



US 20140296081A1

(19) **United States**(12) **Patent Application Publication**
Diehn et al.(10) **Pub. No.: US 2014/0296081 A1**(43) **Pub. Date: Oct. 2, 2014**(54) **IDENTIFICATION AND USE OF
CIRCULATING TUMOR MARKERS**(71) Applicant: **The Board of Trustees of the Leland
Stanford Junior University**, Palo Alto,
CA (US)(72) Inventors: **Maximilian Diehn**, Stanford, CA (US);
Arash Ash Alizadeh, San Mateo, CA
(US); **Aaron M. Newman**, Palo Alto,
CA (US); **Scott V. Bratman**, Palo Alto,
CA (US)(21) Appl. No.: **14/209,807**(22) Filed: **Mar. 13, 2014****Related U.S. Application Data**(60) Provisional application No. 61/798,925, filed on Mar.
15, 2013.**Publication Classification**(51) **Int. Cl.**
G06F 19/22 (2006.01)
C12Q 1/68 (2006.01)
(52) **U.S. Cl.**
CPC **G06F 19/22** (2013.01); **C12Q 1/6886**
(2013.01)
USPC **506/2**; 506/8(57) **ABSTRACT**

Methods for creating a library of recurrently mutated genomic regions and for using the library to analyze cancer-specific and patient-specific genetic alterations in a patient are provided. The methods can be used to measure tumor-derived nucleic acids in patient blood and thus to monitor the progression of disease. The methods can also be used for cancer screening.

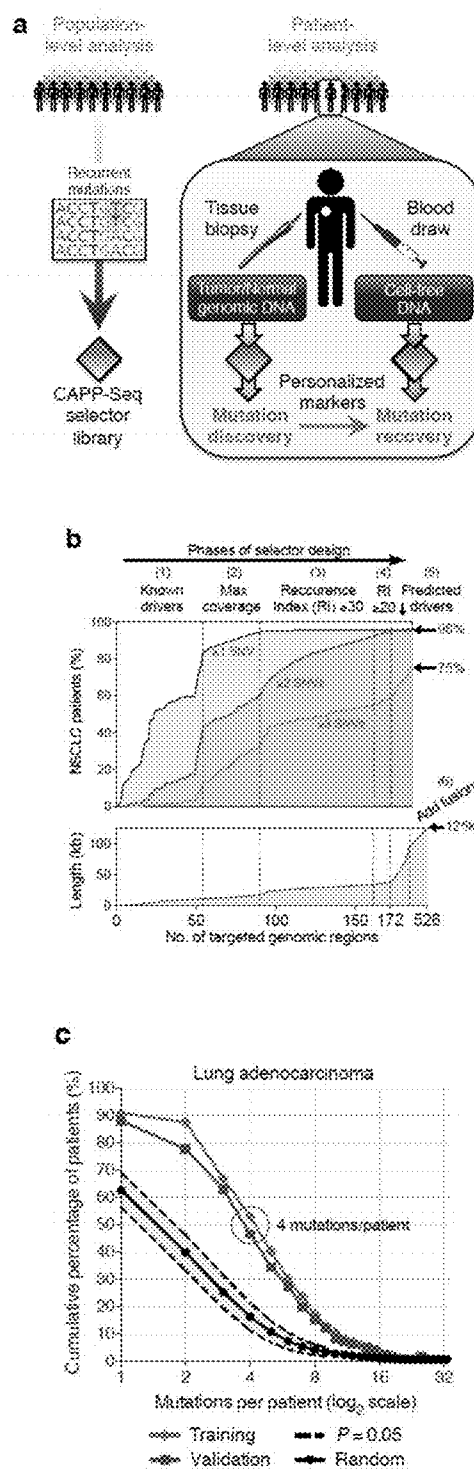


Figure 1

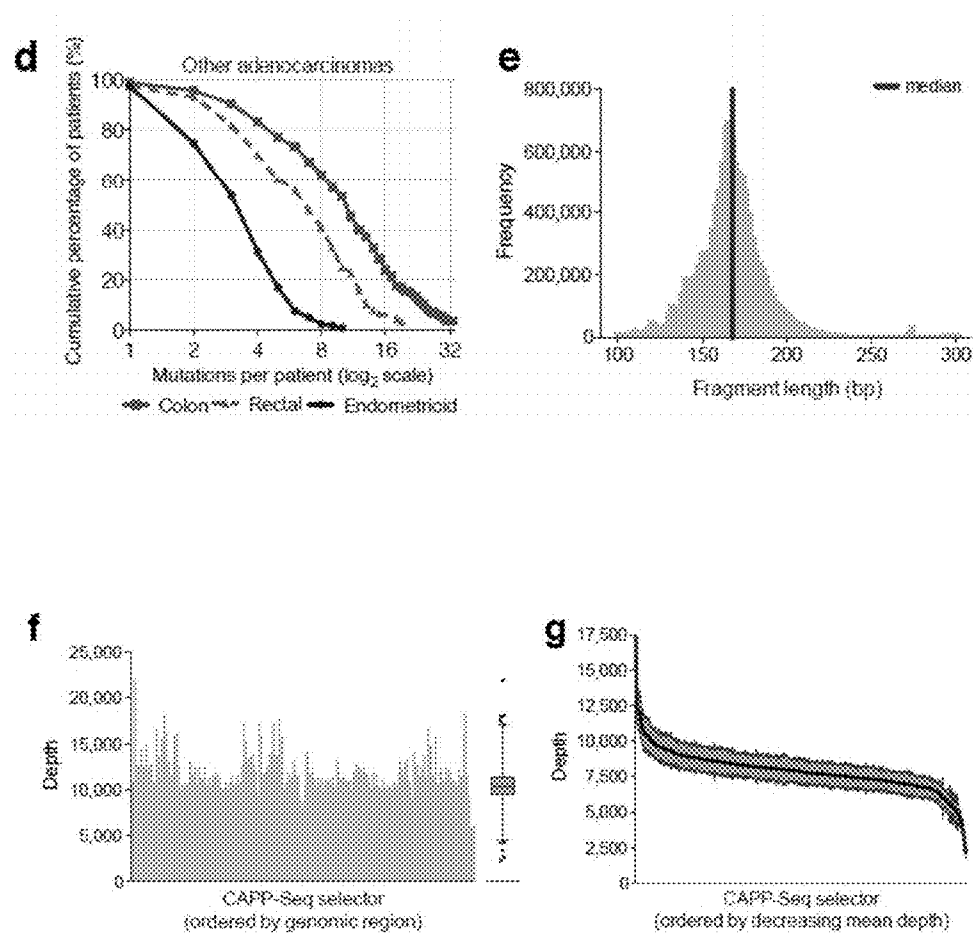


Figure 1 (cont.)

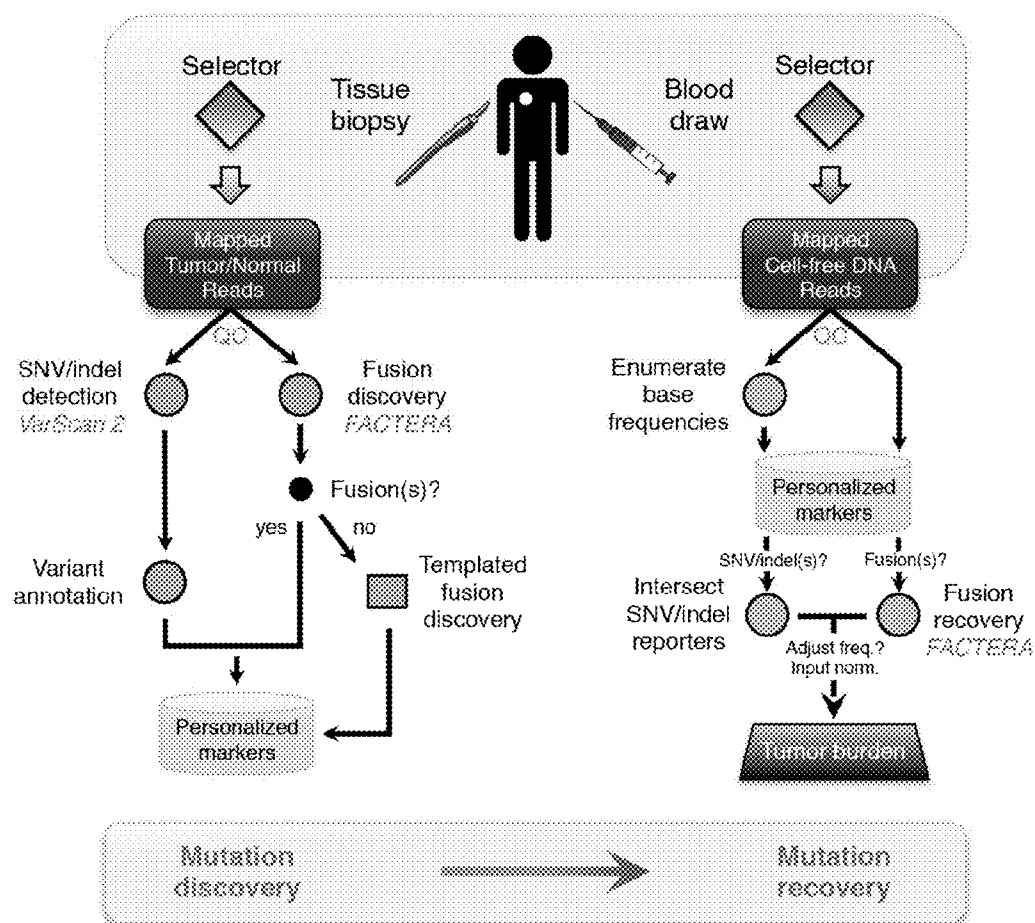


Figure 2

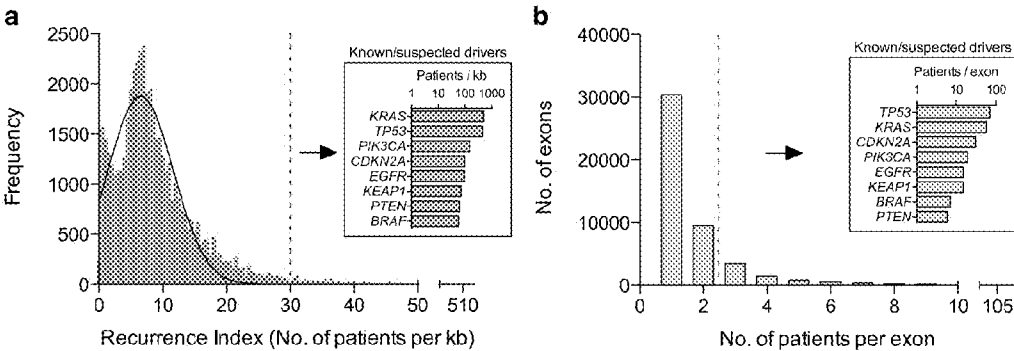


Figure 3

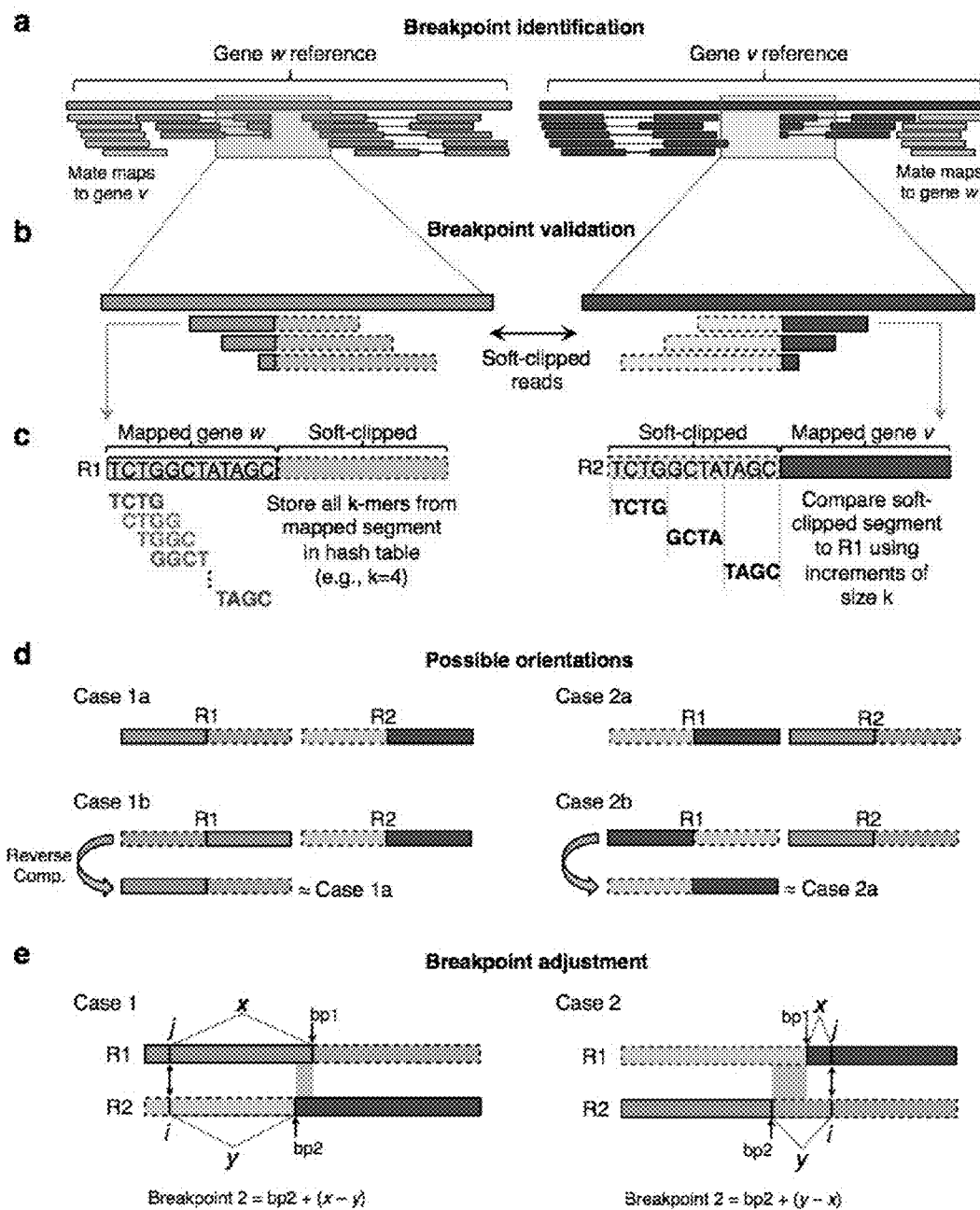


Figure 4

Figure 5

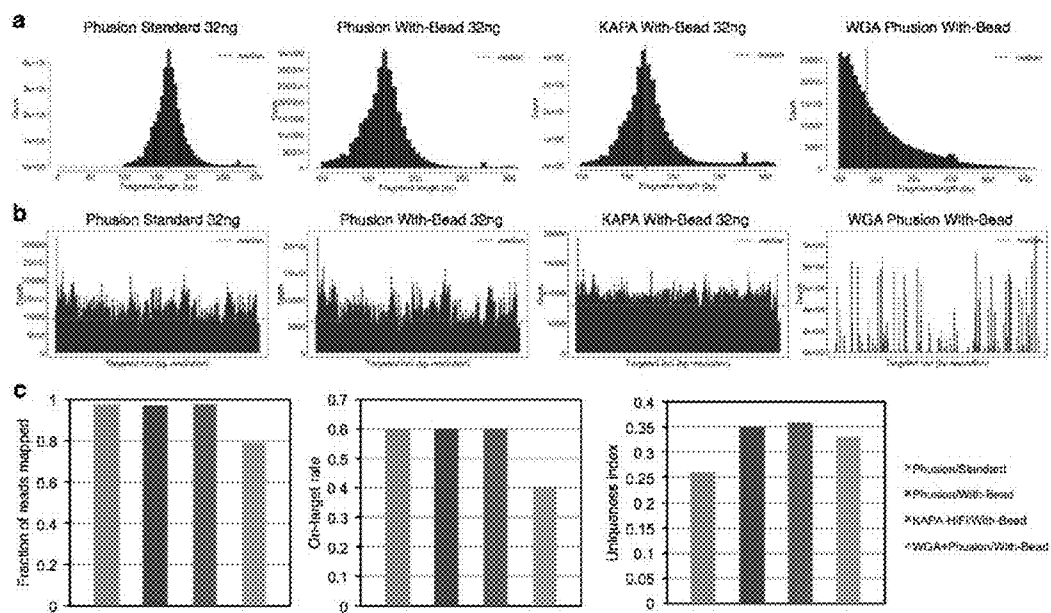


Figure 6

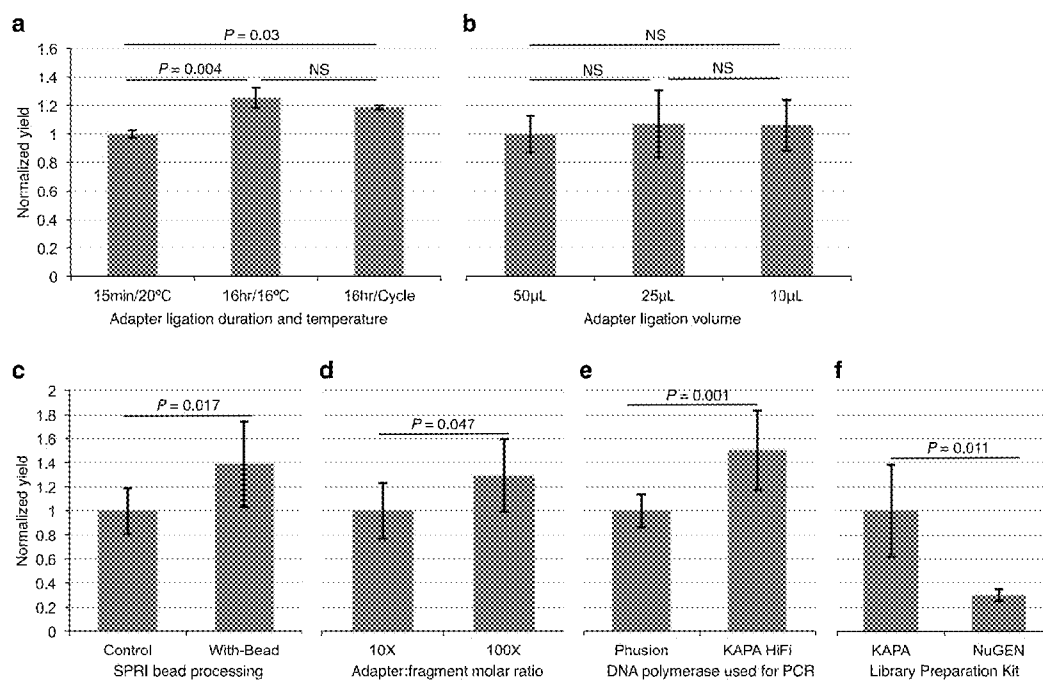


Figure 7

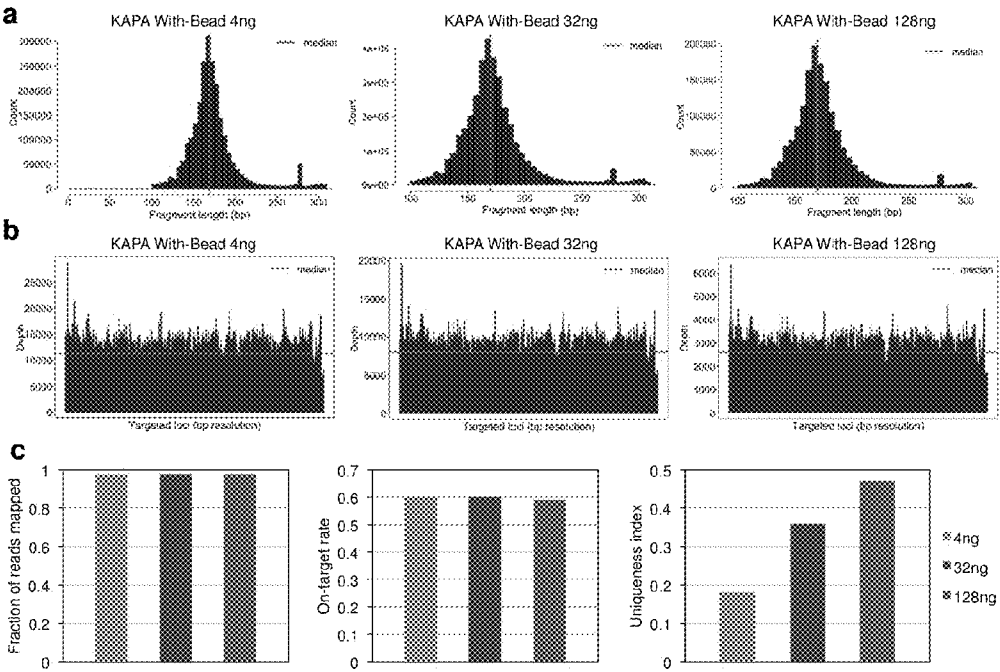


Figure 8

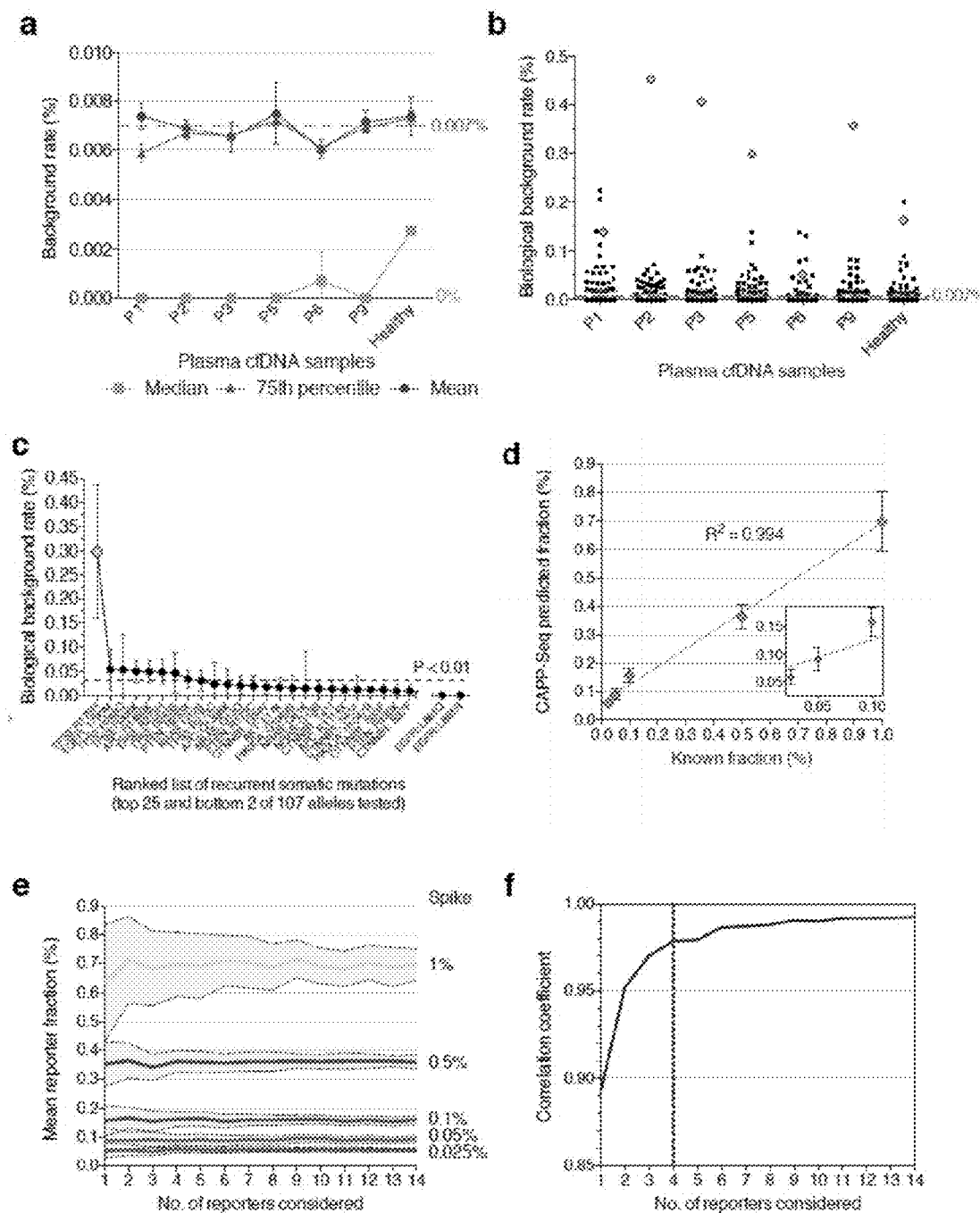


Figure 9

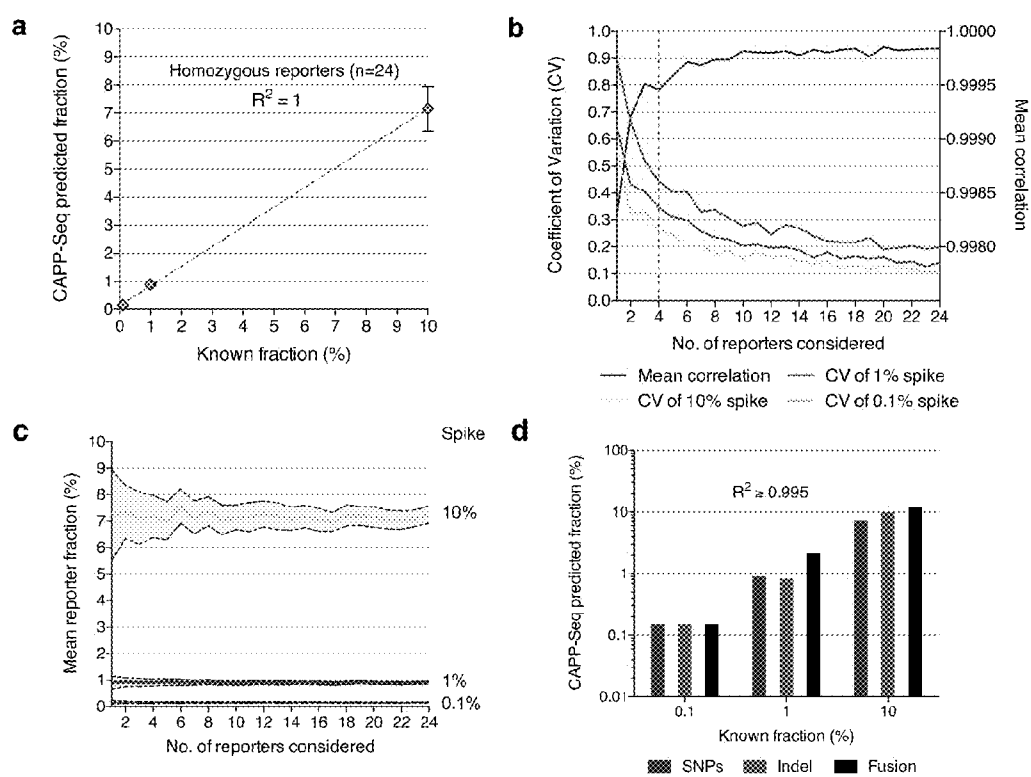


Figure 10

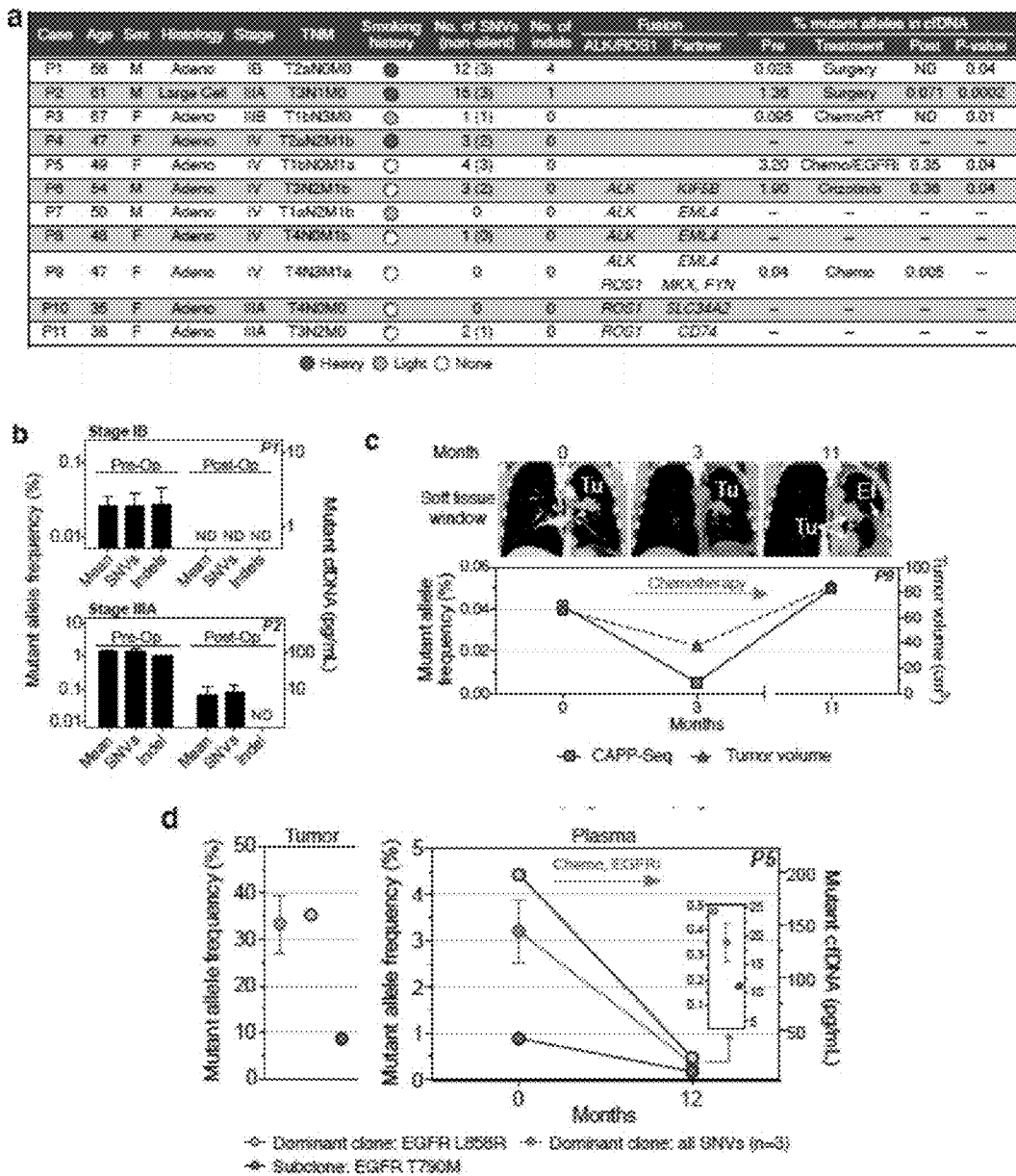


Figure 11

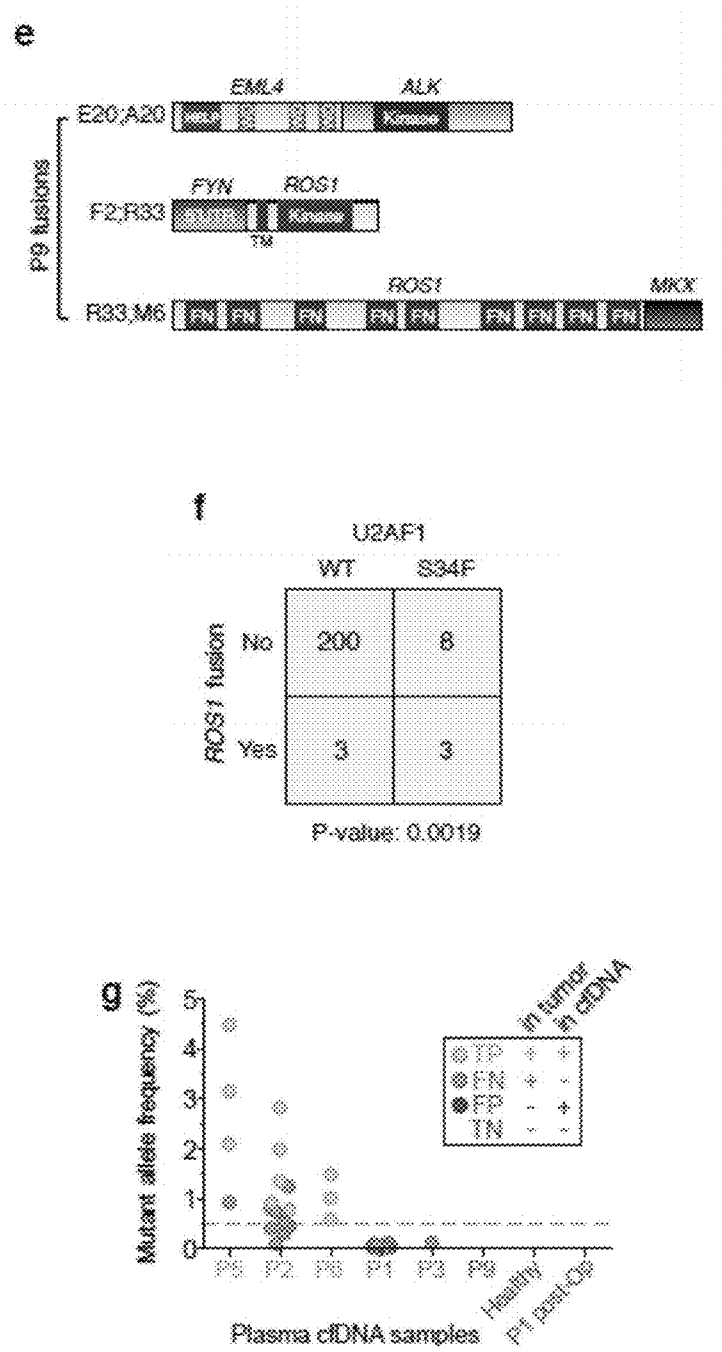


Figure 11 (cont.)

