] It should be appreciated that the data collection device 250 illustrated in Figure 2 is only one example of a multifunction device that may be used for determining a likelihood that a human subject afflicted with a cancer will be responsive to a treatment regimen and/or for identifying an immunotherapy for a cancer using a neoantigen fitness

model, and that the data collection device 250 optionally has more or fewer components than shown, optionally combines two or more components, or optionally has a different configuration or arrangement of the components. The various components shown in Figure 2 are implemented in hardware, software, firmware, or a combination thereof, including one or more signal processing and/or application specific integrated circuits.

[00269] RF (radio frequency) circuitry of network interface 284 receives and sends RF signals, also called electromagnetic signals. In some embodiments, the sequencing reads 206 and HLA type 212 and/or other data is received using this RF circuitry from one or more devices such as nucleic acid sequencers. In some embodiments, the RF circuitry 108 converts electrical signals to/from electromagnetic signals and communicates with communications networks and other communications devices via the electromagnetic signals. The RF circuitry 284 optionally includes well-known circuitry for performing these functions, including but not limited to an antenna system, an RF transceiver, one or more amplifiers, a tuner, one or more oscillators, a digital signal processor, a CODEC chipset, a subscriber identity module (SIM) card, memory, and so forth. RF circuitry 284 optionally communicates with the communication network 106. In some embodiments, the circuitry 284 does not include RF circuitry and, in fact, is connected to the network 106 through one or more hard wires (e.g., an optical cable, a coaxial cable, or the like).

[00270] In some embodiments, the power supply 276 optionally includes a power management system, one or more power sources (e.g., battery, alternating current (AC)), a recharging system, a power failure detection circuit, a power converter or inverter, a power status indicator (e.g., a light-emitting diode (LED)) and any other components associated with the generation, management and distribution of power in portable devices.

[00271] As illustrated in Figure 2, the data collection device 250 preferably comprises an operating system 202 that includes procedures for handling various basic system services. The operating system 202 (e.g., iOS, DARWIN, RTXC, LINUX, UNIX, OS X, WINDOWS, or an embedded operating system such as VxWorks) includes various software components and/or drivers for controlling and managing general system tasks (e.g., memory management, storage device control, power management, etc.) and facilitates communication between various hardware and software components.

[00272] In some embodiments the data collection device 250 is a smart phone. In other embodiments, the data collection device 250 is not a smart phone but rather is a tablet

computer, desktop computer, emergency vehicle computer, or other form or wired or wireless networked device. In some embodiments, the data collection device 250 has any or all of the circuitry, hardware components, and software components found in the data collection device 250 depicted in Figure 2. In the interest of brevity and clarity, only a few of the possible components of the data collection device 250 are shown in order to better emphasize the additional software modules that are installed on the data collection device 250.

[00273] While the system 48 disclosed in Figure 1 can work standalone, in some embodiments it can also be linked with electronic medical records to exchange information in any way.

[00274] 5.3 *Example Methods for Determining a Likelihood that a Human Subject Afflicted with a Cancer will be Responsive to a Treatment Regimen*

[00275] Now that details of a system 48 for determining a likelihood that a human subject afflicted with a cancer will be responsive to a treatment regimen (*e.g.*, where the treatment regimen comprises administering a checkpoint blockade immunotherapy directed to the cancer to the subject) and/or for identifying an immunotherapy for a cancer using a neoantigen fitness model have been disclosed, details regarding a flow chart of processes and features of the system, in accordance with an embodiment of the present disclosure, are disclosed with reference to Figures 3A through 3F. In some embodiments, such processes and features of the system are carried out by the subject assessment module 204 illustrated in Figure 2.

[00276] *Block 302.* Referring to block 302 of Figure 3A, one aspect of the present disclosure provides systems and method for determining a likelihood that a human subject afflicted with a cancer will be responsive to a treatment regimen (*e.g.*, that comprises administering a checkpoint blockade immunotherapy directed to the cancer to the subject) and/or for identifying an immunotherapy for a cancer using a neoantigen fitness model is provided.

[00277] In some embodiments, the checkpoint blockade immunotherapy comprises administering an anti-CTLA-4, anti-PD1 or anti-PD-L1 compound to the cancer subject. *See Jantscheff et al.*, 2016, 76(14) “Anti-PD-1, Anti-PD-L1 and Anti-CTLA-4 checkpoint inhibitor treatment leads to different responses in syngeneic tumor models,” *Cancer Research*, 76(14), DOI: 10.1158/1538-7445.AM2016-3216, which is hereby incorporated by reference for disclosure on Anti-PD-1, Anti-PD-L1 and Anti-CTLA-4 checkpoint inhibitors.

In some embodiments, the checkpoint blockade immunotherapy comprises administering one or more antibodies or other compounds against lymphocyte activation gene-3 (anti-LAG3 compounds), one or more antibodies or other compounds against T cell immunoglobulin and mucin domain-containing protein 3 (anti-TIM-3 compounds), one or more antibodies or other compounds against glucocorticoid-induced TNFR-related protein (anti-GITR compounds), one or more antibodies or other compounds against OX40 (anti-OX40 compounds), one or more antibodies or other compounds against CD40 (anti-CD40 compounds), one or more antibodies or other compounds against T cell immunoreceptor with Ig and ITIM domains (anti-TIGIT compounds), one or more antibodies or other compounds against 4-1BB (anti-4-1BB compounds), one or more antibodies or other compounds against B7 homolog 3 (anti-B7-H3 compounds), one or more antibodies or other compounds against B7 homolog 4 (anti-B7-H4 compounds), or one or more antibodies or other compounds against B- and T-lymphocyte attenuator (anti-BTLA compounds). In some embodiments, the checkpoint blockade immunotherapy comprises administering an anti-CTLA-4, anti-PD1, anti-PD-L1, anti-LAG3, anti-TIM-3, anti-GITR, anti-OX40, anti-CD40, anti-TIGIT, anti-4-1BB, anti-B7-H3, anti-B7-H4, or anti-BTLA compound to the cancer subject (304).

[00278] In some embodiments, the cancer is a carcinoma, a melanoma, a lymphoma/leukemia, a sarcoma, or a neuro-glial tumor (308). In some embodiments, the cancer is lung cancer, pancreatic cancer, colon cancer, stomach or esophagus cancer, breast cancer, ovary cancer, prostate cancer, or liver cancer (310). In some embodiments, the cancer is a solid tumor. In some embodiments, the cancer is a liquid tumor. Non-limiting examples of solid tumor include pancreatic cancer, gastric cancer, bile duct cancer (*e.g.*, cholangiocarcinoma), liver cancer, colorectal cancer, melanoma, lung cancer, and breast cancer. Non-limiting examples of liquid tumor include acute leukemia and chronic leukemia. In certain embodiments, the cancer is pancreatic cancer. In certain embodiments, the pancreatic cancer is pancreatic ductal adenocarcinoma (PDAC).

[00279] *Block 312.* Referring to block 312 of Figure 3A, a plurality of sequencing reads 210 (*e.g.*, whole genome sequencing reads, exome sequencing reads, targeted sequencing reads, *etc.*) is obtained from one or more biological samples 208 from the human cancer subject that is representative of the cancer. In certain embodiments, a biological sample 208 is a biological tissue or fluid. Non-limiting biological samples 208 include bone marrow, blood, blood cells, ascites, (tissue or fine needle) biopsy samples, cell-containing body fluids, free floating nucleic acids, sputum, saliva, urine, cerebrospinal fluid, peritoneal

fluid, pleural fluid, feces, lymph, gynecological fluids, swabs (*e.g.*, skin swabs, vaginal swabs, oral swabs, and nasal swabs), washings or lavages such as a ductal lavages or bronchoalveolar lavages, aspirates, scrapings, specimens (*e.g.*, bone marrow specimens, tissue biopsy specimens, and surgical specimens), feces, other body fluids, secretions, and/or excretions, and cells therefrom, *etc.*

[00280] In some embodiments, the plurality of sequencing reads is whole genome sequencing reads that collectively exhibit an average read depth of less than 200, less than 100, less than 50, less than 40, or less than 20 (314). In some embodiments, the plurality of sequencing reads are whole genome sequencing reads that collectively exhibit an average read depth of between 25 and 60 (316).

[00281] In some embodiments, the plurality of sequencing reads encompasses a subset of the genome of the subject and not the rest of the genome of the subject. In some such embodiments, the plurality of sequencing reads exhibits an average read depth of less than 200, less than 100, less than 50, less than 40, or less than 20 across this subset of the genome of the subject. In some embodiments, the plurality of sequencing reads exhibits an average read depth of between 25 and 60 across this subset of the genome of the subject. In some embodiments, the plurality of sequencing reads comprises whole genome sequencing reads. In some embodiments, the plurality of sequencing reads comprises exome sequencing reads. In some embodiments, the plurality of sequencing reads comprises targeted sequencing reads.

[00282] In some embodiments, the subset of the genome is between one percent and ten percent of a single chromosome of the subject, between five percent and fifteen percent of a single chromosome of the subject, between ten percent and twenty percent of a single chromosome of the subject, between fifteen percent and thirty percent of a single chromosome of the subject, between twenty-five percent and fifty percent of a single chromosome of the subject, between forty-five percent and seventy-five percent of a single chromosome of the subject, or between seventy percent and one hundred percent of a single chromosome of the subject.

[00283] In some embodiments, the subset of the genome is between one percent and ten percent of a two or more chromosomes of the subject, between five percent and fifteen percent of two or more chromosomes of the subject, between ten percent and twenty percent of two or more chromosomes of the subject, between fifteen percent and thirty percent of two or more chromosomes of the subject, between twenty-five percent and fifty percent of two or

more chromosomes of the subject, between forty-five percent and seventy-five percent of two or more chromosomes of the subject, or between seventy percent and one hundred percent of two or more chromosomes of the subject.

[00284] In some embodiments, the subset of the genome is between one percent and ten percent of the genome of the subject, between five percent and fifteen percent of the genome of the subject, between ten percent and twenty percent of the genome of the subject, between fifteen percent and thirty percent of the genome of the subject, between twenty-five percent and fifty percent of the genome of the subject, between forty-five percent and seventy-five percent of the genome of the subject, or between seventy percent and one ninety-nine percent of the genome of the subject.

[00285] *Block 318.* Referring to block 318 of Figure 3A, a human leukocyte antigen (HLA) type of the human cancer subject is determined. In some such embodiments, the HLA type of the human cancer subject is determined from the plurality of sequencing reads (320). In some embodiments, the determining the HLA type of the human cancer subject is determined using a polymerase chain reaction using a biological sample from the cancer subject (322). In some such embodiments, HLA typing is performed using the sequence reads 210 by either low to intermediate resolution polymerase chain reaction-sequence-specific primer (PCR-SSP) method or by high-resolution SeCore HLA sequence-based typing method (HLA-SBT) (INVITROGEN). In some embodiments ATHLATES is used for HLA typing and confirmation. *See* Liu and Duffy *et al.*, 2013, “ATHLATES: accurate typing of human leukocyte antigen through exome sequencing,” *Nucleic Acids Res* 41:e142, which is hereby incorporated by reference

[00286] *Block 324.* Referring to block 324 of Figure 3B, a plurality of clones is determined from the plurality of sequencing reads 210. In some embodiments, the raw sequence reads 210 are processed in order to identify the plurality of clones 216. For instance, in some embodiments, raw sequence data reads 210 are aligned to a reference human genome (*e.g.*, hg19) using an alignment tool such as the Burrows-Wheeler Alignment tool. Base-quality score recalibration, and duplicate-read removal is performed. In some embodiments, this recalibration excludes germline variants, annotation of mutations, and indels as described in Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” *N. Engl. J. Med.* 371, 2189-2199, which is hereby incorporated by reference. In some embodiments, local realignment and quality score recalibration are conducted using the Genome Analysis Toolkit (GATK) according to GATK best practices.

See, DePristo *et al.*, 2011, “A framework for variation discovery and genotyping using next-generation DNA sequencing data,” *Nature Genet.* 43, pp. 491-498; and Van der Auwera *et al.*, 2013, “From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline,” *Curr. Prot. in Bioinformatics* 43, 11.10.1-11.10.33 each of which is hereby incorporated by reference. Further, sequence alignment and mutation identification is performed. In some embodiments, sequence alignment and mutation identification is performed using FASTQ files that are processed to remove any adapter sequences at the end of the reads. In some embodiments, adapter sequences are removed using cutadapt (v1.6). *See* Martin, 2011, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” *EMBnet.journal* 17, pp. 10-12, which is hereby incorporated by reference. Then, resulting files are mapped using a mapping software such as the BWA mapper (bwa mem v0.7.12), (*see* Li and Durbin, 2009, “Fast and accurate short read alignment with Burrows-Wheeler Transform,” *Bioinformatics* 25, pp. 1754-1760, which is hereby incorporated by reference). In some such embodiments, the resulting files (*e.g.* SAM files) are sorted, and read group tags added using the PICARD tools. After sorting in coordinate order, the BAMs are processed with a tool such as PICARD MarkDuplicates. Realignment and recalibration is then carried out in some embodiments, (*e.g.*, with a first realignment using the InDel realigner followed by base quality value recalibration with the BaseQRecalibrator). Once realignment and recalibration have been performed, mutation callers are then used to identify single nucleotide variants. Exemplary mutation callers that can be used include, but are not limited to, Mutect 1.1.4, Somatic Sniper 1.0.4, Varscan 2.3.7, and Strelka 1.013). *See* Wei *et al.*, 2015, “MAC: identifying and correcting annotation for multi-nucleotide variations, *BMC Genomics* 16, p. 569; Snyder and Chan, 2015, “Immunogenic peptide discovery in cancer genomes,” *Curr Opin Genet Dev* 30, pp. 7-16; Nielsen *et al.*, 2003, “Reliable prediction of T-cell epitopes using neural networks with novel sequence representations,” *Protein Sci* 12, pp. 1007-1017; and Shen and Seshan, 2016, “FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing,” *Nucleic Acids Res.* 44, e131, each of which is incorporated by reference.

[00287] In some embodiments, SNVs with an allele read count of less than 4 or with corresponding normal coverage of less than 7 reads are filtered out. In some embodiments, SNVs with an allele read count of less than 7, of less than 5, or of less than 3 or with corresponding normal coverage of less than 12 reads, less than 8 reads, or less than 5 reads are filtered out. *See*, for example, Riaz *et al.*, 2016, “Recurrent SERPINB3 and SERPINB4

mutations in patients who respond to anti-CTLA4 immunotherapy,” *Nat. Genet.* 48, 1327-1329, which is hereby incorporated by reference.

[00288] In some embodiments, the assignment of a somatic mutation to a neoantigen is estimated using a bioinformatics tool such as NAsseek. *See Snyder et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” *N. Engl. J. Med.* 371, pp. 2189-2199, which is hereby incorporated by reference. NAsseek is a computational algorithm that first translates all mutations in exomes to strings of 17 amino acids, for both the wildtype and mutated sequences, with the amino acid resulting from the mutation centrally situated. Secondly, it evaluates putative MHC Class I binding for both wildtype and mutant nonamers using a sliding window method using NetMHC3.4 (on the Internet cbs.dtu.dk/services/NetMHC-3.4/) (*see Andreatta and Nielsen*, 2016, “Gapped sequence alignment using artificial neural networks: application to the MHC class I system,” *Bioinformatics* 32, pp. 511-517, which is hereby incorporated by reference) for patient-specific HLA types, to generate predicted binding affinities for both peptides. NAsseek finally assesses for similarity between nonamers that predicted to be presented by patient-specific MHC Class I. In some embodiments, all nonamers with binding scores below 500 nM are defined as neoantigens. In some embodiments, all nonamers with binding scores below 250 nM are defined as neoantigens. In some embodiments, all nonamers with binding scores below 800 nM are defined as neoantigens.

[00289] In some embodiments, the plurality of clones is 2 or more clones 216, 3 or more clones 216, 4 or more clones 216, 5 or more clones 216, 10 or more clones 216 or 100 or more clones 216. For each respective clone α in the plurality of clones, an initial frequency X_{α} 218 of the respective clone α 216 in the one or more samples 208 is determined.

[00290] In some such embodiments, tumor clones 214 are reconstructed using the PhyloWGS software package. *See Deshwar et al.*, 2015, “PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors,” *Genome Biol.* 16, p. 35, which is hereby incorporated by reference. The trees estimate the nested clonal structure of the tumor and the frequency of each clone, X_{α} . The differences between the high scoring trees are marginal on some datasets, concerning only peripheral clones and small differences in frequency estimates. In some embodiments, the predicted relative size of a cancer population $n(\tau)$ is computed as an averaged prediction over 5 trees with the highest likelihood score, weighting their contribution proportionally to their likelihood. In some embodiments, the predicted relative size of a cancer population $n(\tau)$ is predicted as an

averaged prediction over 5 or more trees, 10 or more trees, 15 or more trees, or 20 or more trees with the highest likelihood score.

[00291] In some such embodiments, to identify tumor clones, input data for the clone determining algorithm is extracted from exome sequencing data: (1) mutation reads obtained from the sequencing reads 210 processed in accordance with the pipeline described above and (2) allele-specific copy-number variant data, obtained with a program such as FACETS v0.5.0. *See*, Shen and Seshan, 2015, “FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing,” *Nucleic Acids Res.* 44, e131, which is hereby incorporated by reference. FACETS clusters mutations into clones by the frequency of their sequence reads 210 and infers possible nesting of clones (ancestral relations) between pairs of clones. Intuitively, an ancestral clone needs to have higher frequency than its derived clone. From this information PhyloWGS reconstructs high likelihood tumor genealogical trees.

[00292] In some embodiments, each clone α 216 in the plurality of clones is uniquely defined by a unique set of somatic mutations (*e.g.*, single nucleotide variant or an indel). In some embodiments, the plurality of clones is determined by a variant allele frequency of each respective somatic mutation in a plurality of somatic mutations determined from the whole-genome sequencing data (326).

[00293] In some embodiments, the plurality of clones is determined by identifying a plurality of inferred copy number variations using the whole-genome sequencing data (328).

[00294] In some embodiments, each clone α in the plurality of clones is uniquely defined by a unique set of somatic mutations. In some embodiments, the plurality of clones is determined by a combination of (i) a variant allele frequency of each respective somatic mutation in the plurality of somatic mutations determined from the whole-genome sequencing data and (ii) an identification of a plurality of inferred copy number variations using the whole-genome sequencing data (330).

[00295] In some embodiments, the plurality of clones consists of two clones (332). In some embodiments, the plurality of clones consists of between two clones and ten clones (334). In some embodiments, the plurality of clones comprises two or more clones, three or more clones, four or more clones, five or more clones, six or more clones, between 3 and 20 clones, or between 5 and 1000 clones.

[00296] In some embodiments, the initial frequency X_{α} of the respective clone α in the one or more samples is determined using the plurality of sequencing reads from the one or more samples from the human cancer subject (336). In some embodiments, the initial clone α frequency is inferred from the sequence reads 210 using techniques disclosed in Deshwar *et al.*, 2015, “PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors,” *Genome Biol.* 16, p. 35, which is hereby incorporated by reference.

[00297] *Block 338.* Referring to block 338 of Figure 3C, for each respective clone α 216 in the plurality of clones, a corresponding clone fitness score 220 of the respective clone 216 is computed, thereby computing a plurality of clone fitness scores, each corresponding clone fitness score computed for a respective clone α by a first procedure.

[00298] In the first procedure, a plurality of neoantigens 222 in the respective clone α 216 is identified (340). Methods for identifying neoantigens in accordance with some embodiments of the disclosure are described above and in Example I. Moreover, in some embodiments, neoantigens are determined by whole exome sequencing, immunoassay, microarray, genome sequencing, RNA sequencing, ELISA, Western Transfer, DNA sequencing, mass spectrometry, or combinations thereof. In some embodiments, a neoantigen is detected by the method described in Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” *N. Engl. J. Med.* 371, pp. 2189-2199; or Rizvi *et al.*, 2015, *Science* 348, pp. 124-128, each of which is hereby incorporated by reference.

[00299] In some embodiments, each neoantigen 222 in the plurality of neoantigens of a clone 216 in the plurality of clones is a nonamer peptide (342). In some embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide that is eight, nine, ten, or eleven residues in length (344). In some embodiments, at least one neoantigen is a neoantigen listed in Table 1 or Table 2. In some embodiments, at least one neoantigen is from MUC16 or a gene listed in Table 2.

[00300] In some embodiments, the method further comprises identifying a population of neoantigens present in the one or more samples by a third procedure comprising: determining a plurality of somatic single nucleotide polymorphisms (SNPs) in the plurality of sequencing reads by comparison of the plurality of sequencing reads (*e.g.*, whole genome sequencing reads, exome sequencing reads, targeted sequencing reads, *etc.*) to a reference

human genome, and evaluating each respective somatic SNP in the plurality of SNPs as a neoantigen candidate by evaluation of a peptide encoded by a portion of one or more sequencing reads in the sequencing reads that includes the respective somatic SNP against a classifier that has been trained to predict peptide binding to class I MHC of the HLA type of the cancer subject, where a neoantigen candidate having a binding score below a threshold value is deemed to be a neoantigen in the population of neoantigens. Further, the identifying the plurality of neoantigens in the respective clone α comprises matching the SNPs in the respective clone α to respective neoantigens in the population of neoantigens. In some such embodiments, the threshold value is 500 nM. In some such embodiments, the threshold value is 1000 nM, 900 nM, 800 nM, 700 nM, 600 nM, 500 nM, 400 nM, 300 nM, 200 nM, or 100 nM. In some such embodiments, the threshold value is 1000 nM or less, 900 nM or less, 800 nM or less, 700 nM or less, 600 nM or less, 500 nM or less, 400 nM or less, 300 nM or less, 200 nM or less, or 100 nM or less.

