Enrich for informative DNA templates.

Sequence DNA templates.

Compare sequences to a reference set to identify DNA sequence polymorphisms.

Example 4. Sorghum Polymorphism Discovery

The technical approaches described above for targeted sequencing and polymorphism discovery adjacent to restriction sites in large genomes was validated on the sorghum genome (800 Mbp). The number of sites across the sorghum genome analyzed for SNP discovery was varied in several ways, as follows: (1) using restriction enzymes with 8- or 6-base recognition sites (FseI, KasI respectively were tested), (2) using restriction enzymes sensitive to DNA methylation (FseI) and a restriction enzyme that is not sensitive to DNA methylation (SphI), and (3) using a 4 bp restriction enzyme or shearing to generate the second end of each amplicon. From this study it was concluded that the use of methylation sensitive restriction enzymes significantly reduced the number of repeat sequences obtained increasing data yield/sequence. It was also found that the use of one targeting enzyme plus DNA shearing (vs. two restriction enzymes) generated a larger number of more uniform size templates increasing data yield/sequence.

FseI/MseI RSL-templates were prepared from BTx623 and IS3620C, and approximately 250,000 RSL-sequences were obtained from each genotype. Among the approximately 11,000 different sequences acquired in the experiment (excluding error containing sequences), approximately 5,000 templates containing unique sequences (found only once per genome) were sequenced 5X or more times from each genotype. Comparison of the sequences from the two genotypes identified 200-400 SNPs/InDels within 27 bp of the FseI restriction site.

In a follow-up experiment, RSL-templates were prepared by digestion of BTx623 and IS3620C genomic DNA with FseI followed by ligation of adapter A. The resulting RFs were sheared and adapter B was ligated to create DNA template. After PCR using primers complementary to sequences in adapter A and adapter B, and enrichment of template of an optimal size, the templates were sequenced using priming sites in adapter A on an ILLUMINA® SGAII™ sequencer. In this experiment, approximately 13,000 different unique 27bp sequences were obtained from both genotypes revealing approximately 1,500 polymorphic sequences.

IDT Exhibit 1020.1562

DNA templates generated using KasI/shearing allowed sequence analysis of approximately 50,000 different unique RSL-tags at 5X or greater depth through acquisition of about 3,000,000 sequences. It is estimated that this collection of sequences from IS3620C and BTx623 will reveal more than 5,000 SNPs/INDELs adjacent to FseI-sites when the data is fully analyzed.

When the CspCl restriction enzyme is used to digest a random 1,000 Mbp genome sequence, it is predicted that this size genome will contain about 62,500 sites for CspCl and generate 125,000 DNA templates. The CspCl enzyme is not methylation sensitive, so nearly all sites would be available for digestion. The resulting small (approximately 34–38bp) DNA fragments may be purified by size selection on agarose gels, blunt ended, and ligated to adapters A and B (one of which containing the sequencing primer binding site plus an indexing sequence). PCR amplification will differentially amplify RSL-tags flanked by two different adapters (due to suppression PCR) and these may be loaded directly onto the sequencer or further purified prior to sequencing as necessary. Because primers are ligated to the termini at random, both strands of each DNA template will be sequenced, thereby eliminating or reducing the extent of the increased sequencing error rates that tend to occur towards the 3'-end of each ILLUMINA® sequencing run. If CspCI digestion cuts a random sequence 1,000 Mbp genome at 62,500 sites creating 125,000 RSL-templates, and 50% of the 33 bp sequences derived from the resulting RSL-templates are unique, and if there is one polymorphism per 1,000 bp in a comparison of two genotypes, then analysis of this set of RSL-templates by sequencing would reveal approximately 2,062 polymorphic sequences. The subset of RSL-tags corresponding to unique sequences spanning polymorphic sites may then be mapped and used for genotyping, DNA fingerprinting, or haplotype analysis.

Example 5. Identification of RSL polymorphism and design of primers for amplification of the polymorphic region in sorghum.

Specific 30 bp RSL sequences from different genotypes are compared to identify single nucleotide polymorphisms (SNPs) or insertion/deletions (InDels). In this example, illustrated in FIG. 6, a 4 bp InDel polymorphism between BTx623 and IS3620c is identified. A reverse primer is designed downstream of the polymorphism. This primer will be used with an FseI adapter-specific primer to produce a PCR product for sequencing. A) denotes the sequence identifier: (Coordinates on Sorghum

IDT Exhibit 1020.1563

pseudomolecule); (Genotype_Sequence ID_Number of sequences in the sequence contig). B) is the alignment and comparison of BTx623 and IS3620c sequence contigs. The InDel is bolded and underlined. The first three bases in a sequence correspond to the Index Sequence for the particular genotype (BTx623=ATC; IS3620c=GTC). The next six bases (CCGGCC) correspond to the FseI half-site. The Index Sequences are contained within the RSL adapter ligated to the FseI site. In C), the results of a BLAST similarity search is shown, comparing the BTx623 sequence with the assembled Phytozome sorghum pseudomolecule database (available on the world wide web at phytozome.net). A 227bp region, containing the 27 bp ILLUMINA® sequence and 200 bp downstream is identified and downloaded. Using PrimerQuest software (available on the world wide web at idtdna.com), an optimal reverse PCR primer is designed within the 227 bp sequence that will produce a PCR product of approximately 75–200 bp, when used with a forward primer specific for the FseI adapter. In D) the adapter-modified DNA fragment, and representations of the forward and reverse FseI adapter-specific primer are shown. In C) and D), the FseI half site is underlined. The SNP or InDel is bolded and underlined, and the reverse oligonucleotide primer-binding site italicized and underlined.