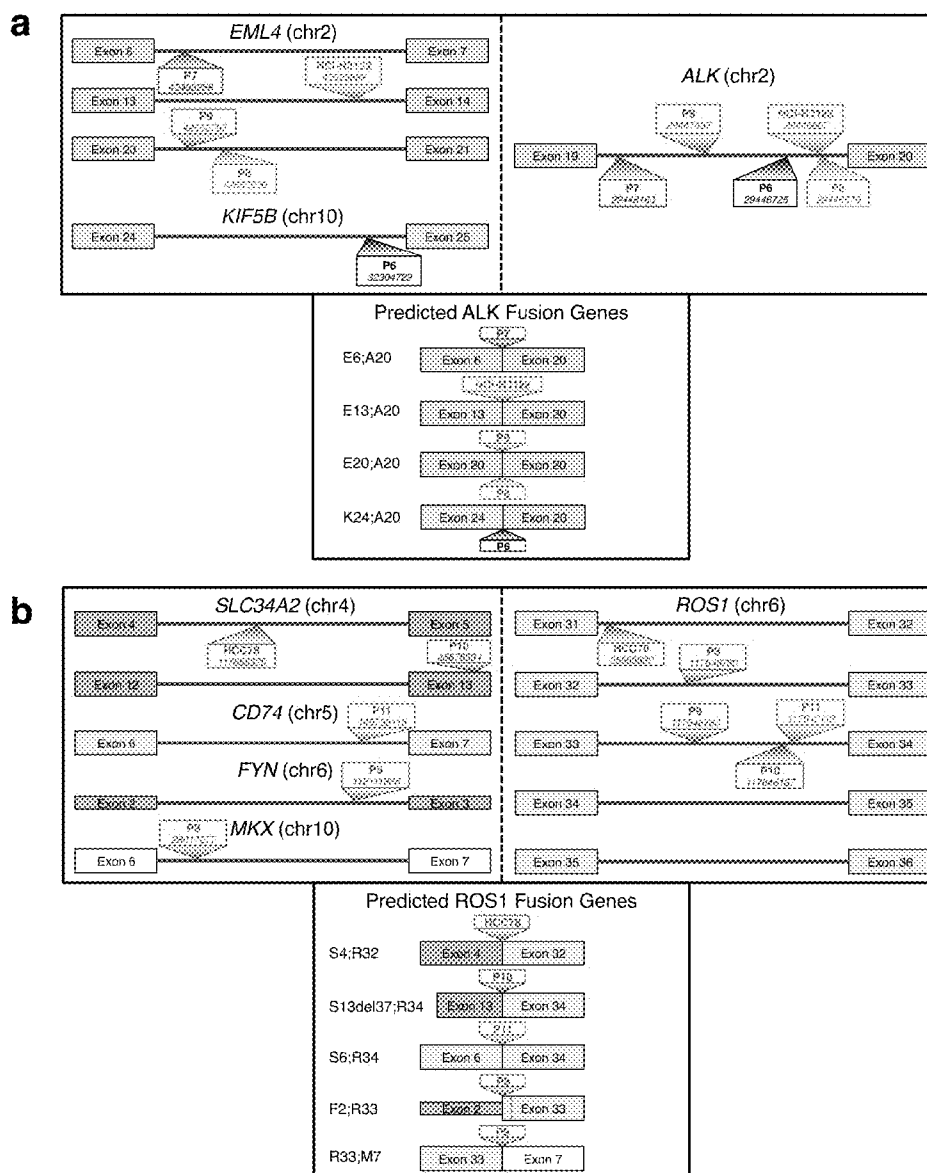


Figure 12

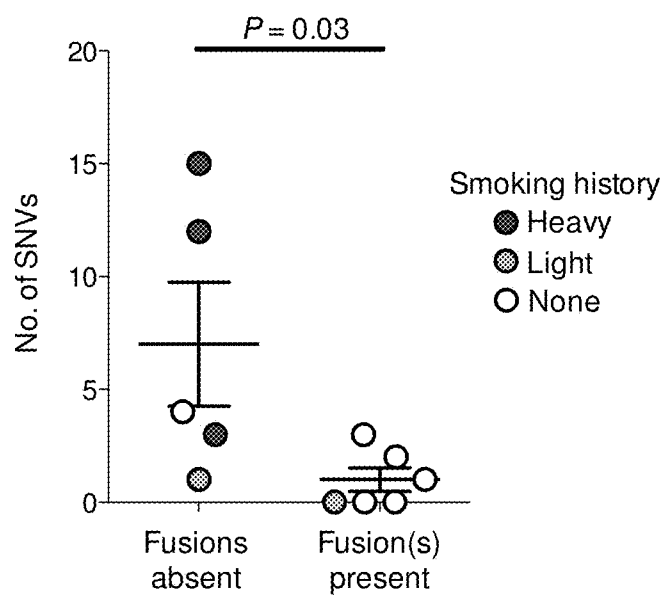


Figure 13

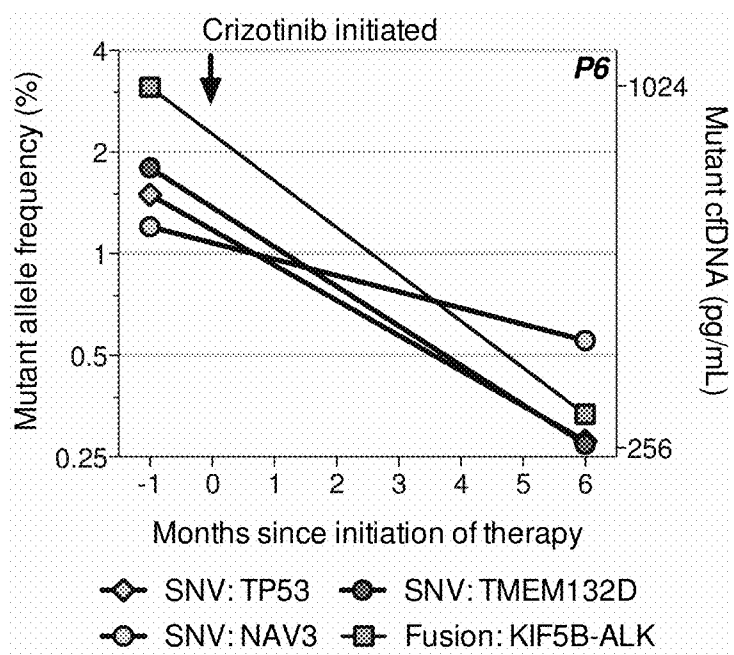


Figure 14

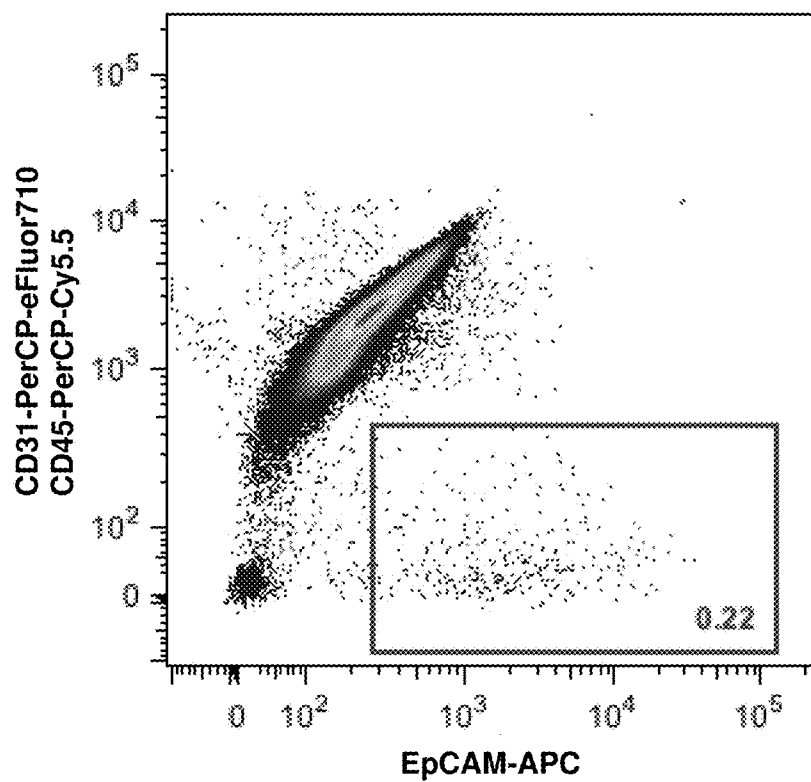


Figure 15

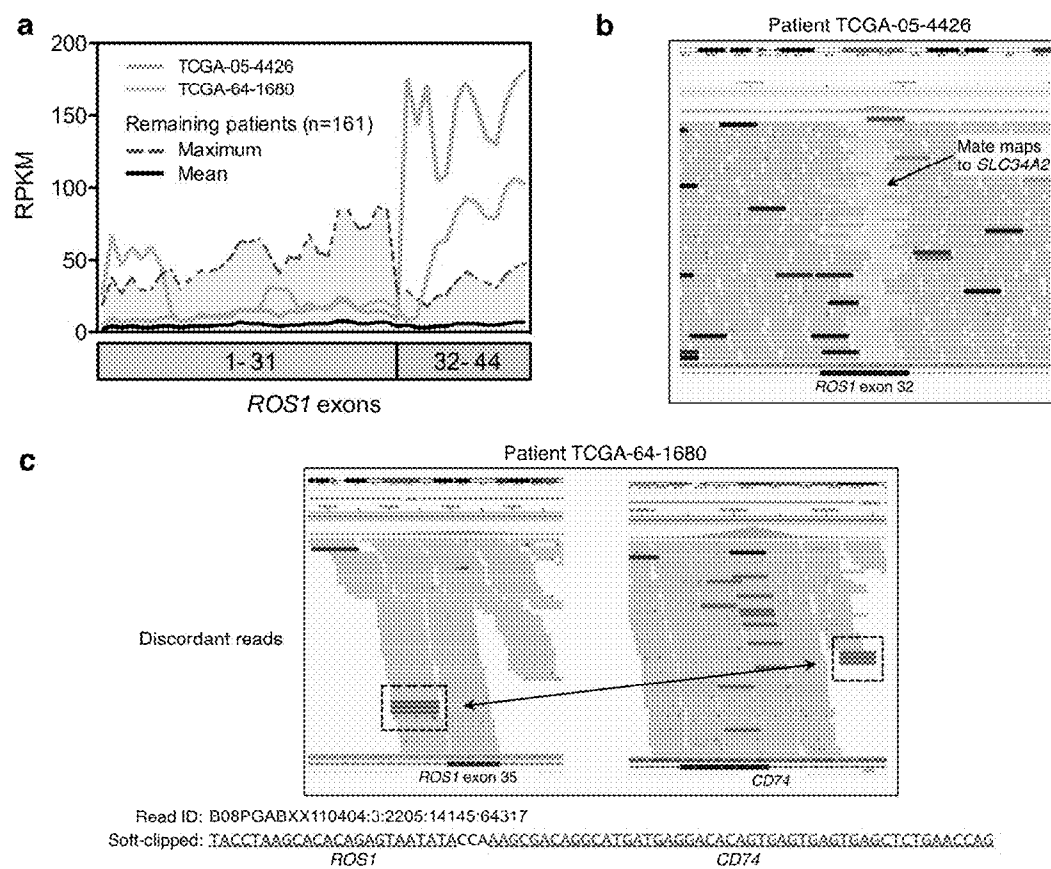


Figure 16

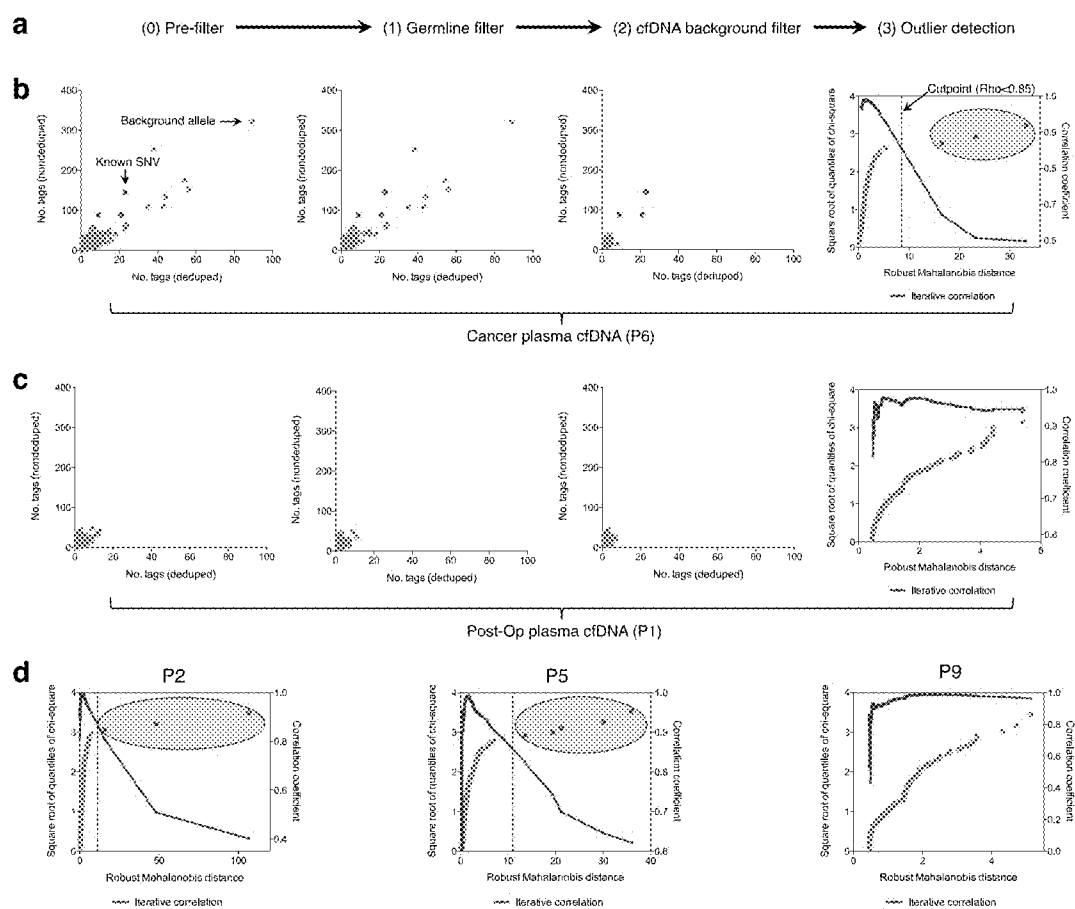


Figure 17

IDENTIFICATION AND USE OF CIRCULATING TUMOR MARKERS

STATEMENT OF GOVERNMENTAL SUPPORT

[0001] This invention was made with government support under grant number W81XWH-12-1-0285 awarded by the Department of Defense. The government has certain rights in the invention.

BACKGROUND OF THE INVENTION

[0002] Analysis of cancer-derived cell-free DNA (cfDNA) has the potential to revolutionize detection and monitoring of cancer. Noninvasive access to malignant DNA is particularly attractive for solid tumors, which cannot be repeatedly sampled without invasive procedures. In non-small cell lung cancer (NSCLC), PCR-based assays have been used previously to detect recurrent point mutations in genes such as KRAS or EGFR in plasma DNA (Taniguchi et al. (2011) *Clin. Cancer Res.* 17:7808-7815; Gautschi et al. (2007) *Cancer Lett.* 254:265-273; Kuang et al. (2009) *Clin. Cancer Res.* 15:2630-2636; Rosell et al. (2009) *N. Engl. J. Med.* 361:958-967), but the majority of patients lack mutations in these genes. Other studies have proposed identifying patient-specific chromosomal rearrangements in tumors via whole genome sequencing (WGS), followed by breakpoint qPCR from cfDNA (Leary et al. (2010) *Sci. Transl. Med.* 2:20ra14; McBride et al. (2010) *Genes Chrom. Cancer* 49:1062-1069). While sensitive, such methods require optimization of molecular assays for each patient, limiting their widespread clinical application. More recently, several groups have reported amplicon-based deep sequencing methods to detect cfDNA mutations in up to 6 recurrently mutated genes (Forshew et al. (2012) *Sci. Transl. Med.* 4:136ra168; Narayan et al. (2012) *Cancer Res.* 72:3492-3498; Kinde et al. (2011) *Proc. Natl Acad. Sci. USA* 108:9530-9535). While powerful, these approaches are limited by the number of mutations that can be interrogated (Rachlin et al. (2005) *BMC Genomics* 6:102) and the inability to detect genomic fusions.

[0003] PCT International Patent Publication No. 2011/103236 describes methods for identifying personalized tumor markers in a cancer patient using “mate-paired” libraries. The methods are limited to monitoring somatic chromosomal rearrangements, however, and must be personalized for each patient, thus limiting their applicability and increasing their cost.

[0004] U.S. Patent Application Publication No. 2010/0041048 A1 describes the quantitation of tumor-specific cell-free DNA in colorectal cancer patients using the “BEAMing” technique (Beads, Emulsion, Amplification, and Magnetics). While this technique provides high sensitivity and specificity, this method is for single mutations and thus any given assay can only be applied to a subset of patients and/or requires patient-specific optimization. U.S. Patent Application Publication No. 2012/0183967 A1 describes additional methods to identify and quantify genetic variations, including the analysis of minor variants in a DNA population, using the “BEAMing” technique.

[0005] U.S. Patent Application Publication No. 2012/0214678 A1 describes methods and compositions for detecting fetal nucleic acids and determining the fraction of cell-free fetal nucleic acid circulating in a maternal sample. While sensitive, these methods analyze polymorphisms occurring between maternal and fetal nucleic acids rather than polymor-

phisms that result from somatic mutations in tumor cells. In addition, methods that detect fetal nucleic acids in maternal circulation require much less sensitivity than methods that detect tumor nucleic acids in cancer patient circulation, because fetal nucleic acids are much more abundant than tumor nucleic acids.

[0006] U.S. Patent Application Publication Nos. 2012/0237928 A1 and 2013/0034546 describe methods for determining copy number variations of a sequence of interest in a test sample comprising a mixture of nucleic acids. While potentially applicable to the analysis of cancer, these methods are directed to measuring major structural changes in nucleic acids, such as translocations, deletions, and amplifications, rather than single nucleotide variations.

[0007] U.S. Patent Application Publication No. 2012/0264121 A1 describes methods for estimating a genomic fraction, for example, a fetal fraction, from polymorphisms such as small base variations or insertions-deletions. These methods do not, however, make use of optimized libraries of polymorphisms, such as, for example, libraries containing recurrently-mutated genomic regions.

[0008] U.S. Patent Application Publication No. 2013/0024127 A1 describes computer-implemented methods for calculating a percent contribution of cell-free nucleic acids from a major source and a minor source in a mixed sample. The methods do not, however, provide any advantages in identifying or making use of optimized libraries of polymorphisms in the analysis.

[0009] PCT International Publication No. WO 2010/141955 A2 describes methods of detecting cancer by analyzing panels of genes from a patient-obtained sample and determining the mutational status of the genes in the panel. The methods rely on a relatively small number of known cancer genes, however, and they do not provide any ranking of the genes according to effectiveness in detection of relevant mutations. In addition, the methods were unable to detect the presence of mutations in the majority of serum samples from actual cancer patients.

[0010] There is thus a need for new and improved methods to detect and monitor tumor-related nucleic acids in cancer patients.

SUMMARY OF THE INVENTION

[0011] The present invention addresses these and other problems by providing novel methods and systems relating to the characterization, diagnosis, and monitoring of cancer. In particular, according to one aspect, the invention provides methods for creating a library of recurrently mutated genomic regions comprising:

[0012] identifying a plurality of genomic regions from a group of genomic regions that are recurrently mutated in a specific cancer;

[0013] wherein the library comprises the plurality of genomic regions;

[0014] the plurality of genomic regions comprises at least 10 different genomic regions; and

[0015] at least one mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

[0016] In specific embodiments of these methods, the plurality of genomic regions comprises at least 25, at least 50, at least 100, at least 150, at least 200, or at least 500 different genomic regions.

[0017] In other specific method embodiments, at least two mutations within the plurality of genomic regions or at least three mutations within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

[0018] In still other specific method embodiments, at least one mutation within the plurality of genomic regions is present in at least 60%, 70%, 80%, 90%, 95%, 98%, 99%, or 99.9% of all subjects with the specific cancer.

[0019] In some embodiments, the identifying step comprises for each genomic region in the plurality of genomic regions, ranking the genomic region to maximize the number of all subjects with the specific cancer having at least one mutation within the genomic region.

[0020] In other embodiments, the identifying step comprises for each genomic region in the plurality of genomic regions, ranking the genomic region to maximize the ratio between the number of all subjects with the specific cancer having at least one mutation within the genomic region and the length of the genomic region.

[0021] In some embodiments, the library comprises a plurality of genomic regions encoding a plurality of driver sequences, more specifically known driver sequences or driver sequences that are recurrently mutated in the specific cancer.

[0022] In some embodiments, the library comprises a plurality of genomic regions that are recurrently rearranged in the specific cancer.

[0023] In preferred embodiments, the specific cancer is a carcinoma, and in more preferred embodiments, the carcinoma is an adenocarcinoma, a non-small cell lung cancer, or a squamous cell carcinoma.

[0024] In specific embodiments, the cumulative length of the plurality of genomic regions is at most 30 Mb, 20 Mb, 10 Mb, 5 Mb, 2 Mb, 1 Mb, 500 kb, 200 kb, 100 kb, 50 kb, 20 kb, or 10 kb.

[0025] In another aspect, the invention provides methods for analyzing a cancer-specific genetic alteration in a subject comprising the steps of:

[0026] obtaining a tumor nucleic acid sample and a genomic nucleic acid sample from a subject with a specific cancer;

[0027] sequencing a plurality of target regions in the tumor nucleic acid sample and in the genomic nucleic acid sample to obtain a plurality of tumor nucleic acid sequences and a plurality of genomic nucleic acid sequences; and

[0028] comparing the plurality of tumor nucleic acid sequences to the plurality of genomic nucleic acid sequences to identify a patient-specific genetic alteration in the tumor nucleic acid sample;

[0029] wherein the plurality of target regions are selected from a plurality of genomic regions that are recurrently mutated in the specific cancer;

[0030] the plurality of genomic regions comprises at least 10 different genomic regions; and

[0031] at least one mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

[0032] In specific embodiments of this aspect of the invention, the plurality of genomic regions comprises at least 25, at least 50, at least 100, at least 150, at least 200, or at least 500 different genomic regions.

[0033] In other specific embodiments, at least two mutations within the plurality of genomic regions or at least three

mutations within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

[0034] In still other specific embodiments, at least one mutation within the plurality of genomic regions is present in at least 60%, 70%, 80%, 90%, 95%, 98%, 99%, or 99.9% of all subjects with the specific cancer.

[0035] In some embodiments, each genomic region in the plurality of genomic regions is identified by ranking the genomic region to maximize the number of all subjects with the specific cancer having at least one mutation within the genomic region.

[0036] In other embodiments, each genomic region in the plurality of genomic regions is identified by ranking the genomic region to maximize the ratio between the number of all subjects with the specific cancer having at least one mutation within the genomic region and the length of the genomic region.

[0037] In some embodiments, the plurality of genomic regions comprises genomic regions encoding a plurality of driver sequences, more specifically known driver sequences or driver sequences that are recurrently mutated in the specific cancer.

[0038] In some embodiments, the plurality of genomic regions comprises genomic regions that are recurrently rearranged in the specific cancer.

[0039] In preferred embodiments, the specific cancer is a carcinoma, and in more preferred embodiments, the carcinoma is an adenocarcinoma, a non-small cell lung cancer, or a squamous cell carcinoma.

[0040] In specific embodiments, the cumulative length of the plurality of genomic regions is at most 30 Mb, 20 Mb, 10 Mb, 5 Mb, 2 Mb, 1 Mb, 500 kb, 200 kb, 100 kb, 50 kb, 20 kb, or 10 kb.

[0041] In some embodiments, the methods further comprising the steps of:

[0042] obtaining a cell-free nucleic acid sample from the subject; and

[0043] identifying the patient-specific genetic alteration in the cell-free nucleic acid sample.

[0044] In specific embodiments, the step of identifying the patient-specific genetic alteration in the cell-free nucleic acid sample comprises sequencing a genomic region comprising the patient-specific genetic alteration in the cell-free sample.

[0045] In other specific embodiments, the step of obtaining a tumor nucleic acid sample and a genomic nucleic acid sample comprises the step of enriching the plurality of target regions in the tumor nucleic acid sample and the genomic nucleic acid sample, and in still more specific embodiments, the enriching step comprises use of a custom library of biotinylated DNA.

[0046] In still other specific embodiments, the step of obtaining a cell-free nucleic acid sample comprises the step of enriching the plurality of target regions in the cell-free nucleic acid sample, and in still more specific embodiments, the enriching step comprises use of a custom library of biotinylated DNA.

[0047] In some embodiments, the methods further comprise the step of quantifying the cancer-specific genetic alteration in the cell-free sample.

[0048] In yet another aspect, the invention provides methods for screening a cancer-specific genetic alteration in a subject comprising the steps of:

[0049] obtaining a cell-free nucleic acid sample from a subject;

[0050] sequencing a plurality of target regions in the cell-free sample to obtain a plurality of cell-free nucleic acid sequences; and

[0051] identifying a cancer-specific genetic alteration in the cell-free sample;

[0052] wherein the plurality of target regions are selected from a plurality of genomic regions that are recurrently mutated in the specific cancer;

[0053] the plurality of genomic regions comprises at least 10 different genomic regions; and

[0054] at least one mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

[0055] In specific embodiments, the plurality of genomic regions comprises at least 25, at least 50, at least 100, at least 150, at least 200, or at least 500 different genomic regions.

[0056] In other specific embodiments, at least two mutations within the plurality of genomic regions or at least three mutations within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

[0057] In still other specific embodiments, at least one mutation within the plurality of genomic regions is present in at least 60%, 70%, 80%, 90%, 95%, 98%, 99%, or 99.9% of all subjects with the specific cancer.

[0058] In particular embodiments, each genomic region in the plurality of genomic regions is identified by ranking the genomic region to maximize the number of all subjects with the specific cancer having at least one mutation within the genomic region.

[0059] In other particular embodiments, each genomic region in the plurality of genomic regions is identified by ranking the genomic region to maximize the ratio between the number of all subjects with the specific cancer having at least one mutation within the genomic region and the length of the genomic region.

[0060] In still other particular embodiments, the plurality of genomic regions comprises genomic regions encoding a plurality of driver sequences, and, more particularly, the driver sequences are known driver sequences or are recurrently mutated in the specific cancer.

[0061] In yet still other particular embodiments, the plurality of genomic regions comprises genomic regions that are recurrently rearranged in the specific cancer.

[0062] In some embodiments, the specific cancer is a carcinoma, including, for example, an adenocarcinoma, a non-small cell lung cancer, or a squamous cell carcinoma.

[0063] In specific embodiments, the cumulative length of the plurality of genomic regions is at most 30 Mb, 20 Mb, 10 Mb, 5 Mb, 2 Mb, 1 Mb, 500 kb, 200 kb, 100 kb, 50 kb, 20 kb, or 10 kb.

[0064] In other specific embodiments, the step of obtaining a cell-free nucleic acid sample comprises the step of enriching the plurality of target regions in the cell-free nucleic acid sample, and, in some embodiments, the enriching step comprises use of a custom library of biotinylated DNA.

BRIEF DESCRIPTION OF THE DRAWINGS

[0065] FIG. 1. Development of CAncer Personalized Profiling by Deep Sequencing (CAPP-Seq). (a) Schematic depicting design of CAPP-Seq selectors and their application for assessing circulating tumor DNA. (b) Multi-phase design of the NSCLC CAPP-Seq selector. (c) Analysis of the number of SNVs per lung adenocarcinoma covered by the NSCLC CAPP-Seq selector in the TCGA WES cohort (Training;

N=229) and an independent lung adenocarcinoma WES data set (Validation; N=183) (Imielinski et al. (2012) *Cell* 150: 1107-1120). (d) Number of SNVs per patient identified by the NSCLC CAPP-Seq selector in WES data from three adenocarcinomas from TCGA, colon (COAD), rectal (READ), and endometrioid (UCEC) cancers. (e-f) Quality parameters from a representative CAPP-Seq analysis of plasma cfDNA, including length distribution of sequenced cfDNA fragments (e), and depth of sequencing coverage across all genomic regions in the selector (f). (g) Variation in sequencing depth across cfDNA samples from 4 patients.

[0066] FIG. 2. CAPP-Seq computational pipeline. Major steps of the bioinformatics pipeline for mutation discovery and quantitation in plasma are schematically illustrated.

[0067] FIG. 3. Statistical enrichment of recurrently mutated NSCLC exons captures known drivers.

[0068] FIG. 4. Development of the FACTERA algorithm. Major steps used by FACTERA (see Detailed Methods) to precisely identify genomic breakpoints from aligned paired-end sequencing data are anecdotally illustrated using two hypothetical genes, w and v. (a) Improperly paired, or “discordant,” reads (indicated in yellow) are used to locate genes involved in a potential fusion (in this case, w and v). (b) Because truncated (i.e., soft-clipped) reads may indicate a fusion breakpoint, any such reads within genomic regions delineated by w and v are also further analyzed. (c) Consider soft-clipped reads, R1 and R2, whose non-clipped segments map to w and v, respectively. If R1 and R2 derive from a fragment encompassing a true fusion between w and v, then the mapped portion of R1 should match the soft-clipped portion of R2, and vice versa. This is assessed by FACTERA using fast k-mer indexing and comparison. (d) Four possible orientations of R1 and R2 are depicted. However, only Cases 1a and 2a can generate valid fusions (see Detailed Methods). Thus, prior to k-mer comparison (panel c), the reverse complement of R1 is taken for Cases 1b and 2b, respectively, converting them into Cases 1a and 2a. (e) In some cases, short sequences immediately flanking the breakpoint are identical, preventing unambiguous determination of the breakpoint. Let iterators i and j denote the first matching sequence positions between R1 and R2. To reconcile sequence overlap, FACTERA arbitrarily adjusts the breakpoint in R2 (i.e., bp2) to match R1 (i.e., bp1) using the sequence offset determined by differences in distance between bp2 and i, and bp1 and j. Two cases are illustrated, corresponding to sequence orientations described in (d).

[0069] FIG. 5. Application of FACTERA to NSCLC cell lines NCI-H3122 and HCC78, and Sanger-validation of breakpoints. (a) Pile-up of a subset of soft-clipped reads mapping to the EML4-ALK fusion identified in NCI-H3122 along with the corresponding Sanger chromatogram. (b) Same as (a), but for the SLC34A2-ROS1 translocation identified in HCC78.

[0070] FIG. 6. Improvements in CAPP-Seq performance with optimized library preparation procedures.

[0071] FIG. 7. Optimizing allele recovery from low input cfDNA during Illumina library preparation.

[0072] FIG. 8. CAPP-Seq performance with various amounts of input cfDNA.

[0073] FIG. 9. Analysis of CAPP-Seq background, allele detection threshold, and linearity. (a) Analysis of background rate for 6 NSCLC patient plasma samples and a healthy individual (Detailed Methods). (b) Analysis of biological background in (a) focusing on 107 recurrent somatic muta-

tions from a previously reported SNaPshot panel (Su et al. (2011) *J. Mol. Diagn.* 13:74-84). Mutations found in a given patient's tumor were excluded. The mean frequency for each patient (horizontal red line) was within confidence limits of the mean background limit of 0.007% (horizontal blue line; panel a). A single outlier mutation (TP53 R175H) is indicated by an orange diamond. (c) Individual mutations from (b) ranked by most to least recurrent, according to median frequency across the 7 samples. (d) Dilution series analysis of expected versus observed frequencies of mutant alleles using CAPP-Seq. Dilution series were generated by spiking fragmented HCC78 DNA into control cfDNA. (e) Analysis of the effect of the number of SNVs considered on the estimates of fractional abundance (95% confidence intervals shown in gray). (f) Analysis of the effect of the number of SNVs considered on the mean correlation coefficient between expected and observed cancer fractions (blue dashed line) using data from panel (d). 95% confidence intervals are shown for (a)-(c). Statistical variation for (d) is shown as s.e.m.

[0074] FIG. 10. Empirical spiking analysis of CAPP-Seq using two NSCLC cell lines. (a) Expected and observed (by CAPP-Seq) fractions of NCI-H3122 DNA spiked into control HCC78 DNA are linear for all fractions tested (0.1%, 1%, and 10%; $R^2=1$). (b) Using data from (a), analysis of the effect of the number of SNVs considered on the estimates of fractional abundance (95% confidence intervals shown in gray). (c) Analysis of the effect of the number of SNVs considered on the mean correlation coefficient and coefficient of variation between expected and observed cancer fractions (dashed lines) using data from panel (a). (d) Expected and observed fractions of the EML4-ALK fusion present in HCC78 are linear ($R^2=0.995$) over all spiking concentrations tested (see FIG. 5(b) for breakpoint verification). The observed EML4-ALK fractions were normalized based on the relative abundance of the fusion in 100% H3122 DNA (see Detailed Methods for details). Moreover, a single heterozygous insertion (indel) discovered within the selector space of NCI-H3122 (chr7: 107416855, +T) was concordant with defined concentrations (shown are observed fractions adjusted for zygosity).

[0075] FIG. 11. Application of CAPP-Seq for noninvasive detection and monitoring of circulating tumor DNA. (a) Characteristics of 11 patients included in this study (Table 3). P-values reflect a two-sided paired t-test for patients with reporter SNVs detected at both time points; other p-values were determined as described in Methods. ND, mutant DNA was not detected above background. Dashes, plasma sample not available. Smoking history, ≥ 20 pack years (heavy), > 0 pack years (light). (b-d) Disease monitoring using CAPP-Seq. Mutant allele frequencies (left y-axis) and absolute concentrations (right y-axis) are shown. The lower limit of detection (defined in FIG. 2(a)-(b)) is indicated by the dashed lines. (b) Pre- and post-surgery circulating tumor DNA levels quantified by CAPP-Seq in a Stage IB and a Stage IIIA NSCLC patient. Complete resections were achieved in both cases. (c) Disease burden changes in response to chemotherapy in a Stage IV NSCLC patient with three rearrangement breakpoints identified by CAPP-Seq. Tumor volume based on CT measurements and CAPP-Seq mutant allele frequencies are shown. Tu, tumor; Ef, pleural effusion. (d) Detection and monitoring of a subclonal EGFR T790M resistance mutation in a patient with Stage IV NSCLC. The fractional abundance of the dominant clone and T790M-containing clone are shown in the primary tumor (left) and plasma samples (right). (e) Predicted transcripts of three fusion genes detected in case

P9. (f) Statistically significant co-occurrence of ROS1 fusions and U2AF1 S34F mutations in NSCLC ($P=0.0019$; two-sided Fisher's exact test). (g) Exploratory analysis of the potential application of CAPP-Seq for cancer screening. Pre-treatment plasma samples from panel (a) and a plasma sample from a healthy individual were examined for the presence of mutant allele outliers without knowledge of the primary tumor mutations (see Detailed Methods). Error bars represent s.e.m.

[0076] FIG. 12. Base-pair resolution breakpoint mapping for all patients and cell lines enumerated by FACTERA. Gene fusions involving ALK (a) and ROS1 (b) are graphically depicted. Schematics in the top panels indicate the exact genomic positions (HG19 NCBI Build 37.1/GRCh37) of the breakpoints in ALK, ROS1, EML4, KIF5B, SLC34A2, CD74, MKX, and FYN. Bottom panels depict exons flanking the predicted gene fusions with notation indicating the 5' fusion partner gene and last fused exon followed by the 3' fusion partner gene and first fused exon. For example, in S13del37; R34 exons 1-13 of SLC34A2 (excluding the 3' 37 nucleotides of exon 13) are fused to exons 34-43 of ROS1. Exons in FYN are from its 5'UTR and precede the first coding exon. The green dotted line in the predicted FYN-ROS1 fusion indicates the first in-frame methionine in ROS1 exon 33, which preserves an open reading frame encoding the ROS1 kinase domain. All rearrangements were each independently confirmed by PCR and/or FISH.

[0077] FIG. 13. Presence of fusions is inversely related to the number of SNVs detected by CAPP-Seq. For each patient listed in FIG. 11(a) the number of identified SNVs versus the presence or absence of detected genomic fusions are plotted. The shading of the symbols is identical to FIG. 11(a), and indicates smoking history. Statistical significance was determined using a two-sided Wilcoxon rank sum test, and error bars indicate s.e.m.

[0078] FIG. 14. Different types of reporters are similarly useful for disease monitoring. Three SNVs and an ALK translocation identified in patient 6 are concordant at each time point, showing a comparable drop in fractional abundance after treatment with the ALK kinase inhibitor Crizotinib. Due to small differences in measured allele frequencies at each time point, linear regression was used to fit all allele frequencies to their adjusted mutant cfDNA concentrations ($R^2=0.93$). Thus, the scale on the right y-axis is interpolated. To accurately quantify disease burden, translocation and SNV frequencies were adjusted based on differences in zygosity and sequencing depth in the tumor sample (see Detailed Methods).

[0079] FIG. 15. Flow cytometry-analysis of P9 pleural effusion. Flow cytometry of cryopreserved cells from a pleural effusion revealed only 0.22% of cells stained positive for the epithelial marker, EpCAM, and negative for the lineage markers CD31 (endothelial cells) and CD45 (immune cells). FACS was used to enrich tumor cells and analysis of tumor-enriched genomic DNA identified 3 fusions (FIG. 11(e)), while unsorted low purity tumor specimen hampered de novo fusion discovery using FACTERA (Detailed Methods).

[0080] FIG. 16. Analysis of RNA-Seq data from lung adenocarcinoma patients in TCGA identifies 2 candidate cases with ROS1 rearrangements. (a) ROS1 fusions are known to result in over-expression of the C-terminal kinase domain, and breakpoints typically occur downstream of exon 31 (Bergethon et al. (2012) *J. Clin. Oncol.* 30:863-870; Rikova et al. (2007) *Cell* 131:1190-1203; Takeuchi et al.