[00301] Continuing with the first procedure, a recognition potential fitness potential of each respective neoantigen in the plurality of neoantigens in the respective clone α is computed by a second procedure (346).

[00302] *Computing an amplitude.* In the second procedure, an amplitude A of the respective neoantigen is computed as a function of the relative major histocompatibility complex (MHC) affinity between the respective wildtype and the mutant peptide (neoantigen) given the HLA type of the subject (348). In other words, the amplitude, A , is the ratio of the relative probability that a neoantigen is bound on class I MHC times the relative probability that a neoantigen's wildtype counterpart is not bound. In some embodiments, the amplitude $A = (P_U^{WT} / P_B^{WT}) \times (P_B^{MT} / P_U^{MT})$, where P_B^{MT} is the binding probability of a neoantigen, P_B^{WT} is the binding probability of its wildtype counterpart, and $P_U^{WT} = 1 - P_B^{WT}$ and $P_U^{MT} = 1 - P_B^{MT}$. As a result, the amplitude, A , rewards cases where the discrimination energy between a mutant and wildtype peptide by the same class I MHC molecule (i.e. the same HLA allele) is large (Storma, 2013, Quantitative Biol. 1, p 115), while the mutant binding energy is kept low. The τ parameter effectively sets this energy scale for dominant neoantigens in a cline when $R = 1$. Assuming similar concentrations for mutant and wildtype peptides, the amplitude is the ratio of wildtype to mutant dissociation constants:

$$A = K_d^{WT} / K_d^{MT}.$$

Negative thymic selection on TCRs is not absolute, but rather “prunes” the repertoire recognizing the self proteome (Yu *et al.*, 2015, *Immunity* 42, p. 929; and Legoux *et al.*, 2015, *Immunity* 43, p. 896). The amplitude A is therefore used as a proxy for the availability of TCRs in the repertoire to recognize a neoantigen. Neoantigens differ from their wildtype peptides by only a single mutation. Given the uniqueness of nonamer sequence in the self-proteome due to finite genome size (SI) it is highly improbable that the mutant peptide would have another 8-mer match in the human proteome, so only the comparison with the respective wildtype peptide is taken into account in some embodiments. We verified that the above is the case for 92% of all neoantigens, with the remainder largely emanating from gene families with many paralogs. The amplitude can be interpreted as a multiplicity of receptors available to cross-reactively recognize a neoantigen.

[00303] In some embodiments, neoantigens with mutations on positions 2 and 9, are excluded. In such embodiments, a high value of amplitude means the wildtype also likely already has hydrophobic residues at the anchor position and could be presented. Since neoantigens differ from their wildtype peptides by a single mutation, and given the uniqueness of nonamer sequences in the proteome, the self-nonamer in the genome with the greatest similarity to a neoantigen is likely to be its wildtype peptide. *See*, for example, Example II below. This implies that a high amplitude usually stands for a self peptide not likely to be abundantly presented by the MHC. Therefore, as its immunogenicity is not mitigated by a homologous self-peptide, the mutant peptide with high affinity is likely to be novel to T-cells.

[00304] In some embodiments, the function of the relative class I MHC affinity of the respective neoantigen and the wildtype counterpart of the respective neoantigen given the HLA type of the human cancer subject is a ratio of the relative class I MHC affinity of the respective neoantigen and the wildtype counterpart of the respective neoantigen given the HLA type of the subject (350).

[00305] In some such embodiments, the MHC presentation is quantified, as amplitude A , using the relative MHC affinity between the wildtype peptide and mutant neoantigen, a ratio used to analyze computational neoantigen predictions. *See* Hundal *et al.*, 2016, “pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens,” *Genome Med.* 8, 1-11, which is hereby incorporated by reference. The relative MHC affinity rewards mutant neoantigens with strong mutant affinities compared to wildtype. Without intending to be

limited to any particular theory, it is posited that the wildtype peptides presented by MHC are potentially subject to tolerance and hence, due to homology, their mutant counterparts may be as well, compromising their immunogenicity.

[00306] In some such embodiments, the function of the relative class I MHC affinity of the respective neoantigen and the wildtype counterpart of the respective neoantigen given the HLA type of the subject is a ratio of: (1) a dissociation constant between the respective neoantigen and the class I MHC presented by the cancer subject given the HLA type of the cancer subject, and (2) a dissociation constant between the wildtype counterpart of the respective neoantigen and the class I MHC presented by the cancer subject given the HLA type of the cancer subject (352).

[00307] In some such embodiments, the dissociation constant between the respective neoantigen and the class I MHC presented by the cancer subject is obtained as output from a first classifier upon inputting into the first classifier the amino acid sequence of the neoantigen. The dissociation constant between the wildtype counterpart of the respective neoantigen and the class I MHC presented by the cancer subject of the HLA type of the subject is obtained as output from the first classifier upon inputting into the first classifier the amino acid sequence of the respective wildtype counterpart of the neoantigen (e.g., the first classifier is specific to the HLA type of the cancer subject and has been trained with the respective class I MHC binding coefficient and sequence data of each peptide epitope in a plurality of epitopes presented by class I MHC in a training population having the HLA type of the subject) (354).

[00308] In some embodiments amplitude A is computed as set forth in Example III. In some embodiments amplitude A is computed by any of the methods disclosed in Łuksza *et al.*, 2017, “A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy,” Nature 551, 517-520.

[00309] Despite their differing only by a single mutation, inferred binding affinities for these peptides can be substantially different (Figures 26-28). Figures 26-28 provide inferred MHC binding affinities of mutant versus wildtype peptides. Neoantigens used in this study are 9-residue long peptides affinities predicted to be less than 500nM by NetMHC3.4 (Andreatta Nielsen, 2016, “Gapped sequence alignment using artificial neural networks: application to the MHC class I system,” Bioinformatics 32, pp. 511-517. Predicted affinities are plotted of mutant peptides, designated K_d^{MT} , versus the predicted dissociation affinities of

the wildtype peptides, which generated them, designated K_d^{WT} . A single point mutation can lead to predicted dissociation constant difference of up to four orders of magnitude.

[00310] Moreover, unlike considering solely mutant or wildtype affinities, the amplitude has consistent predictive value within the disclosed model (Figures 23-25). In Figures 23-25, ranking of fitness models when accounting for tumor subclonal composition is provided, in which fitness models evaluated for the three cohorts, Van Allen *et al.* (n=103), Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” N. Engl. J. Med. 371, pp. 2189-2199 (n=64) and Rizvi *et al.*, 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer, Science 348, pp. 124-128 (n=34) are ranked. The survival prediction of full recognition potential fitness model:

$$F_\alpha = - \max_{i \in \text{Clone } \alpha} (A_i \times R_i)$$

is compared with alternative models: (1) models that eliminate one of the features of the full model, namely A-only model:

$$F_\alpha = - \max_{i \in \text{Clone } \alpha} (A_i)$$

and R -only model:

$$F_\alpha = - \max_{i \in \text{Clone } \alpha} (R_i)$$

models wildtype dissociation constant:

$$F_\alpha = - \max_{i \in \text{Clone } \alpha} K_d^{WT} (\times R_i),$$

and without mutant dissociation constant:

$$F_\alpha = - \max_{i \in \text{Clone } \alpha} \frac{1}{K_d^{MT}} (\times R_i),$$

neoantigen load model:

$$F_\alpha = -L_\alpha$$

an additive neoantigen fitness model which uniformly summates fitness contributions of neoantigens in a clone. Additionally, the model is compared in which alignments to IEDB epitopes are evaluated only on position 3-8, a model that does not implement any filtering of neoantigens on position 2&9, and a model where the R component is evaluated on IEDB assays without positive validation. Finally, an alternative predictive criterion is tested by

using the average fitness over tumor clones instead of $n(\tau)$ to separate patients in survival analysis:

$$n(\tau) = \exp[F_{\tau}].$$

For each model, parameters used for predictions and error bars for these parameters are reported. The predictive power of all models is assessed with survival analysis, separating patients into equal size groups by the median value of $n(\tau)$ or the median value of the average fitness F within the cohort. A logrank test is used, the results of this comparison are reported in Figures 23, 24, and 25. To assign error bars to fluctuations of the log-rank test score a leave-one-out analysis is performed. That is, the survival analysis is repeated for each dataset after leaving out one sample in a cohort and compute standard deviation of the test statistic over all leave-one-out iterations. A fitness model is deemed predictive in some embodiments if it gives patient segregation of highly significant scores in all datasets with the same consistent set of parameters. Only the full neoantigen recognition potential fitness model meets these criteria. The results are highly significant when patient segregation is based on $n(\tau)$ values. The average fitness criterion the equation

$$\langle F \rangle = \sum_{\alpha} X_{\alpha} F_{\alpha}$$

marginally meets the above requirements for predictiveness, but with smaller significance (Figures 23, 24, and 25). The log-rank test scores is reported for all models and the logrank test p-value for models with significant patient segregation ($p < 0.05$).

[00311] Without intending to be limited to any particular theory, an interpretation of the above observation that amplitude has consistent predictive value within the disclosed model is that the amplitude is related to the quantity of TCRs available to recognize the neoantigen. A neoantigen needs to have low dissociation constant (i.e. high binding affinity) to be presented and generate a TCR response. However, if the wildtype peptide also has a low dissociation constant, tolerance mechanisms could have removed wildtype peptide specific TCRs. Due to cross-reactivity, the quantity of mutant specific TCRs could be reduced.

[00312] *Computing a probability of T-cell receptor recognition R.* Referring to block 356 of Figure 3D, in the second procedure, a probability of T-cell receptor recognition R of the respective neoantigen is computed as a probability that the respective neoantigen binds one or more epitopes that are positively recognized by T-cells after class I MHC presentation (356). As such, for TCR-recognition, in some such embodiments, recognition potential of

neoantigens is modeled with positive, class I restricted T-cell antigens from the Immune Epitope Database (IEDB). See, Vita, R. *et al.*, 2014, “The immune epitope database (IEDB) 3.0,” Nucleic Acids Res. 43, D405-D412, which is hereby incorporated by reference. In such embodiments, preexisting host immunity due to this epitope set is not assumed. Rather it is posited that high-scoring neoantigens are more “non-self.” As TCRs have intrinsic biases in their generation probability and can recognize large classes of peptides via cross reactivity (Murugan *et al.*, 2012, “Statistical inference of the generation probability of T-cell receptors from sequence repertoires,” Proc. Natl. Acad. Sci. 109, pp. 16161-16166; and Birnbaum *et al.*, 2014, “Deconstructing the peptide-MHC specificity of T cell recognition,” Cell 157, pp. 1073-1087), such neoantigens would be more likely recognized. In some such embodiments, a thermodynamic model is used to estimate this probability that a given TCR binds a tumor neoantigen. See Berg *et al.*, 1987, “Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters,” J. Mol. Biol. 193, pp. 723-743, which is hereby incorporated by reference. For a given neoantigen 222 with peptide sequence **s** and Epitope Database and Analysis Resource (IEDB) (see Vita *et al.*, 2014, “The immune epitope database (IEDB) 3.0,” Nucleic Acids Res. 43, D405-D412, which is hereby incorporated by reference), epitope with sequence **e**, the alignment score between **s** and **e** estimates the binding free energy between **s** and a TCR that recognizes **e**. Under this assumption, each mutation that changes a residue in **e** into a corresponding residue in **s** in their alignment will increase the binding energy between **s** and the TCR recognizing epitope **e** proportional to the alignment mismatch cost. In some embodiments analysis is restricted to linear epitopes from human infectious diseases that are positively recognized by T-cells after class I MHC presentation. In some embodiments a set of epitopes from the Immune Epitope Database (IEDB), a repository of over 120,000 immune epitopes, is utilized (Vita *et al.*, 2014, “The immune epitope database (IEDB) 3.0,” Nucleic Acids Res. 43, D405-D412). IEDB is a collection of epitope-specific experimental assays – the nature of which can be accessed by various fields (www.iedb.org). Every T cell assay reflects the binding of an epitope-specific TCR to an experimentally tested antigen (Vita *et al.*, 2014, “The immune epitope database (IEDB) 3.0,” Nucleic Acids Res. 43, D405-D412). In this case analysis was restricted to linear epitopes from human infectious diseases studies presented by class I MHC molecules for which there were positive T cell assays. For negative controls, epitopes associated with assays satisfying all of the above fields were

downloaded, except we did not restrict for positive assays. We then excluded those for which we assigned positive by IEDB to create a negative assay list.

[00313] In such an approach, it is assumed that a neoantigen predicted to cross-react with a TCR from this pool of immunogenic epitopes is a neoantigen more likely to be immunogenic itself, as members of the T-cell receptor repertoire both recognize a high number of presented antigens (Mason, 1999, “A very high level of crossreactivity is an essential feature of the T-cell receptor,” *Immunology Today* 19, pp. 395–404; and Sewell, 2012, “Why must T cells be cross-reactive?,” *Nature Rev. Immunol.* 12, pp. 669-677) and have intrinsic biases in their generation probabilities. *See*, Murugan *et al.*, 2012, “Statistical inference of the generation probability of T-cell receptors from sequence repertoires,” *Proc. Natl. Acad. Sci.* 109, pp. 16161-16166.

[00314] The probability a neoantigen is bound by a TCR is given by a nonlinear logistic dependence on sequence alignment scores to the epitope set. This model does not require full 9-amino acid identity of the neoantigen and epitope sequences for recognition. In some embodiments, the total TCR-recognition probability, R , is defined as the probability that neoantigen s is recognized by at least one TCR corresponding to an IEDB epitope.

[00315] In some such embodiments, the probability that the respective neoantigen binds one or more epitopes that are positively recognized by T-cells after class I MHC presentation is determined by a third procedure that comprises (a) selecting a respective epitope e from an epitope database IEDB, where the respective epitope e is positively recognized by T-cells after class I MHC presentation, (b) computing, for the respective epitope e , the probability

$$\text{Pr}_{\text{binding}}(s, e) = \frac{1}{1 + e^{-k(|s,e| - \alpha)}}$$

where, $|s, e|$ is a sequence alignment score between the sequence of the respective neoantigen and the sequence of the respective epitope, and k and α are constants; (c) performing the selecting (a) and the computing (b) for each respective epitope e in a plurality of epitopes in the epitope database IEDB, thereby computing a plurality of probabilities $\text{Pr}_{\text{binding}}(s, e)$; and (d) computing the probability of T-cell receptor recognition R of the respective neoantigen as:

$$R = 1 - \prod_{e \in \text{IEDB}} [1 - \text{Pr}_{\text{binding}}(s, e)],$$

where IEDB is the plurality of epitopes (358). In some such embodiments, the IEDB database is collated as set forth in Example IV.

[00316] In some such embodiments, $|\mathbf{s}, \mathbf{e}|$ is computed as an alignment (*e.g.*, gapless, or an alignment that allows gaps with suitable gap introduction and extension penalties) between the sequence of the respective neoantigen and the sequence of the respective epitope using an amino-acid similarity matrix. In some such embodiments, the amino-acid similarity matrix is a BLOSUM62 matrix. In alternative embodiments, a different amino-acid similarity matrix is used. For instance, in some embodiments, a Blocks Substitutions Matrix (BLOSUM) BLOSUM45 matrix, BLOSUM50 matrix, BLOSUM52 matrix, BLOSUM60 matrix, BLOSUM80 matrix, BLOSUM90 matrix, a Percent Accepted Mutation (PAM) 250 matrix, a PAM200 matrix, a PAM160 matrix, a PAM120 matrix, a PAM100 matrix, or a Gonnet substitution matrix is used. In some embodiments, a matrix disclosed in Pearson, 2013, “Selecting the Right Similarity-Scoring Matrix,” *Curr Protoc. Bioinformatics*, 43: 3.51-3.5.9, ed. Baxevanis *et al.*, which is hereby incorporated by reference, is used.

[00317] In some embodiments, rather than a sequence alignment, a binding energy constant between the neoantigen and receptor is computed using three-dimensional models of the neoantigen and the receptor and an applicable force field such as CHARMM, AMBER GROMACS, or NAMD software packages. *See generally* Adcock & McCammon, 2006, *Chemical Reviews*, 106, 1589, which is hereby incorporated by reference.

[00318] In some embodiments, any of the techniques disclosed for computing R in Łuksza *et al.*, 2017, “A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy,” *Nature* 551, 517-520, is used.

[00319] In some embodiments, a multistate thermodynamic model is used to define R . In this model, sequence similarities are treated as a proxy for binding energies. To assess sequence similarity between a neoantigen with peptide sequence $\mathbf{s}$ and an IEDB epitope $\mathbf{e}$, an alignment (*e.g.*, gapless, or an alignment that allows gaps with suitable gap introduction and extension penalties) between the two sequences is computed with a BLOSUM62 amino-acid similarity matrix (or an equivalent alignment matrix) (*See*, Henikoff and Henikoff, 1992, “Amino acid substitution matrices from protein blocks,” *Proc. Natl. Acad. Sci. USA* 89, pp. 10915–10919) and we denote their alignment scores as $|\mathbf{s}, \mathbf{e}|$. Given these sequence similarities, for a given neoantigen with peptide sequence $\mathbf{s}$, the probability that it is bound by a TCR specific to some epitope $\mathbf{e}$ is computed from the IEDB pool as

$$R = Z(k)^{-1} \sum_{\mathbf{e} \in \text{IEDB}} \exp [-k(a - |\mathbf{s}, \mathbf{e}|)],$$

where a represents the horizontal displacement of the binding curve, k sets the steepness of the curve at a , and

$$Z(k) = 1 + \sum_{\mathbf{e} \in \text{IEDB}} \exp [-k(a - |\mathbf{s}, \mathbf{e}|)]$$

is the partition function over the unbound state and all bound states. In the model, k functions as an inverse temperature and $a - |\mathbf{s}, \mathbf{e}|$ functions as a binding energy. These parameters define the shape of the sigmoid function (Figure 29) and, along with the characteristic time scale τ , are free parameters to be fit in our model. In some such embodiments, the IEDB database is collated as set forth in Example V.

[00320] In some embodiments, the parameters which give consistently informative predictions across all three datasets are $a = 26$ and $k = 4.87$. The logistic function is therefore a strongly nonlinear function of the effective alignment score, $\log(\sum_{\mathbf{e} \in \text{IEDB}} \exp[-k(a - |\mathbf{s}, \mathbf{e}|)])$. The average alignment length corresponding to score 26 is 6.8 for neoantigens in our datasets, but the effective alignment score is occasionally increased by multiple contributions of shorter alignments. Under the interpretation where, for a sufficiently presented neoantigen, A represents the multiplicity of available TCRs and R represents an intrinsic probability of recognition, $A \times R$ represents the effective size of the overall TCR response. We present it as a core quantity that can be modulated by additional environmental factors such as the T-cell infiltration as discussed in Example IV below.

[00321] In some embodiments, the probability that the respective neoantigen binds one or more epitopes that are positively recognized by T-cells after class I MHC presentation is computed as:

$$R = Z(k)^{-1} \sum_{\mathbf{e} \in D} \exp [-k(a - |\mathbf{s}, \mathbf{e}|)],$$

where a is a number that represents a horizontal displacement of a binding curve for the respective neoantigen, k is a number that sets the steepness of the binding curve at a , $Z(k)$ is a partition function over the unbound state and all bound states of the respective neoantigen of the form

$$1 + \sum_{\mathbf{e} \in D} \exp [-k(a - |\mathbf{s}, \mathbf{e}|)]$$

where D is a plurality of epitopes, each respective epitope $\mathbf{e}$ is an epitope from the plurality of epitopes that is positively recognized by T-cells after class I MHC presentation, and $|\mathbf{s}, \mathbf{e}|$ is a

measure of affinity between the respective neoantigen s and the respective epitope e . In some such embodiments, a is set to 26 and k is set to 4.87. In some such embodiments, the measure of affinity $|s, e|$ is computed as a sequence alignment between the sequence of the respective neoantigen s and the sequence of the respective epitope e using an amino-acid similarity matrix. For instance, in some such embodiments, the amino-acid similarity matrix is a Blocks Substitutions Matrix (BLOSUM) BLOSUM45 matrix, BLOSUM50 matrix, BLOSUM52 matrix, BLOSUM60 matrix, BLOSUM80 matrix, BLOSUM90 matrix, a Percent Accepted Mutation (PAM) 250 matrix, a PAM200 matrix, a PAM160 matrix, a PAM120 matrix, a PAM100 matrix, or a Gonnet substitution matrix. In some embodiments, the plurality of epitopes comprises a public database of epitopes that have been recognized by human T-cells from any human, or combination of humans, in a population of humans. In some embodiments, the public database is the immune epitope database 3.0 or equivalent. *See, Vita, R. et al., 2015, "The immune epitope database (IEDB) 3.0," Nucleic Acids Res. 43, D405–D412, which is hereby incorporated by reference.* In some alternative embodiments, the plurality of epitopes consists of epitopes that have been recognized by human T-cells from the human subject. In some embodiments, the plurality of epitopes comprises 1000 epitopes. In some embodiments, the plurality of epitopes comprises 10,000 epitopes. In some embodiments, the plurality of epitopes comprises 100,000 epitopes. In some embodiments, the plurality of epitopes comprises 1×10^6 epitopes.