Example 6. Identification of RSL polymorphism and design of primer for amplification of the polymorphic region in sorghum.

As illustrated in FIG. 7, another 3 bp InDel polymorphism between BTx623 and IS3620c is identified. A) denotes the sequence identifier. B) shows the alignment and comparison of the BTx623 and IS3620c contigs for this sequence. A reverse primer (SEQ ID NO: 14) is designed downstream of the polymorphism, using the genomic sequence shown in C). This primer will be used with an FseI adapter-specific primer to produce a PCR product for sequencing, as described above in Example 5. In C) and D), the FseI half site is underlined. The InDel is bolded and underlined, and the reverse oligonucleotide primer binding site is italicized and underlined.

Example 7. Identification of RSL polymorphism and design of primer for amplification of the polymorphic region in sorghum.

As illustrated in FIG. 8, two SNP polymorphisms between BTx623 and IS3620c are identified. A) denotes the sequence identifier. B) shows the alignment

IDT Exhibit 1020.1564

and comparison of the BTx623 and IS3620c contigs for this sequence. A reverse primer (SEQ ID NO: 18) is designed downstream of the polymorphism, using the genomic sequence shown in C). This primer will be used with an FseI adapter-specific primer to produce a PCR product for sequencing, as described above in Example 5. In C), the FseI half site is underlined. The SNPs are bolded and underlined, and the reverse oligonucleotide primer binding site is italicized and underlined. The adapter-modified DNA fragment, and the forward and reverse FseI adapter-specific primers are similar to those shown in D) of FIG. 6, except that the 30 bp informative DNA sequence is derived from C) of the present example.

Example 8. Identification of RSL polymorphism and design of primer for amplification of the polymorphic region in sorghum.

As illustrated in FIG. 9, another 4 bp InDel polymorphism between BTx623 and IS3620c is identified. A) denotes the sequence identifier. B) shows the alignment and comparison of the BTx623 and IS3620c contigs for this sequence. A reverse primer (SEQ ID NO: 22) is designed downstream of the polymorphism, using the genomic sequence shown in C). This primer will be used with an FseI adapter-specific primer to produce a PCR product for sequencing, as described above in Example 5. In C), the FseI half site is underlined. The InDel is bolded and underlined, and the reverse oligonucleotide primer binding site italicized and underlined. The adapter-modified DNA fragment, and the forward and reverse FseI adapter-specific primers are similar to those shown in D) of FIG. 6, except that the 30 bp informative DNA sequence is derived from C) of the present example.

Example 9. Identification of RSL polymorphism and design of primer for amplification of the polymorphic region in sorghum.

As illustrated in FIG. 10, two SNP polymorphisms between BTx623 and IS3620c are identified. A) denotes the sequence identifier. B) shows the alignment and comparison of the BTx623 and IS3620c contigs for this sequence. A reverse primer (SEQ ID NO: 26) is designed downstream of the polymorphism, using the genomic sequence shown in C). This primer will be used with an FseI adapter-specific primer to produce a PCR product for sequencing, as described above in Example 5. In C), the FseI half site is underlined. The SNPs are bolded and underlined, and the reverse oligonucleotide primer binding site is italicized and

IDT Exhibit 1020.1565

underlined. The adapter-modified DNA fragment, and the forward and reverse FseI adapter-specific primers are similar to those shown in D) of FIG. 6, except that the 30 bp informative DNA sequence is derived from C) of the present example.

Example 10. Identification of RSL polymorphism and design of primer for amplification of the polymorphic region in sorghum.

As illustrated in FIG. 11, two SNP polymorphisms between BTx623 and IS3620c are identified. A) denotes the sequence identifier. B) shows the alignment and comparison of the BTx623 and IS3620c contigs for this sequence. A reverse primer (SEQ ID NO: 30) is designed downstream of the polymorphism, using the genomic sequence shown in C). This primer will be used with an FseI adapter-specific primer to produce a PCR product for sequencing, as described above in Example 5. In C), the FseI half site is underlined. The SNPs are bolded and underlined, and the reverse oligonucleotide primer binding site is italicized and underlined. The adapter-modified DNA fragment, and the forward and reverse FseI adapter-specific primers are similar to those shown in D) of FIG. 6, except that the 30 bp informative DNA sequence is derived from C) of the present example.

Example 11. Identification of RSL polymorphism and design of primer for amplification of the polymorphic region in sorghum.