(2012) *Nat. Med.* 18:378-381). Exon-level RPKM values for ROS1 are plotted for 163 LUAD patients. Two patients (TCGA-05-4426 and TCGA-64-1680) have expression patterns suggestive of ROS1 fusions. (b,c) Pileups of RNA-Seq reads in these two patients illustrate an abundance of reads mapping to regions surrounding ROS1 exon boundaries. Colored reads indicate discordant pairs, consistent with ROS1 fusions. Such pairs map to SLC34A2 for patient TCGA-05-4426 (b) and CD74 for patient TCGA-64-1680 (c). A single soft-clipped RNA-Seq read supports a ROS1-CD74 fusion event in TCGA-64-1680.

[0081] FIG. 17. Non-invasive cancer screening with CAPP-Seq, related to FIG. 11(g). (a) Steps to identify candidate SNVs in plasma cfDNA demonstrated using a patient sample with NSCLC (P6, see Table 3). Following stepwise filtration, outlier detection is applied (Detailed Methods). (b) Same as (a), but using a plasma cfDNA sample from a patient who had their tumor surgically removed. No SNVs are identified, as expected. (c) Three additional representative samples applying retrospective screening to patients analyzed in this study. P2 and P5 samples have confirmed tumor-derived SNVs, while P9 is cancer positive but lacks tumor-derived SNVs. Red points, confirmed tumor-derived SNVs; Green points, background noise.

DETAILED DESCRIPTION OF THE INVENTION

[0082] Tumors continually shed DNA into the circulation, where it is readily accessible. Stroun et al. (1987) *Eur J Cancer Clin Oncol* 23:707-712. Provided herein are methods for the ultrasensitive detection of circulating tumor DNA called Cancer Personalized Profiling by Deep Sequencing (CAPP-Seq). Also provided are methods for creating libraries of recurrently mutated genomic regions used in the CAPP-Seq methods. CAPP-Seq targets hundreds of recurrently mutated genomic regions and simultaneously detects point mutations, insertions/deletions, and rearrangements. CAPP-Seq for non-small cell lung cancer has been demonstrated herein with a design that identified mutations in >95% of tumors. CAPP-Seq accurately quantified circulating tumor DNA from early and advanced stage tumors and identified mutant alleles down to 0.025% with a detection limit of <0.01%. Tumor-derived DNA levels paralleled clinical responses to diverse therapies and CAPP-Seq identified actionable mutations in plasma. Moreover, CAPP-Seq identified significant co-occurrence of ROS1 translocations with U2AF1 splicing factor mutations. Finally, the utility of CAPP-Seq for cancer screening is also described. CAPP-Seq can be routinely applied to noninvasively detect and monitor tumors, thus facilitating personalized cancer therapy.

Methods for Creating Libraries

[0083] According to one aspect of the invention, methods for creating a library of recurrently mutated genomic regions are provided. The methods comprise the step of identifying a plurality of genomic regions from a group of genomic regions that are recurrently mutated in a specific cancer, wherein the library comprises the plurality of genomic regions, the plurality of genomic regions comprises at least 10 different genomic regions, and at least one mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

[0084] It should be understood that the term “library” represents a compilation or collection of individual components.

Thus, a library of recurrently mutated genomic regions is a compilation or collection of recurrently mutated genomic regions. The libraries of the instant disclosure are useful because they include a large number of potentially mutated genomic regions within a minimal length of genomic sequence. Use of these libraries to identify genetic alterations in specific patient samples is particularly advantageous because the libraries do not need to be optimized on a patient-by-patient basis.

[0085] The libraries created according to the instant methods comprise genomic regions that are recurrently mutated in a specific cancer. The identification of these recurrent mutations benefits greatly from the availability of databases such as, for example, The Cancer Genome Atlas (TCGA) and its subsets (<http://cancergenome.nih.gov/>). Such databases serve as the starting point for identifying the recurrently mutated genomic regions of the instant libraries. The databases also provide a sample of mutations occurring within a given percentage of subjects with a specific cancer.

[0086] The libraries created according to the instant methods comprise a plurality of genomic regions, wherein the plurality of genomic regions comprises at least 10 different genomic regions. In some embodiments, the plurality of genomic regions comprises at least 25, at least 50, at least 100, at least 150, at least 200, at least 500, or even more different genomic regions.

[0087] It should be understood that the inclusion of larger numbers of genomic regions generally increases the likelihood that a unique mutation will be identified to distinguish tumor nucleic acid in a subject from the subject's genomic nucleic acid. Including too many genomic regions in the library is not without a cost, however, since the number of genomic regions is directly related to the length of nucleic acids that must be sequenced in the analysis. At the extreme, the entire genome of a tumor sample and a genomic sample could be sequenced, and the resulting sequences could be compared to note any differences. Such a brute force approach is not possible, however, with the vanishingly small quantities of tumor nucleic acid present in a cell-free sample.

[0088] The libraries of the instant disclosure address this problem by identifying genomic regions that are recurrently mutated in a particular cancer, and then ranking those regions to maximize the likelihood that the region will include a distinguishing genetic alteration in a particular tumor. The library of recurrently mutated genomic regions, or “selectors”, can be used across an entire population for a given cancer, and does not need to be optimized for each subject.

[0089] The term “mutation”, as used herein, refers to a genetic alteration in the genome of an organism, specifically to a change in the nucleotide sequence of the organism. Examples of mutations include point mutations, where a single nucleotide is changed in the genome, and larger-scale changes in the genome, such as rearrangements, insertions, deletions, and amplifications. A recurrent mutation is a mutation that has been identified in more than one individual.

[0090] The terms “patient” and “subject” are used interchangeably. These are typically individuals that suffer from the cancer of interest. While the individuals are typically human individuals, the methods and systems of the instant disclosure could also be applied to other species, in particular, to other animal species, for example, livestock animals and pets.

[0091] The libraries of recurrently mutated genomic regions disclosed herein are created for a given type of cancer using one or more of the following design phases:

Phase 1: Identify known “driver” genes, i.e., genes that are known to be mutated frequently in the particular cancer.

Phase 2: Maximize patient coverage by selecting genomic regions that contain recurrent mutations in multiple subjects with the particular cancer and ranking those selections to maximize the number of patients identified by mutations in those regions.

Phases 3 and 4: Further ranking of genomic regions containing recurrent mutations by maximizing the “recurrence index”.

Phase 5: Add genomic regions from genes predicted to harbor “driver” mutations in the particular cancer.

Phase 6: Add genomic regions covering fusions and their flanking regions.

[0092] It should be understood, however, that the above-described phases of selector design are independent of one another and may be applied separately or in a different order within the methods of library creating and still achieve the desired result.

[0093] Application of the above approaches for recurrently mutated genomic regions in non-small cell lung cancer results in the library shown in Table 1. All genomic regions included in the selector, along with their corresponding HUGO gene symbols and genomic coordinates, as well as patient statistics for NSCLC and a variety of other cancers, are shown, organized by selector design phase. The percentage of coverage of NSCLC patients as the Table 1 library was developed is shown in FIG. 1(b). Also shown in the bottom panel of this figure is the cumulative length of genomic regions (in kb) as the library is created according to the above phasing. The three curves in the top panel show percentage coverage of patients with at least one distinguishing mutation between tumor and genomic sequences (≥ 1 SNVs), at least two distinguishing mutations between tumor and genomic sequences (≥ 2 SNVs), and at least three distinguishing mutations between tumor and genomic sequences (≥ 3 SNVs). As is apparent from these graphs, the library created according to the instant methods identifies genomic regions that are highly likely to include identifiable mutations in tumor sequences. This library includes a relatively small total number of genomic regions and thus a relatively short cumulative length of genomic regions and yet provides a high overall coverage of likely mutations in a population. The library does not, therefore, need to be optimized on a patient-by-patient basis. The relatively short cumulative length of genomic regions also means that the analysis of cancer-derived cell-free DNA using these libraries is highly sensitive and allows the sequencing of this DNA to a great depth.

[0094] Accordingly, the libraries of recurrently mutated genomic regions created using the instant methods comprise a plurality of genomic regions that are recurrently mutated in a specific cancer, and the plurality of genomic regions comprises at least 10 different genomic regions. In some embodiments, the plurality of genomic regions comprises at least 25 different genomic regions. In some embodiments, the plurality of genomic regions comprises at least 50 different genomic regions. In some embodiments, the plurality of genomic regions comprises at least 100 different genomic regions. In some embodiments, the plurality of genomic regions comprises at least 150 different genomic regions. In some embodiments, the plurality of genomic regions com-

prises at least 200 different genomic regions. In some embodiments, the plurality of genomic regions comprises at least 500 different genomic regions or even more.

[0095] In some embodiments, the plurality of genomic regions comprises at most 5000 different genomic regions. In some embodiments, the plurality of genomic regions comprises at most 2000 different genomic regions. In some embodiments, the plurality of genomic regions comprises at most 1000 different genomic regions. In some embodiments, the plurality of genomic regions comprises at most 500 different genomic regions. In some embodiments, the plurality of genomic regions comprises at most 200 different genomic regions. In some embodiments, the plurality of genomic regions comprises at most 150 different genomic regions. In some embodiments, the plurality of genomic regions comprises at most 100 different genomic regions. In some embodiments, the plurality of genomic regions comprises at most 50 different genomic regions or even fewer.

[0096] Importantly, the libraries of recurrently mutated genomic regions created according to the instant methods enable the identification of patient- and tumor-specific mutations within the genomic regions in a high percentage of subjects. Specifically, in these libraries, at least one mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer. In some embodiments, at least two mutations within the plurality of genomic regions are present in at least 60% of all subjects with the specific cancer. In specific embodiments, at least three mutations, or even more, within the plurality of genomic regions are present in at least 60% of all subjects with the specific cancer.

[0097] In some embodiments, in the libraries of recurrently mutated genomic regions created according to these methods, at least one mutation within the plurality of genomic regions is present in at least 60%, 70%, 80%, 90%, 95%, 98%, 99%, 99.9% or even higher percentages of all subjects with the specific cancer.

[0098] In specific embodiments, at least two mutations within the plurality of genomic regions are present in at least 60%, 70%, 80%, 90%, 95%, 98%, 99%, 99.9% or even higher percentages of all subjects with the specific cancer.

[0099] In more specific embodiments, at least three mutations, or even more, within the plurality of genomic regions are present in at least 60%, 70%, 80%, 90%, 95%, 98%, 99%, 99.9% or even higher percentages of all subjects with the specific cancer.

[0100] As previously noted, the cumulative length of genomic regions in the libraries of recurrently mutated genomic regions created according to the instant methods are relatively short, thus minimizing sequencing costs associated with the analytical methods relying on these libraries and maximizing their sensitivity. In some embodiments, the cumulative length of genomic regions is at most 30 megabases (Mb). In some embodiments, the cumulative length of genomic regions is at most 20 Mb, 10 Mb, 5 Mb, 2 Mb, or 1 Mb. In some embodiments, the cumulative length of genomic regions is at most 500 kilobases (kb), 200 kb, 100 kb, 50 kb, 20 kb, 10 kb, or even fewer.

[0101] In some embodiments, the library of recurrently mutated genomic regions created according to the instant methods comprises the genomic regions displayed in Table 1, or a subset of those genomic regions.

[0102] The instant methods include the step of identifying a plurality of genomic regions from a group of genomic

regions that are recurrently mutated in a specific cancer. As noted elsewhere, the libraries are particularly useful in methods for analyzing cancer-specific gene alterations in solid tumors, because those alterations can be detected in cell-free nucleic acids present in blood samples. Accordingly, the libraries created according to these methods include genomic regions that are recurrently mutated in a solid tumor. In some embodiments, the solid tumor is a carcinoma. In specific embodiments, the carcinoma is an adenocarcinoma, a non-small cell lung cancer, or a squamous cell carcinoma. The methods are also applicable to genomic regions that are recurrently mutated in other cancers, however. Specifically, the other cancer may be, for example, a sarcoma, a leukemia, a lymphoma, or a myeloma.

Systems

[0103] The methods for creating a library of recurrently mutated genomic regions, as disclosed herein, are typically implemented by a programmed computer system. Therefore, according to another aspect, the instant disclosure provides computer systems for creating a library of recurrently mutated genomic regions. Such systems comprise at least one processor and a non-transitory computer-readable medium storing computer-executable instructions that, when executed by the at least one processor, cause the computer system to carry out the above-described methods for creating a library.

Methods for Analyzing Genetic Alterations

[0104] The libraries created according to the above-described methods are useful in the analysis of genetic alterations, particularly in comparing tumor and genomic sequences in a patient with cancer. As shown in FIG. 2, a tissue biopsy sample from the patient may be used to discover mutations in the tumor by sequencing the genomic regions of the selector library in tumor and genomic nucleic acid samples and comparing the results. Because the selector libraries are designed to identify mutations in tumors from a large percentage of all patients, it is not necessary to optimize the library for each patient.

[0105] Accordingly, in this aspect of the invention, methods are provided for analyzing a cancer-specific genetic alteration in a subject comprising the steps of:

[0106] obtaining a tumor nucleic acid sample and a genomic nucleic acid sample from a subject with a specific cancer;

[0107] sequencing a plurality of target regions in the tumor nucleic acid sample and in the genomic nucleic acid sample to obtain a plurality of tumor nucleic acid sequences and a plurality of genomic nucleic acid sequences; and

[0108] comparing the plurality of tumor nucleic acid sequences to the plurality of genomic nucleic acid sequences to identify a patient-specific genetic alteration in the tumor nucleic acid sample.

[0109] In these methods, the plurality of target regions are selected from a plurality of genomic regions that are recurrently mutated in the specific cancer; the plurality of genomic regions comprises at least 10 different genomic regions; and at least one mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer. More specifically, the plurality of target regions may correspond to the plurality of genomic regions found in the libraries of recurrently mutated genomic regions created using the above-described methods. In other words, in vari-

ous embodiments, the number of different genomic regions in the plurality of genomic regions, the number of mutations within the plurality of genomic regions that are present in a specific percentage of all subjects with the specific cancer, the percentage of all subjects with the specific cancer with at least one mutation within the plurality of genomic regions, the specific composition of the plurality of genomic regions, the types of cancer, and the cumulative length of the plurality of genomic regions have the values disclosed above for the methods of creating a library.

[0110] In some embodiments, the plurality of target regions used in the methods for analyzing a cancer-specific genetic alteration in a subject corresponds to the library of recurrently mutated genomic regions displayed in Table 1, or a subset of those genomic regions.

[0111] It should be understood that the step of obtaining a tumor nucleic acid sample and a genomic nucleic acid sample from a subject with a specific cancer may occur in a single step or in separate steps. For example, it may be possible to obtain a single tissue sample from a patient, for example from a biopsy sample, that includes both tumor nucleic acids and genomic nucleic acids. It is also within the scope of this step to obtain the tumor nucleic acid sample and the genomic nucleic acid sample from the subject in separate samples, in separate tissues, or even at separate times.

[0112] The step of obtaining a tumor nucleic acid sample and a genomic nucleic acid sample from a subject with a specific cancer may also include the process of extracting a biological fluid or tissue sample from the subject with the specific cancer. These particular steps are well understood by those of ordinary skill in the medical arts, particularly by those working in the medical laboratory arts.

[0113] The step of obtaining a tumor nucleic acid sample and a genomic nucleic acid sample from a subject with a specific cancer may additionally include procedures to improve the yield or recovery of the nucleic acids in the sample. For example, the step may include laboratory procedures to separate the nucleic acids from other cellular components and contaminants that may be present in the biological fluid or tissue sample. As noted, such steps may improve the yield and/or may facilitate the sequencing reactions.

[0114] It should also be understood that the step of obtaining a tumor nucleic acid sample and a genomic nucleic acid sample from a subject with a specific cancer may be performed by a commercial laboratory that does not even have direct contact with the subject. For example, the commercial laboratory may obtain the nucleic acid samples from a hospital or other clinical facility where, for example, a biopsy or other procedure is performed to obtain tissue from a subject. The commercial laboratory may thus carry out all the steps of the instantly-disclosed methods at the request of, or under the instructions of, the facility where the subject is being treated or diagnosed.

Methods for Screening

[0115] The methods of the instant invention may also be applied to the detection of cancer in a patient, where there is no prior knowledge of the presence of a tumor in the patient.

Accordingly, in this aspect of the invention are provided methods for screening a cancer-specific genetic alteration in a subject comprising the steps of:

[0116] obtaining a cell-free nucleic acid sample from a subject;

[0117] sequencing a plurality of target regions in the cell-free sample to obtain a plurality of cell-free nucleic acid sequences; and

[0118] identifying a cancer-specific genetic alteration in the cell-free sample.

[0119] In these methods, the plurality of target regions are selected from a plurality of genomic regions that are recurrently mutated in the specific cancer. In some embodiments, the plurality of genomic regions comprises at least 10 different genomic regions, and at least one mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer. More specifically, the plurality of target regions may correspond to the plurality of genomic regions found in the libraries of recurrently mutated genomic regions created using the above-described methods. In other words, in various embodiments, the number of different genomic regions in the plurality of genomic regions, the number of mutations within the plurality of genomic regions that are present in a specific percentage of all subjects with the specific cancer, the percentage of all subjects with the specific cancer with at least one mutation within the plurality of genomic regions, the specific composition of the plurality of genomic regions, the types of cancer, and the cumulative length of the plurality of genomic regions have the values disclosed above for the methods of creating a library.

[0120] In some embodiments, the plurality of target regions used in the methods for screening a cancer-specific genetic alteration in a subject corresponds to the library of recurrently mutated genomic regions displayed in Table 1, or a subset of those genomic regions.

[0121] It will be readily apparent to one of ordinary skill in the relevant arts that other suitable modifications and adaptations to the methods and applications described herein may be made without departing from the scope of the invention or any embodiment thereof. Having now described the present invention in detail, the same will be more clearly understood by reference to the following Examples, which are included herewith for purposes of illustration only and are not intended to be limiting of the invention.

Examples

Noninvasive and Ultrasensitive Quantitation of Circulating Tumor DNA by Hybrid Capture and Deep Sequencing

[0122] To overcome the limitations of prior methods, an ultrasensitive and specific strategy for analysis of cancer-derived cfDNA (Cancer Personalized Profiling by Deep Sequencing (CAPP-Seq)) that can simultaneously detect single nucleotide variants (SNVs), insertions/deletions (indels), and rearrangements, without the need for patient-specific optimization has been developed. CAPP-Seq employs an adaptable “selector” to enrich recurrently mutated regions in the cancer of interest using a custom library of biotinylated DNA oligonucleotides (Ng et al. (2010) *Nat. Genetics* 42:30-

35). To use CAPP-Seq for monitoring circulating tumor DNA, this selector is typically applied first to matched tumor and normal genomic DNA to identify a patient’s cancer-specific genetic aberrations and then directly to cfDNA in order to quantify these mutations (FIG. 1a and FIG. 2).

[0123] The design of an NSCLC CAPP-Seq selector is shown in FIG. 1(b). Phase 1: Genomic regions harboring known/suspected driver mutations in NSCLC. Phases 2-4: Addition of exons containing recurrent SNVs using WES data from lung adenocarcinomas and squamous cell carcinomas from TCGA (N=407). Regions were selected iteratively to maximize the number of mutations per tumor while minimizing selector size. Recurrence index=total unique patients with mutations covered per kb of exon. Phases 5-6: Exons of predicted NSCLC drivers (Ding et al. (2008) *Nature* 455: 1069-1075; Youn & Simon (2011) *Bioinformatics* 27:175-181) and introns/exons harboring breakpoints in rearrangements involving ALK, ROS1, and RET were added. Bottom: increase of selector length during each design phase. FIG. 1(c) shows an analysis of the number of SNVs per lung adenocarcinoma covered by the NSCLC CAPP-Seq selector in the TCGA WES cohort (Training; N=229) and an independent lung adenocarcinoma WES data set (Validation; N=183) (Imielinski et al. (2012) *Cell* 150:1107-1120). Results are compared to selectors randomly sampled from the exome ($P < 1.0 \times 10^{-6}$) for the difference between random selectors and the NSCLC CAPP-Seq selector). FIG. 1(d) shows the number of SNVs per patient identified by the NSCLC CAPP-Seq selector in WES data from three adenocarcinomas from TCGA, colon (COAD), rectal (READ), and endometrioid (UCEC) cancers. FIGS. 1(e) and (f) show quality parameters from a representative CAPP-Seq analysis of plasma cfDNA, including length distribution of sequenced cfDNA fragments 1(e), and depth of sequencing coverage across all genomic regions in the selector 1(f). FIG. 1(g) illustrates the variation in sequencing depth across cfDNA samples from 4 patients. The envelope above and below the solid line represents s.e.m. FIG. 2 illustrates the CAPP-Seq computational pipeline. See Detailed Methods section for details.

[0124] For the initial implementation of CAPP-Seq we focused on NSCLC, although our approach is generalizable to any cancer for which a comprehensive list of recurrent mutations has been identified. We employed a multi-phase approach to design a NSCLC-specific selector, aiming to identify genomic regions recurrently mutated in this disease (FIG. 1b, Table 1, and Methods). We began by including exons covering recurrent mutations in potential driver genes from the Catalogue of Somatic Mutations in Cancer (COSMIC) database (Forbes et al. (2010) *Nucleic Acids Res.* 38:D652-657) as well as other sources (Ding et al. (2008) *Nature* 455:1069-1075; Youn & Simon (2011) *Bioinformatics* 27:175-181) (e.g. KRAS, EGFR, TP53). Next, using whole exome sequencing (WES) data from 407 NSCLC patients profiled by The Cancer Genome Atlas (TCGA), an iterative algorithm was applied to maximize the number of mutations per patient while minimizing selector size. The approach relied on a recurrence index that identified known driver mutations as well as uncharacterized genes that are frequently mutated and are therefore likely to be involved in NSCLC pathogenesis (FIG. 3 and Table 1).

TABLE 1

Recurrently mutated genomic regions in NSCLC.												
Design phase	Selector design			Genomic region					Coverage (unique LUAD & SCC patients; n = 407)			
	Regions covered	Genes covered	Length (bp)	Gene	Chr	Start (bp)	End (bp)	Length (bp)	Patients covered	Patients gained	No. patients per exon	RI
Known drivers	1	1	130	AKT1	chr14	105246424	105246553	130	1	1	1	7.7
Known drivers	2	2	250	BRAF	chr7	140453074	140453192	120	9	8	8	66.7
Known drivers	3	2	369	BRAF	chr7	140481375	140481493	119	16	7	7	58.8
Known drivers	4	3	677	CDKN2A	chr9	21970900	21971207	308	46	30	30	97.4
Known drivers	5	3	1029	CDKN2A	chr9	21974475	21974826	352	53	7	7	19.9
Known drivers	6	4	1258	CTNNB1	chr3	41266016	41266244	229	57	4	6	26.2
Known drivers	7	5	1382	EGFR	chr7	55241613	55241736	124	58	1	3	24.2
Known drivers	8	5	1482	EGFR	chr7	55242414	55242513	100	65	7	8	80.0
Known drivers	9	5	1669	EGFR	chr7	55248985	55249171	187	69	4	5	26.7
Known drivers	10	5	1826	EGFR	chr7	55259411	55259567	157	81	12	14	89.2
Known drivers	11	6	1926	ERBB2	chr17	37880164	37880263	100	81	0	0	0.0
Known drivers	12	6	2113	ERBB2	chr17	37880978	37881164	187	85	4	4	21.4
Known drivers	13	7	2293	HRAS	chr11	533765	533944	180	87	2	3	16.7
Known drivers	14	7	2405	HRAS	chr11	534211	534322	112	90	3	3	26.8
Known drivers	15	8	2583	KEAP1	chr19	10599867	10600044	178	93	3	3	16.9
Known drivers	16	8	2790	KEAP1	chr19	10600323	10600529	207	108	15	15	72.5
Known drivers	17	8	3477	KEAP1	chr19	10602252	10602938	687	128	20	25	36.4
Known drivers	18	8	4117	KEAP1	chr19	10610070	10610709	640	141	13	18	28.1
Known drivers	19	8	4285	KEAP1	chr19	10597327	10597494	168	143	2	2	11.9
Known drivers	20	9	4465	KRAS	chr12	25380167	25380346	180	147	4	4	22.2
Known drivers	21	9	4577	KRAS	chr12	25398207	25398318	112	191	44	56	500.0
Known drivers	22	10	4789	MEK1	chr15	66727364	66727575	212	191	0	0	0.0
Known drivers	23	11	4931	MET	chr7	116411902	116412043	142	193	2	2	14.1
Known drivers	24	12	5199	NFE2L2	chr2	178098732	178098998	268	212	19	31	115.7
Known drivers	25	13	5417	NOTCH1	chr9	139396723	139396940	218	212	0	1	4.6
Known drivers	26	13	5850	NOTCH1	chr9	139399124	139399556	433	212	0	0	0.0
Known drivers	27	13	7339	NOTCH1	chr9	139390522	139392010	1489	214	2	3	2.0
Known drivers	28	13	7489	NOTCH1	chr9	139397633	139397782	150	214	0	0	0.0
Known drivers	29	14	7669	NRAS	chr1	115256420	115256599	180	217	3	5	27.8
Known drivers	30	14	7781	NRAS	chr1	115258670	115258781	112	217	0	0	0.0
Known drivers	31	15	7907	PIK3CA	chr3	178935997	178936122	126	225	8	19	150.8
Known drivers	32	15	8179	PIK3CA	chr3	178951881	178952152	272	228	3	4	14.7
Known drivers	33	16	8259	PTEN	chr10	89624226	89624305	80	229	1	1	12.5
Known drivers	34	16	8345	PTEN	chr10	89653781	89653866	86	229	0	0	0.0
Known drivers	35	16	8391	PTEN	chr10	89685269	89685314	46	231	2	3	65.2
Known drivers	36	16	8436	PTEN	chr10	89690802	89690846	45	231	0	0	0.0
Known drivers	37	16	8676	PTEN	chr10	89692769	89693008	240	234	3	5	20.8
Known drivers	38	16	8819	PTEN	chr10	89711874	89712016	143	235	1	3	21.0
Known drivers	39	16	8987	PTEN	chr10	89717609	89717776	168	238	3	6	35.7
Known drivers	40	16	9213	PTEN	chr10	89720650	89720875	226	239	1	3	13.3
Known drivers	41	17	9504	STK11	chr19	1206912	1207202	291	240	1	4	13.7
Known drivers	42	17	9589	STK11	chr19	1218415	1218498	85	241	1	2	23.5
Known drivers	43	17	9680	STK11	chr19	1219322	1219412	91	242	1	1	11.0
Known drivers	44	17	9814	STK11	chr19	1220371	1220504	134	242	0	4	29.9
Known drivers	45	17	9952	STK11	chr19	1220579	1220716	138	242	0	4	29.0
Known drivers	46	17	10081	STK11	chr19	1221211	1221339	129	242	0	4	31.0
Known drivers	47	17	10140	STK11	chr19	1221947	1222005	59	242	0	0	0.0
Known drivers	48	17	10329	STK11	chr19	1222983	1223171	189	242	0	0	0.0
Known drivers	49	17	10524	STK11	chr19	1226452	1226646	195	242	0	0	0.0
Known drivers	50	18	10662	TP53	chr17	7577018	7577155	138	264	22	56	405.8
Known drivers	51	18	10773	TP53	chr17	7577498	7577608	111	286	22	50	450.5
Known drivers	52	18	10887	TP53	chr17	7578176	7578286	114	300	14	39	342.1
Known drivers	53	18	11167	TP53	chr17	7579311	7579590	280	312	12	31	110.7
Known drivers	54	18	11352	TP53	chr17	7578370	7578554	185	340	28	68	367.6
Max coverage	55	19	11472	REG1B	chr2	79313937	79314056	120	341	1	10	83.3
Max coverage	56	20	11527	TPTE	chr21	10970008	10970062	55	343	2	4	72.7
Max coverage	57	21	11641	CSMD3	chr8	113246593	113246706	114	345	2	8	70.2
Max coverage	58	21	11749	TP53	chr17	7573926	7574033	108	348	3	9	83.3
Max coverage	59	22	11861	FAM135B	chr8	139151228	139151339	112	350	2	8	71.4
Max coverage	60	23	11950	U2AF1	chr21	44524424	44524512	89	351	1	5	56.2
Max coverage	61	24	12084	THSD7A	chr7	11501637	11501770	134	352	1	9	67.2
Max coverage	62	25	12257	MLL3	chr7	151962122	151962294	173	353	1	11	63.6
Max coverage	63	26	12339	EYA4	chr6	133849862	133849943	82	354	1	5	61.0
Max coverage	64	27	12505	HCN1	chr5	45267190	45267355	166	355	1	9	54.2
Max coverage	65	28	12590	AKR1B10	chr7	134222945	134223029	85	357	2	5	58.8
Max coverage	66	29	12692	SLC6A5	chr11	20668379	20668480	102	358	1	5	49.0
Max coverage	67	30	12801	DPP10	chr2	116525872	116525980	109	360	2	6	55.0
Max coverage	68	31	12894	SCN7A	chr2	167327124	167327216	93	361	1	4	43.0
Max coverage	69	32	12988	SNTG1	chr8	51621445	51621538	94	362	1	5	53.2

TABLE 1-continued

Recurrently mutated genomic regions in NSCLC.												
Max coverage	70	33	13093	VPS13A	chr9	79946925	79947029	105	363	1	5	47.6
Max coverage	71	34	13240	IL1RAPL1	chrX	29938065	29938211	147	364	1	7	47.6
Max coverage	72	35	13408	CTNNA2	chr2	80085138	80085305	168	365	1	8	47.6
Max coverage	73	35	13598	CSMD3	chr8	113323206	113323395	190	366	1	9	47.4
Max coverage	74	36	13705	FAM5C	chr1	190203501	190203607	107	367	1	5	46.7
Max coverage	75	37	13813	CACNA1E	chr1	181708282	161708389	108	368	1	4	37.0
Max coverage	76	38	14528	KRTAP5-5	chr11	1651070	1651784	715	371	3	31	43.4
Max coverage	77	39	14650	PDE1C	chr7	31864480	31864601	122	372	1	5	41.0
Max coverage	78	40	14772	RYR2	chr1	237808626	237808747	122	373	1	5	41.0
Max coverage	79	41	14896	NRXN1	chr2	50733632	50733755	124	374	1	5	40.3
Max coverage	80	42	15021	COL19A1	chr6	70637800	70637924	125	375	1	5	40.0
Max coverage	81	42	15349	CSMD3	chr8	113697634	113697961	328	376	1	13	39.6
Max coverage	82	43	15551	LRP1B	chr2	141665445	141665646	202	377	1	7	34.7
Max coverage	83	44	15709	GKN2	chr2	69173435	69173592	158	378	1	6	38.0
Max coverage	84	45	16031	CD5L	chr1	157805624	157805945	322	379	1	12	37.3
Max coverage	85	46	16250	SPTA1	chr1	158627266	158627484	219	380	1	8	36.5
Max coverage	86	47	16392	DHX9	chr1	182812428	182812569	142	381	1	5	35.2
Max coverage	87	48	16535	ADAMTS20	chr12	43858393	43858535	143	382	1	5	35.0
Max coverage	88	49	16707	NLRP4	chr19	56382192	56382363	172	382	0	6	34.9
Max coverage	89	50	17199	CDH18	chr5	19473334	19473825	492	384	2	17	34.6
Max coverage	90	51	17344	MYH2	chr17	10450791	10450935	145	386	2	5	34.5
RI ≥ 30	91	52	18281	OR5L2	chr11	55594694	55595630	937	386	0	30	32.0
RI ≥ 30	92	53	19317	OR4A15	chr11	55135359	55136394	1036	386	0	32	30.9
RI ≥ 30	93	54	20245	OR6F1	chr1	247875130	247876057	928	386	0	26	28.0
RI ≥ 30	94	55	21176	OR4C6	chr11	55432642	55433572	931	387	1	27	29.0
RI ≥ 30	95	56	22224	OR2T4	chr1	248524882	248525929	1048	387	0	33	31.5
RI ≥ 30	96	56	23342	FAM5C	chr1	190067147	190068264	1118	387	0	35	31.3
RI ≥ 30	97	57	23598	PSG2	chr19	43575851	43576106	256	387	0	9	35.2
RI ≥ 30	98	58	23797	ITM2A	chrX	78618438	78618636	199	387	0	6	30.2
RI ≥ 30	99	59	24062	TNN	chr1	175092535	175092799	265	387	0	12	45.3
RI ≥ 30	100	60	24206	GATA3	chr10	8105958	8106101	144	387	0	3	20.8
RI ≥ 30	101	60	24369	HCN1	chr5	45461947	45462109	183	387	0	5	30.7
RI ≥ 30	102	61	24503	OCA2	chr15	28211835	28211968	134	387	0	6	44.8
RI ≥ 30	103	61	24686	CTNNA2	chr2	80816428	80816610	183	387	0	5	27.3
RI ≥ 30	104	62	24863	CNTN5	chr11	99715818	99715994	177	387	0	5	33.9
RI ≥ 30	105	63	25755	POM121L12	chr7	53103364	53104255	892	387	0	28	31.4
RI ≥ 30	106	64	25945	LRRC7	chr1	70225887	70226076	190	387	0	5	26.3
RI ≥ 30	107	65	26165	CNTNAP5	chr2	125530375	125530594	220	387	0	8	36.4
RI ≥ 30	108	66	26313	SLC4A10	chr2	162751188	162751335	148	387	0	5	33.8
RI ≥ 30	109	67	26412	SETD2	chr3	47142947	47143045	99	387	0	3	30.3
RI ≥ 30	110	68	26744	GFRAL	chr6	55216050	55216381	332	387	0	10	30.1
RI ≥ 30	111	69	26837	SORCS3	chr10	106927015	106927107	93	388	1	3	32.3
RI ≥ 30	112	70	27359	POTEG	chr14	19553416	19553937	522	388	0	17	32.6
RI ≥ 30	113	71	27489	F9	chrX	138630521	138630650	130	389	1	4	30.8
RI ≥ 30	114	72	27583	SLC26A3	chr7	107416896	107416989	94	389	0	2	21.3
RI ≥ 30	115	73	27753	UNC5D	chr8	35806044	35606213	170	389	0	5	29.4
RI ≥ 30	116	74	27860	PDE4DIP	chr1	144882775	144882881	107	389	0	4	37.4
RI ≥ 30	117	75	27943	MRPL1	chr4	78870950	78871032	83	389	0	4	48.2
RI ≥ 30	118	76	28013	COL25A1	chr4	109784474	109784543	70	389	0	3	42.9
RI ≥ 30	119	76	28161	SPTA1	chr1	158650372	158650519	148	389	0	5	33.8
RI ≥ 30	120	77	28309	TNR	chr1	175331798	175331945	148	369	0	5	33.8
RI ≥ 30	121	78	28491	GALNT13	chr2	155157921	155158102	182	389	0	6	33.0
RI ≥ 30	122	79	28618	EIF3E	chr8	109241298	109241424	127	389	0	5	39.4
RI ≥ 30	123	80	28691	SLC5A1	chr22	32445929	32446001	73	389	0	4	54.8
RI ≥ 30	124	81	28757	COASY	chr17	40717000	40717065	66	389	0	3	45.5
RI ≥ 30	125	82	28930	TBX15	chr1	119467268	119467440	173	389	0	7	40.5
RI ≥ 30	126	83	29099	PYHIN1	chr1	158908869	158909037	169	389	0	6	35.5
RI ≥ 30	127	84	29164	PSG5	chr19	43690493	43690557	65	389	0	3	46.2
RI ≥ 30	128	85	29262	BTRC	chr10	103290993	103291090	98	389	0	2	20.4
RI ≥ 30	129	86	29394	MDGA2	chr14	47324226	47324357	132	389	0	4	30.3
RI ≥ 30	130	87	29454	GUCY1A3	chr4	156629387	156629446	60	389	0	2	33.3
RI ≥ 30	131	88	29570	HGF	chr7	81386504	81386619	116	389	0	4	34.5
RI ≥ 30	132	89	29656	TIMD4	chr5	156346467	156346552	86	389	0	3	34.9
RI ≥ 30	133	90	29844	AK5	chr1	77752625	77752812	188	389	0	6	31.9
RI ≥ 30	134	91	30077	ODZ3	chr4	183245173	183245405	233	389	0	7	30.0
RI ≥ 30	135	92	30177	COL5A2	chr2	189927897	189927996	100	389	0	3	30.0
RI ≥ 30	136	93	30299	NTM	chr11	132180005	132180126	122	389	0	4	32.8
RI ≥ 30	137	94	30426	LTBP1	chr2	33500031	33500157	127	389	0	5	39.4
RI ≥ 30	138	95	30587	PRSS1	chr7	142458405	142458565	161	389	0	5	31.1
RI ≥ 30	139	95	30794	CDKN2A	chr9	21971001	21971207	207	389	0	26	125.6
RI ≥ 30	140	96	30922	CNGB3	chr8	87738758	87738885	128	389	0	4	31.3
RI ≥ 30	141	97	31049	SI	chr3	164777689	164777815	127	389	0	4	31.5
RI ≥ 30	142	97	31135	SI	chr3	164767578	164767663	86	389	0	4	46.5
RI ≥ 30	143	98	31320	TMEM132D	chr12	129822176	129822362	185	389	0	6	32.4
RI ≥ 30	144	99	31429	ASTN1	chr1	176998769	176998877	109	389	0	3	27.5