[00322] In some embodiments, the probability that the respective neoantigen binds one or more epitopes that are positively recognized by T-cells after class I MHC presentation is computed as:

$$R = Z(k)^{-1} \sum_{t \in F} \exp [-k(a - |s, t|)],$$

where a is a number that represents a horizontal displacement of a binding curve for the respective neoantigen, k is a number that sets the steepness of the binding curve at a , $Z(k)$ is a partition function over the unbound state and all bound states of the respective neoantigen of the form

$$1 + \sum_{t \in F} \exp [-k(a - |s, t|)]$$

where F is a plurality of T-cell receptor sequences, each respective T-cell receptor t is a T-cell receptor from the plurality of T-cell receptor sequences F , and $|s, t|$ is a measure of affinity between the respective neoantigen s and the respective T-cell receptor t . In some

such embodiments, the plurality of T-cell receptors is a database of T-cell receptors drawn from a population of humans. In alternative embodiments, the plurality of T-cell receptors is drawn exclusively from the subject. In some embodiments, the plurality of T-cell receptors comprises 1000 T-cell receptors. In some embodiments, the plurality of T-cell receptors comprises 10,000 T-cell receptors. In some embodiments, the plurality of T-cell receptors comprises 100,000 T-cell receptors. In some embodiments the plurality of T-cell receptors comprises 1×10^6 T-cell receptors.

[00323] Referring to block 360 of Figure 3E, in the second procedure, the recognition potential of the respective neoantigen is computed as a function of (i) the amplitude A of the respective neoantigen and (ii) the probability of T-cell receptor recognition R of the respective neoantigen. As used herein, the term is used broadly to mean any linear or nonlinear function of (i) the amplitude A of the respective neoantigen and (ii) the probability of T-cell receptor recognition R of the respective neoantigen. Such a function may further include additional variables or constants in addition to (i) the amplitude A of the respective neoantigen and (ii) the probability of T-cell receptor recognition R of the respective neoantigen. Checkpoint blockade exposes cancer cells to strong immune pressure on their neoantigens and thereby reduces their reproductive success. The fitness of a cancer cell in a genetic clone α is its expected replication rate, i.e.

$$\frac{dN_\alpha}{d\tau} = F_\alpha N_\alpha$$

where N_α is the population size of clone α and F_α is that clone's fitness. Checkpoint-blockade immunotherapy introduces a strong selection challenge, which is expected to overshadow pre-therapy fitness effects in a productive response. For a given clone α the dynamics of its absolute size are hence given by $N_\alpha(\tau) = N_\alpha(0)\exp(F_\alpha\tau)$, and the total cancer cell population size is computed as a sum over its clones:

$$N(\tau) = \sum_{\alpha} N_\alpha(\tau) = \sum_{\alpha} N_\alpha(0)\exp(F_\alpha\tau).$$

The absolute size $N(\tau)$ is an effective population size, the number of cells estimates to have generated the observed clonal diversity. In some embodiments, the measure of survival used is the evolved relative population size $n(\tau) = N(\tau)/N(0)$, which compares the predicted evolved population size after a characteristic dimensionless time scale of evolution τ to the initial pretreatment effective size $N(0)$, the assumption being that successful responders to therapy will have their future effective cancer cell population size more strongly suppressed.

The initial clone α frequency is denoted $X_\alpha = N_\alpha(0)/N(0)$, these frequencies are inferred from bulk exome reads from a tumor sample. See Deshwar *et al.*, 2015, “PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors,” *Genome Biol.* 16, p. 35, which is hereby incorporated by reference. Hence, to compute $n(\tau)$ only estimates of the initial frequencies and fitness values for each clone are required, as shown in the following equation:

$$n(\tau) = \sum_{\alpha} X_{\alpha} \exp(F_{\alpha} \tau),$$

the absolute population size measurements are not needed. The hypothesis that due to the unleashing of a T-cell mediated immune response by checkpoint-blockade immunotherapy, the deleterious effects due to recognition of neoantigens are a dominant effect, and tumors with the greatest degree of selective immune challenge are better responders to therapy. In the above equation, the predicted relative size $n(\tau)$ of a cancer cell population in a tumor is computed as a weighted sum over its genetic clones, where F_{α} is the fitness and X_{α} is the initial frequency of clone α and τ is a characteristic evolutionary time scale. For each such model defined, its homogenous structure equivalent can be defined by assuming the tumor is strictly clonal with all neoantigens in the same clone at frequency 1.

[00324] Without intending to be limited to any particular theory, intuitively, patients with less immunologically fit tumors will have more significant population size reductions and, hence, improved response to therapy.

[00325] In accordance with block 360 of Figure 3E, the disclosed approach quantifies two factors that determine immunogenicity of a neoantigen: an amplitude determined by MHC-presentation, A 226, and the probability of TCR-recognition, R 232. The product of these two factors, $A \times R$, is labeled the neoantigen recognition potential 224 of the neoantigen. In some embodiments, this recognition potential is computed using the scripts illustrated in Figures 37 through 40 in accordance with the present disclosure. In some embodiments, the subject assessment module 204 includes the scripts illustrated in Figures 37 through 40 in accordance with the present disclosure or equivalents thereof.

[00326] *Block 362.* Referring to block 362, the first procedure continues by determining the corresponding clone fitness score 220 of the respective clone α 216 as an aggregate of the neoantigen recognition potential 224 across the plurality of neoantigens 222 in the respective clone α . Next, as illustrated in Figure 5, the total fitness F_{α} 234 for cancer cells in a tumor clone is computed by aggregating over the fitness scores 220 due to its

neoantigens 222. Specifically, the fitness is modeled for a given clone α by the recognition potential of the immunodominant neoantigen:

$$F_{\alpha} = - \max_{i \in \text{Clone } \alpha} (A_i \times R_i)$$

where index i iterates over all neoantigens 222 in clone α 216. That is, in such embodiments, the fitness of a given clone α is modeled by the recognition potential of the immunodominant neoantigen. Taking the best score within a clone is consistent with the notions of heterologous immunity and immunodominance – that a small set of antigens drive the immune response, whereas summing over neoantigens would imply a more uniform distribution of contributions. In such embodiments, the full form of the predicted relative cancer cell population size is given by

$$n(\tau) = \sum_{\alpha} X_{\alpha} \exp[- \max_{i \in \text{Clone } \alpha} (A_i \times R_i) \tau].$$

In some embodiments, a more general model for aggregating neoantigen fitness effects within a clone is used:

$$n(\tau, \beta) = \sum_{\alpha} X_{\alpha} \exp \left[\sum_{i \in \text{Clone } \alpha} \frac{\exp(-\beta f_i)}{Z(\beta)} f_i \tau \right].$$

where $f_i = -A_i \times R_i$ and $Z(\beta) = \sum_{i \in \text{Clone } \alpha} \exp(-\beta f_i)$. In addition to the above equation for $n(\tau)$, which corresponds to the limit $\beta \rightarrow \infty$, we show the case where $\beta = 0$ (uniform summation over all neoantigens (Figures 23-25). In that sense the above equation for $n(\tau, \beta)$ represents a general mathematical framework for weighing neoantigen contributions, with weights reflecting the probability of their productive recognition. The choice of β could be informed by additional data sources or defined in a clone specific manner, and it would then become an additional model parameter (or parameters). Taking the highest score within a clone as in the above equation for $n(\tau)$ is consistent with notions of immunodominance – that a relatively small set of antigens drive the immune response.

[00327] In some embodiments, the aggregate of the neoantigen recognition potentials across the plurality of neoantigens in the respective clone α is computed as a summation of the recognition potential of each respective neoantigen in the plurality of neoantigens in accordance with the general model for fitness of the clone (216):

$$F_{\alpha} = - \text{Ag}_{i \in \text{Clone } \alpha} (A_i \times R_i)$$

where Ag is an aggregation function over recognition potential fitness effects of neoantigens within a clone, such as a summation over all neoantigens in the clone (Figure 3E, 366), a predetermined subset of the neoantigens in the clone (Figure 3E, 368), or a nonlinear combination of the neoantigens in the clone (Figure 3E, 370). For instance, in some embodiments, the subset of the neoantigens in the plurality of neoantigens constitutes a predetermined number of neoantigens in the plurality of neoantigens that have the top recognition potential for the respective clone α (e.g., the top two neoantigens, the top three neoantigens, the top four neoantigens, *etc.*). In some embodiments, the aggregate of the neoantigen recognition potentials across the plurality of neoantigens in the respective clone α is computed as a nonlinear combination of the recognition potential of all or a subset of the neoantigens in the plurality of neoantigens (370).

[00328] In some embodiments, the aggregate of the neoantigen recognition potentials across the plurality of neoantigens in the respective clone α is computed by any of the methods disclosed in Łuksza *et al.*, 2017, “A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy,” *Nature* 551, 517-520.

[00329] *Block 372.* Referring to block 372 of Figure 3F, a total fitness 234 for the one or more samples is computed as a sum of the clone fitness scores 220 across the plurality of clones, where each clone fitness score 220 is weighted by the initial frequency X_α 218 of the corresponding clone α , and the total fitness 234 quantifies the likelihood that the human subject afflicted with the cancer will be responsive to the treatment regimen. In some such embodiments, the computation of the total fitness 234 for the one or more samples 208 is a sum of the clone fitness scores 220 across the plurality of clones 216. Accordingly, in some embodiments total fitness 234 is computed as $n(\tau)$, the predicted future size of a cancer cell population in a tumor relative to its effective size at the start of therapy as a weighted sum over its genetic clones:

$$n(\tau) = \sum_{\alpha} X_{\alpha} \exp(F_{\alpha} \tau),$$

where τ is a characteristic evolutionary time scale, X_{α} is the initial frequency X_{α} 218 of the corresponding clone α , F_{α} is the clone fitness score 220 of the corresponding clone α , and the summation of this equation is across all clones 216 identified in the one or more biological samples 208 (Figure 3F, 374). In some such embodiments, τ is between 0.0 and 0.5. In some such embodiments, τ is 0.06. In some such embodiments, a lower total fitness score is

associated with (a) a higher likelihood that the cancer subject will be responsive to the immunotherapy and (b) a longer term survival of the cancer patient (Figure 3F, 374).

[00330] *5.4 Example Methods for Identifying an Immunotherapy for a Cancer.*

[00331] Another aspect of the present disclosure provides methods for identifying an immunotherapy for a cancer. In some embodiments, the cancer is a solid tumor. In some embodiments the cancer is a liquid tumor. Non-limiting examples of solid tumors include pancreatic cancer, gastric cancer, bile duct cancer (*e.g.*, cholangiocarcinoma), liver cancer, colorectal cancer, melanoma, lung cancer, and breast cancer. Non-limiting examples of liquid tumors include acute leukemia and chronic leukemia.

[00332] In the disclosed methods, a plurality of sequencing reads 210 (*e.g.*, whole genome sequencing reads, exome sequencing reads, targeted sequencing reads, *etc.*) is obtained from one or more biological samples 208 from a human cancer subject that is representative of the cancer. A human leukocyte antigen (HLA) type of the human cancer subject 212 is determined from the plurality of sequencing reads. A plurality of clones 216 is determined (*e.g.*, from the sequencing reads). For each respective clone α 216 in the plurality of clones, an initial frequency X_{α} 218 of the respective clone α in the one or more samples is determined from the plurality of sequencing reads. Further, for each respective clone α in the plurality of clones, a corresponding clone fitness score 220 of the respective clone is computed, thereby computing a plurality of clone fitness scores, each corresponding clone fitness score computed for a respective clone α by a first procedure.

[00333] In the first procedure, a plurality of neoantigens 222 in the respective clone α are identified. For instance, in some embodiments this is done using the sequence reads 210 as follows. Raw sequence data reads are aligned to a reference human genome (*e.g.*, hg19) using an alignment tool such as the Burrows-Wheeler Alignment tool. In some embodiments, base-quality score recalibration, and duplicate-read removal is performed, with exclusion of germline variants, annotation of mutations, and indels as described in Snyder *et al.*, 2014 “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” *N. Engl. J. Med.* 371, pp. 2189-2199, which is hereby incorporated by reference. Local realignment and quality score recalibration are conducted using, for example the Genome Analysis Toolkit (GATK) according to GATK best practices. *See* DePristo *et al.*, 2011, “A framework for variation discovery and genotyping using next-generation DNA sequencing data,” *Nature*

Genet. 43, 491-498; and Van der Auwera *et al.*, 2013, "From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline," Curr. Prot. in Bioinformatics 43, 11.10.1-11.10.33, each of which is hereby incorporated by reference. For sequence alignment and mutation identification, adapter sequences at the end of the reads are removed. In some embodiments this is done using cutadapt (v1.6). Martin, 2011, "Cutadapt removes adapter sequences from high-throughput sequencing reads," EMBnet.journal 17, pp. 10-12, which is hereby incorporated by reference. In some embodiments, the files were then mapped using the BWA mapper (bwa mem v0.7.12) (Li and Durbin, 2009, "Fast and accurate short read alignment with Burrows-Wheeler Transform," Bioinformatics 25, pp. 1754-1760, hereby incorporated by reference), the SAM files sorted, and read group tags added using the PICARD tools. After sorting in coordinate order, the BAMs are processed with PICARD MarkDuplicates. First realignment was carried out using the InDel realigner followed by base quality value recalibration with the BaseQRecalibrator. Next, mutations callers are used to identify single nucleotide variants (SNVs). Exemplary base callers that are used in some embodiments are a combination of four different mutation callers (Mutect 1.1.4, Somatic Sniper 1.0.4, Varscan 2.3.7, and Strelka 1.013).³³⁻³⁶. See Wei *et al.*, 2015, "MAC: identifying and correcting annotation for multi-nucleotide variations, BMC Genomics 16, p. 569; Snyder and Chan, 2015, "Immunogenic peptide discovery in cancer genomes," Curr Opin Genet Dev 30, pp. 7-16; Nielsen *et al.*, 2003, "Reliable prediction of T-cell epitopes using neural networks with novel sequence representations," Protein Sci 12, pp. 1007-1017; and Shen and Seshan, 2016, "FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing," Nucleic Acids Res. 44, e131, each of which is incorporated by reference. In some embodiments SNVs with an allele read count of less than four or with corresponding normal coverage of less than 7 reads were filtered out in accordance with the criteria set forth in Riaz *et al.*, 2016, "Recurrent SERPINB3 and SERPINB4 mutations in patients who respond to anti-CTLA4 immunotherapy," Nat. Genet. 48, pp. 1327-1329, which is hereby incorporated by reference. In some embodiments, the assignment of a somatic mutation to a neoantigen is estimated using a previously described bioinformatics tool called NAsseek. See Snyder *et al.*, 2014, "Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma," N. Engl. J. Med. 371, pp. 2189-2199, which is hereby incorporated by reference. NAsseek is a computational algorithm that first translates all mutations in exomes to strings of 17 amino acids, for both the wildtype and mutated sequences, with the amino acid resulting from the mutation centrally situated. Secondly, it evaluates putative MHC Class I binding for both wildtype and mutant nonamers using a

sliding window method using NetMHC3.4 (*See* Andreatta and Nielsen, 2016, “Gapped sequence alignment using artificial neural networks: application to the MHC class I system,” *Bioinformatics* 32, pp. 511-517, hereby incorporated by reference) for patient-specific HLA types, to generate predicted binding affinities for both peptides. NAsseek finally assesses for similarity between nonamers that predicted to be presented by patient-specific MHC Class I. In some embodiments all nonamers with binding scores below 500 nM are defined as neoantigens.

[00334] Next, a recognition potential 224 of each respective neoantigen 222 in the plurality of neoantigens in the respective clone α 216 is then computed by a second procedure. In the second procedure, an amplitude A 226 of the respective neoantigen as a function of the relative major histocompatibility complex (MHC) affinity of the respective neoantigen and the wildtype counterpart of the respective neoantigen given the HLA type of the subject is computed. Further, a probability of T-cell receptor recognition R 232 of the respective neoantigen as a probability that the respective neoantigen binds one or more epitopes that are positively recognized by T-cells after class I MHC presentation is computed. In some embodiments, the recognition potential 224 of the respective neoantigen 222 is computed as the product of the amplitude A 226 of the respective neoantigen and the probability of T-cell receptor recognition R 232 of the respective neoantigen.

[00335] In the first procedure, the corresponding clone fitness score 220 of the respective clone α is determined as an aggregate of the neoantigen recognition potentials across the plurality of neoantigens in the respective clone α .

[00336] In the disclosed first procedure, a first neoantigen 222 is selected from a plurality of neoantigens for a respective clone α 216 in the plurality of respective clones based upon the recognition potential 224 of the first neoantigen 222 as the immunotherapy for the cancer.

[00337] In some embodiments, the first procedure is repeated for a plurality of human cancer subjects across a plurality of HLA types and the first neoantigen 222 is selected on the basis of the recognition potential of the first neoantigen 222 across the plurality of HLA types.

[00338] In some embodiments, the first procedure is repeated for a plurality of human cancer subjects and the first neoantigen 222 is selected on the basis of the recognition potential of the first neoantigen 222 across the plurality of human cancer subject.

[00339] In some embodiments, the cancer is a carcinoma, a melanoma, a lymphoma/leukemia, a sarcoma, or a neuro-glial tumor. In some embodiments, the cancer is lung cancer, pancreatic cancer, colon cancer, stomach or esophagus cancer, breast cancer, ovary cancer, prostate cancer, or liver cancer.

[00340] In some embodiments, each clone α 216 in the plurality of clones is uniquely defined by a unique set of somatic mutations, and the plurality of clones is determined by a variant allele frequency of each respective somatic mutation in a plurality of somatic mutations determined from the whole-genome sequencing data. For instance, in some embodiments, the somatic mutation is a single nucleotide variant or an indel.

[00341] In some embodiments, the plurality of clones is determined by identifying a plurality of inferred copy number variations using the whole-genome sequencing data.

[00342] In some embodiments, each clone α 216 in the plurality of clones is uniquely defined by a unique set of somatic mutations, and the plurality of clones is determined by a combination of (i) a variant allele frequency of each respective somatic mutation in the plurality of somatic mutations determined from the whole-genome sequencing data and (ii) an identification of a plurality of inferred copy number variations using the whole-genome sequencing data.

[00343] In some embodiments, the plurality of sequencing reads (*e.g.*, whole genome sequencing reads, exome sequencing reads, targeted sequencing reads, *etc.*) exhibits an average read depth of less than 40. In some embodiments, the plurality of sequencing reads exhibits an average read depth of between 25 and 60.

[00344] In some embodiments, the plurality of sequencing reads encompasses a subset of the genome of the subject and not the rest of the genome of the subject. In some such embodiments, the plurality of sequencing reads encompass exhibits an average read depth of less than 200, less than 100, less than 50, less than 40, or less than 20 across this subset of the genome of the subject. In some embodiments, the plurality of whole genome sequencing reads exhibits an average read depth of between 25 and 60 across this subset of the genome of the subject. In some embodiments, the plurality of sequencing reads comprises whole genome sequencing reads. In some embodiments, the plurality of sequencing reads comprises exome sequencing reads. In some embodiments, the plurality of sequencing reads comprises targeted sequencing reads.

[00345] In some embodiments, the subset of the genome is between one percent and ten percent of a single chromosome of the subject, between five percent and fifteen percent of a single chromosome of the subject, between ten percent and twenty percent of a single chromosome of the subject, between fifteen percent and thirty percent of a single chromosome of the subject, between twenty-five percent and fifty percent of a single chromosome of the subject, between forty-five percent and seventy-five percent of a single chromosome of the subject, or between seventy percent and one hundred percent of a single chromosome of the subject.

[00346] In some embodiments, the subset of the genome is between one percent and ten percent of a two or more chromosomes of the subject, between five percent and fifteen percent of two or more chromosomes of the subject, between ten percent and twenty percent of two or more chromosomes of the subject, between fifteen percent and thirty percent of two or more chromosomes of the subject, between twenty-five percent and fifty percent of two or more chromosomes of the subject, between forty-five percent and seventy-five percent of two or more chromosomes of the subject, or between seventy percent and one hundred percent of two or more chromosomes of the subject.

[00347] In some embodiments, the subset of the genome is between one percent and ten percent of the genome of the subject, between five percent and fifteen percent of the genome of the subject, between ten percent and twenty percent of the genome of the subject, between fifteen percent and thirty percent of the genome of the subject, between twenty-five percent and fifty percent of the genome of the subject, between forty-five percent and seventy-five percent of the genome of the subject, or between seventy percent and one ninety-nine percent of the genome of the subject.

[00348] In some embodiments, each neoantigen 222 in the plurality of neoantigens of a clone 216 in the plurality of clones is a nonamer peptide. In some embodiments, each neoantigen 222 in the plurality of neoantigens of a clone 216 in the plurality of clones is a peptide that is eight, nine, ten, or eleven residues in length. In certain embodiments each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide that is 3-30 amino acids, *e.g.*, about 3-5, about 5-15 (*e.g.*, about 8-11, about 5-10, or about 10-15), about 15-20, about 20-25, or about 20-30 amino acids, in length. In certain embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide that is about 8-11 amino acids in length. In certain embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide that is about 3,

about 4, about 5, about 6, about 7, about 8, about 9, about 10, about 11, about 12, about 13, about 14, about 15, about 16, about 17, about 18, about 19, about 20, about 21, about 22, about 23, about 24, about 25, about 26, about 27, about 28, about 29, or about 30 amino acids in length. In certain embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide that is at least about 3, at least about 5, or at least about 8 amino acids in length. In certain embodiments, each neoantigen in the plurality of neoantigens of a clone in the plurality of clones is a peptide is less than about 30, less than about 20, less than about 15, or less than about 10 amino acids in length.