As illustrated in FIG. 12, three SNP polymorphisms between BTx623 and IS3620c are identified. A) denotes the sequence identifier. B) shows the alignment and comparison of the BTx623 and IS3620c contigs for this sequence. A reverse primer (SEQ ID NO: 34) is designed downstream of the polymorphism, using the genomic sequence shown in C). This primer will be used with an FseI adapter-specific primer to produce a PCR product for sequencing, as described above in Example 5. In C), the FseI half site is underlined. The SNPs are bolded and underlined, and the reverse oligonucleotide primer binding site is italicized and underlined. The adapter-modified DNA fragment, and the forward and reverse FseI adapter-specific primers are similar to those shown in D) of FIG. 6, except that the 30 bp informative DNA sequence is derived from C) of the present example.

IDT Exhibit 1020.1566

Example 12. Identification of RSL polymorphism and design of primer for amplification of the polymorphic region in sorghum.

As illustrated in FIG. 13, a three SNP polymorphism between BTx623 and IS3620c is identified. A) denotes the sequence identifier. B) shows the alignment and comparison of the BTx623 and IS3620c contigs for this sequence. A reverse primer (SEQ ID NO: 38) is designed downstream of the polymorphism, using the genomic sequence shown in C). This primer will be used with an FseI adapter-specific primer to produce a PCR product for sequencing, as described above in Example 5. In C), the FseI half site is underlined. The SNPs are bolded and underlined, and the reverse oligonucleotide primer binding site is italicized and underlined. The adapter-modified DNA fragment, and the forward and reverse FseI adapter-specific primers are similar to those shown in D) of FIG. 6, except that the 30 bp informative DNA sequence is derived from C) of the present example.

Example 13. Identification of RSL polymorphism and design of primer for amplification of the polymorphic region in sorghum.

As illustrated in FIG. 14, a 1bp InDel and one SNP polymorphism between BTx623 and IS3620c are identified. A) denotes the sequence identifier. B) shows the alignment and comparison of the BTx623 and IS3620c contigs for this sequence. A reverse primer (SEQ ID NO: 42) is designed downstream of the polymorphism, using the genomic sequence shown in C). This primer will be used with an FseI adapter-specific primer to produce a PCR product for sequencing, as described above in Example 5. In C), the FseI half site is underlined. The SNP and InDel are bolded and underlined, and the reverse oligonucleotide primer binding site is italicized and underlined. The adapter-modified DNA fragment, and the forward and reverse FseI adapter-specific primers are similar to those shown in D) of FIG. 6, except that the 30 bp informative DNA sequence is derived from C) of the present example.

Example 14. Identification of RSL polymorphism and design of primer for amplification of the polymorphic region in sorghum.

As illustrated in FIG. 15, three SNPs and a 1bp InDel polymorphism between BTx623 and IS3620c are identified. A) denotes the sequence identifier. B) shows the alignment and comparison of the BTx623 and IS3620c contigs for this sequence. A reverse primer (SEQ ID NO: 46) is designed downstream of the polymorphism, using

IDT Exhibit 1020.1567

the genomic sequence shown in C). This primer will be used with an FseI adapter-specific primer to produce a PCR product for sequencing, as described above in Example 5. In C), the FseI half site is underlined. The SNPs and InDel are bolded and underlined, and the reverse oligonucleotide primer binding site is italicized and underlined. The adapter-modified DNA fragment, and the forward and reverse FseI adapter-specific primers are similar to those shown in D) of Fig. 6, except that the 30 bp informative DNA sequence is derived from C) of the present example.

Example 15. Rice Polymorphism Discovery

DNA template generation using combinations of restriction enzymes and adapter ligation, as described in Example 1, above, has been tested *in silico* and in the laboratory based on the rice genome sequence containing about 400,000,000 bp. DNA templates generated by the described technique were sequenced by using the Roche 454 Genome Sequencer 20 System. The approximately 250,000 template sequences generated per run from several sequencing runs were analyzed and compared to results predicted *in silico*. The results confirmed the feasibility of using restriction enzymes/adapter ligation for the reproducible generation of DNA templates for high throughput targeted DNA sequencing and resequencing using the Roche 454 sequencer.

Example 16. High Throughput Multiplex Sequencing of Informative DNA Templates.

In one prospective haplotype scale application, 10,000 informative RSL-templates are targeted for selective amplification, capture and genotyping analysis across 100 accessions of a species germplasm. A 10X depth of sequence analysis of the amplified, enriched informative DNA templates will require the acquisition of 10,000,000 sequences on the ILLUMINA® sequencer. The ILLUMINA® SGAII™ is capable of sequencing approximately 50 million templates per run or approximately 6.25 million per channel. The required 10,000,000 sequences may be distributed across several channels of the ILLUMINA® sequencer with sequencing done in parallel with other samples utilizing unique indexing sequences to assign the sequences to their accession of origin. Genomic DNA from 100 different accessions may be digested with KasI, CspCl, or any other suitable targeting restriction enzyme, followed by ligation of adapters as described above (and illustrated in box 2 of FIG. **2**). The resulting indexed DNA templates are then pooled prior to amplification and enrichment of the informative DNA templates.