TABLE 1-continued

Recurrently mutated genomic regions in NSCLC.											
RI ≥ 30	145	100	31571	SAGE1	chrX	134987410	134987551	142	389	0	6 42.3
RI ≥ 30	146	100	31709	THSD7A	chr7	11464322	11464459	138	389	0	5 36.2
RI ≥ 30	147	101	31907	ADAMTS12	chr5	33683963	33684160	198	389	0	6 30.3
RI ≥ 30	148	101	32090	NRXN1	chr2	50463926	50464108	183	389	0	8 43.7
RI ≥ 30	149	101	32294	CSMD3	chr8	113562899	113563102	204	389	0	7 34.3
RI ≥ 30	150	101	32414	CSMD3	chr8	113364644	113364763	120	389	0	5 41.7
RI ≥ 30	151	102	32504	EPB41L4B	chr9	112018415	112018504	90	389	0	2 22.2
RI ≥ 30	152	103	32687	POLR3B	chr12	106820974	106821136	163	389	0	4 24.5
RI ≥ 30	153	104	32873	ATP10B	chr5	160097469	180097674	208	389	0	7 34.0
RI ≥ 30	154	105	33001	CSMD1	chr8	3165216	3165343	128	389	0	4 31.3
RI ≥ 30	155	106	33164	FBN2	chr5	127648325	127648487	163	389	0	5 30.7
RI ≥ 30	156	107	33252	EXOC5	chr14	57684699	57684786	88	389	0	2 22.7
RI ≥ 30	157	108	33315	ANKRD30A	chr10	37440987	37441049	63	389	0	3 47.6
RI ≥ 30	158	109	33414	TRIML1	chr4	189065189	189065287	99	389	0	4 40.4
RI ≥ 30	159	109	33538	SPTA1	chr1	158631076	158631199	124	389	0	4 32.3
RI ≥ 30	160	110	33699	POLDIP2	chr17	26684313	26684473	161	389	0	5 31.1
RI ≥ 30	161	111	33863	KLHL1	chr13	70314525	70314688	164	389	0	5 30.5
RI ≥ 20	162	112	34454	TRIM58	chr1	248039201	248039791	591	389	0	14 23.7
RI ≥ 20	163	113	34563	GRIA3	chrX	122537262	122537370	109	389	0	3 27.5
RI ≥ 20	164	114	34777	CNOT4	chr7	135048605	135048818	214	389	0	5 23.4
RI ≥ 20	165	115	34947	NAV3	chr12	78582388	78582557	170	389	0	4 23.5
RI ≥ 20	166	115	35975	NAV3	chr12	78400198	78401225	1028	389	0	22 21.4
RI ≥ 20	167	116	36354	TRPC5	chrX	111195270	111195648	379	389	0	8 21.1
RI ≥ 20	168	117	36480	LRRC2	chr3	46592956	46593081	126	389	0	3 23.8
RI ≥ 20	169	118	36726	ADAMTS16	chr5	5239793	5240038	246	389	0	6 24.4
RI ≥ 20	170	119	36869	ACER2	chr9	19424697	19424839	143	389	0	3 21.0
RI ≥ 20	171	120	37103	AMOT	chrX	112024113	112024346	234	389	0	5 21.4
RI ≥ 20	172	121	37215	OBP2A	chr9	138439716	138439827	112	389	0	3 26.8
Predicted drivers	173	122	38109	INHBA	chr7	41729247	41730140	894	389	0	17 19.0
Predicted drivers	174	122	38498	INHBA	chr7	41739584	41739972	389	389	0	3 7.7
Predicted drivers	175	123	38605	EPHA5	chr4	66189831	66189937	107	389	0	3 28.0
Predicted drivers	176	123	38762	EPHA5	chr4	66197690	66197846	157	389	0	2 12.7
Predicted drivers	177	123	38957	EPHA5	chr4	66201649	66201843	195	389	0	2 10.3
Predicted drivers	178	123	39108	EPHA5	chr4	66213771	66213921	151	389	0	3 19.9
Predicted drivers	179	123	39319	EPHA5	chr4	66217106	66217316	211	389	0	4 19.0
Predicted drivers	180	123	39420	EPHA5	chr4	66218740	66218840	101	389	0	2 19.8
Predicted drivers	181	123	39607	EPHA5	chr4	66230734	66230920	187	389	0	3 16.0
Predicted drivers	182	123	39734	EPHA5	chr4	66231649	66231775	127	389	0	3 23.6
Predicted drivers	183	123	39835	EPHA5	chr4	66233058	66233158	101	389	0	2 19.8
Predicted drivers	184	123	39936	EPHA5	chr4	66242698	66242798	101	389	0	0 0.0
Predicted drivers	185	123	40040	EPHA5	chr4	66270091	66270194	104	389	0	2 19.2
Predicted drivers	186	123	40201	EPHA5	chr4	66280001	66280161	161	389	0	1 6.2
Predicted drivers	187	123	40327	EPHA5	chr4	66286158	66286283	126	389	0	0 0.0
Predicted drivers	188	123	40664	EPHA5	chr4	66356094	66356430	337	389	0	5 14.8
Predicted drivers	189	123	40821	EPHA5	chr4	66361105	66361261	157	389	0	1 6.4
Predicted drivers	190	123	41486	EPHA5	chr4	66467358	86468022	665	389	0	6 9.0
Predicted drivers	191	123	41588	EPHA5	chr4	66509062	66509163	102	389	0	0 0.0
Predicted drivers	192	123	41770	EPHA5	chr4	66535279	66535460	182	389	0	1 5.5
Predicted drivers	193	124	41871	EPHA3	chr3	89156892	89156992	101	389	0	0 0.0
Predicted drivers	194	124	41973	EPHA3	chr3	89176340	89176441	102	389	0	2 19.6
Predicted drivers	195	124	42635	EPHA3	chr3	89259009	89259670	662	389	0	6 9.1
Predicted drivers	196	124	42792	EPHA3	chr3	89390065	89390221	157	389	0	4 25.5
Predicted drivers	197	124	43129	EPHA3	chr3	89390904	89391240	337	389	0	3 8.9
Predicted drivers	198	124	43255	EPHA3	chr3	89444986	89445111	126	389	0	2 15.9
Predicted drivers	199	124	43445	EPHA3	chr3	89448467	89448656	190	389	0	1 5.3
Predicted drivers	200	124	43549	EPHA3	chr3	89456418	89456521	104	389	0	0 0.0
Predicted drivers	201	124	43651	EPHA3	chr3	89457198	89457299	102	389	0	0 0.0
Predicted drivers	202	124	43778	EPHA3	chr3	89462290	89462416	127	389	0	3 23.6
Predicted drivers	203	124	43965	EPHA3	chr3	89468354	89468540	187	389	0	1 5.3
Predicted drivers	204	124	44066	EPHA3	chr3	89478236	89478336	101	389	0	0 0.0
Predicted drivers	205	124	44277	EPHA3	chr3	89480299	89480509	211	389	0	4 19.0
Predicted drivers	206	124	44428	EPHA3	chr3	89498374	89498524	151	389	0	1 6.6
Predicted drivers	207	124	44623	EPHA3	chr3	89499326	89499520	185	389	0	2 10.3
Predicted drivers	208	124	44780	EPHA3	chr3	89521613	89521769	157	389	0	3 19.1
Predicted drivers	209	124	44887	EPHA3	chr3	89528546	89528652	107	389	0	1 9.3
Predicted drivers	210	125	44989	PTPRD	chr9	8317857	8317958	102	389	0	2 19.6
Predicted drivers	211	125	45126	PTPRD	chr9	8319830	8319966	137	389	0	0 0.0
Predicted drivers	212	125	45282	PTPRD	chr9	8331581	8331736	156	389	0	1 6.4
Predicted drivers	213	125	45409	PTPRD	chr9	8338921	8339047	127	389	0	2 15.7
Predicted drivers	214	125	45537	PTPRD	chr9	8340342	8340469	128	389	0	1 7.8
Predicted drivers	215	125	45717	PTPRD	chr9	8341089	8341268	180	389	0	0 0.0
Predicted drivers	216	125	46004	PTPRD	chr9	8341692	8341978	287	389	0	2 7.0
Predicted drivers	217	125	46160	PTPRD	chr9	8375935	8376090	156	389	0	1 6.4
Predicted drivers	218	125	46281	PTPRD	chr9	8376606	8376726	121	389	0	1 8.3
Predicted drivers	219	125	46458	PTPRD	chr9	8389231	8389407	177	389	0	0 0.0

TABLE 1-continued

Recurrently mutated genomic regions in NSCLC.												
Predicted drivers	220	125	46583	PTPRD	chr9	8404536	8404660	125	389	0	0	0.0
Predicted drivers	221	125	46684	PTPRD	chr9	8436590	8436690	101	389	0	1	9.9
Predicted drivers	222	125	46785	PTPRD	chr9	8437168	8437268	101	389	0	0	0.0
Predicted drivers	223	125	46899	PTPRD	chr9	8449724	8449837	114	389	0	3	26.3
Predicted drivers	224	125	47001	PTPRD	chr9	8454536	8454637	102	389	0	0	0.0
Predicted drivers	225	125	47163	PTPRD	chr9	8460410	8460571	162	389	0	5	18.5
Predicted drivers	226	125	47374	PTPRD	chr9	8465465	8465675	211	389	0	6	28.4
Predicted drivers	227	125	47476	PTPRD	chr9	8470989	8471090	102	389	0	1	9.8
Predicted drivers	228	125	47737	PTPRD	chr9	8484118	8484378	261	389	0	5	19.2
Predicted drivers	229	125	47839	PTPRD	chr9	8485226	8485327	102	389	0	0	0.0
Predicted drivers	230	125	48428	PTPRD	chr9	8485761	8436349	589	389	0	4	6.8
Predicted drivers	231	125	48547	PTPRD	chr9	8492861	8492979	119	389	0	1	8.4
Predicted drivers	232	125	48649	PTPRD	chr9	8497204	8497305	102	389	0	1	9.8
Predicted drivers	233	125	48844	PTPRD	chr9	8499646	8499840	195	389	0	2	10.3
Predicted drivers	234	125	49151	PTPRD	chr9	8500753	8501059	307	389	0	3	9.8
Predicted drivers	235	125	49297	PTPRD	chr9	8504260	8504405	146	389	0	1	6.8
Predicted drivers	236	125	49432	PTPRD	chr9	8507300	8507434	135	389	0	1	7.4
Predicted drivers	237	125	50015	PTPRD	chr9	8517847	8518429	583	389	0	9	15.4
Predicted drivers	238	125	50286	PTPRD	chr9	8521276	8521546	271	389	0	5	18.5
Predicted drivers	239	125	50387	PTPRD	chr9	8523468	8523568	101	389	0	1	9.9
Predicted drivers	240	125	50499	PTPRD	chr9	8524924	8525035	112	389	0	1	8.9
Predicted drivers	241	125	50600	PTPRD	chr9	8526585	8526685	101	389	0	0	0.0
Predicted drivers	242	125	50702	PTPRD	chr9	8527298	8527399	102	389	0	2	19.6
Predicted drivers	243	125	50892	PTPRD	chr9	8528590	8528779	190	389	0	4	21.1
Predicted drivers	244	125	51035	PTPRD	chr9	8633316	8633458	143	389	0	2	13.6
Predicted drivers	245	125	51182	PTPRD	chr9	8636698	8636644	147	389	0	2	13.6
Predicted drivers	246	125	51283	PTPRD	chr9	8733761	8733861	101	389	0	0	0.0
Predicted drivers	247	126	51507	KDR	chr4	55946107	55946330	224	389	0	1	4.5
Predicted drivers	248	126	51608	KDR	chr4	55948115	55948215	101	389	0	0	0.0
Predicted drivers	249	126	51709	KDR	chr4	55948702	55948802	101	389	0	2	19.8
Predicted drivers	250	126	51862	KDR	chr4	55953773	55953925	153	389	0	3	19.6
Predicted drivers	251	126	51969	KDR	chr4	55955034	55955140	107	389	0	2	18.7
Predicted drivers	252	126	52070	KDR	chr4	55955540	55955640	101	389	0	0	0.0
Predicted drivers	253	126	52183	KDR	chr4	55955857	55955969	113	389	0	1	8.8
Predicted drivers	254	126	52307	KDR	chr4	55956122	55956245	124	389	0	0	0.0
Predicted drivers	255	126	52408	KDR	chr4	55958782	55958882	101	389	0	2	19.8
Predicted drivers	256	128	52563	KDR	chr4	55960968	55961122	155	389	0	2	12.9
Predicted drivers	257	126	52665	KDR	chr4	55961737	55961838	102	389	0	2	19.6
Predicted drivers	258	126	52780	KDR	chr4	55962395	55962509	115	389	0	1	8.7
Predicted drivers	259	126	52886	KDR	chr4	55963828	55963933	106	389	0	3	28.3
Predicted drivers	260	126	53023	KDR	chr4	55964303	55964439	137	389	0	0	0.0
Predicted drivers	261	126	53131	KDR	chr4	55964863	55964970	108	389	0	2	18.5
Predicted drivers	262	126	53264	KDR	chr4	55968063	55968195	133	389	0	1	7.5
Predicted drivers	263	126	53412	KDR	chr4	55968528	55968675	148	389	0	2	13.5
Predicted drivers	264	126	53755	KDR	chr4	55970809	55971151	343	389	0	5	14.6
Predicted drivers	265	126	53865	KDR	chr4	55971998	55972107	110	389	0	2	18.2
Predicted drivers	266	126	53990	KDR	chr4	55972853	55972977	125	389	0	1	8.0
Predicted drivers	267	126	54148	KDR	chr4	55973903	55974060	158	389	0	2	12.7
Predicted drivers	268	126	54313	KDR	chr4	55976569	55976733	165	389	0	2	12.1
Predicted drivers	269	126	54429	KDR	chr4	55976820	55976935	116	389	0	1	8.6
Predicted drivers	270	126	54608	KDR	chr4	55979470	55979648	179	389	0	2	11.2
Predicted drivers	271	128	54749	KDR	chr4	55980292	55980432	141	389	0	0	0.0
Predicted drivers	272	126	54919	KDR	chr4	55981040	55981209	170	389	0	1	5.9
Predicted drivers	273	126	55051	KDR	chr4	55981447	55981578	132	389	0	4	30.3
Predicted drivers	274	126	55249	KDR	chr4	55984770	55984967	198	389	0	0	0.0
Predicted drivers	275	126	55350	KDR	chr4	55987260	55987360	101	389	0	1	9.9
Predicted drivers	276	126	55452	KDR	chr4	55991376	55991477	102	389	0	0	0.0
Predicted drivers	277	127	55639	NTRK3	chr15	88420165	88420351	187	389	0	0	0.0
Predicted drivers	278	127	55799	NTRK3	chr15	88423500	88423659	160	389	0	1	6.3
Predicted drivers	279	127	55900	NTRK3	chr15	88428895	88428995	101	389	0	0	0.0
Predicted drivers	280	127	56145	NTRK3	chr15	88472421	88472665	245	389	0	1	4.1
Predicted drivers	281	127	56319	NTRK3	chr15	88476242	88476415	174	389	0	4	23.0
Predicted drivers	282	127	56451	NTRK3	chr15	88483853	88483984	132	389	0	1	7.6
Predicted drivers	283	127	56571	NTRK3	chr15	88522575	88522694	120	389	0	0	0.0
Predicted drivers	284	127	56707	NTRK3	chr15	88524456	88524591	136	389	0	0	0.0
Predicted drivers	285	127	56897	NTRK3	chr15	88576087	88576276	190	389	0	2	10.5
Predicted drivers	286	127	57001	NTRK3	chr15	88669501	88669604	104	389	0	3	28.8
Predicted drivers	287	127	57103	NTRK3	chr15	88670374	88670475	102	389	0	0	0.0
Predicted drivers	288	127	57204	NTRK3	chr15	88671903	88672003	101	389	0	0	0.0
Predicted drivers	289	127	57502	NTRK3	chr15	88678331	88678628	298	389	0	7	23.5
Predicted drivers	290	127	57645	NTRK3	chr15	88679129	88679271	143	389	0	1	7.0
Predicted drivers	291	127	57789	NTRK3	chr15	88679697	88679840	144	389	0	2	13.9
Predicted drivers	292	127	57948	NTRK3	chr15	88680634	88680792	159	389	0	0	0.0
Predicted drivers	293	127	58050	NTRK3	chr15	88690549	88690650	102	389	0	0	0.0
Predicted drivers	294	127	58151	NTRK3	chr15	88726634	88726734	101	389	0	1	9.9

TABLE 1-continued

Recurrently mutated genomic regions in NSCLC.												
Predicted drivers	295	127	58253	NTRK3	chr15	88727442	88727543	102	389	0	1	9.8
Predicted drivers	296	126	58391	RB1	chr13	48878048	48878185	138	389	0	0	0.0
Predicted drivers	297	128	56519	RB1	chr13	48881415	48881542	128	389	0	3	23.4
Predicted drivers	298	128	58636	RB1	chr13	48916734	48916850	117	389	0	1	8.5
Predicted drivers	299	128	58757	RB1	chr13	48919215	48919335	121	389	0	1	8.3
Predicted drivers	300	128	58859	RB1	chr13	48921929	48922030	102	389	0	0	0.0
Predicted drivers	301	128	58960	RB1	chr13	48923075	48923175	101	389	0	0	0.0
Predicted drivers	302	128	59072	RB1	chr13	48934152	48934283	112	389	0	2	17.9
Predicted drivers	303	128	59216	RB1	chr13	48936950	48937093	144	389	0	0	0.0
Predicted drivers	304	128	59317	RB1	chr13	48939018	48939118	101	389	0	0	0.0
Predicted drivers	305	128	59428	RB1	chr13	48941629	48941739	111	389	0	3	27.0
Predicted drivers	306	128	59529	RB1	chr13	48942651	48942751	101	389	0	0	0.0
Predicted drivers	307	128	59630	RB1	chr13	48947534	48947634	101	389	0	2	19.8
Predicted drivers	308	128	59748	RB1	chr13	48951053	48951170	118	389	0	0	0.0
Predicted drivers	309	128	59850	RB1	chr13	48953707	48953808	102	389	0	2	19.6
Predicted drivers	310	128	59951	RB1	chr13	48954154	48954254	101	389	0	0	0.0
Predicted drivers	311	128	60053	RB1	chr13	48954288	48954389	102	389	0	1	9.8
Predicted drivers	312	128	60251	RB1	chr13	48955382	48955579	198	389	0	0	0.0
Predicted drivers	313	128	60371	RB1	chr13	49027128	49027247	120	389	0	0	0.0
Predicted drivers	314	128	60518	RB1	chr13	49030339	49030485	147	389	0	3	20.4
Predicted drivers	315	128	60665	RB1	chr13	49033823	49033969	147	389	0	1	6.8
Predicted drivers	316	128	60771	RB1	chr13	49037866	49037971	106	389	0	0	0.0
Predicted drivers	317	128	60886	RB1	chr13	49039133	49039247	115	389	0	1	8.7
Predicted drivers	318	128	61051	RB1	chr13	49039340	49039504	165	389	0	2	12.1
Predicted drivers	319	128	61153	RB1	chr13	49047460	49047561	102	389	0	0	0.0
Predicted drivers	320	128	61297	RB1	chr13	49050836	49050979	144	389	0	0	0.0
Predicted drivers	321	128	61398	RB1	chr13	49051465	49051565	101	389	0	0	0.0
Predicted drivers	322	128	61499	RB1	chr13	49054120	49054220	101	389	0	0	0.0
Predicted drivers	323	129	61946	ERBB4	chr2	212248339	212248785	447	389	0	3	6.7
Predicted drivers	324	129	62245	ERBB4	chr2	212251577	212251875	299	389	0	3	10.0
Predicted drivers	325	129	62346	ERBB4	chr2	212252643	212252743	101	389	0	0	0.0
Predicted drivers	326	129	62518	ERBB4	chr2	212285165	212285336	172	389	0	2	11.6
Predicted drivers	327	129	62619	ERBB4	chr2	212286730	212286830	101	389	0	1	9.9
Predicted drivers	328	129	62787	ERBB4	chr2	212288879	212289026	148	389	0	1	6.8
Predicted drivers	329	129	62868	ERBB4	chr2	212293120	212293220	101	389	0	0	0.0
Predicted drivers	330	129	63025	ERBB4	chr2	212295669	212295825	157	389	0	2	12.7
Predicted drivers	331	129	63212	ERBB4	chr2	212426627	212426813	187	389	0	1	5.3
Predicted drivers	332	129	63312	ERBB4	chr2	212483901	212484000	100	389	0	0	0.0
Predicted drivers	333	129	63436	ERBB4	chr2	212488646	212488769	124	389	0	0	0.0
Predicted drivers	334	129	63570	ERBB4	chr2	212495186	212495319	134	389	0	0	0.0
Predicted drivers	335	129	63672	ERBB4	chr2	212522465	212522566	102	389	0	2	19.6
Predicted drivers	336	129	63828	ERBB4	chr2	212530047	212530202	156	389	0	1	6.4
Predicted drivers	337	129	63929	ERBB4	chr2	212537885	212537985	101	389	0	1	9.9
Predicted drivers	338	129	64063	ERBB4	chr2	212543776	212543909	134	389	0	1	7.5
Predicted drivers	339	129	64264	ERBB4	chr2	212566691	212566891	201	389	0	2	10.0
Predicted drivers	340	129	64366	ERBB4	chr2	212568823	212568924	102	389	0	0	0.0
Predicted drivers	341	129	64467	ERBB4	chr2	212570029	212570129	101	389	0	1	9.8
Predicted drivers	342	129	64595	ERBB4	chr2	212576774	212576901	128	389	0	1	7.8
Predicted drivers	343	129	64710	ERBB4	chr2	212578259	212578373	115	389	0	1	8.7
Predicted drivers	344	129	64853	ERBB4	chr2	212587117	212587259	143	389	0	0	0.0
Predicted drivers	345	129	64973	ERBB4	chr2	212589800	212589919	120	389	0	2	16.7
Predicted drivers	348	129	65074	ERBB4	chr2	212615346	212615446	101	389	0	0	0.0
Predicted drivers	347	129	65210	ERBB4	chr2	212652749	212652884	136	389	0	1	7.4
Predicted drivers	348	129	65398	ERBB4	chr2	212812154	212812341	188	390	1	4	21.3
Predicted drivers	349	129	65551	ERBB4	chr2	212989476	212989628	153	390	0	2	13.1
Predicted drivers	350	129	65652	ERBB4	chr2	213403163	213403263	101	390	0	0	0.0
Predicted drivers	351	130	65754	NTRK1	chr1	156785575	156785676	102	390	0	0	0.0
Predicted drivers	352	130	65868	NTRK1	chr1	156811872	156811985	114	390	0	0	0.0
Predicted drivers	353	130	66061	NTRK1	chr1	156830726	156830938	213	390	0	0	0.0
Predicted drivers	354	130	66183	NTRK1	chr1	156834132	156834233	102	390	0	1	9.8
Predicted drivers	355	130	66284	NTRK1	chr1	156834505	156834605	101	390	0	0	0.0
Predicted drivers	356	130	66386	NTRK1	chr1	156836685	156836786	102	390	0	0	0.0
Predicted drivers	357	130	66533	NTRK1	chr1	156837895	156838041	147	390	0	1	6.8
Predicted drivers	358	130	66677	NTRK1	chr1	156838296	156838439	144	390	0	0	0.0
Predicted drivers	359	130	66811	NTRK1	chr1	156841414	156841547	134	390	0	0	0.0
Predicted drivers	360	130	67139	NTRK1	chr1	156843424	156843751	328	390	0	1	3.0
Predicted drivers	361	130	67240	NTRK1	chr1	156844133	156844233	101	390	0	0	0.0
Predicted drivers	362	130	67341	NTRK1	chr1	156844340	156844440	101	390	0	0	0.0
Predicted drivers	363	130	67445	NTRK1	chr1	156844697	156844800	104	390	0	0	0.0
Predicted drivers	364	130	67593	NTRK1	chr1	156845311	156845458	148	390	0	2	13.5
Predicted drivers	365	130	67725	NTRK1	chr1	156845871	156846002	132	390	0	3	22.7
Predicted drivers	366	130	67899	NTRK1	chr1	156846191	156846364	174	390	0	2	11.5
Predicted drivers	367	130	68141	NTRK1	chr1	156848913	156849154	242	390	0	4	16.5
Predicted drivers	368	130	68301	NTRK1	chr1	156849790	156849949	160	390	0	0	0.0
Predicted drivers	369	130	68488	NTRK1	chr1	156851248	156851434	187	390	0	0	0.0

TABLE 1-continued

Recurrently mutated genomic regions in NSCLC.												
Predicted drivers	370	131	68589	NF1	chr17	29422307	29422407	101	390	0	0	0.0
Predicted drivers	371	131	68734	NF1	chr17	29483000	29483144	145	390	0	0	0.0
Predicted drivers	372	131	68835	NF1	chr17	29486019	29486119	101	390	0	1	9.9
Predicted drivers	373	131	69027	NF1	chr17	29490203	29490394	192	390	0	1	5.2
Predicted drivers	374	131	89135	NF1	chr17	29496908	29497015	108	390	0	1	9.3
Predicted drivers	375	131	69236	NF1	chr17	29508423	29508523	101	390	0	0	0.0
Predicted drivers	376	131	69337	NF1	chr17	29508715	29508815	101	390	0	0	0.0
Predicted drivers	377	131	69496	NF1	chr17	29509525	29509683	159	390	0	1	6.3
Predicted drivers	378	131	69671	NF1	chr17	29527439	29527613	175	390	0	3	17.1
Predicted drivers	379	131	69795	NF1	chr17	29528054	29528177	124	390	0	0	0.0
Predicted drivers	380	131	69897	NF1	chr17	29528415	29528516	102	390	0	0	0.0
Predicted drivers	381	131	70030	NF1	chr17	29533257	29533389	133	390	0	0	0.0
Predicted drivers	382	131	70166	NF1	chr17	29541468	29541603	136	390	0	1	7.4
Predicted drivers	383	131	70281	NF1	chr17	29546022	29546136	115	390	0	1	8.7
Predicted drivers	384	131	70423	NF1	chr17	29548867	29549008	142	390	0	1	7.0
Predicted drivers	385	131	70548	NF1	chr17	29550461	29550585	125	390	0	0	0.0
Predicted drivers	386	131	70705	NF1	chr17	29552112	29552268	157	390	0	0	0.0
Predicted drivers	387	131	70956	NF1	chr17	29553452	29553702	251	390	0	1	4.0
Predicted drivers	386	131	71057	NF1	chr17	29554222	29554322	101	390	0	0	0.0
Predicted drivers	389	131	71158	NF1	chr17	29554532	29554632	101	390	0	1	9.9
Predicted drivers	390	131	71600	NF1	chr17	29556042	29556483	442	390	0	2	4.5
Predicted drivers	391	131	71741	NF1	chr17	29556852	29556992	141	390	0	1	7.1
Predicted drivers	392	131	71865	NF1	chr17	29557277	29557400	124	390	0	1	8.1
Predicted drivers	393	131	71966	NF1	chr17	29557851	29557951	101	390	0	0	0.0
Predicted drivers	394	131	72084	NF1	chr17	29559090	29559207	118	390	0	0	0.0
Predicted drivers	395	131	72267	NF1	chr17	29559717	29559899	183	390	0	2	10.9
Predicted drivers	396	131	72480	NF1	chr17	29560019	29560231	213	390	0	1	4.7
Predicted drivers	397	131	72643	NF1	chr17	29562628	29562790	163	390	0	2	12.3
Predicted drivers	398	131	72748	NF1	chr17	29562935	29563039	105	390	0	0	0.0
Predicted drivers	399	131	72885	NF1	chr17	29576001	29576137	137	390	0	0	0.0
Predicted drivers	400	131	72987	NF1	chr17	29579936	29580037	102	390	0	0	0.0
Predicted drivers	401	131	73147	NF1	chr17	29585361	29585520	160	390	0	0	0.0
Predicted drivers	402	131	73248	NF1	chr17	29588048	29586148	101	390	0	1	9.9
Predicted drivers	403	131	73396	NF1	chr17	29587386	29587533	148	390	0	2	13.5
Predicted drivers	404	131	73544	NF1	chr17	29588728	29588875	148	390	0	0	0.0
Predicted drivers	405	131	73656	NF1	chr17	29592246	29592357	112	390	0	0	0.0
Predicted drivers	406	131	74090	NF1	chr17	29652837	29653270	434	390	0	2	4.6
Predicted drivers	407	131	74432	NF1	chr17	29654516	29654857	342	390	0	3	8.8
Predicted drivers	408	131	74636	NF1	chr17	29657313	29657516	204	390	0	2	9.8
Predicted drivers	409	131	74831	NF1	chr17	29661855	29662049	195	390	0	3	15.4
Predicted drivers	410	131	74973	NF1	chr17	29663350	29683491	142	390	0	2	14.1
Predicted drivers	411	131	75254	NF1	chr17	29663652	29663932	281	390	0	0	0.0
Predicted drivers	412	131	75470	NF1	chr17	29664385	29664600	216	390	0	1	4.6
Predicted drivers	413	131	75571	NF1	chr17	29664817	29664917	101	390	0	1	9.9
Predicted drivers	414	131	75687	NF1	chr17	29665042	29665157	116	390	0	0	0.0
Predicted drivers	415	131	75790	NF1	chr17	29665721	29665823	103	390	0	2	19.4
Predicted drivers	416	131	75932	NF1	chr17	29667522	29667663	142	390	0	1	7.0
Predicted drivers	417	131	76060	NF1	chr17	29670026	29670153	128	390	0	2	15.6
Predicted drivers	418	131	76193	NF1	chr17	29676137	29676269	133	390	0	2	15.0
Predicted drivers	419	131	76330	NF1	chr17	29677200	29677336	137	390	0	0	0.0
Predicted drivers	420	131	76489	NF1	chr17	29679274	29679432	159	390	0	2	12.6
Predicted drivers	421	131	76613	NF1	chr17	29683477	29683600	124	390	0	0	0.0
Predicted drivers	422	131	76745	NF1	chr17	29683977	29684108	132	390	0	1	7.6
Predicted drivers	423	131	76847	NF1	chr17	29684286	29684387	102	390	0	1	9.8
Predicted drivers	424	131	76991	NF1	chr17	29685497	29685640	144	390	0	1	6.9
Predicted drivers	425	131	77093	NF1	chr17	29685959	29686060	102	390	0	0	0.0
Predicted drivers	426	131	77311	NF1	chr17	29687504	29687721	216	390	0	0	0.0
Predicted drivers	427	131	77455	NF1	chr17	29701030	29701173	144	390	0	1	6.9
Predicted drivers	428	132	77621	APC	chr5	112043414	112043579	166	390	0	0	0.0
Predicted drivers	429	132	77757	APC	chr5	112090587	112090722	136	390	0	0	0.0
Predicted drivers	430	132	77859	APC	chr5	112102014	112102115	102	390	0	1	9.8
Predicted drivers	431	132	78062	APC	chr5	112102885	112103087	203	390	0	2	9.9
Predicted drivers	432	132	78172	APC	chr5	112111325	112111434	110	390	0	1	9.1
Predicted drivers	433	132	78287	APC	chr5	112116486	112116600	115	390	0	0	0.0
Predicted drivers	434	132	78388	APC	chr5	112128134	112128234	101	390	0	0	0.0
Predicted drivers	435	132	78494	APC	chr5	112136975	112137080	106	390	0	0	0.0
Predicted drivers	436	132	78594	APC	chr5	112151191	112151290	100	390	0	0	0.0
Predicted drivers	437	132	78974	APC	chr5	112154662	112155041	380	390	0	1	2.6
Predicted drivers	438	132	79075	APC	chr5	112157590	112157690	101	390	0	0	0.0
Predicted drivers	439	132	79216	APC	chr5	112162804	112162944	141	390	0	0	0.0
Predicted drivers	440	132	79317	APC	chr5	112163614	112163714	101	390	0	0	0.0
Predicted drivers	441	132	79435	APC	chr5	112164552	112164669	118	390	0	2	16.9
Predicted drivers	442	132	79651	APC	chr5	112170647	112170862	216	390	0	0	0.0
Predicted drivers	443	132	86226	APC	chr5	112173249	112179823	6575	391	1	23	3.5
Predicted drivers	444	133	86327	ATM	chr11	108098337	108096437	101	391	0	0	0.0