[00349] In some embodiments, the method further comprises identifying a population of neoantigens present in the one or more samples by a third procedure in which a plurality of somatic single nucleotide polymorphisms (SNPs) in the plurality of sequencing reads is determined by comparison of the plurality of sequencing reads to a reference human genome. In the third procedure, each respective somatic SNP in the plurality of SNPs is evaluated as a neoantigen 222 candidate by evaluation of a peptide encoded by a portion of one or more sequencing reads in the sequencing reads that includes the respective somatic SNP against a classifier that has been trained to predict peptide binding to class I MHC of the HLA type of the cancer subject, where a neoantigen 222 candidate having a binding score below a threshold value (*e.g.*, 500 nM) is deemed to be a neoantigen 222 in the population of neoantigens. Further, when this third procedure is used, the identifying the plurality of neoantigens in the respective clone α 216 comprises matching the SNPs in the respective clone α 216 to respective neoantigens in the population of neoantigens.

[00350] In some embodiments, the HLA type determination of the human cancer subject is made from the plurality of sequencing reads. In some embodiments, the HLA type determination of the human cancer subject is made using a polymerase chain reaction using a biological sample from the cancer subject.

[00351] In some embodiments, the plurality of clones consists of two clones. In some embodiments, the plurality of clones consists of between two clones and ten clones. In some embodiments, the initial frequency X_{α} 218 of the respective clone α in the one or more samples is determined using the plurality of sequencing reads from the one or more samples from the human cancer subject.

[00352] In some embodiments, the function of the relative class I MHC affinity of the respective neoantigen 222 and the wildtype counterpart of the respective neoantigen 222

given the HLA type of the subject is a ratio of the relative class I MHC affinity of the respective neoantigen 222 and the wildtype counterpart of the respective neoantigen 222 given the HLA type of the subject.

[00353] In some embodiments, the function of the relative class I MHC affinity of the respective neoantigen 222 and the wildtype counterpart of the respective neoantigen 222 given the HLA type of the human cancer subject is a ratio of: (1) a dissociation constant between the respective neoantigen 222 and the class I MHC presented by the cancer subject given the HLA type of the cancer subject, and (2) a dissociation constant between the wildtype counterpart of the respective neoantigen 222 and the class I MHC presented by the cancer subject given the HLA type of the cancer subject. In some such embodiments, the dissociation constant between the respective neoantigen 222 and the class I MHC presented by the cancer subject is obtained as output from a first classifier upon inputting into the first classifier the amino acid sequence of the neoantigen 222, and the dissociation constant between the wildtype counterpart of the respective neoantigen 222 and the class I MHC presented by the cancer subject of the HLA type of the subject is obtained as output from the first classifier upon inputting into the first classifier the amino acid sequence of the respective wildtype counterpart of the neoantigen 222. In some such embodiments, the first classifier is specific to the HLA type of the cancer subject and has been trained with the respective class I MHC binding coefficient and sequence data of each peptide epitope in a plurality of epitopes presented by class I MHC in a training population having the HLA type of the subject.

[00354] In some embodiments, the probability that the respective neoantigen 222 binds one or more epitopes that are positively recognized by T-cells after class I MHC presentation is determined by a third procedure that comprises (a) selecting a respective epitope **e** from an epitope database IEDB, where the respective epitope **e** is positively recognized by T-cells after class I MHC presentation, (b) computing, for the respective epitope **e**, the probability

$$\text{Pr}_{\text{binding}}(\mathbf{s}, \mathbf{e}) = \frac{1}{1 + e^{-k(|\mathbf{s}, \mathbf{e}| - \alpha)}}$$

where $|\mathbf{s}, \mathbf{e}|$ is a sequence alignment score between the sequence of the respective neoantigen 222 and the sequence of the respective epitope, and k and α are constants, and (c) performing the selecting (a) and the computing (b) for each respective epitope **e** in a plurality of epitopes in the epitope database IEDB, thereby computing a plurality of probabilities $\text{Pr}_{\text{binding}}(\mathbf{s}, \mathbf{e})$, and (d) computing the probability of T-cell receptor recognition R of the respective neoantigen 222 as:

$$R = 1 - \prod_{e \in \text{IEDB}} [1 - \text{Pr}_{\text{binding}}(\mathbf{s}, \mathbf{e})],$$

where IEDB the plurality of epitopes. In some such embodiments, a is set to 23 and k is set to 1. In some such embodiments, $|\mathbf{s}, \mathbf{e}|$ is computed as an alignment (e.g., gapless, or an alignment that allows gaps with suitable gap introduction and extension penalties) between the sequence of the respective neoantigen 222 and the sequence of the respective epitope using an amino-acid similarity matrix (e.g., a BLOSUM62 matrix).

[00355] In some embodiments, the first neoantigen 222 from a plurality of neoantigens for a respective clone α 216 in the plurality of respective clones is selected when it has a recognition potential 224 that is lower than the recognition potential of other neoantigens in each plurality of neoantigens for each respective clone α 216 in the plurality of respective clones of the subject.

Exemplification

EXAMPLE I

[00356] *Computational identification of neoantigens.* Neoantigens from the three datasets were inferred using a consistent pipeline established at Memorial Sloan Kettering Cancer Center. Raw sequence data reads were aligned to the reference human genome (hg19) using the Burrows-Wheeler Alignment tool. Base-quality score recalibration, and duplicate-read removal were performed, with exclusion of germline variants, annotation of mutations, and indels as previously described. See Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” N. Engl. J. Med. 371, 2189-2199, which is hereby incorporated by reference. Local realignment and quality score recalibration were conducted using the Genome Analysis Toolkit (GATK) according to GATK best practices. See, DePristo *et al.*, 2011, “A framework for variation discovery and genotyping using next-generation DNA sequencing data,” Nature Genet. 43, pp. 491-498; and Van der Auwera *et al.*, 2013, “From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline,” Curr. Prot. in Bioinformatics 43, 11.10.1-11.10.33 each of which is hereby incorporated by reference. For sequence alignment and mutation identification, the FASTQ files were processed to remove any adapter sequences at the end of the reads using cutadapt (v1.6). See Martin, 2011, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” EMBnet.journal 17, pp. 10-12, which is hereby incorporated by reference. The files were then mapped using the BWA mapper (bwa mem

v0.7.12), (*see* Li and Durbin, 2009, “Fast and accurate short read alignment with Burrows-Wheeler Transform,” *Bioinformatics* 25, pp. 1754-1760, which is hereby incorporated by reference) the SAM files sorted, and read group tags added using the PICARD tools. After sorting in coordinate order, the BAM’s were processed with PICARD MarkDuplicates. First realignment was carried out using the InDel realigner followed by base quality value recalibration with the BaseQRecalibrator.

[00357] A combination of four different mutation callers (Mutect 1.1.4, Somatic Sniper 1.0.4, Varscan 2.3.7, and Strelka 1.013) were used to identify single nucleotide variants (SNVs). *See* Wei *et al.*, 2015, “MAC: identifying and correcting annotation for multi-nucleotide variations, *BMC Genomics* 16, p. 569; Snyder and Chan, 2015, “Immunogenic peptide discovery in cancer genomes,” *Curr Opin Genet Dev* 30, pp. 7-16; Nielsen *et al.*, 2003, “Reliable prediction of T-cell epitopes using neural networks with novel sequence representations,” *Protein Sci* 12, pp. 1007-1017; and Shen and Seshan, 2016, “FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing,” *Nucleic Acids Res.* 44, e131, each of which is incorporated by reference. As previously described, SNVs with an allele read count of less than 4 or with corresponding normal coverage of less than 7 reads were filtered out. *See* Riaz *et al.*, 2016, “Recurrent SERPINB3 and SERPINB4 mutations in patients who respond to anti-CTLA4 immunotherapy,” *Nat. Genet.* 48, 1327-1329, which is hereby incorporated by reference.

[00358] The assignment of a somatic mutation to a neoantigen was estimated using a previously described bioinformatics tool called NASeek. *See* Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” *N. Engl. J. Med.* 371, pp. 2189-2199, which is hereby incorporated by reference. NASeek is a computational algorithm that first translates all mutations in exomes to strings of 17 amino acids, for both the wildtype and mutated sequences, with the amino acid resulting from the mutation centrally situated. Secondly, it evaluates putative MHC Class I binding for both wildtype and mutant nonamers using a sliding window method using NetMHC3.4 (on the Internet cbs.dtu.dk/services/NetMHC-3.4/) (*see* Andreatta and Nielsen, 2016, “Gapped sequence alignment using artificial neural networks: application to the MHC class I system,” *Bioinformatics* 32, pp. 511-517, which is hereby incorporated by reference) for patient-specific HLA types, to generate predicted binding affinities for both peptides. Prediction values are given in nM IC₅₀ values and are trained on nonamer peptides like those used in the disclosed analysis. NASeek finally assesses for similarity between nonamers that are

predicted to be presented by patient-specific MHC Class I. All nonamers with inferred affinities below 500 nM are defined as neoantigens.

[00359] *Clonal tree construction.* Tumor clones are reconstructed using the PhyloWGS software package. See, Deshwar *et al.*, 2015, “PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors,” *Genome Biol.* 16, 35, which is hereby incorporated by reference. The input data for the algorithm is extracted from exome sequencing data: (1) mutation reads obtained with the pipeline described above, and (2) allele-specific copy-number variant data, obtained with FACETS v0.5.0. Shen and Seshan, 2016, “FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing,” *Nucleic Acids Res.* 44, e131. The package clusters mutations into clones by the frequency of their reads and it infers possible nesting of clones (ancestral relations) between pairs of clones. Intuitively, an ancestral clone needs to have higher frequency than its derived clone. From this information PhyloWGS reconstructs high likelihood tumor genealogical trees.

[00360] In accordance with the present disclosure, the predicted relative size $n(\tau)$ of a cancer cell population in a tumor as a weighted sum over its genetic clones is calculated as:

$$n(\tau) = \sum_{\alpha} X_{\alpha} \exp(F_{\alpha} \tau),$$

where F_{α} is the fitness and X_{α} is the initial frequency of clone α and τ is a characteristic evolutionary time scale. The disclosed fitness model:

$$F_{\alpha} = - \max_{i \in \text{Clone } \alpha} (A_i \times R_i),$$

where index i iterates over all neoantigens in clone α was applied to three datasets: two melanoma patient cohorts treated with anti-CTLA4 (Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” *N. Engl. J. Med.* 371, pp. 2189-2199; and Van Allen *et al.*, 2015, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma,” *Science* 350, pp. 207-211, each of which is hereby incorporated by reference) and one lung tumor cohort treated with anti-PD1 (Rizvi *et al.*, 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer, *Science* 348, pp. 124-128, which is hereby incorporated by reference).

[00361] Efficiency is assessed using overall survival of patients from the beginning of immunotherapy. Neoantigen amino-acid anchor positions, 2 and 9, for the majority of HLA types, are constrained as reflected by non-informative MHC affinity amplitudes illustrated in

Figure 6(A). As illustrated in Figure 6 (B), positions 2 and 9 are highly constrained by a bias to be hydrophobic. As illustrated in Figure 31, neoantigens have decreased amino-acid diversity at these positions. *See also*, Lehman *et al.*, 2016, “Fundamental amino acid mass distributions and entropy costs in proteomes,” J. Theor. Biol. 410, pp. 119-124, which is hereby incorporated by reference. In Figure 31, the violin plots represent data density at a given value on a vertical axis. Figure 31A illustrates that neoantigens coming from mutations at position 2 or 9 tend to have wildtype peptides with larger predicted affinities. In particular, this is magnified if the corresponding wildtype residue is non-hydrophobic. Figure 31B illustrates that those biases are reflected in a wider distribution of amplitudes for wildtype peptides with non-hydrophobic residues at positions 2 and 9. Figure 31C illustrates that the Shannon entropy of amino acid diversity by position in neoantigens, shown for all distinct HLA-types and computed based on neoantigens across all datasets. Positions 2 and 9 have lower entropy than other residues. Other sites have the same entropy as the overall proteome (Lehmann *et al.*, 2016, “Fundamental amino acid mass distributions and entropy costs in proteomes,” J. Theor. Biol. 410, pp. 119-124) and are therefore unconstrained. Five HLA with non-canonical entropy profiles are singled out in the plot. These HLA types contributed only five informative neoantigens across all datasets and therefore are not treated differentially in our model. Hence, neoantigens with mutations at positions 2 and 9 were excluded from predictions with the model in this example. Amino-acid diversity on remaining positions is unconstrained as illustrated in Figure 6B.

[00362] Amino acid diversity at i^{th} position in a neoantigen is defined as e^{H_i} , where H_i is Shannon entropy of amino acid usage at this position, i.e.

$$H_i = -\sum_{j=1}^{20} f(a_{ij}) \log(f(a_{ij})),$$

where $f(a_{ij})$ is frequency of the i -th position in all neoantigens in a group. Inferred neoantigens are nonamers, so i ranges in value from 1 to 9. The diversity of neoantigens at a given site were compared to the values found in the human proteome in Lehman *et al.*, 2016, “Fundamental amino acid mass distributions and entropy costs in proteomes,” J. Theor. Biol. 410, pp. 119-124, which is hereby incorporated by reference. To calculate the expected number of words in the proteome the frequency of amino acids from Lehman, *et al* is utilized. The entropy associated with the frequency of amino acids in the human genome is computed as:

$$H(a) = -\sum_{j=1}^{20} f(a_j) \log(f(a_j)),$$

where $f(a_j)$ is the frequency of the j -th amino acid in the human genome. The expected number of words of length n is therefore $e^{nH(a)}$. This value is compared to the observed number of words of length n in the reference proteome for GRCh38.p7 using an entropy of 2.90 (Lehman *et al.*, 2016, “Fundamental amino acid mass distributions and entropy costs in proteomes,” J. Theor. Biol. 410, pp. 119-124). Finite genome size exhausts word usage between 5 and 6-mers. By 9-mer length words the ratio of observed to expected words is approximately 0.000052.

[00363] R , the recognition potential 224 of a neoantigen with a TCR-pool defined as the probability that a neoantigen cross-reacts with at least one TCR corresponding to a known immunogenic epitope:

$$R = 1 - \prod_{\mathbf{e} \in \text{IEDB}} [1 - \text{Pr}_{\text{binding}}(\mathbf{s}, \mathbf{e})]$$

where $\mathbf{s}$ is the peptide sequence of the neoantigen and $\mathbf{e}$ is a set of epitopes of the neoantigen given by the Immune Epitope Database and Analysis Resource. See Vita *et al.*, 2014, “The immune epitope database (IEDB) 3.0,” Nucleic Acids Res. 43, D405-D412, which is hereby incorporated by reference. The probability that a TCR for a given epitope binds a given neoantigen is defined by a two-state thermodynamic model with logistic shape. In this model sequence alignment is used as a proxy for binding energy. To assess homology between a neoantigen with peptide sequence $\mathbf{s}$ and an IEDB epitope $\mathbf{e}$, we compute an alignment (*e.g.*, gapless, or an alignment that allows gaps with suitable gap introduction and extension penalties) between the two sequences with a BLOSUM62 amino-acid similarity matrix. See Berg and von Hippel, 1987, “Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters,” J. Mol. Biol. 193, pp. 723-743, which is hereby incorporated by reference. For an alignment score, $|\mathbf{s}, \mathbf{e}|$, the binding probability is computed as

$$\text{Pr}_{\text{binding}}(\mathbf{s}, \mathbf{e}) = \frac{1}{1 + e^{-k(|\mathbf{s}, \mathbf{e}| - a)}}$$

where a represents the horizontal displacement of the binding curve (midpoint) and k sets the steepness of the curve at a . Model parameters a and k of the logistic binding function describes the probability of binding between neoantigens and epitope-specific TCRs.

[00364] To choose model parameters a and k for this equation, in some embodiments the log-rank-test scores of patient segregation as a function of these parameters was

investigated. The survival analysis is performed by splitting patient cohort into *high* and *low fitness* groups by the median cohort value of $n(\tau)$, the predicted relative cancer cell population size at a characteristic time τ (we discuss the choice of τ below). The survival score landscapes (Fig. 7 and Fig. 17) appear to be consistent between the datasets, with an optimal value of parameter a around 23 and parameter k living on a trivial axis above value 1, suggesting strong nonlinear fitness dependence on the sequence alignment score. Parameters that optimize the log-rank-test score in the largest dataset in the study, the melanoma anti-CTLA4 cohort from Van Allen *et al.*, 2015, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma,” Science 350, pp. 207-211, hereby incorporated by reference were chosen.

[00365] Parameter τ , a characteristic evolutionary time scale for a patient cohort, is a finite value at which cancer populations from tumors are expected to have responded to therapy. Intuitively, this is the time at which samples are expected to have a resolved heterogeneity, with their highest fitness clone dominating the evolutionary dynamics. In other words, the parameter τ in the equation:

$$n(\tau) = \sum_{\alpha} X_{\alpha} \exp(F_{\alpha} \tau)$$

sets the characteristic time scale of response to therapy. At this time, clones with dominant neoantigens having amplitudes larger than $1/\tau$ will have been suppressed. In some embodiments, to estimate τ , in the survival analysis the samples are split by the median cohort value $n(\tau)$ at a specified time scale τ . Intuitively, this time should be set to a finite value at which the tumors are expected to have responded to therapy. At this value of τ the clonal heterogeneity of tumors is supposed to have decreased, with the highest fitness clone dominating in the population. For one tumor this time scale is inversely proportional to the standard deviation of intra-tumor fitness (i.e. of the order of $1/\sigma(F)$), where

$$\sigma^2(F) = \sum_{\alpha} X_{\alpha} F_{\alpha}^2 - \left(\sum_{\alpha} X_{\alpha} F_{\alpha} \right)^2$$

In each cohort the interval of characteristic times of heterogeneous samples was determined. See, Figures 11, 12, and 13, the dependence of the prediction power on τ was tested by performing log-rank tests. See, Figures 14, 15, and 16. The optimal values of τ in each cohort belonged to a relatively wide interval. The consistent broadness of these intervals

suggests low sensitivity of predictive power on τ . Moreover, the parameter intervals giving highly significant patient segregation were also consistent between the cohorts.

[00366] Heterogeneous samples were selected with criterion $e^{H_F} \geq 2$, where H_F is clonal fitness entropy defined as

$$H_F = - \sum_{\beta} Y_{\beta} \log Y_{\beta},$$

where the frequencies of clones with the same fitness are added together and denoted as Y_{β} . The index β then refers to all clones with a given fitness.

[00367] Turning to Figure 17, the survival landscape is defined by the log-rank test score as a function of the model parameters for the logistic curve shape, e.g. the midpoint (a) and steepness (k). The locally smoothed landscape is plotted for the Figure 17 (A) Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” N. Engl. J. Med. 371, pp. 2189-2199, and Figure 17 (B) Rizvi *et al.*, 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer, Science 348, pp. 124-128, datasets. An X marks the optimal parameters from Van Allen *et al.*, 2015, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma,” Science 350, pp. 207-211., $a = 23$ and $k = 1$ (cf. Figure 7), which are used to derive survival curves for these two datasets and are at high score regions of the landscapes.

[00368] In some embodiments, the survival log-rank test score is maximized to fit the binding curve parameters to the data on the largest dataset for Van Allen *et al.* (Van Allen *et al.*, 2015, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma,” Science 350, 207-211, which is hereby incorporated by reference) (103 metastatic patients). The same parameters are then used for validation in the two smaller datasets from Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” N. Engl. J. Med. 371, pp. 2189-2199, and Rizvi *et al.*, 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer, Science 348, pp. 124-128, (64 and 34 patients respectively). See Figure 7 and Figures 18-21. The parameters from Van Allen *et al.*, yield consistently high survival log-rank scores in all three datasets. When using these logistic function parameters in all three datasets, the binding probability of 0.5 is obtained by alignments of average length of 6.55 amino acids; it is 6.98 amino acids for binding probability above 0.95.

[00369] The predicted evolutionary dynamics of tumors naturally separates long- and short-term survivors (therapy responders and non-responders) in the datasets, using patient classifications defined in the original studies, as illustrated in Figures 8, 9, and 10. Long-term survivors (patients with survival time longer than two years in the Van Allen *et al.* and Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” N. Engl. J. Med. 371, pp. 2189-2199 datasets, and one year in Rizvi *et al.*, 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer, Science 348, pp. 124-128 dataset) are predicted to have faster decreasing relative population sizes $n(\tau)$ across a broad ranges of τ values. The fitness of a cancer cell in a genetic clone α is its expected replication rate, i.e.

$$\frac{dN_\alpha}{d\tau} = F_\alpha N_\alpha$$

where N_α is the population size of clone α . Checkpoint-blockade immunotherapy introduces a strong selection challenge, which is expected to overshadow pre-therapy fitness effects in a productive response. For a given clone α the dynamics of its absolute size are hence given by $N_\alpha(\tau) = N_\alpha(0)\exp(F_\alpha\tau)$, and the total cancer cell population size is computed as a sum over its clones:

$$N(\tau) = \sum_\alpha N_\alpha(\tau) = \sum_\alpha N_\alpha(0)\exp(F_\alpha\tau).$$

As the diagnostic of survival the relative population size $n(\tau) = N(\tau)/N(0)$ is used, which compares the predicted evolved population size after a characteristic time scale of evolution τ (discussed below) to the initial pretreatment size $N(0)$. The initial clone α frequency is denoted $X_\alpha = N_\alpha(0)/N(0)$, these frequencies are inferred from bulk exome reads from a tumor sample. See Deshwar *et al.*, 2015, “PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors,” Genome Biol. 16, p. 35, which is hereby incorporated by reference. Hence, to compute $n(\tau)$ only estimates of the initial frequencies and fitness values for each clone are required, as shown in the following equation

$$n(\tau) = \sum_\alpha X_\alpha \exp(F_\alpha\tau),$$

Thus, the absolute population size measurements are not needed. In this equation, and, referring to Figure 4, the predicted relative size $n(\tau)$ of a cancer cell population in a tumor is computed as a weighted sum over its genetic clones, where F_α is the fitness and X_α is the initial frequency of clone α and τ is a characteristic evolutionary time scale. For each such

model defined, its homogenous structure equivalent can be defined by assuming the tumor is strictly clonal with all neoantigens in the same clone at frequency 1.