IDT Exhibit 1020.1568

A potential advantage of the proposed approach, in addition to its lower cost, is the procedural flexibility and low barrier to entry. Many individual investigators will be able to obtain genotyping information at various depths depending on the requirements of the selected application. This approach to DNA template production employing oligonucleotide-bead- or microarray-based capture and enrichment of informative templates, or targeted PCR amplification of informative templates, and high throughput sequencing is potentially applicable to any genome of any size, providing great flexibility in design. It is especially suitable for applications in which higher genotyping resolution is required (*e.g.*, for haplotyping).

Without further elaboration, it is believed that one skilled in the art can, using the description herein, utilize the present invention to its fullest extent. The embodiments described herein are to be construed as illustrative and not as constraining the remainder of the disclosure in any way whatsoever. While the preferred embodiments of the invention have been shown and described, many variations and modifications thereof can be made by one skilled in the art without departing from the spirit and teachings of the invention. For example, the order of execution of the steps of a disclosed process may be varied, in some embodiments, from the order given. Accordingly, the scope of protection is not limited by the description set out above, but is only limited by the claims, including all equivalents of the subject matter of the claims. The disclosures of all patents, patent applications and publications cited herein are hereby incorporated herein by reference, to the extent that they provide procedural or other details consistent with and supplementary to those set forth herein.

* * *

All of the methods disclosed and claimed herein can be made and executed without undue experimentation in light of the present disclosure. While the compositions and methods of this invention have been described in terms of preferred embodiments, it will be apparent to those of skill in the art that variations may be applied to the methods and in the steps or in the sequence of steps of the method described herein without departing from the concept, spirit and scope of the invention. More specifically, it will be apparent that certain agents which are both chemically and physiologically related may be substituted for the agents described herein while the same or similar results would be achieved. All such similar substitutes and modifications apparent to those skilled in the art are deemed to be within the spirit, scope and concept of the invention as defined by the appended claims.

IDT Exhibit 1020.1569

CLAIMS

Claim 1. A method of generating informative DNA templates comprising:

(a) obtaining fragmented genomic DNA sequences from a first individual comprising a plurality of polymorphic sequences;

(b) ligating at least a first adapter to said fragmented genomic DNA sequences, to provide a plurality of adapter-modified informative DNA templates and adapter-modified non-informative DNA templates, wherein a subset of said informative DNA templates comprises a polymorphic sequence; and

(c) selecting adapter-modified informative DNA templates to obtain an enriched mixture of adapter-modified informative DNA templates.

Claim 2. The method of claim 1, wherein selecting adapter-modified informative DNA templates of step (c) comprises hybridization-based selection of said adapter-modified informative DNA templates.

Claim 3. The method of claim 1, wherein selecting adapter-modified informative DNA templates of step (c) comprises targeted PCR amplification of said adapter-modified informative DNA templates.

Claim 4. The method of claim 2, wherein said hybridization-based selection comprises hybridizing said adapter-modified informative DNA templates to solid matrix-attached oligonucleotides to form hybridized complexes, wherein said hybridized complexes exclude said non-informative DNA templates.

Claim 5. The method of claim 4, wherein said solid matrix comprises a plurality of beads.

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

(d) removing at least a portion of the non-hybridized non-informative DNA templates from said hybridized complexes; and

(e) releasing the informative DNA templates from the hybridized complexes, to obtain an enriched mixture of adapter-modified informative DNA templates.

IDT Exhibit 1020.1570

Claim 7. The method of claim 4, wherein forming said hybridized complexes in step (c) comprises hybridizing oligonucleotides to complementary sequences in the informative DNA templates.

Claim 8. The method of claim 7, wherein each said oligonucleotide is about 17–60 nucleotides in length.

Claim 9. The method of claim 7, wherein each said oligonucleotide is complementary to a known informative DNA template sequence corresponding to a polymorphic site in the genome.

Claim 10. The method of claim 6, wherein said oligonucleotides comprise a modification to allow oligonucleotide:informative template hybrids to be separated from non-hybridized template.

Claim 11. The method of claim 10, wherein said modification comprises incorporation of a terminal biotin.

Claim 12. The method of claim 3, wherein said targeted PCR amplification of informative templates comprises using a first primer complementary to said first adapter, and a plurality of second primers complementary to unique sequences in said informative DNA templates, wherein each of said plurality of second primers is designed such that the resulting amplified DNA templates comprise informative DNA sequences.

Claim 13. The method of claim 1, wherein said ligating of step (b) comprises ligating a second adapter to a terminus of each said fragmented genomic DNA sequences opposite said first adapter, to provide said plurality of adapter-modified informative DNA templates and adapter-modified non-informative DNA templates.

Claim 14. The method of claim 13, further comprising selective amplification of said DNA templates containing said first adapter and second adapter.