TABLE 1-continued

Recurrently mutated genomic regions in NSCLC.										
Predicted drivers	445	133	86441 ATM	chr11	108098502	108098615	114	391	0	1 8.8
Predicted drivers	446	133	86588 ATM	chr11	108099904	108100050	147	391	0	0 0.0
Predicted drivers	447	133	86754 ATM	chr11	108106396	108106561	168	391	0	0 0.0
Predicted drivers	448	133	86921 ATM	chr11	108114679	108114845	167	391	0	0 0.0
Predicted drivers	449	133	87161 ATM	chr11	108115514	108115753	240	391	0	1 4.2
Predicted drivers	450	133	87326 ATM	chr11	108117690	108117854	165	391	0	0 0.0
Predicted drivers	451	133	87497 ATM	chr11	108119659	108119829	171	391	0	1 5.8
Predicted drivers	452	133	87870 ATM	chr11	108121427	108121799	373	391	0	0 0.0
Predicted drivers	453	133	88066 ATM	chr11	108122563	108122758	196	391	0	0 0.0
Predicted drivers	454	133	88187 ATM	chr11	108123541	108123641	101	391	0	1 9.9
Predicted drivers	455	133	88394 ATM	chr11	108124540	108124766	227	391	0	0 0.0
Predicted drivers	456	133	88521 ATM	chr11	108126941	108127067	127	391	0	1 7.9
Predicted drivers	457	133	88648 ATM	chr11	108128207	108128333	127	391	0	0 0.0
Predicted drivers	458	133	88749 ATM	chr11	108129707	108129807	101	391	0	0 0.0
Predicted drivers	459	133	88922 ATM	chr11	108137897	108138069	173	391	0	1 5.8
Predicted drivers	460	133	89123 ATM	chr11	108139136	108139336	201	391	0	0 0.0
Predicted drivers	461	133	89225 ATM	chr11	108141781	108141882	102	391	0	0 0.0
Predicted drivers	462	133	89382 ATM	chr11	108141977	108142133	157	391	0	0 0.0
Predicted drivers	463	133	89483 ATM	chr11	108143246	108143346	101	391	0	0 0.0
Predicted drivers	464	133	89615 ATM	chr11	108143448	108143579	132	391	0	1 7.6
Predicted drivers	465	133	89734 ATM	chr11	108150217	108150335	119	391	0	0 0.0
Predicted drivers	466	133	89909 ATM	chr11	108151721	108151895	175	391	0	0 0.0
Predicted drivers	467	133	90080 ATM	chr11	108153436	108153606	171	391	0	2 11.7
Predicted drivers	468	133	90328 ATM	chr11	108154953	108155200	248	391	0	1 4.0
Predicted drivers	469	133	90445 ATM	chr11	108158326	108158442	117	391	0	0 0.0
Predicted drivers	470	133	90573 ATM	chr11	108159703	108159830	128	391	0	1 7.8
Predicted drivers	471	133	90774 ATM	chr11	108160328	108160528	201	391	0	1 5.0
Predicted drivers	472	133	90950 ATM	chr11	108163345	108163520	176	391	0	0 0.0
Predicted drivers	473	133	91116 ATM	chr11	108164039	108164204	166	391	0	0 0.0
Predicted drivers	474	133	91250 ATM	chr11	108165653	108165786	134	391	0	0 0.0
Predicted drivers	475	133	91351 ATM	chr11	108168011	108168111	101	391	0	1 9.9
Predicted drivers	476	133	91524 ATM	chr11	108170440	108170612	173	391	0	1 5.8
Predicted drivers	477	133	91667 ATM	chr11	108172374	108172516	143	391	0	0 0.0
Predicted drivers	478	133	91845 ATM	chr11	108173579	108173756	178	391	0	0 0.0
Predicted drivers	479	133	92024 ATM	chr11	108175401	108175579	179	391	0	2 11.2
Predicted drivers	480	133	92125 ATM	chr11	108178617	108178717	101	391	0	0 0.0
Predicted drivers	481	133	92282 ATM	chr11	108180886	108181042	157	391	0	0 0.0
Predicted drivers	482	133	92383 ATM	chr11	108183131	108183231	101	391	0	1 9.9
Predicted drivers	483	133	92485 ATM	chr11	108186543	108186644	102	391	0	0 0.0
Predicted drivers	484	133	92589 ATM	chr11	108186737	108186840	104	391	0	1 9.6
Predicted drivers	485	133	92739 ATM	chr11	108188099	108188248	150	391	0	0 0.0
Predicted drivers	486	133	92845 ATM	chr11	108190680	108190785	106	391	0	0 0.0
Predicted drivers	487	133	92966 ATM	chr11	108192027	108192147	121	391	0	0 0.0
Predicted drivers	488	133	93202 ATM	chr11	108196036	108196271	236	391	0	1 4.2
Predicted drivers	489	133	93371 ATM	chr11	108196784	108196952	169	391	0	0 0.0
Predicted drivers	490	133	93486 ATM	chr11	108198371	108198485	115	391	0	0 0.0
Predicted drivers	491	133	93705 ATM	chr11	108199747	108199965	218	391	0	1 4.6
Predicted drivers	492	133	93914 ATM	chr11	108200940	108201148	209	391	0	0 0.0
Predicted drivers	493	133	94029 ATM	chr11	108202170	108202284	115	391	0	0 0.0
Predicted drivers	494	133	94189 ATM	chr11	108202605	108202764	160	391	0	0 0.0
Predicted drivers	495	133	94329 ATM	chr11	106203488	108203627	140	391	0	0 0.0
Predicted drivers	496	133	94431 ATM	chr11	108204603	108204704	102	391	0	1 9.8
Predicted drivers	497	133	94573 ATM	chr11	108205695	108205836	142	391	0	3 21.1
Predicted drivers	498	133	94691 ATM	chr11	108206571	108206688	118	391	0	1 8.5
Predicted drivers	499	133	94842 ATM	chr11	108213948	108214098	151	391	0	0 0.0
Predicted drivers	500	133	95009 ATM	chr11	108216469	108216635	167	391	0	0 0.0
Predicted drivers	501	133	95111 ATM	chr11	108217998	108218099	102	391	0	1 9.8
Predicted drivers	502	133	95227 ATM	chr11	108224492	108224607	116	391	0	1 8.6
Predicted drivers	503	133	95328 ATM	chr11	108225519	108225619	101	391	0	0 0.0
Predicted drivers	504	133	95466 ATM	chr11	108235808	108235945	138	391	0	1 7.2
Predicted drivers	505	133	95651 ATM	chr11	108236051	108236235	185	391	0	2 10.8
Predicted drivers	506	134	95753 FGFR4	chr5	176516598	176516699	102	391	0	0 0.0
Predicted drivers	507	134	960718 FGFR4	chr5	176517390	176517654	265	391	0	1 3.8
Predicted drivers	508	134	96120 FGFR4	chr5	176517735	176517836	102	391	0	1 9.8
Predicted drivers	509	134	96288 FGFR4	chr5	176517938	176518105	168	391	0	0 0.0
Predicted drivers	510	134	96413 FGFR4	chr5	176518685	176518809	125	391	0	0 0.0
Predicted drivers	511	134	96605 FGFR4	chr5	176519321	176519512	192	391	0	0 0.0
Predicted drivers	512	134	96745 FGFR4	chr5	176519646	176519785	140	391	0	0 0.0
Predicted drivers	513	134	97160 FGFR4	chr5	176520138	176520552	415	391	0	2 4.8
Predicted drivers	514	134	97283 FGFR4	chr5	176520654	176520776	123	391	0	0 0.0
Predicted drivers	515	134	97395 FGFR4	chr5	176522330	176522441	112	391	0	1 8.9
Predicted drivers	516	134	97587 FGFR4	chr5	176522533	176522724	192	391	0	0 0.0
Predicted drivers	517	134	97711 FGFR4	chr5	176523057	176523180	124	391	0	0 0.0
Predicted drivers	518	134	97813 FGFR4	chr5	176523272	176523373	102	391	0	0 0.0
Predicted drivers	519	134	97952 FGFR4	chr5	176523604	176523742	139	391	0	0 0.0

TABLE 1-continued

Recurrently mutated genomic regions in NSCLC.												
Predicted drivers	520	134	98059	FGFR4	chr5	176524292	176524398	107	391	0	0	0.0
Predicted drivers	521	134	98210	FGFR4	chr5	176524527	176524677	151	391	0	0	0.0
Add fusions	522	135	100435	ALK	chr2	29446207	29448431	2225	—	—	—	—
Add fusions	523	136	117908	ROS1	chr6	117641031	117658503	17473	—	—	—	—
Add fusions	524	137	123433	RET	chr10	43606655	43612179	5525	—	—	—	—
Add fusions	525	138	123876	POGFR4	chr4	55140698	55141140	443	—	—	—	—
Add fusions	526	139	125384	FGFR1	chr8	38275746	38277253	1508	—	—	—	—
Design phase	Coverage (unique LUAD & SCC patients; n = 407)				Coverage (all LUAD & SCC samples; n = 419)							
	No. pa- tients w/1 SNV	% pa- tients ≥1 SNV	% pa- tients ≥2 SNVs	% pa- tients ≥3 SNVs	Samples covered	Samples gained	No. samples per exon	RI	No. sam- ples w/1 SNV	% sam- ples ≥1 SNV	% sam- ples ≥2 SNVs	% sam- ples ≥3 SNVs
Known drivers	1	0.25	0.00	0.00	1	1	1	7.7	1	0.24	0.00	0.00
Known drivers	9	2.21	0.00	0.00	11	10	10	83.3	11	2.63	0.00	0.00
Known drivers	16	3.93	0.00	0.00	18	7	7	58.8	18	4.30	0.00	0.00
Known drivers	46	11.30	0.00	0.00	48	30	30	97.4	48	11.46	0.00	0.00
Known drivers	53	13.02	0.00	0.00	55	7	7	19.9	55	13.13	0.00	0.00
Known drivers	55	14.00	0.49	0.00	59	4	6	26.2	57	14.08	0.48	0.00
Known drivers	54	14.25	0.98	0.00	60	1	3	24.2	56	14.32	0.95	0.00
Known drivers	60	15.97	1.23	0.00	67	7	8	80.0	62	15.99	1.19	0.00
Known drivers	64	16.95	1.23	0.25	71	4	5	26.7	66	16.95	1.19	0.24
Known drivers	74	19.90	1.72	0.25	84	13	15	95.5	77	20.05	1.67	0.24
Known drivers	74	19.90	1.72	0.25	84	0	0	0.0	77	20.05	1.67	0.24
Known drivers	78	20.88	1.72	0.25	88	4	4	21.4	81	21.00	1.67	0.24
Known drivers	79	21.38	1.87	0.25	90	2	3	16.7	82	21.48	1.91	0.24
Known drivers	82	22.11	1.97	0.25	93	3	3	26.8	85	22.20	1.91	0.24
Known drivers	85	22.85	1.97	0.25	96	3	3	16.3	88	22.91	1.91	0.24
Known drivers	100	26.54	1.97	0.25	111	15	15	72.5	103	26.49	1.91	0.24
Known drivers	117	31.45	2.70	0.74	131	20	25	36.4	120	31.26	2.63	0.72
Known drivers	126	34.64	3.69	0.98	145	14	19	29.7	130	34.81	3.58	0.95
Known drivers	128	35.14	3.69	0.98	147	2	2	11.9	132	35.08	3.58	0.95
Known drivers	132	36.12	3.69	0.98	151	4	4	22.2	136	36.04	3.58	0.95
Known drivers	164	46.93	6.63	0.98	196	45	57	508.9	169	46.78	6.44	0.95
Known drivers	164	46.93	6.63	0.98	196	0	0	0.0	169	46.78	6.44	0.95
Known drivers	166	47.42	6.63	0.98	198	2	2	14.1	171	47.26	6.44	0.95
Known drivers	174	52.09	9.34	0.98	217	19	31	115.7	179	51.79	9.07	0.95
Known drivers	173	52.09	9.58	0.98	217	0	1	4.6	178	51.79	9.31	0.95
Known drivers	173	52.09	9.58	0.98	217	0	0	0.0	178	51.79	9.31	0.95
Known drivers	174	52.58	9.83	0.98	219	2	3	2.0	179	52.27	9.55	0.95
Known drivers	174	52.58	9.83	0.98	219	0	0	0.0	179	52.27	9.55	0.95
Known drivers	175	53.32	10.32	0.98	222	3	5	27.8	180	52.98	10.02	0.95
Known drivers	175	53.32	10.32	0.98	222	0	0	0.0	180	52.98	10.02	0.95
Known drivers	174	55.28	12.53	1.47	230	8	19	150.8	179	54.89	12.17	1.43
Known drivers	176	56.02	12.78	1.47	233	3	4	14.7	181	55.61	12.41	1.43
Known drivers	177	56.27	12.78	1.47	234	1	1	12.5	182	55.85	12.41	1.43
Known drivers	177	56.27	12.78	1.47	234	0	0	0.0	182	55.85	12.41	1.43
Known drivers	178	56.76	13.02	1.47	236	2	3	65.2	183	56.32	12.65	1.43
Known drivers	178	56.76	13.02	1.47	236	0	0	0.0	183	56.32	12.65	1.43
Known drivers	179	57.49	13.51	1.47	239	3	5	20.8	184	57.04	13.13	1.43
Known drivers	179	57.74	13.76	1.72	240	1	3	21.0	184	57.28	13.37	1.67
Known drivers	179	58.48	14.50	1.72	243	3	6	35.7	184	58.00	14.08	1.67
Known drivers	179	58.72	14.74	1.97	244	1	3	13.3	184	58.23	14.32	1.91
Known drivers	179	58.97	14.99	2.46	245	1	4	13.7	184	58.47	14.56	2.39
Known drivers	179	59.21	15.23	2.46	246	1	2	23.5	184	58.71	14.80	2.39
Known drivers	180	59.46	15.23	2.46	247	1	1	11.0	185	58.95	14.80	2.39
Known drivers	177	59.46	15.97	2.70	247	0	4	29.9	182	58.95	15.51	2.63
Known drivers	174	59.46	16.71	2.95	247	0	4	29.0	179	58.95	16.23	2.86
Known drivers	171	59.46	17.44	3.19	247	0	4	31.0	176	58.95	16.95	3.10
Known drivers	171	59.46	17.44	3.19	247	0	0	0.0	178	58.95	16.95	3.10
Known drivers	171	59.46	17.44	3.19	247	0	0	0.0	176	58.95	16.95	3.10
Known drivers	171	59.46	17.44	3.19	247	0	0	0.0	176	58.95	16.95	3.10
Known drivers	168	64.86	23.59	5.16	269	22	58	420.3	171	64.20	23.39	5.01
Known drivers	167	70.27	29.24	6.14	292	23	51	459.5	171	69.69	28.88	5.97
Known drivers	164	73.71	33.42	8.11	306	14	39	342.1	168	73.03	32.94	7.88
Known drivers	164	76.66	36.36	9.58	319	13	32	114.3	169	76.13	35.80	9.31
Known drivers	167	83.54	42.51	12.04	347	28	69	373.0	171	82.62	42.00	11.69
Max coverage	163	83.78	43.73	12.78	349	2	11	91.7	168	83.29	43.20	12.41
Max coverage	165	84.28	43.73	13.02	352	3	5	90.9	171	84.01	43.20	12.65
Max coverage	164	84.77	44.47	13.76	354	2	10	87.7	169	84.49	44.15	13.60
Max coverage	164	85.50	45.21	14.50	357	3	9	83.3	169	85.20	44.87	14.32
Max coverage	162	86.00	46.19	14.99	360	3	9	80.4	168	85.92	45.82	14.80
Max coverage	163	86.24	46.19	15.72	362	2	6	67.4	170	86.40	45.82	15.51

TABLE 1-continued

Recurrently mutated genomic regions in NSCLC.												
Max coverage	161	86.49	46.93	16.46	363	1	9	67.2	168	86.63	46.54	16.23
Max coverage	160	86.73	47.42	17.69	364	1	11	63.6	167	86.37	47.02	17.42
Max coverage	161	86.98	47.42	18.43	365	1	5	61.0	168	87.11	47.02	18.14
Max coverage	161	87.22	47.67	19.16	366	1	10	60.2	168	87.35	47.26	18.85
Max coverage	163	87.71	47.67	19.66	368	2	5	58.8	170	87.83	47.26	19.33
Max coverage	163	87.96	47.91	20.15	369	1	6	58.8	170	88.07	47.49	20.05
Max coverage	164	88.45	48.16	20.39	371	2	6	55.0	171	88.54	47.73	20.29
Max coverage	164	88.70	48.40	20.64	372	1	5	53.8	170	88.78	48.21	20.53
Max coverage	163	88.94	48.89	20.64	373	1	5	53.2	169	89.02	48.69	20.53
Max coverage	162	89.19	49.39	20.88	374	1	5	47.6	168	89.26	49.16	20.76
Max coverage	161	89.43	49.88	21.87	375	1	7	47.6	167	89.50	49.64	21.72
Max coverage	161	89.68	50.12	22.85	376	1	8	47.6	167	89.74	49.88	22.67
Max coverage	160	89.93	50.61	23.83	377	1	9	47.4	166	89.98	50.36	23.63
Max coverage	159	90.17	51.11	24.32	378	1	5	46.7	165	90.21	50.84	24.11
Max coverage	158	90.42	51.60	24.57	379	1	5	46.3	163	90.45	51.55	24.34
Max coverage	152	91.15	53.81	26.78	382	3	32	44.8	157	91.17	53.70	26.73
Max coverage	153	91.40	53.81	27.03	383	1	5	41.0	158	91.41	53.70	28.97
Max coverage	153	91.65	54.05	27.03	384	1	5	41.0	158	91.85	53.94	26.97
Max coverage	152	91.89	54.55	27.52	385	1	5	40.3	157	91.89	54.42	27.45
Max coverage	152	92.14	54.79	28.01	386	1	5	40.0	157	92.12	54.65	27.92
Max coverage	151	92.38	55.28	28.99	387	1	13	39.6	156	92.36	55.13	28.88
Max coverage	150	92.63	55.77	29.48	388	1	8	39.6	155	92.60	55.61	29.59
Max coverage	149	92.87	56.27	29.98	389	1	6	38.0	154	92.84	56.09	30.07
Max coverage	147	93.12	57.00	30.96	390	1	12	37.3	152	93.08	56.80	31.03
Max coverage	144	93.37	57.99	30.96	391	1	8	36.5	149	93.32	57.76	31.03
Max coverage	143	93.61	58.48	31.20	392	1	5	35.2	148	93.56	58.23	31.26
Max coverage	144	93.86	58.48	31.20	393	1	5	35.0	149	93.79	58.23	31.26
Max coverage	143	93.86	58.72	31.94	394	1	6	34.9	150	94.03	58.23	31.98
Max coverage	140	94.35	59.95	32.68	396	2	17	34.6	147	94.51	59.43	32.70
Max coverage	142	94.84	59.95	32.92	398	2	5	34.5	149	94.99	59.43	32.94
RI ≥ 30	134	94.84	61.92	35.63	398	0	30	32.0	141	94.99	61.34	35.56
RI ≥ 30	126	94.84	63.88	37.59	398	0	34	32.8	133	94.99	63.25	37.71
RI ≥ 30	121	94.84	65.11	38.33	398	0	28	30.2	127	94.99	64.68	38.42
RI ≥ 30	117	95.09	66.34	39.80	399	1	28	30.1	123	95.23	65.87	39.86
RI ≥ 30	113	95.09	67.32	42.01	399	0	33	31.5	119	95.23	66.83	42.00
RI ≥ 30	109	95.09	68.30	43.24	399	0	36	32.2	115	95.23	67.78	43.20
RI ≥ 30	105	95.09	69.29	43.24	399	0	9	35.2	111	95.23	68.74	43.20
RI ≥ 30	102	95.09	70.02	43.49	399	0	6	30.2	108	95.23	69.45	43.44
RI ≥ 30	99	95.09	70.76	43.73	399	0	12	45.3	105	95.23	70.17	43.68
RI ≥ 30	97	95.09	71.25	43.73	399	0	5	34.7	102	95.23	70.88	43.68
RI ≥ 30	94	95.09	71.99	44.23	399	0	5	30.7	99	95.23	71.80	44.15
RI ≥ 30	91	95.09	72.73	44.23	399	0	7	52.2	96	95.23	72.32	44.15
RI ≥ 30	88	95.09	73.46	44.23	399	0	6	32.8	93	95.23	73.03	44.15
RI ≥ 30	85	95.09	74.20	44.23	399	0	6	33.9	90	95.23	73.75	44.15
RI ≥ 30	82	95.09	74.94	45.21	399	0	29	32.5	87	95.23	74.46	45.11
RI ≥ 30	80	95.09	75.43	45.45	399	0	6	31.6	84	95.23	75.18	45.35
RI ≥ 30	77	95.09	76.17	45.70	399	0	8	36.4	81	95.23	75.89	45.58
RI ≥ 30	75	95.09	76.66	45.70	399	0	5	33.8	79	95.23	76.37	45.58
RI ≥ 30	73	95.09	77.15	45.95	399	0	3	30.3	77	95.23	76.85	45.82
RI ≥ 30	71	95.09	77.64	45.95	399	0	11	33.1	75	95.23	77.33	45.82
RI ≥ 30	70	95.33	78.13	45.95	400	1	3	32.3	74	95.47	77.80	45.82
RI ≥ 30	68	95.33	78.62	47.17	400	0	17	32.6	72	95.47	78.28	47.02
RI ≥ 30	67	95.58	79.12	47.17	401	1	4	30.8	71	95.70	78.76	47.02
RI ≥ 30	67	95.58	79.12	47.42	401	0	3	31.9	69	95.70	79.24	47.02
RI ≥ 30	65	95.58	79.61	47.42	401	0	6	35.3	67	95.70	79.71	47.02
RI ≥ 30	63	95.58	80.10	47.42	401	0	4	37.4	65	95.70	80.19	47.02
RI ≥ 30	61	95.58	80.59	47.42	401	0	4	48.2	63	95.70	80.67	47.02
RI ≥ 30	59	95.58	81.08	47.42	401	0	3	42.9	61	95.70	81.15	47.02
RI ≥ 30	57	95.58	81.57	47.42	401	0	5	33.8	59	95.70	81.62	47.02
RI ≥ 30	56	95.58	81.82	47.42	401	0	7	47.3	57	95.70	82.10	47.26
RI ≥ 30	54	95.58	82.31	47.42	401	0	6	33.0	55	95.70	82.58	47.26
RI ≥ 30	52	95.58	82.80	47.67	401	0	5	39.4	53	95.70	83.05	47.49
RI ≥ 30	51	95.58	83.05	47.67	401	0	4	54.8	52	95.70	83.29	47.49
RI ≥ 30	51	95.58	83.05	48.16	401	0	3	45.5	51	95.70	83.53	47.73
RI ≥ 30	50	95.58	83.29	48.65	401	0	7	40.5	50	95.70	83.77	48.21
RI ≥ 30	49	95.58	83.54	48.89	401	0	6	35.5	49	95.70	84.01	48.45
RI ≥ 30	48	95.58	83.78	48.89	401	0	3	46.2	46	95.70	84.25	48.45
RI ≥ 30	47	95.58	84.03	48.89	401	0	3	30.6	47	95.70	84.49	48.45
RI ≥ 30	46	95.58	84.28	48.89	401	0	4	30.3	46	95.70	84.73	48.45
RI ≥ 30	45	95.58	84.52	48.89	401	0	3	50.0	45	95.70	84.96	48.45
RI ≥ 30	44	95.58	84.77	49.14	401	0	4	34.5	44	95.70	85.20	48.69
RI ≥ 30	43	95.58	85.01	49.14	401	0	3	34.9	43	95.70	85.44	48.69
RI ≥ 30	42	95.58	85.26	49.63	401	0	6	31.9	42	95.70	85.68	49.16
RI ≥ 30	41	95.58	85.50	50.61	401	0	7	30.0	41	95.70	85.92	50.12
RI ≥ 30	40	95.58	85.75	50.86	401	0	3	30.0	40	95.70	86.16	50.36

TABLE 1-continued

Recurrently mutated genomic regions in NSCLC.												
RI ≥ 30	39	95.58	86.00	50.86	401	0	4	32.8	39	95.70	86.40	50.36
RI ≥ 30	38	95.58	86.24	51.11	401	0	5	39.4	38	95.70	86.63	50.60
RI ≥ 30	37	95.58	86.49	51.35	401	0	5	31.1	37	95.70	86.87	50.84
RI ≥ 30	36	95.58	86.73	51.60	401	0	26	125.6	36	95.70	87.11	51.07
RI ≥ 30	35	95.58	86.98	51.60	401	0	4	31.3	35	95.70	87.35	51.07
RI ≥ 30	34	95.58	87.22	51.84	401	0	4	31.5	34	95.70	87.59	51.31
RI ≥ 30	33	95.58	87.47	52.09	401	0	4	46.5	33	95.70	87.83	51.55
RI ≥ 30	32	95.58	87.71	52.09	401	0	6	32.4	32	95.70	88.07	51.55
RI ≥ 30	31	95.58	87.96	52.09	401	0	4	36.7	31	95.70	88.31	51.55
RI ≥ 30	30	95.58	88.21	52.33	401	0	6	42.3	30	95.70	88.54	51.79
RI ≥ 30	29	95.58	88.45	52.33	401	0	5	36.2	29	95.70	88.76	51.79
RI ≥ 30	28	95.58	88.70	52.58	401	0	6	30.3	28	95.70	89.02	52.03
RI ≥ 30	27	95.58	88.94	52.83	401	0	8	43.7	27	95.70	89.26	52.27
RI ≥ 30	26	95.58	89.19	52.83	401	0	7	34.3	26	95.70	89.50	52.27
RI ≥ 30	25	95.58	89.43	53.07	401	0	5	41.7	25	95.70	89.74	52.51
RI ≥ 30	24	95.58	89.68	53.07	401	0	3	33.3	24	95.70	89.96	52.51
RI ≥ 30	23	95.58	89.93	53.56	401	0	5	30.7	23	95.70	90.21	53.22
RI ≥ 30	22	95.58	90.17	53.56	401	0	7	34.0	22	95.70	90.45	53.22
RI ≥ 30	21	95.58	90.42	53.81	401	0	4	31.3	21	95.70	90.69	53.46
RI ≥ 30	20	95.58	90.66	53.81	401	0	5	30.7	20	95.70	90.93	53.46
RI ≥ 30	19	95.58	90.91	53.81	401	0	3	34.1	19	95.70	91.17	53.46
RI ≥ 30	18	95.58	91.15	54.05	401	0	3	47.6	18	95.70	91.41	53.70
RI ≥ 30	17	95.58	91.40	54.30	401	0	4	40.4	17	95.70	91.65	53.94
RI ≥ 30	16	95.58	91.65	54.55	401	0	4	32.3	16	95.70	91.89	54.18
RI ≥ 30	15	95.58	91.89	54.55	401	0	5	31.1	15	95.70	92.12	54.18
RI ≥ 30	14	95.58	92.14	54.55	401	0	6	36.6	14	95.70	92.36	54.18
RI ≥ 20	12	95.58	92.63	55.53	401	0	14	23.7	12	95.70	92.84	55.13
RI ≥ 20	11	95.58	92.87	55.53	401	0	3	27.5	11	95.70	93.08	55.13
RI ≥ 20	10	95.58	93.12	55.77	401	0	5	23.4	10	95.70	93.32	55.37
RI ≥ 20	9	95.58	93.37	56.27	401	0	4	23.5	9	95.70	93.56	55.85
RI ≥ 20	8	95.58	93.61	57.00	401	0	22	21.4	8	95.70	93.79	56.56
RI ≥ 20	7	95.58	93.86	57.49	401	0	8	21.1	7	95.70	94.03	57.04
RI ≥ 20	6	95.58	94.10	57.74	401	0	3	23.8	6	95.70	94.27	57.28
RI ≥ 20	5	95.58	94.35	57.99	401	0	6	24.4	5	95.70	94.51	57.52
RI ≥ 20	4	95.58	94.59	57.99	401	0	4	28.0	4	95.70	94.75	57.52
RI ≥ 20	3	95.58	94.84	58.23	401	0	6	25.6	3	95.70	94.99	57.76
RI ≥ 20	2	95.58	95.09	58.23	401	0	3	26.8	2	95.70	95.23	57.76
Predicted drivers	2	95.58	95.09	58.97	401	0	17	19.0	2	95.70	95.23	56.47
Predicted drivers	2	95.58	95.09	59.46	401	0	3	7.7	2	95.70	95.23	58.95
Predicted drivers	2	95.58	95.09	59.46	401	0	3	28.0	2	95.70	95.23	58.95
Predicted drivers	2	95.58	95.09	59.46	401	0	2	12.7	2	95.70	95.23	58.95
Predicted drivers	2	95.58	95.09	59.71	401	0	2	10.3	2	95.70	95.23	59.19
Predicted drivers	2	95.58	95.09	59.71	401	0	3	19.9	2	95.70	95.23	59.19
Predicted drivers	2	95.58	95.09	59.95	401	0	4	19.0	2	95.70	95.23	59.43
Predicted drivers	2	95.58	95.09	60.44	401	0	2	19.8	2	95.70	95.23	59.90
Predicted drivers	2	95.58	95.09	60.44	401	0	4	21.4	2	95.70	95.23	59.90
Predicted drivers	2	95.58	95.09	60.93	401	0	3	23.6	2	95.70	95.23	60.38
Predicted drivers	2	95.58	95.09	60.93	401	0	2	19.8	2	95.70	95.23	60.38
Predicted drivers	2	95.58	95.09	60.93	401	0	0	0.0	2	95.70	95.23	60.38
Predicted drivers	2	95.58	95.09	60.93	401	0	2	19.2	2	95.70	95.23	60.38
Predicted drivers	2	95.58	95.09	60.93	401	0	1	6.2	2	95.70	95.23	60.38
Predicted drivers	2	95.58	95.09	60.93	401	0	0	0.0	2	95.70	95.23	60.38
Predicted drivers	2	95.58	95.09	60.93	401	0	5	14.8	2	95.70	95.23	60.38
Predicted drivers	2	95.58	95.09	60.93	401	0	1	6.4	2	95.70	95.23	60.38
Predicted drivers	2	95.58	95.09	60.93	401	0	6	9.0	2	95.70	95.23	60.38
Predicted drivers	2	95.58	95.09	60.93	401	0	0	0.0	2	95.70	95.23	60.38
Predicted drivers	2	95.58	95.09	60.93	401	0	1	5.5	2	95.70	95.23	60.38
Predicted drivers	2	95.58	95.09	60.93	401	0	0	0.0	2	95.70	95.23	60.38
Predicted drivers	2	95.58	95.09	60.93	401	0	2	19.6	2	95.70	95.23	60.38
Predicted drivers	2	95.58	95.09	60.93	401	0	6	9.1	2	95.70	95.23	60.38
Predicted drivers	2	95.58	95.09	61.18	401	0	4	25.5	2	95.70	95.23	60.62
Predicted drivers	2	95.58	95.09	61.43	401	0	3	8.9	2	95.70	95.23	60.86
Predicted drivers	2	95.58	95.09	61.67	401	0	2	15.9	2	95.70	95.23	61.10
Predicted drivers	2	95.58	95.09	61.92	401	0	1	5.3	2	95.70	95.23	61.34
Predicted drivers	2	95.58	95.09	61.92	401	0	0	0.0	2	95.70	95.23	61.34
Predicted drivers	2	95.58	95.09	61.92	401	0	0	0.0	2	95.70	95.23	61.34
Predicted drivers	2	95.58	95.09	61.92	401	0	3	23.6	2	95.70	95.23	61.34
Predicted drivers	2	95.58	95.09	61.92	401	0	1	5.3	2	95.70	95.23	61.34
Predicted drivers	2	95.58	95.09	61.92	401	0	0	0.0	2	95.70	95.23	61.34
Predicted drivers	2	95.58	95.09	61.92	401	0	5	23.7	2	95.70	95.23	61.34
Predicted drivers	2	95.58	95.09	61.92	401	0	1	6.6	2	95.70	95.23	61.34
Predicted drivers	2	95.58	95.09	62.16	401	0	2	10.3	2	95.70	95.23	61.58
Predicted drivers	2	95.58	95.09	62.65	401	0	3	19.1	2	95.70	95.23	62.05
Predicted drivers	2	95.58	95.09	62.65	401	0	1	9.3	2	95.70	95.23	62.05
Predicted drivers	2	95.58	95.09	62.65	401	0	2	19.6	2	95.70	95.23	62.05

TABLE 1-continued

Recurrently mutated genomic regions in NSCLC.												
Predicted drivers	2	95.58	95.09	62.65	401	0	0	0.0	2	95.70	95.23	62.05
Predicted drivers	2	95.58	95.09	62.65	401	0	1	6.4	2	95.70	95.23	62.05
Predicted drivers	2	95.58	95.09	62.65	401	0	2	15.7	2	95.70	95.23	62.05
Predicted drivers	2	95.58	95.09	62.65	401	0	1	7.8	2	95.70	95.23	62.05
Predicted drivers	2	95.58	95.09	62.65	401	0	0	0.0	2	95.70	95.23	62.05
Predicted drivers	2	95.58	95.09	62.65	401	0	2	7.0	2	95.70	95.23	62.05
Predicted drivers	2	95.58	95.09	62.65	401	0	1	6.4	2	95.70	95.23	62.05
Predicted drivers	2	95.58	95.09	62.65	401	0	1	8.3	2	95.70	95.23	62.05
Predicted drivers	2	95.58	95.09	62.65	401	0	0	0.0	2	95.70	95.23	62.05
Predicted drivers	2	95.58	95.09	62.65	401	0	0	0.0	2	95.70	95.23	62.05
Predicted drivers	2	95.58	95.09	62.65	401	0	1	9.9	2	95.70	95.23	62.05
Predicted drivers	2	95.58	95.09	62.65	401	0	0	0.0	2	95.70	95.23	62.05
Predicted drivers	2	95.58	95.09	62.90	401	0	3	26.3	2	95.70	95.23	62.29
Predicted drivers	2	95.58	95.09	62.90	401	0	0	0.0	2	95.70	95.23	62.29
Predicted drivers	2	95.58	95.09	62.90	401	0	4	24.7	2	95.70	95.23	62.29
Predicted drivers	2	95.58	95.09	62.90	401	0	7	33.2	2	95.70	95.23	62.29
Predicted drivers	2	95.58	95.09	62.90	401	0	1	9.8	2	95.70	95.23	62.29
Predicted drivers	2	95.58	95.09	62.90	401	0	5	19.2	2	95.70	95.23	62.29
Predicted drivers	2	95.58	95.09	62.90	401	0	0	0.0	2	95.70	95.23	62.29
Predicted drivers	2	95.58	95.09	63.14	401	0	5	8.5	2	95.70	95.23	62.77
Predicted drivers	2	95.58	95.09	63.14	401	0	1	8.4	2	95.70	95.23	62.77
Predicted drivers	2	95.58	95.09	63.14	401	0	1	9.8	2	95.70	95.23	62.77
Predicted drivers	2	95.58	95.09	63.14	401	0	2	10.3	2	95.70	95.23	62.77
Predicted drivers	2	95.58	95.09	63.14	401	0	3	9.8	2	95.70	95.23	62.77
Predicted drivers	2	95.58	95.09	63.14	401	0	1	6.8	2	95.70	95.23	62.77
Predicted drivers	2	95.58	95.09	63.14	401	0	1	7.4	2	95.70	95.23	62.77
Predicted drivers	2	95.58	95.09	63.88	401	0	9	15.4	2	95.70	95.23	63.48
Predicted drivers	2	95.58	95.09	64.13	401	0	5	18.5	2	95.70	95.23	63.72
Predicted drivers	2	95.58	95.09	64.37	401	0	1	9.9	2	95.70	95.23	63.96
Predicted drivers	2	95.58	95.09	64.37	401	0	1	8.9	2	95.70	95.23	63.96
Predicted drivers	2	95.58	95.09	64.37	401	0	0	0.0	2	95.70	95.23	63.96
Predicted drivers	2	95.58	95.09	64.37	401	0	2	19.6	2	95.70	95.23	63.96
Predicted drivers	2	95.58	95.09	64.62	401	0	4	21.1	2	95.70	95.23	64.20
Predicted drivers	2	95.58	95.09	64.86	401	0	3	21.0	2	95.70	95.23	64.44
Predicted drivers	2	95.58	95.09	64.86	401	0	2	13.6	2	95.70	95.23	64.44
Predicted drivers	2	95.58	95.09	64.86	401	0	0	0.0	2	95.70	95.23	64.44
Predicted drivers	2	95.58	95.09	64.86	401	0	1	4.5	2	95.70	95.23	64.44
Predicted drivers	2	95.58	95.09	64.86	401	0	0	0.0	2	95.70	95.23	64.44
Predicted drivers	2	95.58	95.09	64.86	401	0	2	19.8	2	95.70	95.23	64.44
Predicted drivers	2	95.58	95.09	64.86	401	0	3	19.6	2	95.70	95.23	64.44
Predicted drivers	2	95.58	95.09	64.86	401	0	2	18.7	2	95.70	95.23	64.44
Predicted drivers	2	95.58	95.09	64.86	401	0	0	0.0	2	95.70	95.23	64.44
Predicted drivers	2	95.58	95.09	64.86	401	0	1	8.8	2	95.70	95.23	64.44
Predicted drivers	2	95.58	95.09	64.86	401	0	0	0.0	2	95.70	95.23	64.44
Predicted drivers	2	95.58	95.09	64.86	401	0	2	19.8	2	95.70	95.23	64.44
Predicted drivers	2	95.58	95.09	64.86	401	0	2	12.9	2	95.70	95.23	64.44
Predicted drivers	2	95.58	95.09	64.86	401	0	3	29.4	2	95.70	95.23	64.44
Predicted drivers	2	95.58	95.09	65.11	401	0	1	8.7	2	95.70	95.23	64.68
Predicted drivers	2	95.58	95.09	65.11	401	0	3	28.3	2	95.70	95.23	64.68
Predicted drivers	2	95.58	95.09	65.11	401	0	0	0.0	2	95.70	95.23	64.68
Predicted drivers	2	95.58	95.09	65.36	401	0	2	18.5	2	95.70	95.23	64.92
Predicted drivers	2	95.58	95.09	65.36	401	0	1	7.5	2	95.70	95.23	64.92
Predicted drivers	2	95.58	95.09	65.36	401	0	2	13.5	2	95.70	95.23	64.92
Predicted drivers	2	95.58	95.09	65.36	401	0	5	14.6	2	95.70	95.23	64.92
Predicted drivers	2	95.58	95.09	66.36	401	0	2	18.2	2	95.70	95.23	64.92
Predicted drivers	2	95.58	95.09	65.36	401	0	1	8.0	2	95.70	95.23	64.92
Predicted drivers	2	95.58	95.09	65.36	401	0	2	12.7	2	95.70	95.23	64.92
Predicted drivers	2	95.58	95.09	65.36	401	0	2	12.1	2	95.70	95.23	64.92
Predicted drivers	2	95.58	95.09	65.36	401	0	1	8.6	2	95.70	95.23	64.92
Predicted drivers	2	95.58	95.09	65.36	401	0	2	11.2	2	95.70	95.23	64.92
Predicted drivers	2	95.58	95.09	65.36	401	0	0	0.0	2	95.70	95.23	64.92
Predicted drivers	2	95.58	95.09	65.36	401	0	1	5.9	2	95.70	95.23	64.92
Predicted drivers	2	95.58	95.09	65.36	401	0	4	30.3	2	95.70	95.23	64.92
Predicted drivers	2	95.58	95.09	65.36	401	0	0	0.0	2	95.70	95.23	64.92
Predicted drivers	2	95.58	95.09	65.36	401	0	1	9.9	2	95.70	95.23	64.92
Predicted drivers	2	95.58	95.09	65.36	401	0	0	0.0	2	95.70	95.23	64.92
Predicted drivers	2	95.58	95.09	65.36	401	0	0	0.0	2	95.70	95.23	64.92
Predicted drivers	2	95.58	95.09	65.60	401	0	1	6.3	2	95.70	95.23	65.16
Predicted drivers	2	95.58	95.09	65.60	401	0	0	0.0	2	95.70	95.23	65.16
Predicted drivers	2	95.58	95.09	65.60	401	0	2	8.2	2	95.70	95.23	65.16
Predicted drivers	2	95.58	95.09	65.60	401	0	4	23.0	2	95.70	95.23	65.16
Predicted drivers	2	95.58	95.09	65.60	401	0	1	7.6	2	95.70	95.23	65.16
Predicted drivers	2	95.58	95.09	65.60	401	0	0	0.0	2	95.70	95.23	65.16
Predicted drivers	2	95.58	95.09	65.60	401	0	0	0.0	2	95.70	95.23	65.16
Predicted drivers	2	95.58	95.09	65.60	401	0	2	10.5	2	95.70	95.23	65.16