[00370] Moreover, the neoantigen recognition potential fitness model

$$F_{\alpha} = - \max_{i \in \text{Clone } \alpha} (A_i \times R_i),$$

results in highly significant separation of patients in survival analysis of all three datasets (Figures 11-16). The median value of $n(\tau)$ was used to separate patients into high and low predicted response groups. Using the median as opposed to an optimized threshold (*See*, Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” N. Engl. J. Med. 371, pp. 2189-2199; Van Allen *et al.*, 2015, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma,” Science 350, pp. 207-211; and McGranahan *et al.*, 2016, “Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade,” Science 351, pp. 1463-1469, each of which is hereby incorporated by reference) prevents overfitting and allows for robust validation. Log-rank test p -values are $p = 0.04$ for the Van Allen *et al.* (Figure 11), $p = 0.0026$ for Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” N. Engl. J. Med. 371, pp. 2189-2199 (Figure 12), and $p = 0.00624$ for Rizvi *et al.*, 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer, Science 348, pp. 124-128 (Figure 13). For comparison, a model considering only total neoantigen burden, is significant only for Rizvi *et al.*, 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer, Science 348, pp. 124-128 ($p=0.007$), when also using unsupervised median partitioning of patients. An alternative neoantigen load model that only accounts for clonal structure was also used. Again, only the Rizvi *et al.*, 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer,” Science 348, pp. 124-128 cohort has a significant patient survival separation ($p=0.03$).

[00371] *Parameter training.* In some embodiments, the model has two other free parameters: the midpoint and steepness defining R . In some embodiments, these parameters are inferred by maximizing the survival log-rank test score on independent training data. In some embodiments, to choose model parameters a and k in the equation:

$$R = Z(k)^{-1} \sum_{e \in \text{IEDB}} \exp [-k(a - |\mathbf{s}, \mathbf{e}|)],$$

and the characteristic time τ at which the prediction is evaluated in the equations:

$$F_{\alpha} = - \max_{i \in \text{Clone } \alpha} (A_i \times R_i)$$

and

$$n(\tau) = \sum_{\alpha} X_{\alpha} \exp[- \max_{i \in \text{Clone } \alpha} (A_i \times R_i) \tau]$$

the parameters that maximize log-rank-test scores of survival analysis on patient cohorts are identified. The survival analysis is performed by splitting patient cohort by the median value of $n(\tau)$ into *high* and *low fitness* groups. For each of the three cohorts, a parameter training is performed on independent data: the melanoma cohorts are used to train parameters for each other by using the maximal score of one to define parameters for the other, and both melanoma cohorts and maximization of their total log-rank test score is used to train parameters for the lung cohort. To infer consistent parameters between all datasets, the total log-rank test score over the three cohorts is maximized.

[00372] For a given training the optimal parameters $\hat{\Theta} = [\alpha, k, \tau]$ is computed, as an average $\hat{\Theta} = \langle \Theta \rangle_w$ over a distribution $w(\Theta)$ defined by the log-rank test score landscape on this set

$$w(\Theta) = Z^{-1}(\lambda) \exp[\lambda(S_{\max} - S(\Theta))],$$

where $Z(\lambda)$ is the probability distribution normalization constant, $S(\Theta)$ is the value of the log-rank test score with parameters Θ , and $S_{\max}$ is the maximal score value obtained over all possible parameters. The weight parameter λ is chosen such that the total statistical weight of the suboptimal parameter region is less than 0.01, the suboptimal scores are those less than $\max(3.841, S_{\max} - 2)$ (where 3.841 is the score value corresponding to 5% significance level of the log-rank test score). Using a smooth local neighborhood of parameters around the optimal values prevents over-fitting on a potentially rugged score landscape. For each individual parameter, the error bars reported in Figures 23-25, and Figures 32A, 32B, and 32C are computed as standard deviation using marginalized probability distribution $w(\Theta)$ for this parameter.

[00373] The survival score landscape, as illustrated in Figure 33, are consistent between the datasets. The optimal value of parameter α , the midpoint of the logistic binding function, is around 26 and parameter k , the steepness of the logistic function, lives on a trivial axis above value 4, suggesting strong nonlinear fitness dependence on the sequence alignment score.

[00374] We use the Snyder melanoma cohort with 64 patients to train parameters for the 103 metastatic patients in the Van Allen cohort and vice versa; we use the total score of both melanoma cohorts to train parameters for the smaller lung cancer cohort from Rizvi *et al.* with 34 patients (Methods). For each cohort, significant stratification of patients is obtained: log-rank test p-values are $p=0.0049$ for the Van Allen *et al.*, $p=0.0026$ for Snyder *et al.*, and $p=0.0062$ for Rizvi *et al.* (Figures 23-25). The parameters thereby obtained are consistent between datasets and mutually included within each other's error bars (Figures 23-25). A joint optimization of the cumulative log-rank test score of the three cohorts was further performed, obtaining a single set of parameters with predictions highly stable around these values (Figure 33). The alignment threshold parameter is consistently set to 26 (Figures 23-25), which in the datasets is obtained by alignments of average length of 6.8 amino-acids, just above the length of peptide motifs one would expect the TCR repertoire to discriminate. There are approximately 10^8 unique T-cell receptors in a given human (Arstila *et al.*, 1999, "A direct estimate of the human $\alpha\beta$ T cell receptor diversity," Science 286, pp. 958–961), and, moreover, the genome wide entropy of amino acid usage is approximately 2.9023. Therefore, one expects the length, L , of words TCRs can typically discriminate to be given by $10^8 \approx e^{2.90L}$ on average (as opposed to say 20^L if one assumed uniform genome amino acid usage). Solving for this length yields $L \approx 6.35$. The slope parameter is set to 4.87 defining a strongly nonlinear dependence on alignment score, with the recognition probability dropping below 0.01 for score 25 and reaching above 0.99 at score 27 (Figure 29). The τ parameter is set to 0.09, meaning clones with amplitudes larger than 11.1 are, on average, suppressed at prediction time. At these consistent parameters, separation of patients does not change for Van Allen *et al.* and Rizvi *et al.* (log-rank score increases by less than 1 unit, $p=0.004$ for Van Allen *et al.* and $p=0.0062$ for Rizvi *et al.*), and it improves to $p=0.00026$ for Snyder *et al.* (Figures 11-16). Patient segregation by $n(\tau)$ evaluated at infinitesimally small τ (equivalent to average tumor fitness over clones) is also significant (Figures 23-25, and 33), suggesting predictive power depends more on the model's ability to capture immune interactions than the duration of evolutionary projections. The performance of the model deteriorates when we disrupt the biological relevance of input data. When using the IEDB epitopes not supported by positive T-cell assays, the model loses predictive ability in both melanoma cohorts (Figures 23-25, and 30A-30F). Similarly, the model generally does not give significant patient separations when using neoantigens derived with randomized patient

HLA types (Figure 35, Example VI). The success of the disclosed model strongly depends on the joint contribution of A and R in the equation:

$$F_{\alpha} = - \max_{i \in \text{Clone } \alpha} (A_i \times R_i)$$

Partial models with only one component were constructed and the same training and validation procedure was repeated, with survival analysis separating patients into equal size groups as in the full model (Figs. 11-16, and 23-25). In all datasets, partial models have lower log-rank scores than the full model and neither A nor R -only models result in significant segregation for any cohort. We also compare our full model with a neoantigen load model, which assigns a uniform fitness cost to each neoantigen. This model does not significantly separate patients by median in either cohort (Figures 11-16, and 23-25).

[00375] *Model components.* The success of the disclosed neoantigen recognition potential fitness model depends on the joint contribution of two fitness components, the MHC presentation amplitude A and TCR recognition probability R in the equation:

$$F_{\alpha} = - \max_{i \in \text{Clone } \alpha} (A_i \times R_i)$$

where index i iterates over all neoantigens in respective clone α .

[00376] *MHC amplitude A .* The amplitude due to the dissociation constant between a neoantigen and its wildtype peptide is defined as:

$$A = K_D^{WT} / K_D^{MT}.$$

This is an approximate form derived with the use of equilibrium kinetics, where the concentration of peptide bound to MHC is given by their individual concentrations and inferred binding constant K_D , derived from NetMHC. *See, Andreatta et al.*, 2016, “Gapped sequence alignment using artificial neural networks: application to the MHC class I system, *Bioinformatics* 32, pp. 511-517, which is hereby incorporated by reference. The underlying dependencies are:

$$A = [\text{MHC: neoantigen}]^{MT} / [\text{MHC: neoantigen}]^{WT} \\ = (K_D^{WT} [\text{MHC}]^{MT} [\text{neoantigen}]^{MT}) / (K_D^{MT} [\text{MHC}]^{WT} [\text{neoantigen}]^{WT}),$$

where $[\text{MHC: neoantigen}]^{MT}$ is the concentration of the mutant form of the neoantigen to MHC, with the WT superscript representing the same quantity for the wild-type peptide. The above quantity is assumed to be dominated by the ratio of dissociation constants which

derives the formula for A in the equation $A = K_D^{WT}/K_D^{MT}$. A standard cutoff, $K_D^{MT} < 500$ nM, for the inferred mutant dissociation constant is used. See, Andreatta *et al.*, 2016, “Gapped sequence alignment using artificial neural networks: application to the MHC class I system, *Bioinformatics* 32, pp. 511-517, which is hereby incorporated by reference.

[00377] The predictive power of amplitude A , as illustrated in Figures 11-16, can be related to the fact that this quantity reflects the relative concentration of mutant to wildtype peptide and therefore the likelihood that the mutant peptide would be presented versus its wildtype peptide. As neoantigens with mutations on positions 2 and 9, are excluded, a high value of amplitude means the wildtype also likely already has hydrophobic residues at the anchor position and could be presented. Since neoantigens differ from their wildtype peptides by a single mutation, and given the uniqueness of nonamer sequences in the proteome, as illustrated in Figure 22, the self-nonamer in the genome with the greatest similarity to a neoantigen is likely to be its wildtype peptide. This was verified to be the case for 92% of all neoantigens, with the remainder largely emanating from gene families with many paralogs (Example II), implying a high amplitude usually stands for a self peptide not likely to be abundantly presented by the MHC. Therefore, as its immunogenicity is not mitigated by a homologous self-peptide, the mutant peptide with high affinity is likely to be novel to T-cells.

[00378] *TCR-recognition.* R is modeled as the recognition potential 224 of a neoantigen with a TCR-pool defined as the probability that a neoantigen cross-reacts with at least one TCR corresponding to a known immunogenic epitope. The recognition potential 224 of neoantigen is profiled *in silico* with a set of epitopes given by the Immune Epitope Database and Analysis Resource (IEDB). See Vita *et al.*, 2014, “The immune epitope database (IEDB) 3.0,” *Nucleic Acids Res.* 43, D405-D412, which is hereby incorporated by reference. Only IEDB epitopes that are positively recognized by T-cells after class I MHC presentation are used on the basis that a neoantigen that is predicted to cross-react with a TCR from this pool of immunogenic epitopes is a neoantigen more likely to be immunogenic itself.

[00379] The probability that a TCR for a given epitope binds a given neoantigen is defined by a two-state thermodynamic model with logistic shape. In this model we use sequence alignment as a proxy for binding energy. See Berg and von Hippel, 1987, “Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and

application to operators and promoters, J. Mol. Biol. 193, pp. 723-743, which is hereby incorporated by reference.

[00380] To assess homology between a neoantigen with peptide sequence **s** and an IEDB epitope **e**, an alignment (*e.g.*, gapless, or an alignment that allows gaps with suitable gap introduction and extension penalties) between the two sequences is computed with a BLOSUM62 amino-acid similarity matrix. *See* Henikoff and Henikoff, 1992, “Amino acid substitution matrices from protein blocks,” Proc. Natl. Acad. Sci. USA 89, pp. 10915–10919, which is hereby incorporated by reference. For an alignment score, $|s, e|$, the binding probability is computed as

$$\text{Pr}_{\text{binding}}(s, e) = \frac{1}{1 + e^{-k(|s, e| - a)}},$$

where a represents the horizontal displacement of the binding curve and k sets the steepness of the curve at a . These are two free parameters to be fit in the model. The parameters that are used in predictions in the instant example are $a=23$ and $k=1$; these parameters give binding probability $\text{Pr}_{\text{binding}}(s, e) = 0.5$ at alignment score $|s, e|=23$; the probability drops to below 0.05 at $|s, e|=20$ and reaches value of above 0.95 at $|s, e|=26$. See Figure 7. The corresponding alignment score span of 6 is close to the average identity match score in the BLOSUM62 matrix (5.64). The average alignment length corresponding to score 26 is 6.98 amino acids in the datasets and it is 6.55 for binding probability 0.5. The logistic function is therefore a strongly nonlinear function of the alignment score, where a mismatch on 1-2 positions can decide about lack of binding between the neoantigen and the epitope specific TCR.

[00381] For a given neoantigen **s** we calculate the probability it is recognized by a TCR within a repertoire as the probability it cross-reacts with at least one IEDB epitope:

$$R = 1 - \prod_{e \in \text{IEDB}} [1 - \text{Pr}_{\text{binding}}(s, e)].$$

[00382] The model

$$F_{\alpha} = - \max_{i \in \text{Clone } \alpha} (A_i \times R_i)$$

is decomposed by removing each of the components one at a time (Figures 14, 15, and 16). The MHC-only model, achieved by fixing $R_i = 1$, results in consistently worse segregation of patients (not significant in Snyder *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” N. Engl. J. Med. 371, pp. 2189-2199, decreased

significance in Van Allen, *et. al.*, and Rizvi *et al.*, 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer, *Science* 348, pp. 124-128, $p=0.033$ and $p=0.013$ respectively). The TCR-recognition-only model, achieved by fixing $A_i = 1$, does not result in a significant segregation in any cohort.

[00383] It is of interest to assess the clonal structure of a tumor when trying to identify dominant neoantigens. The importance of tumor clonal structure in our identification of dominant neoantigens. In all data sets, the full clonal model performs significantly better than an alternative model assuming homogenous tumor structure (Figures 11-16). Clonality appears less crucial in partial models, which have either marginal or no statistical significance (Figures 11-16, 32A, 32B, and 32C). Moreover, the full model is predictive independent of other clinical correlates (Proportional Hazard model, Figure 36).

[00384] The disclosed framework allows for straightforward incorporation of information about the tumor’s microenvironment. For the cohort from Van Allen *et al.*, gene expression data is available on 40 patients and local cytolytic activity is significantly associated with benefit ($p=0.04$, Example IV, below), as also observed in the original study by Van Allen *et al.*, 2015, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma,” *Science* 350, pp. 207-211. As a proof of principle, cytolytic score (Rooney *et al.*, 2015, “Molecular and genetic properties of tumors associated with local immune cytolytic activity,” *Cell* 160, pp. 48-61) was incorporated as an amplitude multiplying the T-cell recognition probability. Its inclusion improves predictions on these 40 patients, as assessed with survival analysis, ($p=0.043$ and $p=0.0025$ respectively, Figure 34).

[00385] In a broader context, the model suggests strong similarities in the evolution of cancers and fast-evolving pathogens. In both systems, immune interactions govern the dynamics of a genetically heterogeneous population; fitness models can predict important aspects of these dynamics, as recently shown for human influenza A. *See*, Łuksza and Lässig, 2014, “Predictive fitness model for influenza,” *Nature* 507, pp. 57-61, which is hereby incorporated by reference. Yet there are important differences. Influenza evolution is determined by antigenic similarity with previous strains in the same lineage whereas cancer cells acquire somatic mutations in a large set of proteins. Hence, their immune interactions are distributed in a larger and less homogenous antigenic space. The fitness effects of these interactions have a specific interpretation: they capture neoantigen “non-selfness.” The model formalizes what makes a tumor immunologically different from its hosts, analogously to models for innate recognition of non-self nucleic acids. *See*, Tanne *et al.*, 2015,

Distinguishing the immunostimulatory properties of noncoding RNAs expressed in cancer cells.” Proc.Natl. Acad. Sci. USA 112, pp. 5154-15159, which is hereby incorporated by reference.

[00386] The disclosed approach naturally extends to other fitness effects, such as positive selection due to acquisition of driver mutations, the impact of additional components in the microenvironment, or the hypothesized role of the microbiome. *See, Vétizou et al.*, 2015, “Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota.” *Science* 350, pp. 1079–1084; and Dubin *et al.*, 2016, “Intestinal microbiome analyses identify melanoma patients at risk for checkpoint-blockade-induced colitis.” *Nat. Commun.* 7, p. 10391 (2016). Further advances in predicting proteosomal processing (Abelin *et al.*, 2017, Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction.” *Immunity* 46, pp. 315-326) and stability (Ströner *et al.*, 2016, “Targeting of cancer neoantigens with donor-derived T cell receptor repertoires,” *Science* 352, pp. 1337-1341) of neoantigen-MHC binding could improve predictions. The disclosed approach is also useful in studies on acquired resistance to therapy. Moreover, this insight may be crucial for understanding when cross-reactivity with self-peptides may result in side effects. *See Johnson et al.* 2016, “Fulminant myocarditis with combination immune checkpoint blockade, *New Engl. J. Med.* 375, pp. 1749-1755; and Hofmann *et al.*, 2016, “Cutaneous, gastrointestinal, hepatic, endocrine, and renal side-effects of anti-PD-1 therapy,” *European J. Cancer* 60, 190-209, each of which is hereby incorporated by reference. Because the proposed neoantigen recognition potential fitness model is based on specific interactions underlying the presentation and recognition of neoantigens, it may also inform the choice of therapeutic targets for tumor vaccine design.

EXAMPLE II

[00387] *Identification of closest nonamers in human proteome to neoantigens.* This example illustrates identification of closest nonamers in human proteome to neoantigens. The wildtype (WT, 9 matches) and mismatched type (MT, 8 matches, 1 mismatch) 9-mer peptides to all proteins in the current human reference genome (GRCh38.p7) with at least 8 out of 9 matches and no gaps (allowing only mismatches) were mapped using LAST (version 819) (*see Kielbasa et al.*, 2011, “Adaptive seeds tame genomic sequence comparison,” *Genome Res.* 21, pp. 487-493, which is hereby incorporated by reference) with the following parameters:

- `blast -f BlastTab -j1 -r2 -q1 -e15 -y2 -m1000000000 -l4 -L4 -P0`

where 9-mer mapping with at most one mismatch is guaranteed to have a matching 4-mer word.

[00388] One expects the mutated peptide to only map to the same location as the wildtype peptide, wildtype mapping exactly (9 matches) and MT mapping with one mismatch (8 matches). The expected case is that the wildtype peptide maps to the proteome exactly and the MT peptide maps to the proteome with one mismatch and only to the loci wildtype peptide maps to.

[00389] This rule can be violated in the following cases, sorted from the most to the least severe:

[00390] 1. WT peptide does not map to the proteome exactly. Some possible reasons are: a difference in the reference assemblies used for mutation calling and peptide mapping, a germline mutation mistakenly identified as somatic, or a difference between the patient genome and the reference genome used for alignments.

[00391] 2. WT peptide maps to the proteome exactly (9 matches), MT peptide maps to the proteome exactly (9 matches) but to a different locus.

[00392] 3. WT peptide maps to the proteome exactly, MT peptide maps to the proteome with one mismatch; however, MT peptide maps with one mismatch to the subjects WT does not map exactly.

[00393] 4. WT peptide maps to the proteome exactly, MT peptide maps to the proteome with one mismatch; however, MT peptide maps with one mismatch to a different locus on the gene WT maps to.

[00394] Each peptide was examined for the worst possible scenario, going from category 1 to 4 in the list. Category 1 indicates a difference in the reference genome. Categories 2-4 typically are due to mutations that occur in repetitive gene families with many paralogs. Once a peptide was identified as belonging to any category, it was excluded from further considerations. In this way, the numbers of peptides in each category add up to the total number of peptides. Below is a summary for the different datasets utilized in this study:

[00395] Van Allen, *et al.*, 2015, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma, Science 350, pp. 207-211, hereby incorporated by reference:

[00396] 39373 total peptides,

- [00397] (category 1)
- [00398] 42 WT unmapped, leaving 39331
- [00399] 36783 expected peptides (93.42%),
- [00400] (category 2) 387 have 9 matches in MT,
- [00401] (category 3) 2076 have other alignments, and
- [00402] (category 4) 85 have other alignments to the same subject;
- [00403] Snyder, *et al.*, 2014, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma, N. Engl. J. Med. 371, pp. 2189-2199, hereby incorporated by reference:
- [00404] 29781 total peptides,
- [00405] (category 1)
- [00406] 35 WT unmapped, leaving 29746
- [00407] 27674 expected peptides (92.93%),
- [00408] (category 2) 361 have 9 matches in MT,
- [00409] (category 3) 1644 have other alignments, and
- [00410] (category 4) 67 have other alignments to the same subject; and
- [00411] Rizvi, *et al.*, 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer,” Science 348, pp. 124-128 (2015), hereby incorporated by reference:
- [00412] 5581 total peptides,
- [00413] (category 1)
- [00414] 6 WT unmapped, leaving 5575
- [00415] 5125 expected peptides (91.83%),
- [00416] (category 2) 105 have 9 matches in MT,
- [00417] (category 3) 323 have other alignments, and
- [00418] (category 4) 22 have other alignments to the same subject.