IDT Exhibit 1020.1571

Claim 15. The method of claim 1, further comprising selecting adapter-modified informative DNA templates based on the length of the DNA templates.

Claim 16. The method of claim 15, wherein adapter-modified informative DNA templates of about 50 to 200 base pairs are selected.

Claim 17. The method of claim 15, wherein adapter-modified informative DNA templates of about 400 to 600 base pairs are selected.

Claim 18. The method of claim 1, further comprising:

(d) subjecting the enriched mixture of adapter-modified informative DNA templates to DNA sequencing to obtain the sequences of said informative DNA templates.

Claim 19. The method of claim 18, wherein said DNA sequencing is high-throughput sequencing on an ILLUMINA® SGAII™, ABI DNA SOLID™ System or Roche 454 GS–FLX™ Genome Analysis System.

Claim 20. The method of claim 18, further comprising:

(e) comparing the sequences of the informative DNA templates to at least one set of reference genomic DNA sequences.

Claim 21. The method of claim 20, wherein at least one set of reference genomic DNA sequences is obtained from at least one other individual.

Claim 22. The method of claim 1, wherein at least said first adapter comprises an indexing sequence that can be correlated to said first individual.

Claim 23. The method of claim 1, further comprising, prior to (c):

(b_1) preparing adapter-modified DNA templates from a plurality individuals wherein each said adapter-modified DNA template from each respective individual contains a unique indexing sequence; and

IDT Exhibit 1020.1572

(b_2) combining said adapter-modified DNA templates obtained from all said individuals, to obtain a pool of adapter-modified DNA templates.

Claim 24. The method of claim 1, wherein obtaining fragmented genomic DNA sequences of step (a) comprises subjecting a genomic DNA sequence to digestion by at least a first restriction enzyme or to mechanical shearing.

Claim 25. The method of claim 24, wherein obtaining fragmented genomic DNA sequences of step (a) comprises digesting a genomic DNA sequence with at least a first restriction enzyme, and said ligating comprises ligating said first adapter and a second adapter to said fragmented genomic DNA sequences, to provide said adapter-modified DNA templates, wherein said first and second adapters are attached to the restriction termini of each said adapter-modified DNA fragments.

Claim 26. The method of claim 24, wherein, said first restriction enzyme is sensitive to DNA methylation.

Claim 27. The method of claim 26, further comprising separately digesting a sample of said genomic DNA from said first individual with a second restriction enzyme that is not sensitive to DNA methylation.

Claim 28. The method of claim 1, wherein each said plurality of polymorphic sequences comprise a polymorphism selected from the group consisting of a single nucleotide polymorphism, a simple sequence repeat polymorphism, a sequence inversion polymorphism, a nucleotide insertion polymorphism and a nucleotide deletion polymorphism.

Claim 29. A reagent for selecting informative DNA templates, comprising a plurality of oligonucleotides attached to a solid matrix, wherein each of said oligonucleotides is in the range of about 17-60 nucleotides in length and is complementary to a genomic sequence located 600 nucleotides or less from a polymorphic sequence.

IDT Exhibit 1020.1573

Claim 30. The reagent of claim 29, wherein each of said oligonucleotides is complementary to a genomic sequence located 50 to 200 or 400 to 600 nucleotides from a polymorphic sequence.

Claim 31. The reagent of claim 29, wherein said solid matrix comprises a plurality of beads.

Claim 32. A hybrid DNA complex comprising:

the reagent of claim 29; and

a plurality of adapter-modified informative DNA templates hybridized to said matrix-attached oligonucleotides, wherein each said informative DNA template comprises at least one polymorphic sequence.

Claim 33. The hybrid DNA complex of claim 32, wherein said adapter-modified informative DNA templates are from a single individual.

Claim 34. The hybrid DNA complex of claim 32, wherein said adapter-modified informative DNA templates are from a plurality of individuals, and said adapter includes a unique indexing sequence for matching each said adapter-modified informative DNA templates to a specific individual.

Claim 35. A method of marker-assisted selection comprising:

(a) obtaining fragmented genomic DNA sequences from a plurality individual plants or plant cells, to provide a plurality of fragmented genomic DNA sequences comprising a plurality of polymorphic sequences at least one of which is linked to a trait of interest;

(b) ligating at least a first adapter to said fragmented DNA sequences, to provide a plurality of adapter-modified informative DNA templates and adapter-modified non-informative DNA templates, wherein each of said informative DNA templates comprises a polymorphic sequence and wherein said first adapter comprises an index sequence that can be correlated to genomic DNA of an individual;

(c) selecting adapter-modified informative DNA templates to obtain an enriched mixture of adapter-modified informative DNA templates;

IDT Exhibit 1020.1574

(d) subjecting the enriched mixture of adapter-modified informative DNA templates to DNA sequencing to obtain the sequences of said informative DNA templates; and

(e) selecting an individual plant or plant cell based on the presence of at least one polymorphism in an informative DNA template wherein the polymorphism is linked to a trait of interest.