TABLE 1-continued

Recurrently mutated genomic regions in NSCLC.												
Predicted drivers	2	95.58	95.09	66.09	401	0	3	28.8	2	95.70	95.23	65.63
Predicted drivers	2	95.58	95.09	66.09	401	0	0	0.0	2	95.70	95.23	65.63
Predicted drivers	2	95.58	95.09	66.09	401	0	0	0.0	2	95.70	95.23	65.63
Predicted drivers	2	95.58	95.09	66.09	401	0	8	26.8	2	95.70	95.23	65.63
Predicted drivers	2	95.58	95.09	66.34	401	0	1	7.0	2	95.70	95.23	65.87
Predicted drivers	2	95.58	95.09	66.34	401	0	2	13.9	2	95.70	95.23	65.87
Predicted drivers	2	95.58	95.09	66.34	401	0	0	0.0	2	95.70	95.23	65.87
Predicted drivers	2	95.58	95.09	66.34	401	0	0	0.0	2	95.70	95.23	65.87
Predicted drivers	2	95.58	95.09	66.58	401	0	1	9.9	2	95.70	95.23	66.11
Predicted drivers	2	95.58	95.09	66.83	401	0	1	9.8	2	95.70	95.23	66.35
Predicted drivers	2	95.58	95.09	66.83	401	0	0	0.0	2	95.70	95.23	66.35
Predicted drivers	2	95.58	95.09	67.57	401	0	3	23.4	2	95.70	95.23	67.06
Predicted drivers	2	95.58	95.09	67.57	401	0	1	8.5	2	95.70	95.23	67.06
Predicted drivers	2	95.58	95.09	67.57	401	0	1	8.3	2	95.70	95.23	67.06
Predicted drivers	2	95.58	95.09	67.57	401	0	0	0.0	2	95.70	95.23	67.06
Predicted drivers	2	95.58	95.09	67.57	401	0	0	0.0	2	95.70	95.23	67.06
Predicted drivers	2	95.58	95.09	67.57	401	0	2	17.9	2	95.70	95.23	67.06
Predicted drivers	2	95.58	95.09	67.57	401	0	0	0.0	2	95.70	95.23	67.06
Predicted drivers	2	95.58	95.09	67.57	401	0	0	0.0	2	95.70	95.23	67.06
Predicted drivers	2	95.58	95.09	68.06	401	0	3	27.0	2	95.70	95.23	67.54
Predicted drivers	2	95.58	95.09	68.06	401	0	0	0.0	2	95.70	95.23	67.54
Predicted drivers	2	95.58	95.09	68.06	401	0	2	19.8	2	95.70	95.23	67.54
Predicted drivers	2	95.58	95.09	68.06	401	0	0	0.0	2	95.70	95.23	67.54
Predicted drivers	2	95.58	95.09	68.06	401	0	2	19.6	2	95.70	95.23	67.54
Predicted drivers	2	95.58	95.09	68.06	401	0	0	0.0	2	95.70	95.23	67.54
Predicted drivers	2	95.58	95.09	68.06	401	0	1	9.8	2	95.70	95.23	67.54
Predicted drivers	2	95.58	95.09	68.06	401	0	0	0.0	2	95.70	95.23	67.54
Predicted drivers	2	95.58	95.09	68.06	401	0	0	0.0	2	95.70	95.23	67.54
Predicted drivers	2	95.58	95.09	68.30	401	0	3	20.4	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	1	6.8	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	0	0.0	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	1	8.7	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	2	12.1	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	0	0.0	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	0	0.0	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	0	0.0	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	3	6.7	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	3	10.0	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	0	0.0	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	2	11.6	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	1	9.9	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	1	6.8	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	0	0.0	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	2	12.7	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	1	5.3	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	0	0.0	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	0	0.0	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	0	0.0	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	2	19.6	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	1	6.4	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	1	9.9	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	1	7.5	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	2	10.0	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	0	0.0	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	1	9.9	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	1	7.8	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	2	17.4	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	0	0.0	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	2	16.7	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.30	401	0	0	0.0	2	95.70	95.23	67.78
Predicted drivers	2	95.58	95.09	68.55	401	0	1	7.4	2	95.70	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	1	4	21.3	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	2	13.1	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	0	0.0	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	0	0.0	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	0	0.0	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	0	0.0	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	1	9.8	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	0	0.0	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	0	0.0	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	1	6.8	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	0	0.0	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	0	0.0	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	0	0.0	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	1	3.0	3	95.94	95.23	68.02

TABLE 1-continued

Recurrently mutated genomic regions in NSCLC.												
Predicted drivers	3	95.82	95.09	68.55	402	0	0	0.0	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	0	0.0	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	0	0.0	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	2	13.5	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	3	22.7	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	2	11.5	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	4	16.5	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	0	0.0	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	0	0.0	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	0	0.0	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.55	402	0	0	0.0	3	95.94	95.23	68.02
Predicted drivers	3	95.82	95.09	68.80	402	0	1	9.9	3	95.94	95.23	68.26
Predicted drivers	3	95.82	95.09	68.80	402	0	1	5.2	3	95.94	95.23	68.26
Predicted drivers	3	95.82	95.09	68.80	402	0	1	9.3	3	95.94	95.23	68.26
Predicted drivers	3	95.82	95.09	68.80	402	0	0	0.0	3	95.94	95.23	68.26
Predicted drivers	3	95.82	95.09	68.80	402	0	0	0.0	3	95.94	95.23	68.26
Predicted drivers	3	95.82	95.09	69.04	402	0	1	6.3	3	95.94	95.23	68.50
Predicted drivers	3	95.82	95.09	69.29	402	0	3	17.1	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	0	0.0	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	0	0.0	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	0	0.0	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	1	7.4	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	1	8.7	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	1	7.0	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	0	0.0	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	0	0.0	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	1	4.0	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	0	0.0	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	1	9.9	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	2	4.5	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	1	7.1	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	1	8.1	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	0	0.0	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	0	0.0	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	2	10.9	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	1	4.7	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	2	12.3	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	0	0.0	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	0	0.0	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	0	0.0	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	0	0.0	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.29	402	0	0	0.0	3	95.94	95.23	68.74
Predicted drivers	3	95.82	95.09	69.53	402	0	1	9.9	3	95.94	95.23	68.97
Predicted drivers	3	95.82	95.09	69.78	402	0	2	13.5	3	95.94	95.23	69.21
Predicted drivers	3	95.82	95.09	69.78	402	0	0	0.0	3	95.94	95.23	69.21
Predicted drivers	3	95.82	95.09	69.78	402	0	0	0.0	3	95.94	95.23	69.21
Predicted drivers	3	95.82	95.09	69.78	402	0	2	4.6	3	95.94	95.23	69.21
Predicted drivers	3	95.82	95.09	69.78	402	0	3	8.8	3	95.94	95.23	69.21
Predicted drivers	3	95.82	95.09	69.78	402	0	3	14.7	3	95.94	95.23	69.21
Predicted drivers	3	95.82	95.09	70.02	402	0	3	15.4	3	95.94	95.23	69.45
Predicted drivers	3	95.82	95.09	70.27	402	0	2	14.1	3	95.94	95.23	69.69
Predicted drivers	3	95.82	95.09	70.27	402	0	0	0.0	3	95.94	95.23	69.69
Predicted drivers	3	95.82	95.09	70.52	402	0	1	4.6	3	95.94	95.23	69.93
Predicted drivers	3	95.82	95.09	70.52	402	0	1	9.9	3	95.94	95.23	69.93
Predicted drivers	3	95.82	95.09	70.52	402	0	0	0.0	3	95.94	95.23	69.93
Predicted drivers	3	95.82	95.09	70.76	402	0	2	19.4	3	95.94	95.23	70.17
Predicted drivers	3	95.82	95.09	71.01	402	0	1	7.0	3	95.94	95.23	70.41
Predicted drivers	3	95.82	95.09	71.01	402	0	2	15.6	3	95.94	95.23	70.41
Predicted drivers	3	95.82	95.09	71.01	402	0	2	15.0	3	95.94	95.23	70.41
Predicted drivers	3	95.82	95.09	71.01	402	0	0	0.0	3	95.94	95.23	70.41
Predicted drivers	3	95.82	95.09	71.25	402	0	2	12.6	3	95.94	95.23	70.64
Predicted drivers	3	95.82	95.09	71.25	402	0	0	0.0	3	95.94	95.23	70.64
Predicted drivers	3	95.82	95.09	71.25	402	0	1	7.6	3	95.94	95.23	70.64
Predicted drivers	3	95.82	95.09	71.50	402	0	1	9.8	3	95.94	95.23	70.88
Predicted drivers	3	95.82	95.09	71.50	402	0	1	6.9	3	95.94	95.23	70.88
Predicted drivers	3	95.82	95.09	71.50	402	0	0	0.0	3	95.94	95.23	70.88
Predicted drivers	3	95.82	95.09	71.50	402	0	0	0.0	3	95.94	95.23	70.88
Predicted drivers	3	95.82	95.09	71.50	402	0	0	0.0	3	95.94	95.23	70.88
Predicted drivers	3	95.82	95.09	71.50	402	0	1	6.9	3	95.94	95.23	70.88
Predicted drivers	3	95.82	95.09	71.50	402	0	0	0.0	3	95.94	95.23	70.88
Predicted drivers	3	95.82	95.09	71.50	402	0	0	0.0	3	95.94	95.23	70.88
Predicted drivers	3	95.82	95.09	71.50	402	0	1	9.8	3	95.94	95.23	70.88
Predicted drivers	3	95.82	95.09	71.74	402	0	2	9.9	3	95.94	95.23	71.12
Predicted drivers	3	95.82	95.09	71.74	402	0	1	9.1	3	95.94	95.23	71.12
Predicted drivers	3	95.82	95.09	71.74	402	0	0	0.0	3	95.94	95.23	71.12
Predicted drivers	3	95.82	95.09	71.74	402	0	0	0.0	3	95.94	95.23	71.12
Predicted drivers	3	95.82	95.09	71.74	402	0	0	0.0	3	95.94	95.23	71.12

TABLE 1-continued

Recurrently mutated genomic regions in NSCLC.												
Predicted drivers	3	95.82	95.09	71.74	402	0	0	0.0	3	95.94	95.23	71.12
Predicted drivers	3	95.82	95.09	71.74	402	0	1	2.6	3	95.94	95.23	71.12
Predicted drivers	3	95.82	95.09	71.74	402	0	0	0.0	3	95.94	95.23	71.12
Predicted drivers	3	95.82	95.09	71.74	402	0	0	0.0	3	95.94	95.23	71.12
Predicted drivers	3	95.82	95.09	71.74	402	0	0	0.0	3	95.94	95.23	71.12
Predicted drivers	3	95.82	95.09	71.74	402	0	2	16.9	3	95.94	95.23	71.12
Predicted drivers	3	95.82	95.09	71.74	402	0	0	0.0	3	95.94	95.23	71.12
Predicted drivers	4	96.07	95.09	72.97	403	1	23	3.5	4	96.18	95.23	72.32
Predicted drivers	4	96.07	95.09	72.97	403	0	0	0.0	4	96.18	95.23	72.32
Predicted drivers	4	96.07	95.09	72.97	403	0	1	8.8	4	96.18	95.23	72.32
Predicted drivers	4	96.07	95.09	72.97	403	0	0	0.0	4	96.18	95.23	72.32
Predicted drivers	4	96.07	95.09	72.97	403	0	0	0.0	4	96.18	95.23	72.32
Predicted drivers	4	96.07	95.09	72.97	403	0	0	0.0	4	96.18	95.23	72.32
Predicted drivers	4	96.07	95.09	72.97	403	0	1	4.2	4	96.18	95.23	72.32
Predicted drivers	4	96.07	95.09	72.97	403	0	0	0.0	4	96.18	95.23	72.32
Predicted drivers	4	96.07	95.09	72.97	403	0	1	5.8	4	96.18	95.23	72.32
Predicted drivers	4	96.07	95.09	72.97	403	0	0	0.0	4	96.18	95.23	72.32
Predicted drivers	4	96.07	95.09	72.97	403	0	0	0.0	4	96.18	95.23	72.32
Predicted drivers	4	96.07	95.09	72.97	403	0	0	0.0	4	96.18	95.23	72.32
Predicted drivers	4	96.07	95.09	73.22	403	0	1	9.9	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	1	7.9	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	1	5.8	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	1	7.6	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	2	11.7	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	1	4.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	1	7.8	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	1	5.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	1	9.9	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	1	5.8	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	2	11.2	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	1	9.9	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	1	9.6	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	1	4.2	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	1	4.6	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	0	0.0	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.22	403	0	1	9.8	4	96.18	95.23	72.55
Predicted drivers	4	96.07	95.09	73.46	403	0	3	21.1	4	96.18	95.23	72.79
Predicted drivers	4	96.07	95.09	73.46	403	0	1	8.5	4	96.18	95.23	72.79
Predicted drivers	4	96.07	95.09	73.46	403	0	0	0.0	4	96.18	95.23	72.79
Predicted drivers	4	96.07	95.09	73.46	403	0	0	0.0	4	96.18	95.23	72.79
Predicted drivers	4	96.07	95.09	73.46	403	0	1	9.8	4	96.18	95.23	72.79
Predicted drivers	4	96.07	95.09	73.71	403	0	1	8.6	4	96.18	95.23	73.03
Predicted drivers	4	96.07	95.09	73.71	403	0	0	0.0	4	96.18	95.23	73.03
Predicted drivers	4	96.07	95.09	73.96	403	0	1	7.2	4	96.18	95.23	73.27
Predicted drivers	4	96.07	95.09	74.45	403	0	2	10.8	4	96.18	95.23	73.75
Predicted drivers	4	96.07	95.09	74.45	403	0	0	0.0	4	96.18	95.23	73.75
Predicted drivers	4	96.07	95.09	74.45	403	0	1	3.8	4	96.18	95.23	73.75
Predicted drivers	4	96.07	95.09	74.45	403	0	1	9.8	4	96.18	95.23	73.75
Predicted drivers	4	96.07	95.09	74.45	403	0	0	0.0	4	96.18	95.23	73.75
Predicted drivers	4	96.07	95.09	74.45	403	0	0	0.0	4	96.18	95.23	73.75

TABLE 1-continued

Recurrently mutated genomic regions in NSCLC.												
Predicted drivers	4	96.07	95.09	74.45	403	0	0	0.0	4	96.18	95.23	73.75
Predicted drivers	4	96.07	95.09	74.45	403	0	0	0.0	4	96.18	95.23	73.75
Predicted drivers	4	96.07	95.09	74.45	403	0	2	4.6	4	96.18	95.23	73.75
Predicted drivers	4	96.07	95.09	74.45	403	0	0	0.0	4	96.18	95.23	73.75
Predicted drivers	4	96.07	95.09	74.45	403	0	1	8.9	4	96.18	95.23	73.75
Predicted drivers	4	96.07	95.09	74.45	403	0	0	0.0	4	96.18	95.23	73.75
Predicted drivers	4	96.07	95.09	74.45	403	0	0	0.0	4	96.18	95.23	73.75
Predicted drivers	4	96.07	95.09	74.45	403	0	0	0.0	4	96.18	95.23	73.75
Predicted drivers	4	96.07	95.09	74.45	403	0	0	0.0	4	96.18	95.23	73.75
Predicted drivers	4	96.07	95.09	74.45	403	0	0	0.0	4	96.18	95.23	73.75
Predicted drivers	4	96.07	95.09	74.45	403	0	0	0.0	4	96.18	95.23	73.75
Predicted drivers	4	96.07	95.09	74.45	403	0	0	0.0	4	96.18	95.23	73.75
Predicted drivers	4	96.07	95.09	74.45	403	0	0	0.0	4	96.18	95.23	73.75
Add fusions	—	—	—	—	—	—	—	—	—	—	—	—
Add fusions	—	—	—	—	—	—	—	—	—	—	—	—
Add fusions	—	—	—	—	—	—	—	—	—	—	—	—
Add fusions	—	—	—	—	—	—	—	—	—	—	—	—
Add fusions	—	—	—	—	—	—	—	—	—	—	—	—

[0125] FIG. 3 illustrates how the statistical enrichment of recurrently mutated NSCLC exons captures known drivers. Two metrics were employed to prioritize exons with recurrent mutations for inclusion in the CAPP-Seq NSCLC selector. The first, termed Recurrence Index (RI), is defined as the number of unique patients (i.e. tumors) with somatic mutations per kilobase of a given exon and the second metric is based on the minimum number of unique patients (i.e. tumors) with mutations in a given kb of exon. Exons containing at least one non-silent SNV genotyped by TCGA (n=47,769) in a combined cohort of 407 lung adenocarcinoma (LUAD) and squamous cell carcinoma (SCC) patients were analyzed. As shown in FIG. 3(a), known/suspected NSCLC drivers are highly enriched at $RI \geq 30$ (inset), comprising 1.8% (n=861) of analyzed exons. As shown in FIG. 3(b), known/suspected NSCLC drivers are highly enriched at ≥ 3 patients with mutations per exon (inset), encompassing 16% of analyzed exons.

[0126] Approximately 8% of NSCLCs contain clinically actionable rearrangements involving the receptor tyrosine kinases, ALK, ROS1 and RET (Bergethon et al. (2012) *J. Clin. Oncol.* 30:863-870; Kwak et al. (2010) *N. Engl. J. Med.* 363:1693-1703; Pao & Hutchinson (2012) *Nat. Med.* 18:349-351). To utilize the personalized nature and low false detection rate of structural rearrangements (Leary et al. (2010) *Sci. Transl. Med.* 2:20ra14; McBride et al. (2010) *Genes Chrom. Cancer* 49:1062-1069), introns and exons spanning recurrent fusion breakpoints in these genes were included in the final design phase (FIG. 1b). To detect fusions in tumor and plasma DNA, a breakpoint-mapping algorithm called FACTERA was developed (FIG. 4). Application of FACTERA to next generation sequencing (NGS) data from 2 NSCLC cell lines known to harbor fusions with previously uncharacterized breakpoints (Koivunen et al. (2008) *Clin. Cancer Res.* 14:4275-4283; Rikova et al. (2007) *Cell* 131:1190-1203) readily identified the breakpoints in both cases (FIG. 5).

[0127] Collectively, the NSCLC CAPP-Seq selector design targets 521 exons and 13 introns from 139 recurrently mutated genes, in total covering ~125 kb (FIG. 1b). Within this small target (0.004% of the human genome), CAPP-Seq identifies a median of 4 point mutations and covered 96% of

patients with lung adenocarcinoma or squamous cell carcinoma. To validate the number of mutations covered per tumor, we examined the selector region in WES data from an independent cohort of 183 lung adenocarcinoma patients (Imielinski et al. (2012) *Cell* 150:1107-1120). The selector covered 88% of patients with a median of 4 SNVs per patient, thus validating our selector design algorithm ($P < 1.0 \times 10^{-6}$; FIG. 1c). When compared to randomly sampling the exome, regions targeted by CAPP-Seq captured ~4-fold as many mutations per patient (at the median, FIG. 1c). Due to similarities in key oncogenic machinery across cancers (Hanahan & Weinberg (2011) *Cell* 144:646-674), we hypothesized that our NSCLC selector would perform favorably on other carcinomas. Indeed, when applied to TCGA WES data, the selector successfully captured 99% of colon, 98% of rectal, and 97% of endometrioid uterine carcinomas, with a median of 12, 7, and 3 mutations per patient, respectively (FIG. 1d). This demonstrates the value of targeting hundreds of recurrently mutated genomic regions and suggests that a CAPP-Seq selector could be designed to simultaneously cover mutations for a wide variety of human malignancies.

[0128] Using this CAPP-Seq selector, we profiled a total of 52 samples including NSCLC cell lines, primary tumor specimens, peripheral blood leukocytes (PBLs), and cfDNA isolated from plasma of patients with NSCLC before and after various cancer therapies (Table 2). To assess and optimize the performance of CAPP-Seq, we first applied it to cfDNA purified from healthy control plasma. Approximately 60% of reads mapped within the selector target region (Table 2). Sequenced cfDNA fragments had a median length of 169 bp (FIG. 1e), closely corresponding to the length of DNA contained within a chromosome (Fan et al. (2008) *Proc. Natl Acad. Sci. USA* 105:16266-16271). To optimize library preparation from small quantities of cfDNA we explored a variety of modifications to the ligation and post-ligation amplification steps including temperature, incubation time, enzyme source, and “with-bead” clean-up. The optimized protocol increased recovery efficiency by >300% and decreased bias for libraries constructed from as little as 4 ng of cfDNA (FIGS. 6-8). Consequently, fluctuations in sequencing depth were minimal (FIG. 1f-g) and unlikely to impact performance.

TABLE 2

Profile of samples using NSCLC CAPP-Seq selector							
Sample	DNA mass used for library (ng)	Library mass used for capture (ng)	Total reads mapped	Fraction of properly paired reads	Read on-target rate	Median depth	Median fragment length
H3122 0.1% into HCC78	128	111	99.0%	96.8%	69.5%	8688	173
H3122 1% into HCC78	128	111	98.9%	96.7%	69.8%	8657	171
H3122 10% into HCC78	128	111	98.9%	96.5%	69.8%	6890	170
H3122 100%	128	111	99.0%	96.8%	68.6%	6739	174
HCC78 100%	128	111	99.0%	96.9%	69.7%	7602	172
cfDNA 100% 6 cycles	32	83.3	97.5%	86.7%	60.3%	8280	168
HCC78 10% into cfDNA 4 cycles	128	83.3	97.5%	83.3%	59.3%	2682	170
HCC78 10% into cfDNA 8 cycles SigmaWGA	624	83.3	79.5%	72.0%	50.4%	15	158
HCC78 10% into cfDNA 6 cycles	32	83.3	97.7%	87.2%	60.4%	8261	169
HCC78 10% into cfDNA 8 cycles NEBNextOvernightBead	32	83.3	96.9%	91.8%	61.1%	6258	166
HCC78 10% into cfDNA 8 cycles OrigNEBNext 15 minLig	32	83.3	98.0%	93.1%	60.9%	9862	167
HCC78 10% into cfDNA 4 ng 9 cycles	4	83.3	97.6%	87.6%	60.5%	11630	169
P11 PBL	500	83.3	96.7%	93.8%	59.0%	6970	169
P11 Tumor	500	83.3	93.4%	88.3%	61.3%	7700	156
P6 PBL	500	83.3	96.7%	92.6%	67.2%	3848	152
P6 Tumor	1000	83.3	87.0%	81.8%	64.7%	2445	158
P8 PBL	500	83.3	96.9%	93.0%	65.8%	4021	154
P8 Tumor	500	83.3	91.7%	85.4%	63.6%	5331	151
P10 PBL	400	83.3	96.9%	93.6%	65.3%	4572	161
P10 Tumor	500	83.3	94.0%	89.6%	65.1%	5335	157
P7 PBL	500	83.3	97.1%	93.5%	67.1%	3552	155
P7 Tumor	500	83.3	94.1%	89.3%	64.0%	4793	162
HCC78 0.025% into cfDNA	32	83.3	98.2%	87.0%	46.3%	3913	169
HCC78 0.05% into cfDNA	32	83.3	98.1%	86.1%	44.7%	6549	169
HCC78 0.1% into cfDNA	32	83.3	98.4%	88.1%	44.9%	6897	169
HCC78 0.5% into cfDNA	32	83.3	98.8%	89.8%	46.2%	8096	169
HCC78 1% into cfDNA	32	83.3	98.5%	89.8%	46.5%	7779	171
P6-1 cfDNA	17	83.3	98.6%	91.3%	46.4%	11172	166
P6-2 cfDNA	20	83.3	98.5%	92.0%	46.6%	8455	166
P9 PBL	500	83.3	97.0%	94.4%	59.2%	5441	172
P9 Tumor	69	83.3	99.2%	97.3%	55.3%	7312	239
P3 PBL	500	83.3	99.3%	97.8%	57.0%	8838	235
P3 Tumor	500	83.3	99.3%	98.0%	66.0%	9562	204
P2 PBL	500	83.3	99.2%	97.5%	57.7%	7680	235
P2 Tumor	500	83.3	99.0%	97.1%	62.3%	7247	204
P4 PBL	500	83.3	99.1%	96.5%	56.5%	7331	227
P4 Tumor	200	83.3	97.5%	94.1%	60.0%	3968	189
P1 PBL	500	83.3	99.3%	97.1%	57.1%	7336	220
P1 Tumor	500	83.3	94.6%	90.1%	60.9%	976	192
P5 PBL	500	83.3	99.2%	97.2%	58.7%	8155	219
P5 Tumor	100	83.3	98.8%	97.0%	63.5%	6930	187
P9-1 cfDNA	12	83.3	99.1%	84.2%	65.6%	6839	172
P9-2 cfDNA	17	83.3	98.4%	83.9%	65.2%	6043	169
P9-3 cfDNA	16	83.3	99.4%	88.7%	67.6%	8141	167
P3-1 cfDNA	15	83.3	99.2%	86.0%	63.5%	7057	170
P3-2 cfDNA	16	83.3	99.3%	86.5%	63.5%	10089	171
P2-1 cfDNA	13	83.3	99.4%	86.9%	67.3%	6876	172
P2-2 cfDNA	16	83.3	99.5%	96.4%	63.6%	5248	185
P1-1 cfDNA	13	83.3	99.0%	85.0%	64.6%	5079	171
P1-2 cfDNA	7	83.3	99.4%	84.7%	64.1%	6487	172
P5-1 cfDNA	9	83.3	99.3%	87.8%	66.6%	7604	169
P5-2 cfDNA	15	83.3	99.4%	88.0%	67.5%	10451	170

[0129] FIG. 6 illustrates the improvements in CAPP-Seq performance achieved with optimized library preparation procedures. Using 32 ng of input cfDNA from plasma, standard versus “with bead” (Fisher et al. (2011) *Genome biology* 12:R1) library preparation methods were compared, as well as two commercially available DNA polymerases (Phusion and KAPA HiFi). Template pre-amplification by Whole Genome Amplification (WGA) using Degenerate Oligonucleotide PCR (DOP) were also compared. Indices considered for these comparisons included (a) length of the captured cfDNA fragments sequenced, (b) depth and uniformity of sequencing coverage across all genomic regions in the selector, and (c) sequence mapping and capture statistics, includ-

ing uniqueness. Collectively, these comparisons identified KAPA HiFi polymerase and a “with bead” protocol as having most robust and uniform performance.

[0130] FIG. 7 illustrates the optimization of allele recovery from low input cfDNA during Illumina library preparation. Bars reflect the relative yield of CAPP-Seq libraries constructed from 4 ng cfDNA, calculated by averaging quantitative PCR measurements of 4 pre-selected reporters within CAPP-Seq with pre-defined amplification efficiencies. (a) Sixteen hour ligation at 16° C. increases ligation efficiency and reporter recovery. (b) Adapter ligation volume did not have a significant effect on ligation efficiency and reporter recovery. (c) Performing enzymatic reactions “with-bead” to

minimize tube transfer steps increases reporter recovery. (d) Increasing adapter concentration during ligation increases ligation efficiency and reporter recovery. Reporter recovery is also higher when using KAPA HiFi DNA polymerase compared to Phusion DNA polymerase (e) and when using the KAPA Library Preparation Kit with the modifications in a-d compared to the NuGEN SP Ovation Ultralow Library System with automation on a Mondrian SP Workstation (f). Relative reporter abundance was determined by qPCR using the $2^{-\Delta C_T}$ method. All values are mean $\pm$ s.d. N.S., not significant. Based on these results, it was estimated that combining the methodological modifications in a and c-e improves yield in NGS libraries by 3.3-fold.

[0131] FIG. 7 illustrates the performance of CAPP-Seq with various amounts of input cfDNA. (a) Length of the captured cfDNA fragments sequenced. (b) Depth of sequencing coverage across all genomic regions in the selector. (c) Sequence mapping and capture statistics. As expected, more input cfDNA mass correlates with more unique fragments sequenced.

[0132] The detection limit of CAPP-Seq is affected by the absolute number of available cfDNA molecules in a given volume of peripheral blood, as well as PCR and sequencing errors (i.e. “technical” background). The latter primarily affects substitutions/SNVs as opposed to other CAPP-Seq reporters (i.e., indels (Minoche et al. (2011) *Genome Biol.* 12:R112) and rearrangements). Separately, mutant cfDNA could be present in the absence of cancer due to contributions from pre-neoplastic cells from diverse tissues (i.e., “biological” background). The combined background from these sources was measured by assessing the error rate at each nucleotide position across the selector in plasma cfDNA from 6 patients and a healthy individual, excluding tumor-derived mutations. Mean and median background rates of $\sim 0.007\%$ and $\sim 0\%$ (not detected, N.D.) were found, respectively (FIG. 9 (a)). Next, we hypothesized that if significant biological background is present, it should be highest for recurrently mutated positions in cancer driver genes. We therefore analyzed mutation rates of 107 recurrent cancer-associated SNVs (Su et al. (2011) *J. Mol. Diagn.* 13:74-84) in the same 7 plasma samples, again excluding those SNVs found in corresponding tumors. Though the median fractional abundance was comparable ($\sim 0\%$, N.D.), the mean was marginally higher at 0.012% (FIG. 9 (b)). However, only one cancer-associated mutation (TP53 R175H) was detectable in plasma

at levels significantly above global background ($P < 0.01$). Since this allele was detected at a median frequency of $\sim 0.3\%$ across all samples (FIG. 9(c)), we hypothesize that it reflects true biological background and thus excluded it as a potential CAPP-Seq reporter. Collectively, this analysis suggests that biological background is not a significant factor for disease monitoring at the current detection limits of CAPP-Seq.

[0133] Next, the allele frequency detection limit and linearity of CAPP-Seq was benchmarked by spiking defined concentrations of fragmented genomic DNA from a NSCLC cell line into cfDNA from a healthy individual (FIG. 9(d)) or into genomic DNA from a second NSCLC line (FIG. 10(a)). CAPP-Seq accurately detected variants at fractional abundances between 0.025% and 10% with high linearity ($R^2 \geq 0.994$). Analyses of the influence of the number of SNV reporters on error metrics showed only marginal improvements above a threshold of 4 reporters per tumor (FIGS. 9(e)-(f), 10 (b)-(c)), equivalent to the median number of SNVs per NSCLC identified by the NSCLC selector. Finally, whether fusion breakpoints and indels could also serve as linear reporters was tested. It was found that the fractional abundance of these mutations correlated highly with expected concentrations ($R^2 \geq 0.995$; FIG. 10(d)).

[0134] Having designed, optimized, and benchmarked CAPP-Seq, it was applied to the discovery of somatic mutations in tumors collected from a diverse group of NSCLC patients ($n=11$; FIG. 11(a) and Table 3). To test the breakpoint enumeration capability of CAPP-Seq, 6 patients with clinically confirmed fusions were included. These translocations served as positive controls, along with SNVs in other tumors previously identified by clinical assays ($N=9$; Table 3). Tumor samples included formalin fixed surgical or biopsy specimens and pleural fluid. At a mean sequencing depth of $\sim 6,000\times$ in tumor and paired germline samples, CAPP-Seq confirmed all previously identified SNVs and fusions (3 and 8, respectively) and discovered many additional somatic variants (FIG. 11(a) and Table 3). Moreover, CAPP-Seq characterized breakpoints and partner genes at base pair resolution for each of the 8 rearrangements (FIG. 12). Tumors containing fusions were almost exclusively from never smokers and, as expected (Govindan et al. (2012) *Cell* 150:1121-1134), contained fewer SNVs than those lacking fusions (FIG. 13). Excluding patients with fusions ($<10\%$ of the TCGA design cohort), CAPP-Seq identified a median of 4 SNVs per patient as we had predicted (FIG. 1(b)-(c)).

TABLE 3

Characteristics of patients used for noninvasive detection and monitoring of circulating tumor DNA by CAPP-Seq.												
Case	Age	Sex	Histology	Grade and Other Histological Features	TNM Stage	Stage Group	Smoker	Pack- Years	Tumor Source	Germline Source	SNVs by Clinical Assays	Fusions Detected by FISH
P1	66	M	Adeno- carcinoma	Papillary type	T2aN0M0	B	Yes	20	FFPE cores	Frozen PBL		
P2	61	M	Large Cell	NOS	T3N1M0	IIIA	Yes	80	FFPE cores	Frozen PBL		
P3	67	F	Adeno- carcinoma	Acinar type	T1bN3M0	IIIB	Yes	15	FFPE cores	Frozen PBL		
P4	47	F	Adeno- carcinoma	Micropapillary and papillary type	T2aN2M1b	IV	Yes	45	FFPE cores	Frozen PBL	KRAS G13D	
P5	49	F	Adeno- carcinoma	Well differentiated	T1bN0M1a	IV	No	0	FFPE cores	Frozen PBL	EGFR L858R; EGFR T790M	
P6	54	M	Adeno- carcinoma	NOS	T3N2M1b	IV	No	0	Fresh	Frozen PBL		ALK

TABLE 3-continued

Characteristics of patients used for noninvasive detection and monitoring of circulating tumor DNA by CAPP-Seq.												
Case	Age	Sex	Histology	Grade and Other Histological Features	TNM Stage	Stage Group	Smoker	Pack- Years	Tumor Source	Germline Source	SNVs by Clinical Assays	Fusions Detected by FISH
P7	50	M	Adeno-carcinoma	Poorly differentiated	T1aN2M1b	IV	Yes	4	FFPE cores	Frozen PBL		ALK
P8	48	F	Adeno-carcinoma	Mutinous type	T4N0M1b	IV	No	0	FFPE cores	Frozen PBL		ALK
P9	49	M	Adeno-carcinoma	Not otherwise specified (NOS)	T4N3M1a	IV	No	0	Fresh	Frozen PBL		ALK
P10	35	F	Adeno-carcinoma	NOS	T4N0M0	IIIA	No	0	FFPE cores	Frozen PBL		ROS1
P11	38	F	Adeno-carcinoma	Well-to-moderately differentiated	T3N2M0	IIIA	No	0	FFPE cores	Frozen PBL		ROS1

: Related to FIGS. 11 (a) and 14, regarding smoking history, ≥ 20 pack years was considered heavy and >0 pack years was considered light.