EXAMPLE III

[00419] In some embodiments, amplitude A is computed as follows. MHC-binding probabilities are derived from the dissociation constants, which are themselves estimated from computationally predicted binding affinities, as justified in the present disclosure. Affinities are inferred for each peptide sequence and patient HLA type (Andreatta and Nielsen, 2016, *Bioinformatics* 32, p. 511); all mutant peptide sequences considered as neoantigens meet a standard 500 nM cutoff for their affinities. NetMHC 3.4 occasionally predicts affinities with very high values where training may be limited, and creating small denominators that can inflate the amplitude. In melanoma and lung cancer a high mutational burden inflates the frequency of such events. As a remedy, a pseudocount, ε , is introduced so that, for both mutant and wildtype peptides $P_U/P_B \rightarrow (P_U + \varepsilon)/(P_B + \varepsilon)$. In this case the new dissociation constant divided by peptide concentration becomes

$$\frac{K_d/[L] + \varepsilon(1 + K_d/[L])}{1 + \varepsilon(1 + K_d/[L])} \approx \frac{K_d/[L]}{1 + \varepsilon K_d/[L]}$$

for small ε , where K_d was the original dissociation constant and $[L]$ is the peptide concentration. Consequently $1/\varepsilon$ sets a scale at which dissociation constants are not reliable for large K_d at a given concentration. To fix these scales, it is noted that assays to determine dissociation constants for peptide-MHC binding are typically performed at 0.1-1 nM where the ligand concentration is typically small compared to the dissociation constant. *See*, Paul, 2013, *J. Immunol.* 191, p. 5831. In this regime, affinities can be interpreted as dissociation constants and 3687 nM is the outer range of predictability for the assays upon which NetMHC 3.4 is trained at no more than unit peptide concentrations. $\varepsilon/[L]$ is therefore chosen to be $0.0003 \approx 1/3687$ across datasets.

[00420] In embodiments where the affinity is less than 500 nM for the mutant peptide this correction is only relevant for the wildtype peptides. The corrected amplitude then becomes:

$$A \approx \frac{K_d^{WT}}{K_d^{MT}} \cdot \frac{1}{1 + \left(\varepsilon/[L]\right) \cdot K_d^{WT}}$$

The amplitude in this form, combined with the TCR-recognition term, has a high predictive value for patient survival predictions (Figures 8-10), consistently over the three patient cohorts, which is not the case of either the mutant or wildtype dissociation constants on their own. Figures 23-25).

EXAMPLE IV

[00421] *Inclusion of Microenvironment and Proteosomal Processing in the Neoantigen Recognition Potential Fitness Model.* The role of the microenvironment in the likelihood of productive T-cell recognition of tumor neoantigens can be incorporated in a natural manner into the disclosed modeling framework. We utilize the cytolytic score (CYT), the geometric mean of the transcript per kilobase million of perforin and granzyme (Rooney *et al.*, 2015, “Molecular and genetic properties of tumors associated with local immune cytolytic activity,” Cell 160, pp. 48-61. We do so for the 40 patients from the Van Allen, *et al.*, anti-CTLA4 melanoma dataset, which have matched genome and transcriptome sequencing and where CYT had shown predictive value. For this set we also derive the CD8 T-cell fraction using CIBERSORT. *See*, Newman *et al.*, 2015, “Robust enumeration of cell subsets from tissue expression profiles. Nature Methods 12, pp. 453-457). The two values have a Pearson correlation coefficient of 0.938. Given their encapsulation of similar information we used CYT as it had previously been shown to give significant segregation of patient benefit. *See* Van Allen, *et al.*, 2015, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma,” Science 350, pp. 207-211. The score provides an additional amplitude A_{CYT} and the recognition potential becomes $A_{CYT} \times A \times R$. Therefore, the cytolytic score amplifies the recognition potential by the degree of cytolytic activity. We attempted to include proteosomal processing into our model as an additional criterion, as evaluated with NetCHOP. *See*, Nielsen *et al.*, 2005, “The role of the proteasome in generating cytotoxic T cell epitopes: Insights obtained from improved predictions of proteasomal cleavage,” Immunogenetics 57, pp. 33-41. We tested this procedure on the Rizvi *et al.* cohort; however, the imposed stronger filtering of neoantigens leads to the loss of predictive power of the model.

EXAMPLE V

[00422] *IEDB sequences.* The predictive value of R depends on the input set of IEDB sequences. In some embodiments, the set used in the disclosed analysis contained 2552 unique epitopes. We tested how the predictions depend on the content and size of the dataset by performing iterative subsampling of IEDB sequences at frequencies varying from 10% to 90% of the total set size. We repeated the survival analysis and log-rank test score evaluation

(Figures 31-36). For all three datasets removal of sequences has on average a negative impact on their predictive power, which monotonically decreases with the subsampling rate. In the Van Allen *et al.* cohort median performance was below significance already at 70% subsampling and lower, and for Snyder *et al.* and Rizvi *et al.* at 20% and lower. To investigate the biological input associated with the set of curated IEDB sequences that we use, we also evaluated the R component using an alternative set of IEDB sequences, coming from T-cell assays that did not have a positive validation. This is a larger set of 4657 sequences. In the two melanoma datasets, the predictions have gotten worse, not giving significant separation of patients in the survival analysis. This effect was also not due to the different sequence set size – subsampling of sequences did not improve the outcome. While in the Rizvi *et al.* dataset the predictions were still significant, this significance was not supported by consistency between all three datasets which is observed on the IEDB sequence set with positive assays.

EXAMPLE VI

[00423] *Alternative fitness models.* The full neoantigen recognition potential fitness model

$$n(\tau) = \sum_{\alpha} X_{\alpha} \exp\left[- \max_{i \in \text{Clone } \alpha} (A_i \times R_i) \tau\right].$$

is compared to alternative models using model decompositions, where only one component is used

$$F_{\alpha} = - \max_{i \in \text{Clone } \alpha} (A_i),$$

$$F_{\alpha} = - \max_{i \in \text{Clone } \alpha} (R_i).$$

Further, the amplitude $A = K_d^{WT} \times \frac{1}{K_d^{MT}}$ is decomposed and various variants of the model tested, with and without the R component,

$$F_{\alpha} = - \max_{i \in \text{Clone } \alpha} K_d^{WT} (\times R_i),$$

$$F_{\alpha} = - \max_{i \in \text{Clone } \alpha} \frac{1}{K_d^{MT}} (\times R_i).$$

How informative the alignments contributing to the R_i components are is investigated and a model where alignments are restricted to the 6 residues in-between anchor positions, on positions 3-8, is also tested. The loss of predictive power of a model that does not implement any filtering of neoantigens mutated on position 2 and 9 is also demonstrated (see Figure 31).

The problem of choosing the neoantigen aggregating function is reduced to that of model selection. A model where fitness is defined by the total effect of all neoantigens in the clone (which is the limit case of $\beta = 0$ in the equation:

$$n(\tau, \beta) = \sum_{\alpha} X_{\alpha} \exp \left[\sum_{i \in \text{Clone } \alpha}^{\max} \frac{\exp(-\beta f_i)}{Z(\beta)} f_i \tau \right].$$

is also tested as:

$$F_{\alpha} = - \sum_{i \in \text{Clone } \alpha} A_i \times R_i.$$

Finally, a fitness model is formulated that associates a constant fitness cost with each neoantigen,

$$F_{\alpha} = -L_{\alpha},$$

where L_{α} is the number of neoantigens in clone α , referred to as the neoantigen load of clone α .

EXAMPLE VII

[00424] *Homogenous structure models.* For each fitness model, its homogenous structure equivalent can be defined, which assumes a tumor is strictly clonal with all neoantigens in the same clone at frequency 1. In a homogenous model the population size is thus modeled by an exponential,

$$n(\tau) = \exp[F\tau],$$

where F is the fitness function of the homogenous tumor. Since in this model tumors show a constant decay over time, ranking of $n(\tau)$ values for patients is defined only by fitness and does not depend on τ . Therefore, τ is not a free parameter in these models when optimizing log-rank test score in survival analysis.

EXAMPLE VIII

[00425] *Average fitness.* The average fitness of clones was also investigated:

$$\langle F \rangle = \sum_{\alpha} X_{\alpha} F_{\alpha},$$

as a predictive marker for patients and an alternative to $n(\tau)$. The average fitness reflects the rate at which the tumor cell population is decreasing in size at the beginning of therapy. For the purpose of patient ranking, it is equivalent to $n(\tau)$ at infinitesimally small values of the time parameter τ . This is a lower complexity model because τ is not a free parameter.

However, this measure is less robust to outliers - small clones with very low fitness can dominate the average fitness, while the evolutionary projection in $n(\tau)$ removes such effects.

EXAMPLE IX

[00426] *Comparison with Thresholded Neoantigen Load.*

[00427] In the disclosed survival analysis, a standard, non-optimized partitioning of patients into two equally sized groups by the median value of $n(\tau)$ was used. This approach allows for unbiased comparison of models, and assigns a stringent predictive value. The disclosed results do not contradict the earlier reported predictive quality of neoantigen load. Consistent with Snyder *et al.*, a significant split at a threshold value of 100 neoantigens or less is observed. This threshold classifies more than 70% of the patients in a long-term surviving group; separation by total neoantigen load is not significant at lower fractional partitions, including the median. In Van Allen *et al.*, survival analysis was not originally presented and we did not see a significant separation of patients at any possible splitting by a neoantigen load threshold. Finally, the significant separation for the Rizvi *et al.* cohort is observed for the range a 32-50% range of partitions, including by the median (Figs 11-16, and 23-25). It is worth noting that for this cohort we use previously unpublished overall survival data, which differs from the progression free survival data used by the original study (Rizvi, *et al.* 2015, “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer,” Science 348, 124-128). In all cohorts, the disclosed neoantigen recognition potential fitness model and partitioning based on $n(\tau)$ measure give significant separations at a larger range of partitions: 40-60% for the Van Allen *et al.* cohort, above 40% for the Snyder *et al.* cohort and 47-80% for the Rizvi *et al.* cohort.

EXAMPLE X

[00428] *Introduction*

[00429] PDAC represents a class of solid tumors that has remained largely refractory to checkpoint immunotherapy. Although T cell immunity has been linked to this rare outcome⁷⁻¹⁰, the relevant tumor antigens remain unknown. Immune checkpoint blockade can amplify spontaneous T cell recognition of neoantigens that arise in cancer-specific mutations^{11,12}, and elicit responses in heavily mutated tumors¹³⁻¹⁵. However, it is unclear if primary checkpoint blockade resistance in tumors with fewer mutations occurs due to diminished endogenous T cell reactivity to neoantigens.

[00430] *Results*

[00431] To define the significance of neoantigens in these patients, stage-matched cohorts of treatment-naïve, surgically resected, rare long term survivors (median survival 6 years) was compare to short term survivors with more typical poor outcome (median survival 0.8 years) (Figure 42A, Tables 3-6, Figures 41A and 41B).

Table 3 Clinicopathologic characteristics of patients in tissue microarray cohort.

Variable	Short Term (n=45) n (%)	Long Term (n=51) n (%)	P- value
Gender			
Male	24 (53)	26 (51)	NS
Female	21 (47)	25 (49)	
Age			
Median (Range)	78 (54-91)	74 (38-95)	NS
Tumor Location			
Head	32 (71)	37 (73)	NS
Body/Tail	13 (29)	14 (27)	
Procedure			
Distal Pancreatectomy	13 (29)	13 (25)	NS
Pancreaticoduodenectomy	31 (69)	37 (73)	
Total Pancreatectomy	1 (2)	1 (2)	
Pathological Stage			
I	0 (0)	2 (4)	NS
II	38 (84)	47 (92)	
III	4 (9)	2 (4)	
IV*	3 (7)	0 (0)	
pT			
1	0 (0)	0 (0)	NS
2	0 (0)	2 (4)	
3	42 (96)	47 (92)	
4	3 (4)	2 (4)	
pN			
0	11 (24)	21 (41)	NS
1	34 (76)	30 (59)	
pM			
0	42 (93)	51 (100)	NS
1*	3 (7)	0 (0)	
Margin			
Positive	8 (18)	2 (4)	0.03
Negative	37 (82)	49 (96)	
Adjuvant Treatment			
Yes	24 (53)	42 (82)	0.0077
No	20 (45)	9 (18)	
Unknown	1 (2)	0 (0)	

* One patient had liver metastasis noted on final pathology.

One patient had distal pancreatectomy and metastasis to small bowel/mesentery.

One patient had splenic metastasis.

Table 4 Clinicopathologic characteristics of patients 829 in transcriptome profiling cohort.

Variable	Short Term (n=15) n (%)	Long Term (n=15) n (%)	P- value
Gender			
Male	8 (53)	5 (33)	NS
Female	7 (47)	10 (67)	
Age			
Median (Range)	76 (54-84)	65 (51-85)	NS
Tumor Location			
Head	10 (67)	12 (80)	NS
Body/Tail	5 (33)	3 (20)	
Procedure			
Distal Pancreatectomy	5 (33)	3 (20)	NS
Pancreaticoduodenectomy	10 (67)	12 (80)	
Pathological Stage			
I	0 (0)	0 (0)	NS
II	12 (80)	14 (93)	
III	1 (7)	1 (7)	
IV*	2 (13)	0 (0)	
pT			
1	0 (0)	0 (0)	NS
2	0 (0)	0 (0)	
3	13 (87)	14 (93)	
4	2 (13)	1 (7)	
pN			
0	5 (33)	6 (40)	NS
1	10 (67)	9 (60)	
pM			
0	13 (87)	15 (100)	NS
1*	2 (13)	0 (0)	
Margin			
Positive	4 (27)	1 (7)	NS
Negative	11 (73)	14 (93)	
Adjuvant Treatment			
Yes	10 (67)	13 (87)	NS
No	4 (26)	2 (13)	
Unknown	1 (7)	0 (0)	

* One patient had distal pancreatectomy and metastasis to small bowel/mesentery.
One patient had liver metastasis noted on final pathology.

Table 5 Clinicopathologic characteristics 836 of patients in TCR sequencing cohort

Variable	Short Term (n=30) n (%)	LT (n=30) n (%)	P- value
Gender			
Male	17 (57)	12 (40)	NS
Female	13 (43)	18 (60)	
Age			
Median (Range)	73 (45-91)	75 (54-95)	NS
Tumor Location			
Head	17 (57)	23 (77)	NS
Body/Tail	13 (43)	7 (23)	
Procedure			
Distal Pancreatectomy	13 (43)	7 (23)	NS
Pancreaticoduodenectomy	17 (57)	23 (77)	
Pathological Stage			
I	0 (0)	2 (7)	NS
II	26 (87)	27 (90)	
III	1 (3)	1 (3)	
IV*	3 (10)	0 (0)	
pT			
1	0 (0)	2 (7)	NS
2	0 (0)	0 (0)	
3	28 (93)	27 (90)	
4	2 (7)	1 (3)	
pN			
0	14 (47)	17 (57)	NS
1	16 (53)	13 (43)	
pM			
0	28 (93)	30 (100)	NS
1*	2 (7)	0 (0)	
Margin			
Positive	1 (3)	2 (7)	NS
Negative	29 (97)	28 (93)	
Adjuvant Treatment			
Yes	18 (60)	26 (86)	NS
No	9 (30)	2 (7)	
Unknown	3 (10)	2 (7)	

*One patient had liver metastasis noted on final pathology.

One patient had distal pancreatectomy and metastasis to small bowel/mesentery.

Table 6 Clinicopathologic characteristics of patients 843 in whole exome sequencing cohort.

Variable	Short Term (n=32) n (%)	Long Term (n=26) n (%)	P- value
Gender			
Male	15 (47)	8 (31)	NS
Female	17 (53)	18 (69)	
Age			
Median (Range)	73 (48-91)	75 (51-95)	NS
Tumor Location			
Head	18 (56)	21 (81)	NS
Body/Tail	14 (44)	5 (19)	
Procedure			
Distal Pancreatectomy	13 (41)	5 (19)	NS
Pancreaticoduodenectomy	19 (59)	21 (81)	
Pathological Stage			
I	0 (0)	1 (4)	NS
II	28 (88)	24 (92)	
III	1 (3)	1 (4)	
IV*	3 (9)	0 (0)	
pT			
1	0 (0)	1 (4)	NS
2	0 (0)	0 (0)	
3	29 (91)	24 (92)	
4	3 (9)	1 (4)	
pN			
0	9 (28)	11 (42)	NS
1	23 (72)	15 (58)	
pM			
0	30 (94)	26 (100)	NS
1*	2 (6)	0 (0)	
Margin			
Positive	5 (16)	3 (12)	NS
Negative	27 (84)	23 (88)	
Adjuvant Treatment			
Yes	22 (63)	24 (92)	NS
No	9 (28)	2 (8)	
Unknown	1 (9)	0 (0)	

* One patient had a distal pancreatectomy and metastasis to small bowel mesentery.
One patient had liver metastasis noted on final pathology.

Table 7 Clinicopathologic characteristics of patients in matched primary and metastatic tumor cohort

Variable	MUC16 Neoantigenic (n = 2), n (%)	Non- MUC16 Neoantigenic (n = 2), n (%)	P- value
Gender			
Male	0 (0)	1 (50)	NS
Female	2 (100)	1 (50)	
Age			
Median (Range)	71 (57-85)	52 (50-54)	NS
Pathological Stage			
I	0 (0)	0 (0)	NS
II	0 (0)	0 (0)	
III	1 (50)	0 (0)	
IV	1 (50)	2 (100)	
Chemotherapy			
Yes	2 (100)	1 (50)	NS
No	0 (0)	1 (50)	

Table 8

Variable	Long Term (n=7) n (%)	
Gender		
Male	1	(14)
Female	6	(86)
Age		
Median (Range)	73	(60-88)
Tumor Location		
Head	4	(57)
Body/Tail	3	(43)
Procedure		
Distal Pancreatectomy	3	(43)
Pancreaticoduodenectomy	4	(57)
Pathological Stage		
I	1	(14)
II	6	(86)
III	0	(0)
IV	0	(0)
pT		
1	1	(14)
2	0	(0)
3	6	(86)
4	0	(0)
pN		
0	5	(71)
1	2	(29)
pM		
0	7	(100)
1	0	(0)
Margin		
Positive	1	(14)
Negative	6	(86)
Adjuvant Treatment		
Yes	7	(100)
No	0	(0)
Unknown	0	(0)
Recurrence and Survival	Recurrence	Survival (years)
Patient 1	No	10.5
Patient 2	No	9.5
Patient 3	No	11.5
Patient 4	No	11.8
Patient 5	Yes	7.3
Patient 6	Yes	11.8
Patient 7	No	7.7

Clinicopathologic characteristics of very long term pancreatic cancer survivors.

[00432] In Using 9-parameter immunophenotyping with multiplexed immunohistochemical consecutive staining on single slides¹⁶ in tissue microarrays (TMAs), it was found greater densities of CD3⁺CD8⁺ T cells (3-fold), cytolytic CD3⁺CD8⁺Granzyme-B⁺ cells (12-fold), DC-LAMP⁺ mature dendritic cells, FoxP3⁺ regulatory T cells, and CD68⁺

macrophages in tumors of long term survivors, yet no differences in CD20⁺ B cells and MHC-I⁺ cells (Figures 42B-42E, Figure 43). Consistently, immunofluorescent phenotyping also revealed that CD8⁺ T cells were increased and conversely, CD4⁺ T cells were decreased in tumors of long term survivors (Figure 44). Tumor transcriptomic profiling revealed an immunogenic microenvironment in tumors of long term survivors, with upregulation of molecular markers of dendritic cells¹⁷ and antigen reactivity¹⁸⁻²⁰ (PD-1 and TIGIT), with downregulation of immunosuppressive markers (STAT3²¹⁻²²) (Figure 42F). T cell receptor (TCR) V β -chain sequencing demonstrated that intratumoral T cells were increased 5-fold compared to matched adjacent non-tumor pancreatic tissue (11.6% vs. 2.2%), and displayed a markedly polyclonal repertoire (Figure 42G, Figure 45). Strikingly, >94% of intratumoral T cell clones were unique to tumors, consistent with tumor specificity (Figure 42H). In order to further address intratumoral T cell specificity, flow cytometry on freshly isolated intratumoral T cells in unselected patients revealed upregulation of activation and memory markers compared to T cells in matched draining lymph nodes and blood (Figure 42I, Figure 46). Finally, tumors of long term survivors exhibited greater TCR repertoire diversity compared to tumors of short term survivors (Figure 42J). Importantly, the association of activated CD8⁺ T cells and survival was independent of clinicopathologic factors and adjuvant chemotherapy (Figure 42K). Collectively, these data identify an activated, polyclonal, tumor-specific T cell infiltrate in tumors of long term survivors, implying differential antigenic targets.