Claim 36. The method of claim 35, wherein selecting adapter-modified informative DNA templates of step (c) comprises hybridization-based selection of said adapter-modified informative DNA templates.

Claim 37. The method of claim 35, wherein selecting adapter-modified informative DNA templates of step (c) comprises targeted PCR amplification of said adapter-modified informative DNA templates.

Claim 38. The method of claim 35, wherein selecting an individual plant or plant cell comprises selecting a plant cell for regeneration of a plant.

Claim 39. The method of claim 35, wherein the plant or plant cell is a wheat, maize, rye, rice, oat, barley, turfgrass, sorghum, millet, sugarcane, tobacco, tomato, potato, soybean, cotton, canola, sunflower and alfalfa plant or plant cell.

Claim 40. The method of claim 35, wherein the trait is a trait of agronomic interest.

Claim 41. The method of claim 40, wherein the trait of agronomic interest is selected from the group consisting of drought tolerance, enhanced yield, cold tolerance, pest resistance, insect resistance, salt tolerance and herbicide tolerance.

IDT Exhibit 1020.1575

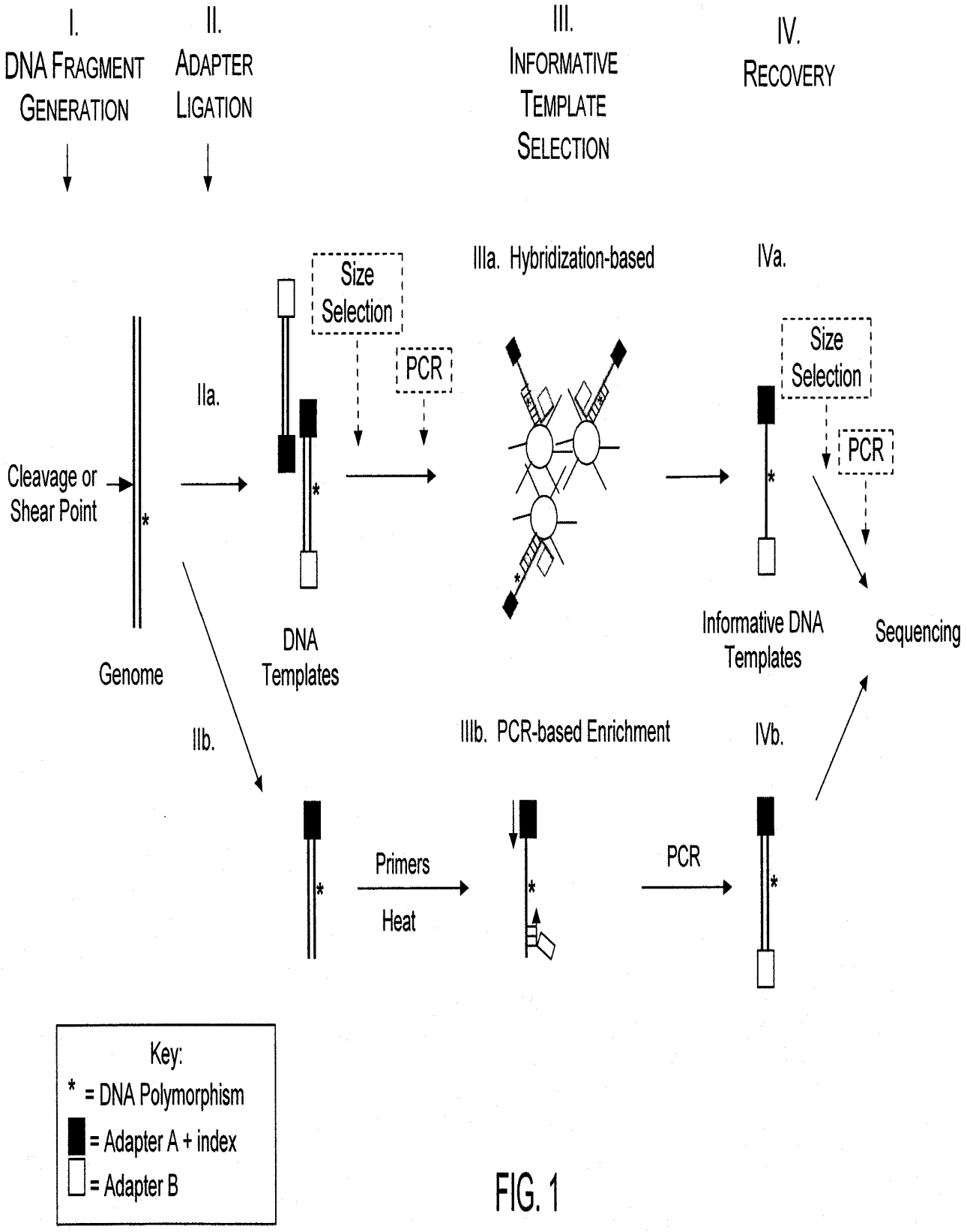

I.
DNA Fragment Generation
II.
Adapter Ligation
III.
Informative Template Selection
IV.
Recovery
IIIa. Hybridization-based
IVa.
Size Selection
PCR
Size Selection
PCR
IIa.
Cleavage or Shear Point
Genome
DNA Templates
Informative DNA Templates
Sequencing
IIb.
IIIb. PCR-based Enrichment
IVb.
Primers
Heat
PCR
Key:
* = DNA Polymorphism
= Adapter A + index
= Adapter B