[0135] To explore the potential clinical utility of CAPP-Seq for disease monitoring and minimal residual disease detection, we next applied CAPP-Seq to serial plasma samples collected from a subset of these same 11 patients (N=6), all of whom had pre- and post-treatment samples available (FIG. 11; Table 4). Starting from ~15 ng of plasma cfDNA (~3 mL of peripheral blood) and sequenced to a mean depth of nearly 8,000 $\times$ (Table 3), CAPP-Seq detected cancer-derived cfDNA in both early and advanced stage patients (Table 4). Among patients with SNV or indel reporters, all showed a significant reduction in cancer cfDNA burden following treatment, consistent with radiographic response assessment by computed tomography (CT) (FIG. 11(a)). These included two patients—one with stage IB adenocarcinoma (P1) and another with stage IIIA large cell carcinoma (P2)—who underwent surgery with complete tumor resection (FIG. 11(b)). Post-treatment cancer-derived cfDNA was undetectable in the Stage I patient but was above background for the Stage IIIA patient suggesting that residual cancer cells remained after surgery even though a complete resection was thought to have been achieved. In a third case (P6), CAPP-

Seq detected 3 SNVs and a KIF5B-ALK fusion, and both mutation types reported similar fractional abundances of mutant cfDNA (FIG. 14). Next, we analyzed a patient with 3 fusions and no detectable SNVs/indels (P9), but from whom 3 serial cfDNA samples were collected. Abundance of fusion product in the plasma was highly correlated with tumor burden and correctly indicated initial response to therapy followed by relapse ($R^2=0.97$; FIG. 11(c)). Finally, in a fifth patient (P5), CAPP-Seq identified a sub-clonal population harboring the T790M EGFR gatekeeper mutation (Kobayashi et al. (2005) *N. Engl. J. Med.* 352:786-792) (FIG. 11(d)). The ratio between clones was identical in the tumor and pre-treatment plasma cfDNA but changed after treatment with cytotoxic chemotherapy followed by a 3rd generation EGFR inhibitor (FIG. 11(d), inset), suggesting that CAPP-Seq can detect clinically relevant subclones and monitor clonal dynamics during therapy. Taken together, these data demonstrate the potential utility of CAPP-Seq as a noninvasive clinical assay for measuring tumor burden in early and advanced stage NSCLC and for monitoring tumor-derived cfDNA during therapy.

TABLE 4

Monitoring of cfDNA in patients using CAPP-Seq.																
Case	Mutant allele	Ref. allele	Chr	Position	Time point 1				Time point 2				Time point 3			
					Mutant allele depth	Total depth	Mutant allele %	Final %	Mutant allele depth	Total depth	Mutant allele %	Final %	Mutant allele depth	Total depth	Mutant allele %	Final %
P1	A	G	chr1	156785560	0	4572	0.000	0.000	3	6202	0.048	0.048	—	—	—	—
P1	T	G	chr1	157806043	0	1838	0.000	0.000	0	2266	0.000	0.000	—	—	—	—
P1	G	C	chr1	248525206	0	2828	0.000	0.000	0	4529	0.000	0.000	—	—	—	—
P1	C	T	chr2	33500291	1	943	0.106	0.106	0	943	0.000	0.000	—	—	—	—
P1	A	C	chr4	55946307	0	6856	0.000	0.000	0	8817	0.000	0.000	—	—	—	—
P1	G	A	chr4	55963949	0	5742	0.000	0.000	0	7335	0.000	0.000	—	—	—	—
P1	A	C	chr4	55968672	0	5856	0.000	0.000	0	7431	0.000	0.000	—	—	—	—
P1	C	T	chr6	117642146	0	5266	0.000	0.000	4	6849	0.058	0.058	—	—	—	—
P1	T	G	chr9	8376700	3	5535	0.054	0.054	0	7322	0.000	0.000	—	—	—	—
P1	T	C	chr9	8733625	1	827	0.121	0.121	0	1398	0.000	0.000	—	—	—	—
P1	T	G	chr10	43611663	0	3722	0.000	0.000	0	4565	0.000	0.000	—	—	—	—
P1	T	G	chr15	88522525	1	4919	0.020	0.020	4	6736	0.059	0.059	—	—	—	—
P1	+G	C	chr17	7578474	0	1762	0.000	0.000	0	2373	0.000	0.000	—	—	—	—
P1	-A	G	chr17	29552244	1	4484	0.022	0.022	0	6485	0.000	0.000	—	—	—	—
P1	+T	C	chr17	29553484	0	3657	0.000	0.000	0	4713	0.000	0.000	—	—	—	—
P1	-T	C	chr17	29592185	3	3694	0.081	0.081	0	3247	0.000	0.000	—	—	—	—
P2	A	C	chr2	50463926	49	6724	0.729	1.457	0	4981	0.000	0.000	—	—	—	—
P2	G	A	chr3	89457148	40	4838	0.827	0.827	0	4311	0.000	0.000	—	—	—	—
P2	T	G	chr3	89468286	5	4667	0.107	0.214	2	3625	0.055	0.110	—	—	—	—

TABLE 4-continued

Monitoring of cfDNA in patients using CAPP-Seq.																
Case	Mutant allele	Ref. allele	Chr	Position	Time point 1				Time point 2				Time point 3			
					Mutant allele depth	Total depth	Mutant allele %	Final %	Mutant allele depth	Total depth	Mutant allele %	Final %	Mutant allele depth	Total depth	Mutant allele %	Final %
P2	T	A	chr3	89480240	15	5073	0.296	0.591	0	4321	0.000	0.000	—	—	—	—
P2	T	A	chr4	66189669	4	950	0.421	0.842	5	1436	0.348	0.696	—	—	—	—
P2	T	G	chr4	66242868	16	2107	0.759	0.759	0	1655	0.000	0.000	—	—	—	—
P2	A	C	chr5	176522747	46	2220	2.072	2.072	0	1377	0.000	0.000	—	—	—	—
P2	C	T	chr6	117648229	70	7819	0.895	1.791	0	5985	0.000	0.000	—	—	—	—
P2	A	C	chr12	78400637	35	7907	0.443	0.885	1	6326	0.016	0.032	—	—	—	—
P2	T	G	chr12	78400910	106	8211	1.291	2.582	1	6289	0.016	0.032	—	—	—	—
P2	T	C	chr17	7577551	112	5629	1.990	1.990	2	3814	0.052	0.052	—	—	—	—
P2	T	G	chr19	1207247	15	1124	1.335	2.669	0	747	0.000	0.000	—	—	—	—
P2	+A	C	chr2	79314100	16	3280	0.488	0.98	0	2390	0.000	0.000	—	—	—	—
P3	A	C	chr17	7578253	6	6345	0.095	0.095	0	8583	0.000	0.000	—	—	—	—
P5	T	C	chr7	55249071	42	4736	0.887	0.887	10	5597	0.179	0.179	—	—	—	—
P5	G	T	chr7	55259515	503	11349	4.432	4.432	58	12222	0.475	0.475	—	—	—	—
P5	A	G	chr11	55135338	86	4063	2.117	2.117	10	4798	0.208	0.208	—	—	—	—
P5	T	C	chr17	7577097	227	7429	3.056	3.056	36	9723	0.370	0.370	—	—	—	—
P6	A	G	chr12	78400791	84	13970	0.601	1.203	28	10128	0.276	0.553	—	—	—	—
P6	T	G	chr12	129822187	78	8680	0.899	1.797	9	6604	0.136	0.273	—	—	—	—
P6	A	G	chr17	7576275	140	9376	1.493	1.493	22	7897	0.279	0.279	—	—	—	—
P6	KIF5B-ALK	—	chr10/chr2	—	28	15006	0.187	3.116	2	9989	0.020	0.334	—	—	—	—
P9	EML4-ALK	—	chr2/chr2	—	0	10688	0.000	0.000	0	13647	0.000	0.000	0	13521	0.000	0.000
P9	FYN-ROS1	—	chr6/chr6	—	0	9261	0.000	0.000	0	6826	0.000	0.000	2	10693	0.019	0.019
P9	ROS1-MKX	—	chr6/chr10	—	10	8029	0.125	0.125	1	6485	0.015	0.015	13	9943	0.131	0.131

Bolded reporters indicate potential homozygous alleles (see Table 3 and Detailed Methods).

Note that mutant cfDNA percentages for P5 were calculated from the 3 SNVs representing the dominant clone (see FIGS. 11 (a) and 11 (d)); EGFR T790M (chr7: 55249071 C→T) was not included.

Final allelic percentages reflect any adjustments made based on estimated zygosity (using inferred homozygous reporters) and/or sequencing coverage. See Detailed Methods for details.

[0136] In addition to its potential clinical utility, CAPP-Seq analysis promises to yield novel biological insights. For example, in one patient's tumor (P9), we identified both a classic EML4-ALK fusion and two previously unreported fusions involving ROS1: FYN-ROS1 and ROS1-MKX (FIG. 11(e), FIG. 15). While the potential function of these novel ROS1 fusions is unknown, to the best of our knowledge this is the first observation of ROS1 and ALK fusions in the same NSCLC patient. All fusions were confirmed by qPCR amplification of genomic DNA, and were independently recovered in plasma samples (Table 4). Separately, among cases with a ROS1 rearrangement, we found an unexpected enrichment for S34F missense mutations in U2AF1, the 35 kD subunit of the U2 spliceosomal complex auxiliary factor. This SNV was initially described as a recurrent heterozygous mutation in myelodysplastic syndrome (Graubert et al. (2012) *Nat. Genet.* 44:53-57; Yoshida et al. (2011) *Nature* 478:64-69). While U2AF1 mutations (Imielinski et al. (2012) *Cell* 150: 1107-1120) and ROS1 translocations (Bergethon et al. (2012) *J. Clin. Oncol.* 30:863-870) were recently reported to occur individually in ~3% and ~1.7% of lung adenocarcinomas, respectively, combining the samples we profiled with publicly available data (Detailed Methods), we observed a significant enrichment for U2AF1 S34F mutations tumors harboring ROS1 fusions (in 3 of 6; P=0.0019; FIG. 11(f), FIG. 16 and Detailed Methods).

[0137] Finally, we explored whether CAPP-Seq analysis of cfDNA could potentially be used for cancer screening. As proof-of-principle, we blinded ourselves to the mutations

present in each patient's tumor and developed a statistical method to test for the presence of cancer DNA in each pre-treatment plasma sample in our cohort (FIG. 17). This method identified mutant DNA in all plasma samples containing tumor-derived mutant alleles above fractional abundances of 0.5%. Mutant DNA below this level could not be detected by our algorithm, but no mutations were falsely called, indicating the high specificity of this approach (FIG. 11(g) and Detailed Methods). Since ~95% of nodules identified in patients at high risk for NSCLC by low-dose CT are false positives (Aberle et al. (2011) *N. Engl. J. Med.* 365:395-409), CAPP-Seq could potentially serve as a complementary non-invasive screening test. However, methodological improvements to further lower the detection threshold will be required to detect early stage tumors.

[0138] In conclusion, we have developed a flexible method for ultrasensitive and specific assessment of circulating tumor DNA. CAPP-Seq overcomes limitations of previously proposed methods for cfDNA analysis by simultaneously measuring multiple types of mutations without patient-specific optimization and by covering mutations in the majority of patients. Moreover, due to multiplexing, CAPP-Seq is highly economical, and per sample costs for plasma cfDNA are expected to drop further as NGS costs continue to fall. Our method has the potential to accelerate the personalized detection, therapy, and monitoring of cancer patients. We anticipate that CAPP-Seq will prove valuable in a variety of clinical settings, including the assessment of cancer DNA in alternative biological fluids and specimens with low cancer cell content.

Methods

Patient Selection

[0139] Between April 2010 and June 2012, patients undergoing treatment for newly diagnosed or recurrent NSCLC were enrolled in a study approved by the Stanford University Institutional Review Board. Enrolled patients had not received blood transfusions within 3 months of blood collection. Patient characteristics are in Table 3.

Sample Collection and Processing

[0140] Peripheral blood from consented patients was collected in EDTA Vacutainer tubes (BD). Blood samples were processed within 3 hours of collection. Plasma was separated by centrifugation at 2,500×g for 10 min, transferred to microcentrifuge tubes, and centrifuged at 16,000×g for 10 min to remove cell debris. The cell pellet from the initial spin was used for isolation of germline genomic DNA from PBLs (peripheral blood leukocytes) with the DNeasy Blood & Tissue Kit (Qiagen). Matched tumor DNA was isolated from FFPE specimens or from the cell pellet of pleural effusions. Genomic DNA was quantified by Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen).

Cell-Free DNA Purification and Quantification

[0141] Cell-free DNA (cfDNA) was isolated from 1-5 mL plasma with the QIAamp Circulating Nucleic Acid Kit (Qiagen). Absolute quantification of purified cfDNA was determined by quantitative PCR (qPCR) using an 81 bp amplicon on chromosome 1 (Fan et al. (2008) *Proc. Natl Acad. Sci. USA* 105:16266-16271) and a dilution series of intact male human genomic DNA (Promega) as a standard curve. Power SyberGreen was used for qPCR on a HT7900 Real Time PCR machine (Applied Biosystems). Standard PCR thermal cycling parameters were used.

Illumina NGS Library Construction

[0142] Indexed Illumina NGS libraries were prepared from cfDNA and shorn tumor, germline, and cell line genomic DNA. For patient cfDNA, 7-32 ng DNA was used for library construction without additional shearing or fragmentation. For tumor, germline, and cell line genomic DNA, 69-1000 ng DNA was sheared prior to library construction with a Covaris S2 instrument using the recommended settings for 200 bp fragments. See Table 2 for details.

[0143] The NGS libraries were constructed using the KAPA Library Preparation Kit (Kapa Biosystems) employing a DNA Polymerase possessing strong 3'-5' exonuclease (or proofreading) activity and displaying the lowest published error rate (i.e. highest fidelity) of all commercially available B-family DNA polymerases (Quail et al. (2012) *Nat. Methods* 9:10-11; Oyola et al. (2012) *BMC Genomics* 13:1). The manufacturer's protocol was modified to incorporate with-bead enzymatic and cleanup steps (Fisher et al. (2011) *Genome Biol.* 12:R1). Briefly, following the end repair reaction, Agencourt AMPure XP beads (Beckman-Coulter) were added to bind and wash the DNA fragments. The DNA was then eluted directly into 50 µL 1× A-tailing buffer containing the A-tailing enzyme. Following the A-tailing reaction, the DNA fragments were forced to bind to the same AMPure XP beads by adding 90 µL (1.8×) of PEG buffer (20% PEG-8000 in 2.5M NaCl). After washing, the DNA was eluted into 50 µL 1× ligation buffer with ligase and 100-fold molar excess of

indexed Illumina TruSeq adapters. Ligation was performed for 16 hours at 16° C. Single-step size selection was performed by adding 40 µL (0.8×) of PEG buffer to enrich for ligated DNA fragments. The ligated fragments were then amplified using 500 nM Illumina backbone oligonucleotides and a variable number of PCR cycles (between 4 and 9) depending on input DNA mass. In order to minimize bias and maximize recovery of GC-rich templates, all PCR reactions were carried out in a BioRad DNA Engine Thermal Cycler with a ramp rate of 2.2° C./sec or an Eppendorf Vapo Protect Mastercycler with the Safe ramp rate setting.

[0144] Library purity and concentration was assessed by spectrophotometer (NanoDrop 2000) and qPCR (KAPA Biosystems), respectively. Fragment length was determined on a 2100 Bioanalyzer using the DNA 1000 Kit (Agilent).

Design of Library for Hybrid Selection

[0145] Custom hybrid selection was performed with the SeqCap EZ Choice Library, v2.0 (Roche NimbleGen). The custom SeqCap library was designed through the NimbleDesign portal (v1.2.R1) using genome build HG19 NCBI Build 37.1/GRCh37 and with Maximum Close Matches set to 1. Input genomic regions were selected according to the most frequently mutated genes and exons in NSCLC. These regions were identified from the COSMIC database, TCGA, and other published sources as described in the Detailed Materials. Final selector coordinates are provided in Table 1.

Hybrid Selection and High Throughput Sequencing

[0146] NimbleGen SeqCap EZ Choice was used according to the manufacturer's protocol with modifications. Between 9 and 12 indexed Illumina libraries were included in a single capture reaction. Prior to hybrid selection, the libraries were quantified with a NanoDrop 2000 spectrophotometer, and 83-111 ng of each library was added (1 µg total DNA per capture reaction). Following hybrid selection, the captured DNA fragments were amplified with 12-to-14 cycles of PCR using 1× KAPA HiFi Hot Start Ready Mix and 2 µM Illumina backbone oligonucleotides in 4-to-6 separate 50 µL reactions. The reactions were then pooled and processed with the QIAquick PCR Purification Kit (Qiagen). Multiplexed libraries were sequenced using 2×100 bp paired-end runs on an Illumina HiSeq 2000.

Mapping and Quality Control of NGS Data

[0147] Paired-end reads were mapped to the hg19 reference genome with BWA 0.6.2 (default parameters) (Li & Durbin (2009) *Bioinformatics* 25:1754-1760), and sorted/indexed with SAMtools (Li et al. (2009) *Bioinformatics* 25:2078-2079). QC was assessed using a custom Perl script to collect a variety of statistics, including mapping characteristics, read quality, and selector on-target rate (i.e., number of unique reads that intersect the selector space divided by all aligned reads), generated respectively by SAMtools flagstat, FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), and BEDTools coverageBed (Quinlan & Hall (2010) *Bioinformatics* 26:841-842). Importantly, we used a custom version of coverageBed modified to count each read at most once. Plots of fragment length distribution and sequence depth/coverage were automatically generated for visual QC assessment. To mitigate the impact of sequencing errors, analyses not involving fusions were restricted to properly

paired reads, and high-quality bases with a Phred quality score of at least 30 ($\leq 0.1\%$ probability of a sequencing error) were further analyzed.

Analysis of Detection Thresholds by CAPP-Seq

[0148] Two dilution series were performed to assess the linearity and accuracy of CAPP-Seq for quantitating tumor-derived cfDNA. In one experiment, shorn genomic DNA from a NSCLC cell line (HCC78) was spiked into cfDNA from a healthy individual, while in a second experiment, shorn genomic DNA from one NSCLC cell line (NCI-H3122) was spiked into shorn genomic DNA from a second NSCLC line (HCC78). A total of 32 ng DNA was used for library construction. Following mapping and quality control, homozygous reporters were identified as alleles unique to each sample with at least 20 $\times$ sequencing depth at an allelic fraction $>80\%$. Fourteen such reporters were identified between HCC78 genomic DNA and plasma cfDNA (FIG. 9 (d), (e)), whereas 24 reporters were found between NCI-H3122 and HCC78 genomic DNA (FIG. 10).

CAPP-Seq Bioinformatics Pipeline

[0149] Details of bioinformatics methods are supplied in the Detailed Methods, and a graphical schematic is provided in FIG. 2. Briefly, for detection of SNVs and indels, we employed VarScan 2 (Koboldt et al. (2012) *Genome Res.* 22:568-576) with strict post-processing filters to improve variant call confidence, and for fusion identification and breakpoint characterization we used a novel algorithm, termed FACTERA (Detailed Methods). To quantify tumor burden in plasma cfDNA, allele frequencies of reporter SNVs/indels were assessed using the output of SAMtools mpileup (Li et al. (2009) *Bioinformatics* 25:2078-2079), and fusions, if detected, were enumerated with FACTERA.

Statistical Analysis

[0150] The NSCLC selector was validated in silico using an independent cohort of lung adenocarcinomas (Imielinski et al. (2012) *Cell* 150:1107-1120) (FIG. 1(c)). To assess statistical significance, we analyzed the same cohort using 10,000 random selectors sampled from the exome, each with an identical size distribution to the CAPP-Seq NSCLC selector. The performance of random selectors had a Gaussian distribution, and p-values were calculated accordingly. Note that all identified somatic lesions were considered in this analysis.

[0151] We used Monte Carlo sampling to estimate the distribution of background alleles across the NSCLC selector (FIG. 9 (a), (c); Detailed Methods). For each plasma sample, background alleles were defined as alleles remaining after exclusion of germline and/or somatic variant calls made by VarScan 2 (Koboldt et al. (2012) *Genome Res.* 22:568-576) (somatic p-value=0.01; otherwise, default parameters), and with a Phred quality score ≥ 30 . To evaluate the impact of reporter number on tumor burden estimates, we also performed Monte Carlo sampling (1,000 $\times$), varying the number of reporters available $\{1, 2, \dots, \max n\}$ in two spiking experiments (FIG. 9 (d)-(f); FIG. 10 (b)-(d)).

[0152] To assess the significance of tumor burden estimates in plasma cfDNA, we compared patient-specific SNV frequencies against the null distribution of background SNVs across the selector. Briefly, patient-specific background was quantified using the method described for FIG. 9 (a) (Detailed Methods), but using the number of SNVs identified in the

patient's tumor. For patients with at least 1 SNV, but no other reporter types, tumor-derived cfDNA was considered not detectable if mean SNV fractions fell below the 95th percentile of background alleles (i.e., $P \geq 0.05$) (FIG. 11 (a)). (Due to the ultra-low false detection rate for indels (Minoche et al. (2011) *Genome Biol.* 12:R112) and fusion breakpoints, these mutation types were considered detected when present with >0 read support.) For patients with detectable disease in only 1 time point, the corresponding empirical p-value is shown in FIG. 11 (a). To assess normality, we analyzed the patient with the most reporter alleles (i.e., P2; FIG. 11 (a)), and found that fractional abundance measurements fit a normal distribution (D'Agostino and Pearson omnibus normality test). Thus, for patients with detectable tumor-derived cfDNA in two time points and with at least 3 cfDNA SNVs/indels, the change in tumor burden was statistically assessed using a two-sided paired t-test. For P9, who lacked reporter SNVs/indels, statistical significance was estimated by correlation of CAPP-Seq measurements with known tumor volume (as measured by CT scans).

[0153] Additional details on cell lines, tumor cell sorting, optimizations of library preparation, mutation/translocation validation, CAPP-Seq design and analytical pipelines including FACTERA translocation detection tool, and additional statistical methods are presented in the Detailed Methods.

Detailed Methods

A. Molecular Biology Methods

A1. Cell Lines

[0154] The lung adenocarcinoma cell lines NCI-H3122 and HCC78 were obtained from ATCC and DSMZ, respectively, and grown in RPMI 1640 with L-glutamine (Gibco) supplemented with 10% fetal bovine serum (Gembio) and 1% penicillin/streptomycin cocktail. Cells were maintained in mid-log-phase growth in a 37° C. incubator with 5% CO₂. Genomic DNA was purified from freshly harvested cells with the DNeasy Blood & Tissue Kit (Qiagen).

A2. Pleural Fluid Processing and Flow Cytometry, and Cell Sorting

[0155] Cells from pleural fluid from patients P9 and P6 were harvested by centrifugation at 300 $\times$ g for 5 min at 4° C. and washed in FACS staining buffer (HBSS+2% heat-inactivated calf serum [HICS]). Red blood cells were lysed with ACK Lysing Buffer (Invitrogen), and clumps were removed by passing through a 100 μ m nylon filter. Filtered cells were spun down and resuspended in staining buffer. While on ice, the cell suspension was blocked for 20 min with 10 μ g/mL rat IgG and then stained for 20 min with APC-conjugated mouse anti-human EpCAM (BioLegend, clone 9C4), PerCP-Cy5.5-conjugated mouse anti-human CD45 (eBioscience, clone 2D1), and PerCP-eFluor710-conjugated mouse anti-human CD31 (eBioscience, clone WM59). After staining, cells were washed and resuspended with staining buffer containing 1 μ g/mL DAPI, analyzed, and sorted with a FACSaria II cell sorter (BD Biosciences). Cell doublets and DAPI-positive cells were excluded from analysis and sorting. CD31⁺CD45⁺EpCAM⁺ cells were sorted into staining buffer, spun down, and flash frozen in liquid nitrogen. DNA was isolated with the QIAamp DNA Micro Kit (Qiagen).

A3. Optimization of NGS Library Preparation from Low Input cfDNA

[0156] Any method for detecting mutant cfDNA relies on its ability to interrogate each cfDNA molecule in the circulation in order to maximize sensitivity. For this reason, we used the QIAamp Circulating Nucleic Acid kit (Qiagen) with carrier RNA as per the manufacturer's protocol to isolate cfDNA. We also took specific steps to improve the Illumina library preparation workflow.

[0157] Protocols for Illumina library construction were compared in a step-wise manner with the goal of (1) optimizing adapter ligation efficiency, (2) reducing the necessary number of PCR cycles following adapter ligation, (3) preserving the naturally occurring size distribution of cfDNA fragments, and (4) minimizing variability in depth of sequencing coverage across all captured genomic regions. Initial optimization was done with NEBNext DNA Library Prep Reagent Set for Illumina (New England BioLabs), which includes reagents for end-repair of the cfDNA fragments, A-tailing, adapter ligation, and amplification of ligated fragments with Phusion High-Fidelity PCR Master Mix. Input was 4 ng cfDNA (obtained from plasma of the same healthy volunteer) for all conditions. Relative allelic abundance in the constructed libraries was assessed by qPCR of 4 genomic loci (Roche NimbleGen: NSC-0237, NSC-0247, NSC-0268, and NSC-0272) and compared by the $2^{-\Delta C_t}$ method.

[0158] Ligations were performed at 20° C. for 15 min (as per the manufacturer's protocol), at 16° C. for 16 hours, or with temperature cycling for 16 hours as previously described (Lund et al. (1996) *Nucl. Acids Res.* 24:800-801). Ligation volumes were varied from the standard (50 μ L) down to 10 μ L while maintaining a constant concentration of DNA ligase, cfDNA fragments, and Illumina adapters. Subsequent optimizations incorporated ligation at 16° C. for 16 hours in 50 μ L reaction volumes.

[0159] Next, we compared standard SPRI bead processing procedures, in which new AMPure XP beads are added after each enzymatic reaction and DNA is eluted from the beads for the next reaction, to with-bead protocol modifications as previously described (Fisher, S. et al. (2011) *Genome Biol.* 12:R1). We compared 2 concentrations of Illumina adapters in the ligation reaction: 12 nM (10-fold molar excess to cfDNA fragments) and 120 nM (100-fold molar excess).

[0160] Using the optimized library preparation procedures, we next compared the NEBNext DNA Library Prep Reagent Set (with Phusion DNA Polymerase) to the KAPA Library Preparation Kit (with KAPA HiFi DNA Polymerase). The KAPA Library Preparation Kit with our modifications was also compared to the NuGEN SP Ovation Ultralow Library System with automation on Mondrian SP Workstation.

A4. Evaluation of Library Preparation Modifications on CAPP-Seq Performance

[0161] We performed CAPP-Seq on 32 ng cfDNA using standard library preparation procedures with the NEBNext kit, or with optimized procedures using either the NEBNext kit or the KAPA Library Preparation Kit. In parallel we performed CAPP-Seq on 4 ng and 128 ng cfDNA using the KAPA kit with our optimized procedures. Indexed libraries were constructed, and hybrid selection was performed in multiplex. The post-capture multiplexed libraries were

amplified with Illumina backbone primers for 14 cycles of PCR and then sequenced on a paired-end 100 bp lane of an Illumina HiSeq 2000.

[0162] We also evaluated CAPP-Seq on ultralow input following whole genome amplification (WGA). For WGA we chose not to use multiple displacement amplification with Φ 29 DNA polymerase due given the small size of cfDNA fragments in plasma (FIG. 1(e)), and due to concern for chimera formation, which would confound analysis of recurrent gene fusions in NSCLC by CAPP-Seq. Instead we used SeqPlex DNA Amplification Kit (Sigma-Aldrich), which employs degenerate oligonucleotide primer PCR. We used the upper limit of input into this kit (1 ng) and performed whole genome amplification according to the manufacturer's protocol. Briefly, 1 ng cfDNA was amplified with real-time monitoring with SYBR Green I (Sigma-Aldrich) on a HT7900 Real Time PCR machine (Applied Biosystems). The amplification was terminated after 17 cycles yielding 2.8 μ g DNA. The primer removal step yielded ~600 ng DNA, and this entire amount was used for library preparation using the NEBNext kit with optimized procedures as described above.

A5. Validation of Variants Detected by CAPP-Seq

[0163] All structural rearrangements and a subset of tumoral SNVs detected by CAPP-Seq were independently confirmed by qPCR and/or Sanger sequencing of amplified fragments. For HCC78, a 120 bp fragment containing the SLC34A2-ROS1 breakpoint was amplified from genomic DNA using the primers: 5'-AGACGGGAGAAAATAG-CACC-3' and 5'-ACCAAGGGTTGCAGAAATCC-3'. A 141 bp fragment containing exon 2 of U2AF1 was amplified using the primers: 5'-CATGTGTTTGATATCTTCCCAGC-3' and 5'-CTGGCTAAACGTCGGTTTATTG-3'. For NCI-H3122, a 143 bp fragment containing the EML4-ALK breakpoint was amplified using the primers: 5'-GAGATG-GAGTTTCACTCTTGTTGC-3' and 5'-GAACCTTTCCAT-CATACTTAGAAATAC-3'. 5 ng genomic DNA was used as template with 250 nM oligos and 1 $\times$ Phusion PCR Master Mix (NEB) in 50 μ L reactions. Products were resolved on 2.5% agarose gel and bands of the expected size were removed. The amplified DNA fragments were purified using the Qiaquick Gel Extraction Kit (Qiagen) and submitted for Sanger sequencing (Elim Biopharm). For P9, genomic DNA breakpoints were confirmed by qPCR using the primers: 5'-TCCATGGAAGCCAGAAC-3' and 5'-ATGCTAAGAT-GTGTCTGTCA-3' for EML4-ALK; 5'-CCTTAACACA-GATGGCTCTTGATGC-3' and 5'-TCCTCTTTCCACCT-TGGCTTTCC-3' for ROS1-MKX; and 5'-GGTTCAGAACTACCAATAACAAG-3' and 5'-ACCT-GATGTGTGACCTGATTGATG-3' for FYN-ROS1. For qPCR, 10 ng of pre-amplified genomic DNA was used as template with 250 nM oligos and 1 $\times$ Power SyberGreen Master Mix in 10 μ L reactions performed in triplicate on a HT7900 Real Time PCR machine (Applied Biosystems). Standard PCR thermal cycling parameters were used. Amplification of amplicons spanning all 3 breakpoints detected in P9 were confirmed in tumor genomic DNA as well as plasma cfDNA, and PBL genomic DNA was used as a negative control. Separately, at least 88% of SNVs and indels detected were bona fide somatic mutations in tumors, as 38 of 46 of them were independently observed above 0.025% allele frequency in plasma cfDNA and/or were independently confirmed by SNaPshot clinical assays.

B. Bioinformatics and Statistical Methods

B1. Analysis of CAPP-Seq Background

[0164] The CAPP-Seq background rate was estimated by Monte Carlo sampling of allelic frequencies across the NSCLC selector (FIG. 9 (a)). Plasma cfDNA samples were pre-filtered to remove all variant calls and dominant alleles. Specifically, for each patient, we excluded germline, loss of heterozygosity (LOH), and/or somatic variant calls made by VarScan 2 (Koboldt et al. (2012) *Genome Res.* 22:568-576) (somatic p-value=0.01; otherwise, default parameters). We sampled 4 random background alleles across this subset of the selector (equal to the median number of SNVs per NSCLC patient detected by CAPP-Seq) and calculated their mean allelic frequency, only considering bases discordant with the prevailing genotype of the plasma sample at those 4 positions. This process was iterated 10,000 times, and mean, median, and 75th percentile statistics were collected. The entire procedure was then repeated for 5 total simulations, shown in FIG. 2a.

[0165] We likewise applied Monte Carlo simulation to estimate the probability of finding a background allele in plasma cfDNA at a given fractional abundance (FIG. 9 (c)). For consistency with the ranking of alleles in FIG. 9 (c), we populated a vector containing the mean background allele frequency for each genomic position across 7 plasma cfDNA samples, each filtered to remove dominant alleles as described above. Alleles were randomly sampled from this vector 10,000 times to identify the allele frequency with an empirical p-value of 0.01.

B2. ROS1 and U2AF1 Co-Association Analysis

B2.1 Assembly of ROS1 and U2AF1 Mutant NSCLC

[0166] We included only cases in which the status of both ROS1 fusion status and U2AF1 S34 mutation was known. There were 163 such cases from TCGA (genotyped for U2AF1 by whole exome sequencing and for ROS1 fusions by RNA-Seq as detailed below), 23 cases from Imielinski et al. (2012) *Cell* 150:1107-1120, 17 cases from Govindan et al. (2012) *Cell* 150:1121-1134, and 13 cases from the present study (11 patients and 2 NSCLC cell lines). U2AF1 S34F mutations were detected in 11 cases (5 from TCGA, 3 from Imielinski et al., 1 from Govindan et al., and 2 from the present study), and ROS1 fusions were detected in 6 cases (2 from TCGA, described below, and 4 from the present study). Significance testing was performed using the Fisher's exact test, and a two-tailed P-value is reported.

B2.2. Analysis of Whole Transcriptome Sequencing Data from TCGA for ROS1 Fusions

[0167] We identified two TCGA lung adenocarcinoma patients, TCGA-05-4426 and TCGA-64-1680, harboring candidate ROS1 fusions (FIG. 16 (a)). Importantly, the latter patient also has the U2AF1 S34F missense mutation reported in this study and in prior literature (see above). To further analyze both patients' putative rearrangements, whole transcriptome RNA-Seq data (.bam files) were obtained using the UCSC GeneTorrent system (<https://cghub.ucsc.edu/downloads.html>) and realigned to hg19 using BWA 0.6.2 using default parameters (Li & Durbin (2009) *Bioinformatics* 25:1754-1760). Importantly, mapped RNA-Seq reads extended significantly past coding regions, allowing for improved assessment of fusion events (FIG. 16 (b), (c)). From

a manual inspection of associated RPKM expression data across ROS1 exons (FIG. 16 (a)), we suspected that breakpoint sites for these fusions may lie directly upstream of ROS1 exons 32 and 35, respectively. Using the Integrated Genome Viewer (IGV) (Robinson et al. (2011) *Nat. Biotechnol.* 29:24-26), we found improperly paired (or discordant) reads near these exons that link ROS1 to its well-described partners, SLC34A2 and CD74, respectively (FIG. 16 (b), (c)). Indeed, by applying FACTERA's templated fusion discovery (detailed below) to patient TCGA-64-1680, we recovered a single read near ROS1 exon 35 that also maps to CD74 (FIG. 16 (c)). Collectively, these data strongly support the existence of expressed ROS1 fusions in these two TCGA patients.

B3. CAPP-Seq Selector Design

[0168] Most human cancers are relatively heterogeneous for somatic mutations in individual genes. Specifically, in most human tumors, recurrent somatic alterations of single genes account for a minority of patients, and only a minority of tumor types can be defined using a small number of recurrent mutations (<5-10) at predefined positions. Therefore, the design of the selector is vital to the CAPP-Seq method because (1) it dictates which mutations can be detected in with high probability for a patient with a given cancer, and (2) the selector size (in kb) directly impacts the cost and depth of sequence coverage. For example, the hybrid selection libraries available in current whole exome capture kits range from 51-71 Mb, providing ~40-60 fold maximum theoretical enrichment versus whole genome sequencing. The degree of potential enrichment is inversely proportional to the selector size such that for a ~100 kb selector, >10,000 fold enrichment should be achievable.

[0169] We employed a six-phase design strategy to identify and prioritize genomic regions for the CAPP-Seq NSCLC selector as detailed below. Three phases were used to incorporate known and suspected NSCLC driver genes, as well as genomic regions known to participate in clinically actionable fusions (phases 1, 5, 6), while another three phases employed an algorithmic approach to maximize both the number of patients covered and SNVs per patient (phases 2-4). The latter relied upon a metric that we termed "Recurrence Index" (RI), defined as the number of NSCLC patients with SNVs that occur within a given kilobase of exonic sequence (i.e., No. of patients with mutations/exon length in kb). RI thus serves to measure patient-level recurrence frequency at the exon level, while simultaneously normalizing for gene/exon size. As a source of somatic mutation data uniformly genotyped across a large cohort of patients, in phases 2-4, we analyzed non-silent SNVs identified in TCGA whole exome sequencing data from 178 patients in the Lung Squamous Cell Carcinoma dataset (SCC) (Hammerman et al. (2012) *Nature* 489:519-525) and from 229 patients in the Lung Adenocarcinoma (LUAD) datasets (TCGA query date was Mar. 13, 2012). Thresholds for each metric (i.e. RI and patients per exon) were selected to statistically enrich for known/suspected drivers in SCC and LUAD data (FIG. 9). RefSeq exon coordinates (hg19) were obtained via the UCSC Table Browser (query date was Apr. 11, 2012).