[00433] It has been posited that PDACs infrequently harbor neoantigens owing to a relatively low somatic mutational prevalence¹². However, these estimates have been confounded by stromal contamination, as more recent sequencing efforts utilizing techniques to maximize tumor DNA capture have identified a higher burden of somatic mutations^{23,24}. To determine the true neoantigen frequency in PDACs, whole exome sequencing on macrodissected tumor islands was performed, enabling isolation of DNA from specimens characterized by a high proportion of tumor cells (Figure 47)^{24,25}. Utilizing a previously developed pipeline for neoantigen prediction^{13,14}, a median of 61 missense mutations with 38 predicted neoantigens per tumor were detected (Figure 48A). Remarkably, patients with the highest predicted neoantigen number and either the greatest CD3⁺CD8⁺, CD3⁺CD8⁺Granzyme-B⁺, or polyclonal T cell repertoire, but not CD4⁺ T cell infiltrates, exhibited the longest survival (median survival not reached). These findings were corroborated using a second, independent, neoantigen prediction algorithm pVAC-Seq²⁶

(Figure 48B, Figure 49). It was observed that no survival differences in patients are stratified by highest mutational load combined with either the highest CD3⁺CD8⁺ infiltrates or repertoire clonality (Figure 50). It was also found that there is no differences in the two cohorts in predicted neoantigen quantity, driver or nonsynonymous mutation frequency, or associations of unique mutations with improved outcome (Figures 49-53)^{24,27}. As an orthogonal test of these findings, unsupervised hierarchical clustering of bulk tumor transcriptomic profiling demonstrated that tumors harboring the highest activated T cell infiltrates and predicted neoantigen number segregated to unique clusters (Figure 48C). Together, these data suggest that neoantigen immunogenicity/quality, and not purely quantity, modulate T cell responses and impact outcome in long term survivors.

[00434] Indeed, recent data has shown that T cell-recognized neoantigens can be selectively lost from the tumor cell population either by mutant allelic loss or overall reduced gene expression. Consistently, genes with high quality neoantigens evidenced a modest trend to lower mRNA expression compared to gene expression in the absence of high quality neoantigens (Figure 69A). To further explore this possibility of in vivo high quality neoantigen immunoediting, neoantigen clonal dynamics was examined on primary to metastatic tumor progression in one patient obtained through rapid autopsy. Of the three clones in the primary tumor, both clones with high quality neoantigens were lost in multiple metastatic samples, in contrast to the clone with a low quality neoantigen which was propagated to multiple metastatic sites (Figure 69B, Table 7). These findings suggest differential immune fitness of clones bearing high versus low quality neoantigens within the same primary tumor.

[00435] We next sought to detect in vivo T cell responses to high quality neoantigens. We identified 7 very long term PDAC survivors (median OS 10.5 years) that normally account for <2% of all PDAC patients (Table 8) and pulsed their peripheral blood mononuclear cells with antigens predicted by the quality model. Remarkably, we observed selective CD8⁺ T cell expansion and degranulation to neopeptides and cross reactive peptides but not WT peptides, with identical clones that significantly expanded with both neopeptides and cross reactive peptides in all patients (Figure 69C and 69D). Strikingly, in 5 of 7 patients, we identified neoantigen and microbial cross reactive peripheral T cell clones that 191 were also present in their respective archival primary tumors. Patient 3, alive and disease free 12 years after primary tumor removal illustrated the most extreme instance – 15 neoantigen and microbial cross reactive T cell clones that persisted in the peripheral blood were found in the

primary tumor, including the top T cell clone at an intratumoral rank frequency of 6.2% (Figure 69D). These results support the idea that the quality model identifies bonafide neoantigens targeted by T cells and that tumor-infiltrating T cells can recognize both cancer neoantigens and homologous non-cancer microbial antigens.

[00436] Next, neoantigen qualities that modulate differential immunogenicity was investigated. The theory of molecular mimicry postulates that TCRs generated against pathogens can cross react against non-pathogenic antigens, including tumor antigens. Cross reactivity between microbial and self-antigens has been documented in autoimmune diseases²⁸. Although early experimental evidence suggests that cross reactivity between microbial and tumor antigens can stimulate immunosurveillance, this conclusion remains speculative²⁹. It was hypothesized that neoantigen homology to microbial epitopes recognized by the human TCR repertoire would serve as a surrogate for differential neoantigen immunogenicity or “non-selfness”. Furthermore, it was theorized that enhanced immunogenicity would impart a negative evolutionary fitness cost to tumor cell clones expressing these neoantigens, thereby inducing immunoediting of these clones. To test this hypothesis, two fitness models were developed: a neoantigen-microbial epitope recognition potential model (also interchangeably termed herein a cross reactivity model), inspired by recent methods predicting human influenza viral evolution for vaccine selection³⁰ and as further disclosed in Łuksza *et al.*, 2017, “A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy,” *Nature* 551, 517-520, and a neoantigen load model, based on neoantigen number (Figure 54A). For the recognition potential model, sequence alignment scores were derived for each neoepitope to microbial epitopes with positive immune assays from the Immune Epitope Database (IEDB), and assigned a non-linear sigmoid dependence of alignment scores to TCR binding probability based on a two-state thermodynamic model. These binding probabilities were amplified by relative wild type and mutant peptide-MHC affinities to calculate recognition potential scores for every neoepitope (Figure 54B). The neoantigen with the maximum recognition potential score within a clone was defined as the immunodominant neoantigen. For the neoantigen load model, the neoantigen score equaled the total number of neoantigens within a clone. Finally, the clonal tree structure for each tumor based on mutant allele frequencies³¹ was recreated, and weighted contributions of each tumor clone to total immunogenicity in each model were summated. Of the resulting fitness functions, the recognition potential model significantly stratified short and long term survivors whereas the neoantigen load model did not (Figure

54C, Figure 56, Figure 54E). These findings were further confirmed by testing a larger cohort of patients unselected by survival (International Cancer Genome Consortium (ICGC); n=166). Consistent with the MSKCC cohort, neoantigen quality but not quantity, was strongly prognostic of survival independent of confounding variables (Figure 54F, Figure 54G), with a stable association with survival in subsampled datasets in both cohorts (Figure 54H). Notably, nearly all tumors with the highest neoantigen load in combination with the most abundant CD8⁺ T cell infiltrates harbored neoantigens with homology to microbial epitopes (Figure 54D). Hence, enhanced neoantigen homology to microbial epitopes can modulate immunogenicity, immunoselection, and outcome in long term survivors.

[00437] Although neoantigen formation is stochastic and private¹², the above outlined data suggested that unique neoantigen qualities might mediate preferential targeting. Therefore, further exploration was done on whether select genetic loci or “immunogenic hotspots” were preferentially enriched for neoantigens. Thresholds were applied to permit sufficient power to detect differences between patients with and without neoantigens in any given gene, and performed locus filtering. Four loci harboring neoantigens were detected in > 15% of all patients, with one locus preferentially enriched in long term survivors – the tumor antigen MUC16, a common ovarian cancer biomarker (CA125), and an established target for T cell immunotherapy³² (Figure 55A). In tumors of long term survivors, a 4-fold higher MUC16 neoantigen frequency (27% vs. 6%) and multiple MUC16 neoantigens were found in the same tumor (maximum 5) whereas non-antigenic MUC16 mutation frequency was no different (Figures 55B-55C, Figures 57-59). Only one patient with MUC16 neoantigens had a hypermutated phenotype (>500 mutations), and exclusion of this patient did not alter the results (Figure 55C). Consistently, the pVAC-Seq pipeline identified MUC16 as the most common genetic locus generating neoantigens, following the most frequently mutated oncogenes in PDAC (KRAS, TP53) (Figure 60). In support of possible in vivo anti-MUC16 immune responses, MUC16 was the most differentially expressed gene in the two cohorts, with long term survivors evidencing significantly lower mRNA (6.6-fold), protein, and mutant allele frequency in non-hypermutated tumors (4-fold) (Figures 55D-55F). It was found no evidence of non-immunogenic MUC16 mutations altering RNA or protein expression. Furthermore, there were no differences in the two cohorts in cell autonomous regulators of MUC16 expression, mediators of MUC16 dependent effects on tumor progression, or expression of other mucins, and tumor antigens (Figures 61-65)³³⁻³⁷. One potential interpretation of these findings is MUC16-neoantigen specific T cell immunity

induces immunoediting of MUC16-expressing clones in primary tumors, curtails the development of metastases, and prolongs survival, given the cell-autonomous roles of MUC16 in promoting metastases^{34,36-38}. This hypothesis is consistent with recent evidence of possible neoantigen immunoediting demonstrated by overall reduced gene expression and mutant allelic loss in the setting of T cell neoantigen reactivity. Notably, MUC16 expression was low yet not absent in tumors of long term survivors, indicating antigen availability (Figure 70). Consistent with possible MUC16 immunoediting, MUC16 neoantigens in primary tumors had complete neoantigenic mutational loss in matched metastases (n = 10) in contrast to MUC16 non-neoantigenic mutations that demonstrated mutation enrichment on metastatic progression (Table 7). MUC16 was also the only locus recurrently harboring neoantigens in both MSKCC and ICGC cohorts, outside of genes expected to do so based on high mutation frequency (oncogenes -KRAS, TP53; largest human gene -TTN). Although the propensity to generate MUC16 neoantigens may be related to its large size, we did not detect trends towards neoantigen formation based on gene size alone across cohorts or pipelines. Our results are also consistent with recent findings demonstrating that MHC-I-restricted peptides derived from selective regions of the human genome. Hence MUC16 is a candidate immunogenic hotspot.

[00438] We next stimulated peripheral blood from 2 long term survivors (both disease-free 8 years following surgery) with predicted MUC16 neoantigens to identify in vivo MUC16- neoantigen specific T cell immunity. In both patients, we observed CD8+ T cell expansion and degranulation, with expanded clones also detected in archival surgically resected primary tumors. We confirmed peripheral blood CD8+ T cell recognition of 2 additional MUC16 neoantigen-MHC complexes using peptide-MHC multimers in HLA-matched healthy donors (Figure 67), consistent with binding of putative MUC16 neoantigens by the human TCR repertoire. Hence we present evidence of in vivo T cell reactivity to neoantigens in the tumor antigen MUC16, with lasting MUC16-specific T cell immunity in PDAC survivors.

[00439] It is important to highlight that MUC16 expression was low yet not absent in tumors of long term survivors, supporting expression of MUC16 neoantigens. To further examine if MUC16 is an immunogenic hotspot in PDAC, neoantigens in a larger set of surgically resected PDAC unselected by survival (International Cancer Genome Consortium, n=169) were predicted²⁵. MUC16 was the only locus recurrently harboring neoantigens in both cohorts, outside of genes expected to do so (mutated oncogenic driver genes -KRAS,

TP53; largest human gene -TTN) (Figures 55G, 55H, 66). Although the propensity to generate neoantigens in MUC16 can be related to its large size, a general trend towards preferential neoantigen formation in genes based on size alone, either across cohorts or pipelines, was not detected. These results are also consistent with recent findings demonstrating that MHC-I restricted peptides derive from selective regions of the human genome³⁹. These data suggest that MUC16 is a candidate immunogenic hotspot in PDAC.

[00440] To identify the presence of *in vivo* MUC16-neoantigen specific T cell immunity in long term survivors, peripheral blood from 2 long term survivors (both disease-free 8 years following surgery) with predicted MUC16 neoantigens was stimulated. In both patients, over 30-fold CD8⁺ T cell expansion, and degranulation, with polyclonal T cell expansion selectively to mutant but not wild type MUC16 nonamers were observed. Remarkably, in both patients, expanding clones in the peripheral blood (defined by unique TCR V β sequences) were also detected in their respective archival primary tumors that were surgically resected 8 years prior, suggesting persistent MUC16-specific anti-tumor T cell immunity (Figures 55I-55K). Additionally, for 3 predicted MUC16 neoantigens, peripheral blood CD8⁺ T cell recognition of neoantigen-MHC complexes using peptide-MHC multimers in HLA-matched healthy donors were confirmed (Figure 55L, 67, 68), consistent with binding of predicted MUC16 neoantigens by the human TCR repertoire. Hence, the study presented evidence of *in vivo* T cell reactivity against mutational neoantigens in the tumor antigen MUC16, with long-lasting MUC16-specific T cell immunity in long term PDAC survivors.

[00441] Materials and Methods

[00442] Patient Samples

[00443] MSKCC PDAC cohort: All tissues were collected at Memorial Sloan Kettering Cancer Center under institutional review board protocols. Informed consent was obtained on all patients. All tumor samples were surgically resected primary pancreatic ductal adenocarcinomas. Patients treated with neoadjuvant therapy were excluded. All tumors were subjected to pathologic re-review and histologic confirmation by two expert PDAC pathologists prior to analyses. Long term survivors were defined as patients with overall survival > 3 years from surgery, short term survivors as patients with survival >3m and < 1 year from surgery to exclude perioperative mortalities.

[00444] ICGC cohort: Clinical characteristics of the ICGC cohort have been previously described²⁵.

[00445] *Tissue Microarray*

[00446] Tissue microarrays were constructed from tumor and adjacent non-tumor cores from formalin-fixed, paraffin embedded tissue blocks in short (n=45 tumors, 5 normals) and long term (n=51 tumors, 5 normals) survivors. Histology sections were reviewed by two expert PDAC pathologists and the most representative areas were selected and marked on H&E slides. 1 mm diameter cores were sampled from three different tumor regions per patient using an automated TMA Grand Master (Perkin Elmer, USA). Five μ m sections were prepared from TMA blocks for immunohistochemistry.

[00447] *Immunohistochemistry*

[00448] Human specific antibodies to MUC16 (clone OCT125, dilution 1:130), WT1 (clone CAN-R9 (IHC)-56-2, dilution 1:30), and Annexin A2 (ab54771, 5 μ g/ml) were purchased from Abcam (MA, USA), antibodies to MUC1 (clone M695, dilution 1:100), and Mesothelin (clone 5B2, dilution 1:50) were purchased from Vector laboratories (CA, USA). Immunohistochemistry was performed using standard techniques. MUC16 expression was scored as described⁴¹. For each core, a cumulative MUC16 expression score was calculated as the product of a score for the frequency of tumor cells expressing MUC16 (0-25%=1; 26-50%=2; 51-75%=3; 76-100%=4) and a score for the intensity of staining (0=negative; 1=weak; 2=moderate; 3=strong). The mean expression score across triplicate cores is reported as the final score for each patient.

[00449] Multiplexed consecutive immunohistochemistry on the same slide was performed as described¹⁶. Tissue microarray (TMA) slides were baked overnight at 37°C. Then, paraffin was removed using xylene and tissue rehydrated prior to incubation in antigen retrieval solution at 95°C for 45 minutes (pH 9 Target Retrieval Solution, Dako). After endogenous peroxidase inhibition and FcR blocking, Granzyme B was stained with anti-Granzyme B monoclonal antibody (clone GrB-7, Dako) for 1 hour at room temperature. After signal amplification with an HRP labeled polymer (Dako), the revelation was done using 3-Amino-9-ethylcarbazole (AEC, Vector Laboratories). Then slides were immersed in hematoxylin, rinsed in distilled water and mounted in aqueous-based mounting median (glycergel, Dako). After imaging using whole slide scanner, the slides were subjected to the Multiplexed Immunohistochemical Consecutive Staining on Single Slide protocol

(MICSSS)³⁴ and stained for T cells (CD3, clone 2GV6, Ventana and CD8, clone C8/144b, Dako), regulatory T cells (FoxP3, clone 236A/E7, Abcam), B cells (CD20, clone L26, Dako), macrophages (CD68, clone KP1, Dako), mature dendritic cells (DC-LAMP, clone 1010E1.01, Novus Biologicals)], MHC class I (HLA-ABC, clone EMR8-5, Abcam) and tumor cells (CK19, clone EP1580Y, Abcam).

[00450] *Immunofluorescence*

[00451] For CD4, FoxP3, and CK19 staining, sections first were incubated with anti-CD4 (Ventana, cat#790-4423, 0.5 ug/ml) for 5 hours, followed by 60 minutes incubation with biotinylated goat anti-rabbit IgG (Vector, cat # PK6101) at 1:200 dilution. The detection was performed with Streptavidin-HRP D (part of DABMap kit, Ventana Medical Systems), followed by incubation with Tyramide Alexa 488 (Invitrogen, cat# T20922) prepared according to manufacturer instruction with predetermined dilutions. Next, slides were incubated with anti-FoxP3 (Abcam, cat#ab20034, 5 ug/ml) for 4 hours, followed by 60 minutes incubation with biotinylated horse anti-mouse IgG (Vector Labs, cat# MKB-22258) at 1:200 dilution. The detection was performed with Streptavidin-HRP D (part of DABMap kit, Ventana Medical Systems), followed by incubation with Tyramide Alexa Fluor 568 (Invitrogen, cat# T20914) prepared according to manufacturer instruction with predetermined dilutions. Finally, sections were incubated with anti-CK19 (Abcam, cat#ab52625, 1ug/ml) for 5 hours, followed by 60 minutes incubation with biotinylated goat anti-rabbit IgG (Vector, cat # PK6101) at 1:200 dilution. The detection was performed with Streptavidin-HRP D (part of DABMap kit, Ventana Medical Systems), followed by incubation with Tyramide Alexa 647 (Invitrogen, cat# T20936) prepared according to manufacturer instruction with predetermined dilutions. After staining slides were counterstained with DAPI (Sigma Aldrich, cat# D9542, 5 ug/ml) for 10 min and coverslipped with Mowiol.

[00452] For CD3, CD8, and CK19 staining, slides first were incubated with anti-CD3 (DAKO, cat#A0452, 1.2 ug/ml) for 4 hours, followed by 60 minutes incubation with biotinylated goat anti-rabbit IgG (Vector Labs, cat# PK6101) at 1:200 dilution. The detection was performed with Streptavidin-HRP D (part of DABMap kit, Ventana Medical Systems), followed by incubation with Tyramide Alexa 488 (Invitrogen, cat# T20922) prepared according to manufacturer instruction with predetermined dilutions. Next, slides were incubated with anti-CD8 (Ventana, cat#790-4460, 0.35ug/ml) for 5 hours, followed by 60 minutes incubation with biotinylated goat anti-rabbit IgG (Vector, cat # PK6101) at 1:200 dilution. The detection was performed with Streptavidin-HRP D (part of DABMap kit,

Ventana Medical Systems), followed by incubation with Tyramide Alexa Fluor 568 (Invitrogen, cat# T20914) prepared according to manufacturer instruction with predetermined dilutions. Finally, sections were incubated with anti-CK19 (Abcam, cat#ab52625, 1ug/ml) for 5 hours, followed by 60 minutes incubation with biotinylated goat anti- rabbit IgG (Vector, cat # PK6101) at 1:200 dilution. The detection was performed with Streptavidin-HRP D (part of DABMap kit, Ventana Medical Systems), followed by incubation with Tyramide Alexa 647 (Invitrogen, cat# T20936) prepared according to manufacturer instruction with predetermined dilutions. After staining slides were counterstained with DAPI (Sigma Aldrich, cat# D9542, 5 ug/ml) for 10 min and coverslipped with Mowiol.

[00453] *Digital Image Processing and Analysis*

[00454] Tissue microarrays (TMAs) for each immunohistochemical stain were individually digitally scanned using Panoramic Flash (3DHistech, Budapest, Hungary) with a 40x/0.95NA objective. Images registration and alignment was performed using Image J (NIH, Bethesda, MD). ROIs were drawn for each core of one scan then transferred to others using CaseViewer (3DHistech). Each region from each scan was exported as tiff images at full resolution (0.243µm/pixel). Images of the same core from multiple scans were stacked together and aligned using Linear Stack Alignment with SIFT algorithm from FIJI/ImageJ (NIH, Bethesda, MD). Once aligned, the RGB images were color deconvoluted to separate AEC and hematoxylin stainings and converted into 8-bit pseudo-fluorescent images. Individual immunohistochemical targets were sequentially assigned to fluorescent channels and subsequently merged. Hematoxylin staining was used to segment and count the number of nucleated cells in the core. After processing the images using background subtraction and median filter, staining was thresholded and split using Biovoxxel Watershed Irregular Features plugin. ROIs were drawn around each cell and matched to the signals from all other AEC stainings to count the number of positive cells for each staining. Total tissue area was measured by setting a very low threshold for hematoxylin images. For quantification, all nucleated cells were identified, followed by an intensity-based threshold determination of each target to identify positive cells. Triplicate cores were quantified followed by determination of the median number of cells per square mm of tissue (Image J, NIH, Bethesda MD). Quantification of cells detected using immunofluorescence was performed in a similar fashion.

[00455] *Nucleic Acid Extraction*

[00456] 10µm slides were cut from OCT embedded frozen tumor and matched normal tissues. Sections were brought to containers with 70% ethanol for OCT removal. Following OCT removal, specimens were dissected for subsequent DNA and RNA extraction. For whole exome sequencing, tumor islands of $\geq 70\%$ cellularity were macrodissected based on expert PDAC pathologic review, and DNA was extracted using the DNA Easy kit. Total RNA from Fresh Frozen OCT embedded tissues was extracted using TRIzol RNA Isolation Reagents (cat# 15596-026, Life Technologies).

[00457] *Transcriptome Analysis*

[00458] Extracted RNA was qualified on Agilent BioAnalyzer and quantified by fluorometry (Ribogreen). Preparation of RNA for whole transcriptome expression analysis was done using the WT Pico Reagent Kit (Affymetrix, CA, USA). Reverse transcription was initiated at the poly-A tail as well as throughout the entire length of RNA to capture both coding and multiple forms of non-coding RNA. RNA amplification was achieved using low-cycle PCR followed by linear amplification using T7 in vitro transcription (IVT) technology. The cRNA was then converted to biotinylated sense strand DNA hybridization targets. Prepared target was hybridized to GeneChip® Human Transcriptome Array 2.0 (Affymetrix, CA, USA). Wash and scan was done using the GeneChip® Hybridization, Wash and Stain Kit using a Fluidics Station 450/250. Arrays were scanned using the GeneChip® Scanner 3000. Data analysis for the array was done using Affymetrix Expression Console™ Software (SST-RMA algorithm to summarize the signal from array probesets). Unsupervised hierarchical clustering was performed using Ward linkage and Euclidean distance (R, v 3.3.0). Clusters were defined using a probe set variance cutoff of 0.3. A dendritic cell signature was defined as previously described, using the genes CCL13, CCL17, CCL22, PPFIBP2, NPR1, HSD11B1, and CD209/DC-SIGN¹⁷.