FIG. 1

1. A DNA fragment is generated with an arbitrary restriction site (YYYYYYY) and by shearing, per Stage I of Fig. 1. The sheared end is end-repaired and a dATP (A) is added at the 3' end to facilitate T/A-ligation with the adapter

5′- **YY**NNNNNNNN...NNNNNNNN**A** -3′
3′- **YYYY**NNNNNNNN...NNNNNNNN -5′

2. After ligation of Adapters A and B to the ends of the DNA fragment (per Stage IIa or IIb of Fig. 1), the resulting modified DNA template is represented as follows:

5'-ACACTCTTTCCCTACACGACGCTCTTCCGATC**TXXXYYY**yyNNNNNNNNN...NNNNNNN*AGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG*
3'-TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAG**A**xxxyyyyyNNNNNNNNN...NNNNNNN*TCTAGCCTTCTCGAGCATACGGCAGAAGACGAAC*

From this mixture of adapter-modified templates, the informative templates are selected, per Stage III of Fig. 1.

3. After the final PCR with Illumina® A and B primers (Stage IV), the modified DNA template is represented as follows:

5'-*AATGATACGGCGACCACCGAGATCTA*CACTCTTTCCCTACACGACGCTCTTCCGATC**T**XXXYYYyyNNNNNNNNN...NNNNNNN*AGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG*
3'-*TTACTATGCCGCTGGTGGCTCTAGAT*GTGAGAAAGGGATGTGCTGCGAGAAGGCTAG**A**xxxyyyyyNNNNNNNNN...NNNNNNN*TCTAGCCTTCTCGAGCATACGGCAGAAGACGAAC*

The amplified templates are then ready to be applied to the flow-cell, and clusters generated, in the Illumina® cluster station.

4. After cluster generation and post-processing, the flow cell is loaded on the SGAII™ sequencer. Sequencing is carried out on the template using the Illumina® Genomic DNA sequencing primer. Sequencing of each DNA template initiates at the index sequence (xxx).

ACACTCTTTCCCTACACGACGCTCTTCCGATC**T** → → → →
3'-*TTACTATGCCGCTGGTGGCTCTAGAT*GTGAGAAAGGGATGTGCTGCGAGAAGGCTAG**A**xxxyyyyyNNNNNNNNN...NNNNNNN*TCTAGCCTTCTCGAGCATACGGCAGAAGACGAAC*