[0170] The following algorithm was used to design the CAPP-Seq selector (parenthetical descriptions match design phases noted in FIG. 1 (b)).

[0171] Phase 1 (Known Drivers)

[0172] Initial seed genes were chosen based on their frequency of mutation in NSCLCs.

[0173] Analysis of COSMIC (v57) (Forbes et al. (2010) *Nucl. Acids Res.* 38:D652-657) identified known driver genes

that are recurrently mutated in $\geq 9\%$ of NSCLC (denominator ≥ 500 cases). Specific exons from these genes were selected based on the pattern of SNVs previously identified in NSCLC. The seed list also included single exons from genes with recurrent mutations that occurred at low frequency but had strong evidence for being driver mutations, such as BRAF exon 15, which harbors V600E mutations in $<2\%$ of NSCLC (Ding et al. (2008) *Nature* 455:1069-1075; Youn & Simon (2011) *Bioinformatics* 27:175-181; Okuda et al. (2008) *Cancer Sci.* 99:2280-2285; Su et al. (2011) *J. Mol. Diagn.* 13:74-84; Tsao et al. (2007) *J. Clin. Oncol.* 25:5240-5247; Chaff et al. (2012) *Mol. Cancer Ther.* 11:485-491; Paik et al. (2011) *J. Clin. Oncol.* 29:2046-2051; Stephens et al. (2004) *Nature* 431:525-526; Jin et al. (2010) *Lung Cancer* 69:279-283; Malanga et al. (2008) *Cell Cycle* 7:665-669).

[0174] Phase 2 (Max. Coverage)

[0175] For each exon with SNVs covering ≥ 5 patients in LUAD and SCC, we selected the exon with highest RI that identified at least 1 new patient when compared to the prior phase. Among exons with equally high RI, we added the exon with minimum overlap among patients already captured by the selector. This was repeated until no further exons met these criteria.

[0176] Phase 3 (RI ≥ 30)

[0177] For each remaining exon with an RI ≥ 30 and with SNVs covering ≥ 3 patients in LUAD and SCC, we identified the exon that would result in the largest reduction in patients with only 1 SNV. To break ties among equally best exons, the exon with highest RI was chosen. This was repeated until no additional exons satisfied these criteria.

[0178] Phase 4 (RI ≥ 20)

[0179] Same procedure as phase 3, but using RI ≥ 20 .

[0180] Phase 5 (Predicted Drivers)

[0181] We included all exons from additional genes previously predicted to harbor driver mutations in NSCLC (Ding et al. (2008) *Nature* 455:1069-1075; Youn & Simon (2011) *Bioinformatics* 27:175-181).

[0182] Phase 6 (Add Fusions)

[0183] For recurrent rearrangements in NSCLC involving the receptor tyrosine kinases ALK, ROS1, and RET, the introns most frequently implicated in the fusion event and the flanking exons were included.

[0184] All exons included in the selector, along with their corresponding HUGO gene symbols and genomic coordinates, as well as patient statistics for NSCLC and a variety of other cancers, are provided in Table 1, organized by selector design phase.

C. CAPP-Seq Computational Pipeline

C1. Mutation Discovery: SNVs/Indels

[0185] For detection of somatic SNV and insertion/deletion events, we employed VarScan 2 (Koboldt et al. (2012) *Genome Res* 22:568-576) (somatic p-value=0.01, minimum variant frequency=5%, and otherwise default parameters). Somatic variant calls (SNV or indel) present at less than 0.5% mutant allelic frequency in the paired normal sample (PBLs), but in a position with at least 1000 $\times$ overall depth in PBLs and 100 $\times$ depth in the tumor, and with at least 1 $\times$ read depth on each strand, were retained (Table 3). While the selector was designed to predominantly capture exons, in practice, it also captures limited sequence content flanking each targeted region. For instance, this phenomenon is the basis for the (thus far) uniformly successful recovery by CAPP-Seq of

fusion partners (which are not included within the selector) for kinase genes such as ALK and ROS1 recurrently rearranged in NSCLC. As such, we also considered variant calls detected within 500 bps of defined selector coordinates. These calls were eliminated if present in non-coding repeat regions, since repeats may confound mapping accuracy. Repeat sequence coordinates were obtained using the RepeatMasker track in the UCSC table browser (hg19). Variant annotation was performed using the SeattleSeq Annotation 137 web server (<http://snp.gs.washington.edu/SeattleSeqAnnotation137/>). Complete details for all identified SNVs and indels are provided in Table 2.

[0186] By manual inspection, two patients (P2 and P6) had SNVs with frequencies consistent with potential heterozygous and homozygous alleles. We labeled these alleles accordingly (Table 3), and based on our assumption of zygosity in these two patients, we adjusted measured fractions of heterozygous reporters in plasma cfDNA to better estimate tumor burden (Table 4).

C2. Mutation Discovery: Fusions

[0187] For practical and robust de novo enumeration of genomic fusion events and breakpoints from paired-end next-generation sequencing data, we developed a novel heuristic approach, termed FACTERA (FACile Translocation Enumeration and Recovery Algorithm). FACTERA has minimal external dependencies, works directly on a preexisting .bam alignment file, and produces easily interpretable output. Major steps of the algorithm are summarized below, and are complemented by a graphical schematic to illustrate key elements of the breakpoint identification process (FIG. 4).

[0188] As input, FACTERA requires a .bam alignment file of paired-end reads produced by BWA (Li & Durbin (2009) *Bioinformatics* 25:1754-1760), exon coordinates in .bed format (e.g., hg19 RefSeq coordinates), and a 0.2 bit reference genome to enable fast sequence retrieval (e.g., hg19). In addition, the analysis can be optionally restricted to reads that overlap particular genomic regions (.bed file), such as the CAPP-Seq selector used in this work.

[0189] FACTERA processes the input in three sequential phases: identification of discordant reads, detection of breakpoints at base pair-resolution, and in silico validation of candidate fusions. Each phase is described in detail below.

C2.1. Identification of Discordant Reads

[0190] To iteratively reduce the sequence space for gene fusion identification, FACTERA, like other algorithms (e.g. BreakDancer (Chen et al. (2009) *Nat. Methods* 6:677-681)), identifies and classifies discordant read pairs. Such reads indicate a nearby fusion event since they either map to different chromosomes or are separated by an unexpectedly large insert size (i.e. total fragment length), as determined by the BWA mapping algorithm. The bitwise flag accompanying each aligned read encodes a variety of mapping characteristics (e.g., improperly paired, unmapped, wrong orientation, etc.) and is leveraged to rapidly filter the input for discordant pairs. The closest exon of each discordant read is subsequently identified, and used to cluster discordant pairs into distinct gene-gene groups, yielding a list of genomic regions R adjacent to candidate fusion sites. For each member gene of a discordant gene pair, the genomic region R_i is defined by taking the minimum of all 3' exon/read coordinates in the cluster, and the maximum of all 5' exon/read coordinates in

the cluster. These regions are used to prioritize the search for breakpoints in the next phase (FIG. 4 (a)).

C2.2 Detection of Breakpoints at Base Pair-Resolution

[0191] Discordant read pairs may be introduced by NGS library preparation and/or sequencing artifacts (e.g., jumping PCR). However, they are also likely to flank the breakpoints of bona fide fusion events. As such, all discordant gene pairs identified in the preceding of one read matches the soft-clipped region of the other, FACTERA records a putative fusion event. To assess inter-read concordance (e.g. see reads 1 and 2 in FIG. 4 (c)), FACTERA employs the following algorithm. The mapped region of read 1 is parsed into all possible subsequences of length k (i.e., k -mers) using a sliding window ($k=10$, by default). Each k -mer, along with its lowest sequence index in read 1, is stored in a hash table data structure, allowing k -mer membership to be assessed in constant time (FIG. 4 (c), left panel). Subsequently, the soft clipped sequence of read 2 is parsed into non-overlapping subsequences of length k , and the hash table is interrogated for matching k -mers (FIG. 4 (c), right panel). If a minimum matching threshold is achieved ($=0.5 \times$ the minimum length of the two compared subsequences), then the two reads are considered concordant. FACTERA will process at most 1000 (by default) putative breakpoint pairs for each discordant gene pair. Moreover, for each gene pair, FACTERA will only compare reads whose orientations are compatible with valid fusions. Such reads have soft-clipped sequences facing opposite directions (FIG. 4 (d), top panel). When this condition is not satisfied, FACTERA uses the reverse complement of read 1 for k -mer analysis (FIG. 4 (d), bottom panel).

[0192] In some instances, genomic subsequences flanking the true breakpoint may be nearly or completely identical, causing the aligned portions of soft-clipped reads to overlap. Unfortunately, this prevents an unambiguous determination of the breakpoint. As such, FACTERA incorporates a simple algorithm to arbitrarily adjust the breakpoint in one read (i.e., read 2) to match the other (i.e., read 1). Depending upon read orientation, there are two ways this can occur, both of which are illustrated in FIG. 4 (e). For each read, FACTERA calculates the distance between the breakpoint and the read coordinate corresponding to the first k -mer match between reads. For example, as anecdotally illustrated in FIG. 4 (e), x is defined as the distance between the breakpoint coordinate of read 1 and the index of the first matching k -mer, j , whereas y denotes the corresponding distance for read 2. The offset is estimated as the difference in distances (x, y) between the two reads (see FIG. 4 (e)).

C2.3. In Silico Validation of Candidate Fusions

[0193] To confirm each candidate breakpoint in silico, FACTERA performs a local realignment of reads against a template fusion sequence (± 500 bp around the putative breakpoint) extracted from the 0.2 bit reference genome. BLAST is currently employed for this purpose, although BLAT or other fast aligners could be substituted. A BLAST database is constructed by collecting all reads that map to each candidate fusion sequence, including discordant reads and soft-clipped reads, as well as all unmapped reads in the original input .bam file. All reads that map to a given fusion candidate with at least 95% identity and a minimum length of 90% of the input read length (by default) are retained, and reads that span or flank the breakpoint are counted. As a final step, output redundan-

cies are minimized by removing fusion sequences within a 20 bp interval of any fusion sequence with greater read support and with the same sequence orientation (to avoid removing reciprocal fusions).

[0194] FACTERA produces a simple output text file, which includes for each fusion sequence, the gene pair, the chromosomal sequence coordinates of the breakpoint, the fusion orientation (e.g., forward-forward or forward-reverse), the genomic sequences within 50 bp of the breakpoint, and depth statistics for reads spanning and flanking the breakpoint. Fusions identified in patients analyzed in this work are provided in Table 3.

C2.4. Experimental Validation of FACTERA

[0195] To experimentally evaluate the performance of FACTERA, we generated NGS data from two NSCLC cell lines, HCC78 (21.5M $\times$ 100 bp paired-end reads) and NCI-H3122 (19.4M $\times$ 100 bp paired-end reads), each of which has a known rearrangement (ROS1 and ALK, respectively) (Bergeth et al. (2012) *J. Clin. Oncol.* 30:863-870; McDermott et al. (2008) *Cancer Res.* 68:3389-3395) with a breakpoint that has, to the best of our knowledge, not been previously published. FACTERA readily revealed evidence for a reciprocal SLC34A2-ROS1 translocation in the former and an EML4-ALK fusion in the latter. Precise breakpoints predicted by FACTERA were experimentally validated by PCR amplification and Sanger sequencing (FIG. 5; see also Validation of Variants Detected by CAPP-Seq). Importantly, FACTERA completed each run in practical time (~ 90 sec), using only a single thread on a hexa-core 3.4 GHz Intel Xeon E5690 chip. These initial results illustrate the utility of FACTERA as part of the CAPP-Seq analysis pipeline.

C2.5. Templated Fusion Discovery

[0196] We implemented a user-directed option to “hunt” for fusions within expected candidate genes. A fusion could be missed by FACTERA if the fusion detection criteria employed by FACTERA are incompletely satisfied—such as if discordant reads, but not soft-clipped reads, are identified—and will most likely occur when fusion allele frequency in the tumor is extremely low. As input, the method is supplied with candidate fusion gene sequences as “baits”. All unmapped and soft-clipped reads in the input .bam file are subsequently aligned to these templates (using blastn) to identify reads that have sufficient similarity to both (for each read, 95% identity, e -value $< 1.0e-5$, and at least 30% of the read length must map to the template, by default). Such reads are output as a list to the user for manual analysis.

[0197] We tested this simple approach on a low purity tumor sample found to harbor an ALK fusion by FISH, but not FACTERA (i.e., case P9). Using templates for ALK and its common fusion partner, ELM4, we identified 4 reads that mapped to both, in a region with an overall depth of $\sim 1900\times$. The estimated allele frequency of 0.21% is strikingly similar to the 0.22% tumor purity measured by FACS (FIG. 15), confirming the utility of the templated fusion discovery method. We subsequently FACS-depleted CD45+ immune populations and re-sequenced this patient’s tumor. In the enriched tumor sample, FACTERA identified the EML4-ALK fusion, along with two novel ROS1 fusions (FIG. 4 (e), Table 3).

C3. Mutation Recovery: SNVs/Indels

[0198] Using a custom Perl script, previously identified reporter alleles were intersected with a SAMtools mpileup file generated for each plasma cfDNA sample, and the number and frequency of supporting reads was calculated for each reporter allele. Only reporters in properly paired reads at positions with at least 500 $\times$ overall depth were considered.

C4. Mutation Recovery: Fusions

[0199] For enumeration of fusion frequency in sequenced plasma DNA, FACTERA executes the last step of the discovery phase (i.e., in silico validation of candidate fusions, above) using the set of previously identified fusion templates. The fusion allele frequency is calculated as α/β , where α is the number of breakpoint-spanning reads, and β is the mean overall depth within a genomic region ± 5 bps around the breakpoint. Regarding the NSCLC selector described in this work, the latter calculation was always performed on the single gene contained in the NSCLC selector library. If both fusion genes are targeted within a selector library, overall depth is estimated by taking the mean depth calculated for both genes.

[0200] Notably, in some cases we observed lower fusion allele frequencies than would be expected for heterozygous alleles (e.g., see cell line fusions in Table 3). This was seen in cell lines, in an empirical spiking experiment, and in one patient's tumor and plasma samples (i.e., P6), and could potentially result from inefficient "pull-down" of fusions whose partners are not represented in the selector. Regardless, fusions are useful reporters—they possess virtually no background signal and show linear behavior over defined concentrations in a spiking experiment (FIG. 10 (d)). Moreover, allelic frequencies in plasma are easily adjusted for such inefficiencies by dividing the measured frequency in plasma by the corresponding frequency in the tumor. In cases where sequenced tumor tissue is impure, tumor content can be estimated using the frequencies of SNVs (or indels) as a reference frame, allowing the fusion fraction to be normalized accordingly (Table 4). As for SNVs/indels, only fusions present in at least one plasma sample were included in calculations of tumor burden.

C5. Screening Plasma cfDNA without Knowledge of Tumor DNA

[0201] We devised the following statistical algorithm as a first step toward non-invasive cancer screening with plasma cfDNA. The method identifies candidate SNVs using iterative models of (i) background noise in paired germline DNA (in this work, PBLs), (ii) base-pair resolution background frequencies in plasma cfDNA across the selector, and (iii) sequencing error in cfDNA. Anecdotal examples are provided in FIG. 17. The algorithm works in four main steps, detailed below.

[0202] As input, the algorithm takes allele frequencies from a single plasma cfDNA sample and analyzes high quality background alleles, defined in a first step for each genomic position as the non-dominant base with highest fractional abundance. Only alleles with depth of at least 500 $\times$ and strand bias <90% (conservative, by default) are analyzed. For consistency with variant calling, we allowed the screening approach to interrogate selector regions within 500 bp of defined coordinates, expanding the effective sequence space from ~ 125 kb to ~ 600 kb.

[0203] Second, the binomial distribution is used to test whether a given input cfDNA allele is significantly different from the corresponding paired germline allele (FIG. 17 (a)-(b)). Here the probability of success is taken to be the frequency of the background allele in PBLs, and the number of trials is the allele's corresponding depth in plasma cfDNA. To avoid contributions from alleles in rare circulating tumor cells that might contaminate PBLs, input alleles with a fractional abundance greater than 0.5% in paired PBLs (by default) or a Bonferroni-adjusted binomial probability greater than 2.08×10^{-8} are not further considered (alpha of 0.05/[~ 600 kb*4 alleles per position]).

[0204] Third, a database of cfDNA background allele frequencies is assembled. Here, we used samples analyzed in the present study (i.e., pre-treatment NSCLC samples and 1 sample from a healthy volunteer), except the input sample is left out to avoid bias. Based on the assumption that all background allele fractions follow a normal distribution, a Z-test is employed to test whether a given input allele differs significantly from typical cfDNA background at the same position (FIG. 17 (a)-(b)). All alleles within the selector are evaluated, and those with an average background frequency of 5% or greater (by default) or a Bonferroni-adjusted single-tailed Z-score <5.6 are not further considered (alpha of 0.05, adjusted as above).

[0205] Finally, candidate alleles are tested for remaining possible sequencing errors. This step leverages the observation that non-tumor variants (i.e., "errors") in plasma cfDNA tend to have a higher duplication rate than bona fide variants detectable in the patient's tumor (data not shown). As such, the number of supporting reads is compared for each input allele between nondeduped (all fragments meeting QC criteria; see Methods) and deduped data (only unique fragments meeting QC criteria). An outlier analysis is then used to distinguish candidate tumor-derived SNVs from remaining background noise (FIG. 17 (a)-(c)). Specifically, to reveal outlier tendency in the data, the square root of the robust distance R_d (Mahalanobis distance) is compared against the square root of the quantiles of a chi-squared distribution C_s . This transformation reveals natural separation between true SNVs and false positives in cancer patients (FIG. 17 (a), (c)), and notably, reveals an absence of outlier structure in patient samples lacking tumor-derived SNVs (FIG. 17 (b), (c)). To automatically call SNVs without prior knowledge, the screening approach iterates through data points by decreasing R_b and recalculating the Pearson's correlation coefficient Rho between R_d and C_s for points 1 to i , where R_{d_i} is the current maximum R_d . The algorithm iteratively reports outliers (i.e., candidate SNVs) until it terminates when $Rho \geq 0.85$.

[0206] Importantly, this approach positively identified 60% of the cancer samples with tumor-derived SNVs analyzed in this study with no false positive calls (FIG. 11 (g)). When corresponding germline DNA from PBLs are unavailable, one can skip the 2nd step in this screening routine. After removal of germline SNVs with an allelic fraction >20%, this modified approach identified no SNVs when applied to a healthy volunteer.

[0207] All patents, patent publications, and other published references mentioned herein are hereby incorporated by reference in their entireties as if each had been individually and specifically incorporated by reference herein.

[0208] While specific examples have been provided, the above description is illustrative and not restrictive. Any one or

more of the features of the previously described embodiments can be combined in any manner with one or more features of any other embodiments in the present invention. Furthermore, many variations of the invention will become apparent

to those skilled in the art upon review of the specification. The scope of the invention should, therefore, be determined by reference to the appended claims, along with their full scope of equivalents.

SEQUENCE LISTING

<160> NUMBER OF SEQ ID NOS: 35

<210> SEQ ID NO 1

<211> LENGTH: 12

<212> TYPE: DNA

<213> ORGANISM: Artificial Sequence

<220> FEATURE:

<223> OTHER INFORMATION: synthetic polynucleotide

<400> SEQUENCE: 1

tctggctata gc 12

<210> SEQ ID NO 2

<211> LENGTH: 101

<212> TYPE: DNA

<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 2

agaaatacta ataaatgat taaagaagggt gtgtctttaa ttgaagcatg atttaaagta 60

aatgcaaagc taaaaatcag accactgcac tccagcctgg g 101

<210> SEQ ID NO 3

<211> LENGTH: 101

<212> TYPE: DNA

<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 3

tactaataaaa atgattaaag aagggtgtgtc tttaattgaa gcatgattta aagtaaattgc 60

aaagctaaaa atcagaccac tgcactccag cctggggaac a 101

<210> SEQ ID NO 4

<211> LENGTH: 101

<212> TYPE: DNA

<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 4

aaatgattaa agaagggtgtg tctttaattg aagcatgatt taaagtaa at gcaaagctaa 60

aaatcagacc actgcactcc agcctgggga acaagagtga a 101

<210> SEQ ID NO 5

<211> LENGTH: 101

<212> TYPE: DNA

<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 5

gtgtgtcttt aattgaagca tgatttaaag taaatgcaaa gctaaaaatc agaccactgc 60

actccagcct ggggaacaag agtgaaaccc catctcaaaa a 101

<210> SEQ ID NO 6

<211> LENGTH: 100

<212> TYPE: DNA

<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 6

-continued

```
gtgtctttaa ttgaagcatg atttaaagta aatgcaaagc taaaaatcag accactgcac    60
tccagcctgg ggaacaagag tgaaacccca tctcaaaaac                            100
```

```
<210> SEQ ID NO 7
<211> LENGTH: 100
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens
```

```
<400> SEQUENCE: 7
```

```
gtctttaatt gaagcatgat ttaaagtaaa tgcaaagcta aaaatcagac cactgcactc    60
cagcctgggg aacaagagtg aaaccccatc tcaaaaacaa                            100
```

```
<210> SEQ ID NO 8
<211> LENGTH: 92
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens
```

```
<400> SEQUENCE: 8
```

```
tgaagcatga tttaaagtaa atgcaaagct aaaaatcaga ccactgcact ccagcctggg    60
gaacaagagt gaaaccccat ctcaaaaaca aa                                    92
```

```
<210> SEQ ID NO 9
<211> LENGTH: 92
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens
```

```
<400> SEQUENCE: 9
```

```
atgatttaaa gtaaatgcaa agctaaaaat cagaccactg cactccagcc tggggaacaa    60
gagtgaaacc ccatctcaaa aacaaacaaa ca                                    92
```

```
<210> SEQ ID NO 10
<211> LENGTH: 92
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens
```

```
<400> SEQUENCE: 10
```

```
agtaaatgca aagctaaaaa tcagaccact gcactccagc ctggggaaca agagtgaaac    60
cccatctcaa aaacaaacaa acaaaacaaa ac                                    92
```

```
<210> SEQ ID NO 11
<211> LENGTH: 100
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens
```

```
<400> SEQUENCE: 11
```

```
atgcaaagct aaaaatcaga ccactgcact ccagcctggg gaacaagagt gaaaccccat    60
ctcaaaaaca aacaaacaaa acaaaacaaa aaaaactaag                            100
```

```
<210> SEQ ID NO 12
<211> LENGTH: 40
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens
```

```
<400> SEQUENCE: 12
```

```
atgcaaagct aaaaatcaga ccactgcact ccagcctggg                            40
```

```
<210> SEQ ID NO 13
```

-continued

<211> LENGTH: 101
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 13

tgtcagagta gtggtggttt ataagacggg agaaaatagc acctcacttc cagaaagctt 60
taagacaaaa ggtgagtact agagtaagat tcagtctcag a 101

<210> SEQ ID NO 14
<211> LENGTH: 101
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 14

gagtagtggt ggtttataag acgggagaaa atagcacctc acttcagaa agctttaaga 60
caaaagggtga gtactagagt aagattcagt ctcagatctg g 101

<210> SEQ ID NO 15
<211> LENGTH: 103
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 15

gtggtggttt ataagacggg agaaaatagc acctcacttc cagaaagctt taagacaaaa 60
ggtgagtact agagtaagat tcagtctcag atctgggtga cac 103

<210> SEQ ID NO 16
<211> LENGTH: 101
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 16

gtttataaga cgggagaaaa tagcacctca cttccagaaa gctttaagac aaaagggtgag 60
tactagagta agattcagtc tcagatctgg gtgacacaaa g 101

<210> SEQ ID NO 17
<211> LENGTH: 101
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 17

ataagacggg agaaaatagc acctcacttc cagaaagctt taagacaaaa ggtgagtact 60
agagtaagat tcagtctcag atctgggtga cacaaaggac c 101

<210> SEQ ID NO 18
<211> LENGTH: 101
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 18

agaaaatagc acctcacttc cagaaagctt taagacaaaa ggtgagtact agagtaagat 60
tcagtctcag atctgggtga cacaaaggac catggatttc t 101

<210> SEQ ID NO 19
<211> LENGTH: 101
<212> TYPE: DNA
<213> ORGANISM: Homo sapiens

-continued

<400> SEQUENCE: 19

aatagcacct cacttcacaga aagctttaag acaaaagggtg agtactagag taagattcag 60
tctcagatct ggggtgacaca aaggaccatg gatttctgca a 101

<210> SEQ ID NO 20

<211> LENGTH: 104

<212> TYPE: DNA

<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 20

acctcacttc cagaaagcct taagacaaaa ggtgagtact agagtaagat tcagtctcag 60
atctgggtga cacaaaggac catggatttc tgcaaccett ggtg 104

<210> SEQ ID NO 21

<211> LENGTH: 101

<212> TYPE: DNA

<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 21

cagaaagcct taagacaaaa ggtgagtact agagtaagat tcagtctcag atctgggtga 60
cacaaaggac catggatttc tgcaaccett ggtgccttct t 101

<210> SEQ ID NO 22

<211> LENGTH: 101

<212> TYPE: DNA

<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 22

aagacaaaag gtgagtacta gagtaagatt cagtctcaga tctgggtgac acaaaggacc 60
atggatttct gcaacccttg gtgccttctt tgggaaccca t 101

<210> SEQ ID NO 23

<211> LENGTH: 40

<212> TYPE: DNA

<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 23

aagacaaaag gtgagtacta gagtaagatt cagtctcaga 40

<210> SEQ ID NO 24

<211> LENGTH: 76

<212> TYPE: DNA

<213> ORGANISM: Homo sapiens

<400> SEQUENCE: 24

tacctaagca cacagagtaa tataccaaag cgacaggcat gatgaggaca cagtgagtga 60
gtgagctctg aaccag 76

<210> SEQ ID NO 25

<211> LENGTH: 20

<212> TYPE: DNA

<213> ORGANISM: Artificial sequence

<220> FEATURE:

<223> OTHER INFORMATION: synthetic polynucleotide

<400> SEQUENCE: 25

agacgggaga aaatagcacc 20

-continued

<210> SEQ ID NO 26
<211> LENGTH: 20
<212> TYPE: DNA
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic polynucleotide

<400> SEQUENCE: 26

accaagggtt gcagaaatcc 20

<210> SEQ ID NO 27
<211> LENGTH: 23
<212> TYPE: DNA
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic polynucleotide

<400> SEQUENCE: 27

catgtgtttg atatcttccc agc 23

<210> SEQ ID NO 28
<211> LENGTH: 22
<212> TYPE: DNA
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic polynucleotide

<400> SEQUENCE: 28

ctggctaacc gtcggtttat tg 22

<210> SEQ ID NO 29
<211> LENGTH: 24
<212> TYPE: DNA
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic polynucleotide

<400> SEQUENCE: 29

gagatggagt ttcactcttg ttgc 24

<210> SEQ ID NO 30
<211> LENGTH: 27
<212> TYPE: DNA
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic polynucleotide

<400> SEQUENCE: 30

gaacctttcc atcatactta gaaatac 27

<210> SEQ ID NO 31
<211> LENGTH: 17
<212> TYPE: DNA
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic polynucleotide

<400> SEQUENCE: 31

tccatggaag ccagaac 17

<210> SEQ ID NO 32
<211> LENGTH: 25
<212> TYPE: DNA

-continued

```

<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic polynucleotide

<400> SEQUENCE: 32

ccttaacaca gatggctctt gatgc                                     25

<210> SEQ ID NO 33
<211> LENGTH: 23
<212> TYPE: DNA
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic polynucleotide

<400> SEQUENCE: 33

tcctctttcc accttggett tcc                                     23

<210> SEQ ID NO 34
<211> LENGTH: 23
<212> TYPE: DNA
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic polynucleotide

<400> SEQUENCE: 34

ggttcagaac taccaataac aag                                     23

<210> SEQ ID NO 35
<211> LENGTH: 24
<212> TYPE: DNA
<213> ORGANISM: Artificial sequence
<220> FEATURE:
<223> OTHER INFORMATION: synthetic polynucleotide

<400> SEQUENCE: 35

acctgatgtg tgacctgatt gatg                                     24

```

What is claimed is:

1. A method for creating a library of recurrently mutated genomic regions comprising:

identifying a plurality of genomic regions from a group of genomic regions that are recurrently mutated in a specific cancer;

wherein the library comprises the plurality of genomic regions;

the plurality of genomic regions comprises at least 10 different genomic regions; and

at least one mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

2. The method of claim 1, wherein the plurality of genomic regions comprises at least 25, at least 50, at least 100, at least 150, at least 200, or at least 500 different genomic regions.

3. The method of claim 1, wherein at least two mutations within the plurality of genomic regions or at least three mutations within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

4. The method of claim 1, wherein at least one mutation within the plurality of genomic regions is present in at least 60%, 70%, 80%, 90%, 95%, 98%, 99%, or 99.9% of all subjects with the specific cancer.

5. The method of claim 1, wherein the identifying step comprises for each genomic region in the plurality of genomic regions, ranking the genomic region to maximize the number of all subjects with the specific cancer having at least one mutation within the genomic region.

6. The method of claim 1, wherein the identifying step comprises for each genomic region in the plurality of genomic regions, ranking the genomic region to maximize the ratio between the number of all subjects with the specific cancer having at least one mutation within the genomic region and the length of the genomic region.

7. The method of claim 1, wherein the library comprises a plurality of genomic regions encoding a plurality of driver sequences.

8. The method of claim 7, wherein the driver sequences are known driver sequences.

9. The method of claim 7, wherein the driver sequences are recurrently mutated in the specific cancer.

10. The method of claim 1, wherein the library comprises a plurality of genomic regions that are recurrently rearranged in the specific cancer.

11. The method of claim 1, wherein the specific cancer is a carcinoma.

12. The method of claim **11**, wherein the carcinoma is an adenocarcinoma, a non-small cell lung cancer, or a squamous cell carcinoma.

13. The method of claim **1**, wherein the cumulative length of the plurality of genomic regions is at most 30 Mb, 20 Mb, 10 Mb, 5 Mb, 2 Mb, 1 Mb, 500 kb, 200 kb, 100 kb, 50 kb, 20 kb, or 10 kb.

14. A method for analyzing a cancer-specific genetic alteration in a subject comprising the steps of:

obtaining a tumor nucleic acid sample and a genomic nucleic acid sample from a subject with a specific cancer;

sequencing a plurality of target regions in the tumor nucleic acid sample and in the genomic nucleic acid sample to obtain a plurality of tumor nucleic acid sequences and a plurality of genomic nucleic acid sequences; and

comparing the plurality of tumor nucleic acid sequences to the plurality of genomic nucleic acid sequences to identify a patient-specific genetic alteration in the tumor nucleic acid sample;

wherein the plurality of target regions are selected from a plurality of genomic regions that are recurrently mutated in the specific cancer;

the plurality of genomic regions comprises at least 10 different genomic regions; and

at least one mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

15. The method of claim **14**, wherein the plurality of genomic regions comprises at least 25, at least 50, at least 100, at least 150, at least 200, or at least 500 different genomic regions.

16. The method of claim **14**, wherein at least two mutations within the plurality of genomic regions or at least three mutations within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

17. The method of claim **14**, wherein at least one mutation within the plurality of genomic regions is present in at least 60%, 70%, 80%, 90%, 95%, 98%, 99%, or 99.9% of all subjects with the specific cancer.

18. The method of claim **14**, wherein each genomic region in the plurality of genomic regions is identified by ranking the genomic region to maximize the number of all subjects with the specific cancer having at least one mutation within the genomic region.

19. The method of claim **14**, wherein each genomic region in the plurality of genomic regions is identified by ranking the genomic region to maximize the ratio between the number of all subjects with the specific cancer having at least one mutation within the genomic region and the length of the genomic region.

20. The method of claim **14**, wherein the plurality of genomic regions comprises genomic regions encoding a plurality of driver sequences.

21. The method of claim **20**, wherein the driver sequences are known driver sequences.

22. The method of claim **20**, wherein the driver sequences are recurrently mutated in the specific cancer.

23. The method of claim **14**, wherein the plurality of genomic regions comprises genomic regions that are recurrently rearranged in the specific cancer.

24. The method of claim **14**, wherein the specific cancer is a carcinoma.

25. The method of claim **24**, wherein the carcinoma is an adenocarcinoma, a non-small cell lung cancer, or a squamous cell carcinoma.

26. The method of claim **14**, wherein the cumulative length of the plurality of genomic regions is at most 30 Mb, 20 Mb, 10 Mb, 5 Mb, 2 Mb, 1 Mb, 500 kb, 200 kb, 100 kb, 50 kb, 20 kb, or 10 kb.

27. The method of any one of claims **14-26**, further comprising the steps of:

obtaining a cell-free nucleic acid sample from the subject; and

identifying the patient-specific genetic alteration in the cell-free nucleic acid sample.

28. The method of claim **27**, wherein the step of identifying the patient-specific genetic alteration in the cell-free nucleic acid sample comprises sequencing a genomic region comprising the patient-specific genetic alteration in the cell-free sample.

29. The method of claim **27**, wherein the step of obtaining a tumor nucleic acid sample and a genomic nucleic acid sample comprises the step of enriching the plurality of target regions in the tumor nucleic acid sample and the genomic nucleic acid sample.

30. The method of claim **29**, wherein the enriching step comprises use of a custom library of biotinylated DNA.

31. The method of claim **27**, wherein the step of obtaining a cell-free nucleic acid sample comprises the step of enriching the plurality of target regions in the cell-free nucleic acid sample.

32. The method of claim **27**, further comprising the step of quantifying the cancer-specific genetic alteration in the cell-free sample.

33. A method for screening a cancer-specific genetic alteration in a subject comprising the steps of:

obtaining a cell-free nucleic acid sample from a subject;

sequencing a plurality of target regions in the cell-free sample to obtain a plurality of cell-free nucleic acid sequences; and

identifying a cancer-specific genetic alteration in the cell-free sample;

wherein the plurality of target regions are selected from a plurality of genomic regions that are recurrently mutated in the specific cancer;

the plurality of genomic regions comprises at least 10 different genomic regions; and

at least one mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

34. The method of claim **33**, wherein the plurality of genomic regions comprises at least 25, at least 50, at least 100, at least 150, at least 200, or at least 500 different genomic regions.

35. The method of claim **33**, wherein at least two mutations within the plurality of genomic regions or at least three mutations within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

36. The method of claim **33**, wherein at least one mutation within the plurality of genomic regions is present in at least 60%, 70%, 80%, 90%, 95%, 98%, 99%, or 99.9% of all subjects with the specific cancer.

37. The method of claim **33**, wherein each genomic region in the plurality of genomic regions is identified by ranking the

genomic region to maximize the number of all subjects with the specific cancer having at least one mutation within the genomic region.

38. The method of claim **33**, wherein each genomic region in the plurality of genomic regions is identified by ranking the genomic region to maximize the ratio between the number of all subjects with the specific cancer having at least one mutation within the genomic region and the length of the genomic region.

39. The method of claim **33**, wherein the plurality of genomic regions comprises genomic regions encoding a plurality of driver sequences.

40. The method of claim **39**, wherein the driver sequences are known driver sequences.

41. The method of claim **39**, wherein the driver sequences are recurrently mutated in the specific cancer.

42. The method of claim **33**, wherein the plurality of genomic regions comprises genomic regions that are recurrently rearranged in the specific cancer.

43. The method of claim **33**, wherein the specific cancer is a carcinoma.

44. The method of claim **43**, wherein the carcinoma is an adenocarcinoma, a non-small cell lung cancer, or a squamous cell carcinoma.

45. The method of claim **33**, wherein the cumulative length of the plurality of genomic regions is at most 30 Mb, 20 Mb, 10 Mb, 5 Mb, 2 Mb, 1 Mb, 500 kb, 200 kb, 100 kb, 50 kb, 20 kb, or 10 kb.

46. The method of claim **33**, wherein the step of obtaining a cell-free nucleic acid sample comprises the step of enriching the plurality of target regions in the cell-free nucleic acid sample.

47. The method of claim **46**, wherein the enriching step comprises use of a custom library of biotinylated DNA.

* * * * *