[00459] *T Cell Receptor Vβ Sequencing*

[00460] Frozen tumor (short term n=30, long term n=30) and paired non-tumor adjacent pancreas tissue (short term n=30, long term n=30) samples were processed (Adaptive Biotechnologies, Seattle, USA). Genomic DNA was extracted according to the manufacturer's instructions (QIAasympyphony, Qiagen, Germany). The quantity and quality of extracted DNA was verified prior to sequencing. Using a standard quantity of input DNA, the TCR Vβ CDR3 regions were amplified and sequenced using the survey multiplexed PCR ImmunoSeq assay. The ImmunoSeq platform combines multiplex PCR with high throughput

sequencing to selectively amplify the rearranged complementarity determining region 3 (CDR3) of the TCR, producing fragments sufficiently long to identify the VDJ region spanning each unique CDR3. 45 forward primers specific for TCR V β gene segments and 13 reverse primers specific to TCR J β gene segments were used (Adaptive Biotechnologies). Read lengths of 156bp were obtained using the Illumina HiSeq System. The ImmunoSeq assay allows for quantitative assessment of both total and unique TCRs in a sample, as it uses a complete synthetic repertoire of TCRs to establish an amplification baseline and adjust the assay chemistry to correct for primer bias. Barcoded, spiked-in synthetic templates were also used to measure and correct for sequencing coverage and residual PCR bias. Output data were then filtered and clustered using the relative frequency ratio between similar clones and a modified nearest-neighbour algorithm, to merge closely related sequences and remove PCR and sequencing errors. The number of rearranged TCRs per diploid genome in the input material (total number of T cells) was estimated as previously described³⁷. Data were analyzed using the ImmunoSeq analyzer tool. The frequency of T cells was determined as the total number of T cells per total number of sequenced cells in the input material. A T cell clone was defined as a T cell with a unique TCR V β CDR3 amino acid sequence. Clonality was defined as (1-normalized entropy). Normalized entropy was calculated as the Shannon entropy divided by the logarithm of the number of unique productive (exonic) TCR sequences. Shannon entropy equals the clonal abundance of all productive TCR sequences in the input material. For in vitro stimulated cells, clones with identical amino acid sequences detected in all four compartments, that expanded on day 22 compared to day 0 (fold change > 2, ≥ 5 templates) are indicated. Data analysis was performed using Adaptive Biotechnologies ImmunoSeq Analyzer (Analyses 3.0, Seattle, Washington).

[00461] *Whole Exome Sequencing*

[00462] For all MSKCC PDAC patients, 500ng of genomic DNA was fragmented to a target size of 150 to 200 bp on the Covaris LE220 system. Barcoded libraries (Kapa Biosystems) were subjected to exon capture by hybridization using the SureSelect Human All Exon 51MB V4 kit (Agilent). DNA libraries were subsequently sequenced on a HiSeq 4000 (Illumina) in a Paired End 100/100, using the TruSeq SBS Kit v3 (Illumina) with a target coverage of 150X for tumor samples and 70X for matched normal (MSKCC Center for Molecular Oncology). Sequence data were demultiplexed using CASAVA, and after removal of adapter sequences using cutadapt (v1.6) reads were aligned to the reference human genome (hg19) using the Burrows-Wheeler Alignment tool (bwa mem v0.7.12).

Duplicate-read removal, InDel realignment and Base Quality Score Recalibration were performed using the Genome Analysis Toolkit (GATK) according to GATK best practices, as previously described¹³. Variants were identified on processed data using mutect, mutect rescue (SNPs) and haplotype caller (insertions/deletions) (Figure 71). A mean unique sequence coverage of 167.45X was achieved for tumor samples and 84.75X for normal samples. All MUC16 mutations were manually reviewed by 3 investigators using the Integrated Genomics Viewer (IGV) v2.3.72. Whole genome and whole exome sequencing for ICGC²⁵ patients has been previously described. For all ICGC samples, BAM files were re-processed and mutations identified as per the above outlined MSKCC protocol.

[00463] *HLA Typing*

[00464] HLA typing for PDAC patients was performed in silico using the tool Short Oligonucleotide Analysis Package-HLA (SOAP).

[00465] *Immunogenicity Predictions of Somatic Mutations*

[00466] MSKCC Pipeline: Immunogenicity of somatic mutations was estimated using a previously described bioinformatics tool called NAsSeek3. Briefly, NAsSeek is a computational algorithm that first translates all mutations in exomes to strings of 17 amino acids, for both the wild type and mutated sequences, with the amino acid resulting from the mutation centrally situated. Secondly, it evaluates putative MHC Class I binding for both wild type and mutant nonamers using a sliding window method using NetMHC3.4 for patient-specific HLA types, to generate predicted binding affinities for both peptides. NAsSeek finally assesses for similarity between nonamers that were predicted to be presented by patient-specific MHC Class I. All mutations with binding scores below 500 nM are defined as neoantigens. As the MSKCC pipeline was, on average, more stringent with respect to the number of neoantigens identified (in comparison to the pVAC-Seq pipeline below), all neoantigen predictions were performed with the MSKCC pipeline unless otherwise specified.

[00467] pVAC-Seq Pipeline: The pVAC-Seq pipeline²⁶ was used with the NetMHCpan binding strength predictor⁴³ to identify neoantigens (< 500nM binding strength). As recommended there, the variant effect predictor from Ensembl⁴⁴ was used to annotate variants for downstream processing by pVAC-Seq.

[00468] *Neoantigen Fitness Modeling*

[00469] The fitness of a clone is defined as

$$\frac{dN_{\alpha}}{dt} = F_{\alpha} N_{\alpha}$$

[00470] where N_{α} is the effective population size of tumor clone α , F_{α} is the fitness function of clone α . The effective population size is the size of a clone estimated from a tumors phylogenetic tree³¹.

[00471] Fitness due to neoantigen recognition potential (interchangeably referred to herein as cross reactivity) is defined as $F_{\alpha} = -C_{\alpha}$, where

$$C_{\alpha} = \max_{i \in \text{Clone } \alpha} (A_i^{MHC} \times R_i),$$

[00472] that is, within a clone α , the maximal product of the amplitude A_i^{MHC} and the recognition potential (cross reactivity) probability R_i for a neoantigen n_i , is probability that neoantigen n_i will be recognized by a T-Cell receptor.

[00473] Fitness in the neoantigen load hypothesis is defined as $F_{\alpha} = -L_{\alpha}$, where L_{α} is the number of neoantigens in clone α .

[00474] For a given neoantigen n_i , the recognition potential (cross reactivity) probability R_i was calculated as the probability that neoantigen n_i would be recognized by at least one T cell receptor specific to a microbial epitope e :

$$R_i = 1 - \prod_{e \in \text{IEDB}} (1 - \text{Pr}_{\text{binding}}(n_i, e)),$$

[00475] where $\text{Pr}_{\text{binding}}(n_i, e)$ estimates the binding probability of neoantigen n_i to a T cell receptor specific to epitope e . Epitopes e were derived from the Immune Epitope Database, restricting the search to all human, infectious disease derived, class-I restricted targets with positive immune assays. The probability of a neoantigen n_i eliciting recognition potential (cross reactivity) is modeled by its alignment to a validated IEDB epitope e via a logistic function of the alignment score between the two peptides, $|n_i, e|$, using the BLOSUM62 alignment matrix:

$$\text{Pr}_{\text{binding}}(n_i, e) = 1/(1 + e^{-k(|n_i, e| - e_0)}),$$

where e_0 represents the horizontal displacement of the binding curve and k sets the slope of the curve at e_0 . These are two free parameters to be fit in our model.

[00476] The amplitude due to relative MHC dissociation constants between a neoantigen and its wildtype counterpart is $A_i^{MHC} \approx K_d^{WT}/K_d^{\text{Mutant}}$. The standard cutoff for

K_d^{Mutant} was used, the mutant dissociation constant, used in the literature, that is $K_d^{\text{Mutant}} < 500 \text{ nM}$.

[00477] For all cases, the neoantigen load without clonal phylogeny was computed, which was the standard benchmark, and the neoantigen load with clonal phylogeny, by taking into account the effective size of clones in which neoantigens were contained. Our results were also compared to using the wild type cross reactivity alone, in which case our MHC amplitude was one, and the cross reactivity without clonality, which essentially just scores the best neoantigen across the tumor.

[00478] The predicted relative reduction in effective total tumor population size at time τ is therefore

$$S(\tau) = \frac{N(\tau)}{N(0)} = \sum_{\alpha} X_{\alpha}(0) \exp(F_{\alpha}\tau)$$

[00479] where $N(0) = \sum_{\alpha} N_{\alpha}(0)$ is the initial total effective population size of all clones within the tumor, and $X_{\alpha}(0) = N_{\alpha}(0)/N(0)$ is frequency of clone α . The parameter $\tau = 0.05$, and is a time-scale fixed across all datasets which empirically corresponds to the time after which the signal from the model degrades. Now $S(\tau)$ is calculated for each sample. Samples were split by the median value of the cohort, with samples below this value designated as a low fitness group, and those above as a high fitness group. Then survival for high versus low fitness groups were compared. Values for the shift parameter that optimized survival were calculated. To test the stability of this choice, the optimal value for subsampled datasets was derived, with subsampling frequencies of 0.8, 0.7, 0.6, 0.5 and 0.4. The optimal value of the parameter obtained on the full dataset, $e_0 = 27$, was within the standard deviation of the median optimal values for all subsampling frequencies (Figure 56).

[00480] Finally, Monte-Carlo cross-validation was performed by randomly dividing the dataset into 50-50 partitions. In each case the optimal parameter was derived on one side of the partition (the training set), and then the p-value was determined between survival curves on the other 50% of samples (the validation set). The procedure was repeated 5000 times and gave the median p-value of 0.05.

[00481] *In Vitro T Cell Assays*

[00482] Fresh blood was collected from two PDAC long term survivors whose tumors were identified based on whole exome sequencing and *in silico* prediction to harbor MUC16

neoantigens. Peripheral-blood mononuclear cells were isolated by density centrifugation over Ficoll-Paque Plus (GE Healthcare). Peptides were generated for MUC16 neoantigens and the corresponding WT nonamers (Peptide 2.0, VA, USA). In vitro peptide stimulation was performed as described with minor modifications³. Briefly, 2×10^6 PBMCs were cultured with mutant or WT peptides (10ug/ml) on day 1. IL-2 (10U/ml) and IL-15 (10ng/ml) were added on day 2 and every subsequent 2-3 days. Mutant and WT peptides were added to respective cultures on days 7, and day 14 for second and third rounds of restimulation. On day 21 and 29, cells were restimulated in the presence of Brefeldin A and Golgiplug (BD Bioscience) for 6 hours and cells were subsequently stained as per manufacturer's instructions.

[00483] *Flow Cytometry*

[00484] Fresh blood and tumor samples from 6 individual patients undergoing elective surgery at Memorial Hospital were collected. Informed consent was obtained according to a Memorial Hospital Institutional Review Board approved protocol. Blood was drawn at the time of surgery, and peripheral-blood mononuclear cells were isolated by density centrifugation over Ficoll-Paque Plus (GE Healthcare). Tumor and draining lymph node tissues were processed immediately after removal from the patient and single-cell suspensions were prepared. To assess if T cells bind in silico predicted neoantigen-HLA complexes, T cells of peripheral blood mononuclear cells (PBMCs) from HLA-specific healthy donors (Precision For Medicine, Frederick, MD) were assessed for binding to *MUC16*-neoantigen-MHC multimers. *MUC16*-MHC-FITC multimers were designed to HLA-B0801, A1101, A2402, and A0301 (Immudex, Copenhagen, Denmark) with nonamer peptide sequences derived based on mutated MUC16 sequences identified on whole exome sequencing that were in-silico predicted to be immunogenic. Single cell PBMC suspensions were surface stained for anti-human CD45, CD3, CD56, CD8, CD4, CD107a, and MHC-multimers according to manufacturer's instructions. Human-specific antibodies used in all flow cytometric phenotyping included CD45 (clone HI30, BioLegend), CD3 (clone OKT3, BioLegend), CD4 (clone SK3, BD Biosciences), CD8 (clone SK1, BioLegend), CD56 (clone B159, BD Biosciences), CD69 (clone FN50, BD Biosciences), CD19 (clone SJ25C1, BD Biosciences), PD1 (clone MIH4, BD Biosciences), CD45RA (clone HI100, BD Biosciences), CD45 RO (clone UCHL1, BD Biosciences) and CD107a (clone H4A3, BD Biosciences). Flow cytometry was performed on an LSRFortessa (BD Biosciences) and data were analyzed using FlowJo Software (Tree Star).

[00485] *Statistics*

[00486] Comparisons between two groups were performed using unpaired two-tailed Mann-Whitney test (unpaired samples), paired two-tailed Mann Whitney test (paired samples), and two-tailed students t-test (normally distributed parameters). Multiple samples were compared using Kruskal-Wallis test (non-grouped) and ANOVA with Tukey's post-test for multiple comparisons (grouped). Survival curves were compared using log-rank test (Mantel-Cox). Categorical variables were compared using chi-square test. All comparison groups had equivalent variances. $P < 0.05$ was considered to be statistically significant.

[00487] Discussion

[00488] The current study in this Example X includes genomic, molecular, and cellular immunoprofiling with neoantigen discovery in rare long term survivors of pancreatic ductal adenocarcinoma (PDAC) (n=82, median survival 6 years), a lethal, checkpoint blockade refractory cancer with few predicted neoantigens¹². Compared to short term survivors with poor outcome, greater cytolytic, polyclonal T cell infiltrates and adaptive immune activation in tumors of long term survivors were detected. Using whole exome sequencing and in silico prediction, it was found that tumors with the highest neoantigen number and the most abundant intratumoral cytolytic T cell infiltrates stratified patients with the longest survival (median survival not reached). To understand the specific neoantigen qualities in long term survivors, a neoantigen immunogenicity fitness model was developed, integrating clonal genealogy, epitope homology, and T cell receptor affinity. A model conferring greater immunogenicity to dominant neoantigens with homology to microbial epitopes identified long term survivors, whereas a model ascribing greater immunogenicity to increasing neoantigen number did not. Consistent with long term survivors harboring immunogenic neoantigens, it was found that long term survivors were 4 times more likely to harbor neoantigens at the *MUC16* locus encoding the tumor antigen CA125. Finally, it was detected that evidence of an anti-*MUC16* immune response in tumors of long term survivors with subsequent long-lasting *MUC16* neoantigen-specific T cell immunity in peripheral blood, identifying *MUC16* as a candidate immunogenic hotspot. The data suggest that even in tumors with fewer mutations such as PDAC, neoantigens can influence protective immunity, with immunodominant neoantigens identified by adaptive immune fitness and immunogenic hotspots. Therefore, clarification of factors determining neoantigen immunogenicity and detection of immunogenic hotspots can facilitate focused therapeutic targeting of neoantigens and inform application of future checkpoint blockade immunotherapies

[00489] The results shed novel insight into the heterogeneous immunobiology of presumed poorly immunogenic and checkpoint blockade refractory tumors such as PDAC, demonstrating that neoantigens can be dominant T cell targets in subsets of long term survivors. It is proposed that neoantigen quality, and not merely quantity, modulate immunogenicity, clonal fitness, and immunoselection during tumor evolution, with neoantigens in immunogenic residues such as MUC16 emerging as apparent hotspots. It is noteworthy that our results do not invoke associations of pre-existing microbial and anti-tumor immunity in long term survivors; instead, our data suggest that microbial homology can serve as an effective surrogate for immunogenic neoantigen qualities. The data presented herein suggest that neoantigen-specific immunity gained during primary tumor outgrowth can be associated with decreased relapse and prolonged survival, comparable to classical murine studies of prior tumor exposure protecting against tumor rechallenge³⁵. These findings have implications for cancer immunotherapy. Specifically, they provide a rationale for development of immunotherapeutic strategies to harness neoantigen-specific immunity in the treatment of checkpoint blockade refractory cancers, and immunogenic hotspot discovery for directed neoantigen targeting.

[00490] For further information, see Balachandran, VP et al., *Nature*, 551(7681):512-16 (2017), the content of which is expressly incorporated herein by reference in its entirety, for all purposes.

REFERENCES CITED AND ALTERNATIVE EMBODIMENTS

- [00491]** 1. Siegel, R.L., Miller, K.D. & Jemal, A. *Cancer Statistics* **65**, 5–29 (2015).
- [00492]** 2. Kleeff, J. *et al.* Pancreatic cancer, *Nature Publishing Group* **2**, 16022 (2016).
- [00493]** 3. Hidalgo, M. Pancreatic cancer, *The New England Journal of Medicine* **362**, 1605–1617 (2010).
- [00494]** 4. Valsecchi, M.E., Diaz-Canton, E., de la Vega, M. & Littman, S. J., “Recent treatment advances and novel therapies in pancreas cancer: a review,” *Journal of Gastrointestinal Cancer* **45**, 190–201 (2014).

- [00495] 5. Royal, R.E. *et al.*, “Phase 2 trial of single agent Ipilimumab (anti-CTLA-4) for locally advanced or metastatic pancreatic adenocarcinoma,” *Journal of Immunotherapy* **33**, 828–833 (2010).
- [00496] 6. Bailey, P. *et al.*, “Genomic analyses identify molecular subtypes of pancreatic cancer,” *Nature* **531**, 47–52 (2016).
- [00497] 7. Ryschich, E. *et al.*, “Control of T-cell-mediated immune response by HLA class I in human pancreatic carcinoma,” *Clin Cancer Res* **11**, 498–504 (2005).
- [00498] 8. Ino, Y. *et al.*, “Immune cell infiltration as an indicator of the immune microenvironment of pancreatic cancer,” *Br J Cancer* **108**, 914–923 (2013).
- [00499] 9. Hiraoka, N. *et al.*, “Intratumoral tertiary lymphoid organ is a favourable prognosticator in patients with pancreatic cancer,” *Br J Cancer* **112**, 1782–1790 (2015).
- [00500] 10. Dal Molin, M. *et al.*, “Very Long-term Survival Following Resection for Pancreatic Cancer Is Not Explained by Commonly Mutated Genes: Results of Whole-Exome Sequencing Analysis,” *Clin Cancer Res* **21**, 1944–1950 (2015).
- [00501] 11. Gubin, M. M. *et al.*, “Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens,” *Nature* **515**, 577–581 (2014).
- [00502] 12. Schumacher, T.N. & Schreiber, R.D., “Neoantigens in cancer immunotherapy,” *Science* **348**, 69–74 (2015).
- [00503] 13. Snyder, A. *et al.*, “Genetic basis for clinical response to CTLA-4 blockade in melanoma,” *N Engl J Med* **371**, 2189–2199 (2014).
- [00504] 14. Rizvi, N.A. *et al.*, “Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer,” *Science* **348**, 124–128 (2015).
- [00505] 15. Van Allen, E. M. *et al.*, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma,” *Science* **350**, 207–211 (2015).
- [00506] 16. Remark, R., Merghoub, T. & Grabe, N., “In-depth tissue profiling using multiplexed immunohistochemical consecutive staining on single slide,” *Science* (2016).
- [00507] 17. Bindea, G. *et al.*, “Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer,” *Immunity* **39**, 782–795 (2013).

- [00508] 18. Gros, A. *et al.*, “PD-1 identifies the patient-specific CD8⁺ tumor-reactive repertoire infiltrating human tumors,” *The Journal of Clinical Investigation* **124**, 2246–2259 (2014).
- [00509] 19. Gros, A. *et al.*, “Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients,” *Nat Med* **22**, 433–438 (2016).
- [00510] 20. Chauvin, J. -M. *et al.*, “TIGIT and PD-1 impair tumor antigen-specific CD8⁺ T cells in melanoma patients,” *The Journal of Clinical Investigation* **125**, 2046–2058 (2015).
- [00511] 21. Yu, H., Pardoll, D. & Jove, R., “STATs in cancer inflammation and immunity: a leading role for STAT3,” *Nat Rev Cancer* **9**, 798–809 (2009).
- [00512] 22. McAllister, F. *et al.*, “Oncogenic Kras activates a hematopoietic-to-epithelial IL-17 signaling axis in preinvasive pancreatic neoplasia,” *Cancer Cell* **25**, 621–637 (2014).
- [00513] 23. Biankin, A.V. *et al.*, “Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes,” *Nature* **491**, 399–405 (2012).
- [00514] 24. Witkiewicz, A.K. *et al.*, “Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets,” *Nat Commun* **6**, 6744 (2015).
- [00515] 25. Bailey, P. *et al.*, “Genomic analyses identify molecular subtypes of pancreatic cancer,” *Nature* **531**, 47–52 (2016).
- [00516] 26. Hundal, J. *et al.*, “pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens,” *Genome Med* **8**, 11 (2016).
- [00517] 27. Sausen, M. *et al.*, “Clinical implications of genomic alterations in the tumour and circulation of pancreatic cancer patients,” *Nat Commun* **6**, 7686 (2015).
- [00518] 28. Zitvogel, L., Ayyoub, M., Routy, B. & Kroemer, G., “Microbiome and Anticancer Immunosurveillance,” *Cell* **165**, 276–287 (2016).
- [00519] 29. Vétizou, M. *et al.*, “Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota,” *Science* **350**, 1079–1084 (2015).
- [00520] 30. Łuksza, M. & Lässig, M., “A predictive fitness model for influenza,” *Nature* **507**, 57–61 (2014).