FIG. 2

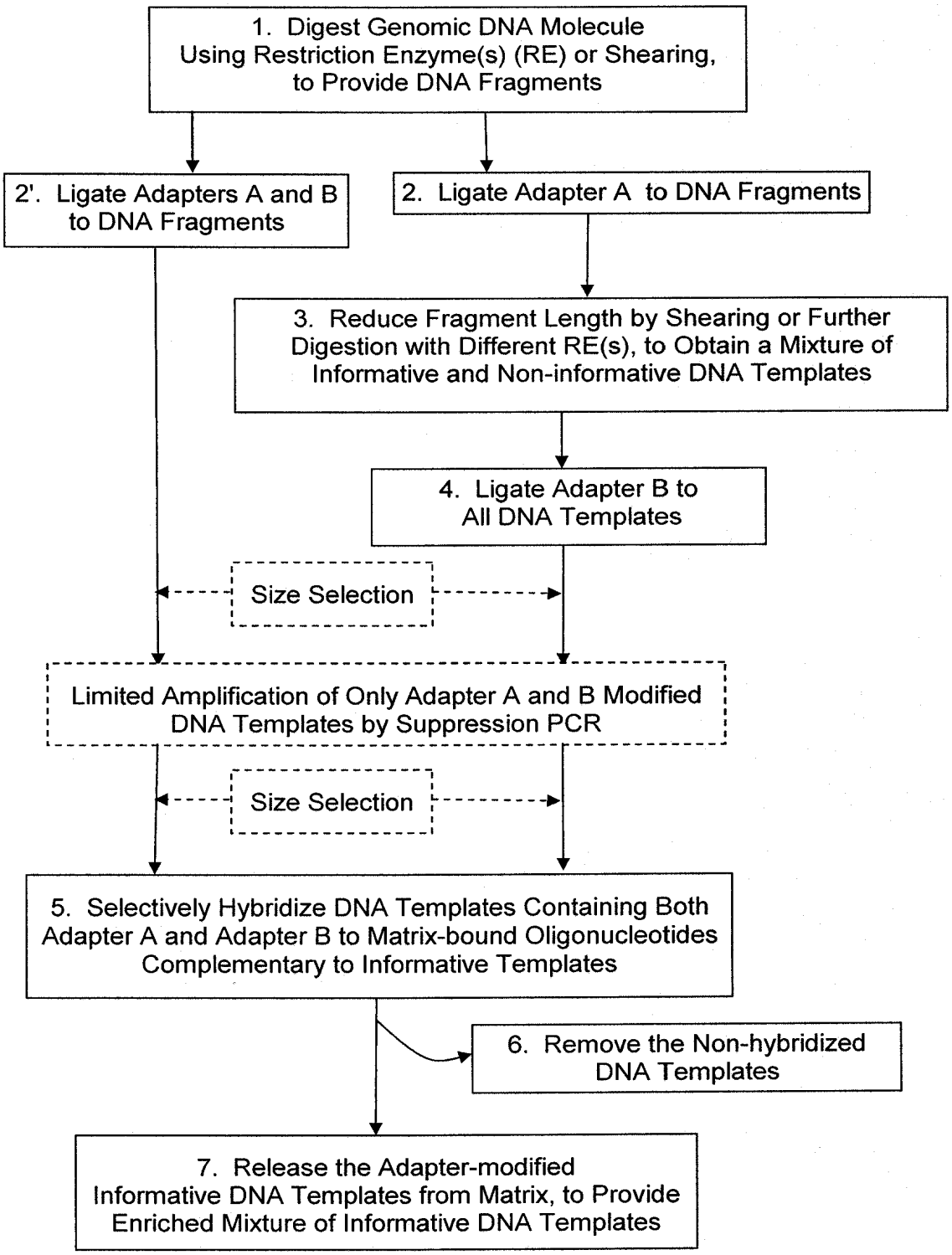

1. Digest Genomic DNA Molecule Using Restriction Enzyme(s) (RE) or Shearing, to Provide DNA Fragments
2'. Ligate Adapters A and B to DNA Fragments
2. Ligate Adapter A to DNA Fragments
3. Reduce Fragment Length by Shearing or Further Digestion with Different RE(s), to Obtain a Mixture of Informative and Non-informative DNA Templates
4. Ligate Adapter B to All DNA Templates
Size Selection
Limited Amplification of Only Adapter A and B Modified DNA Templates by Suppression PCR
Size Selection
5. Selectively Hybridize DNA Templates Containing Both Adapter A and Adapter B to Matrix-bound Oligonucleotides Complementary to Informative Templates
6. Remove the Non-hybridized DNA Templates
7. Release the Adapter-modified Informative DNA Templates from Matrix, to Provide Enriched Mixture of Informative DNA Templates

FIG. 3

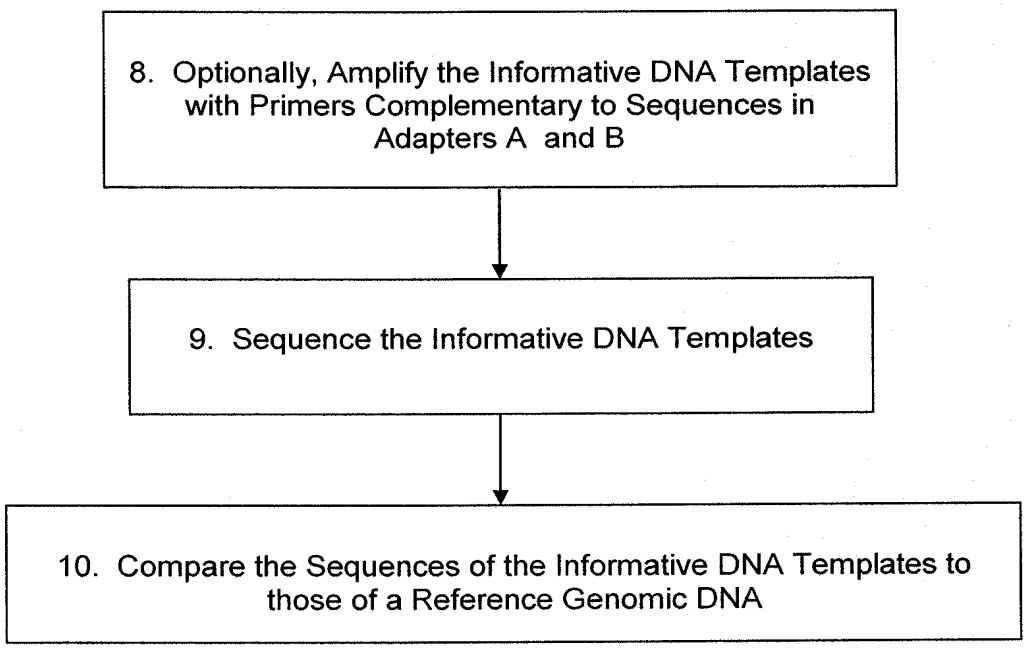

8. Optionally, Amplify the Informative DNA Templates with Primers Complementary to Sequences in Adapters A and B
9. Sequence the Informative DNA Templates
10. Compare the Sequences of the Informative DNA Templates to those of a Reference Genomic DNA

FIG. 4

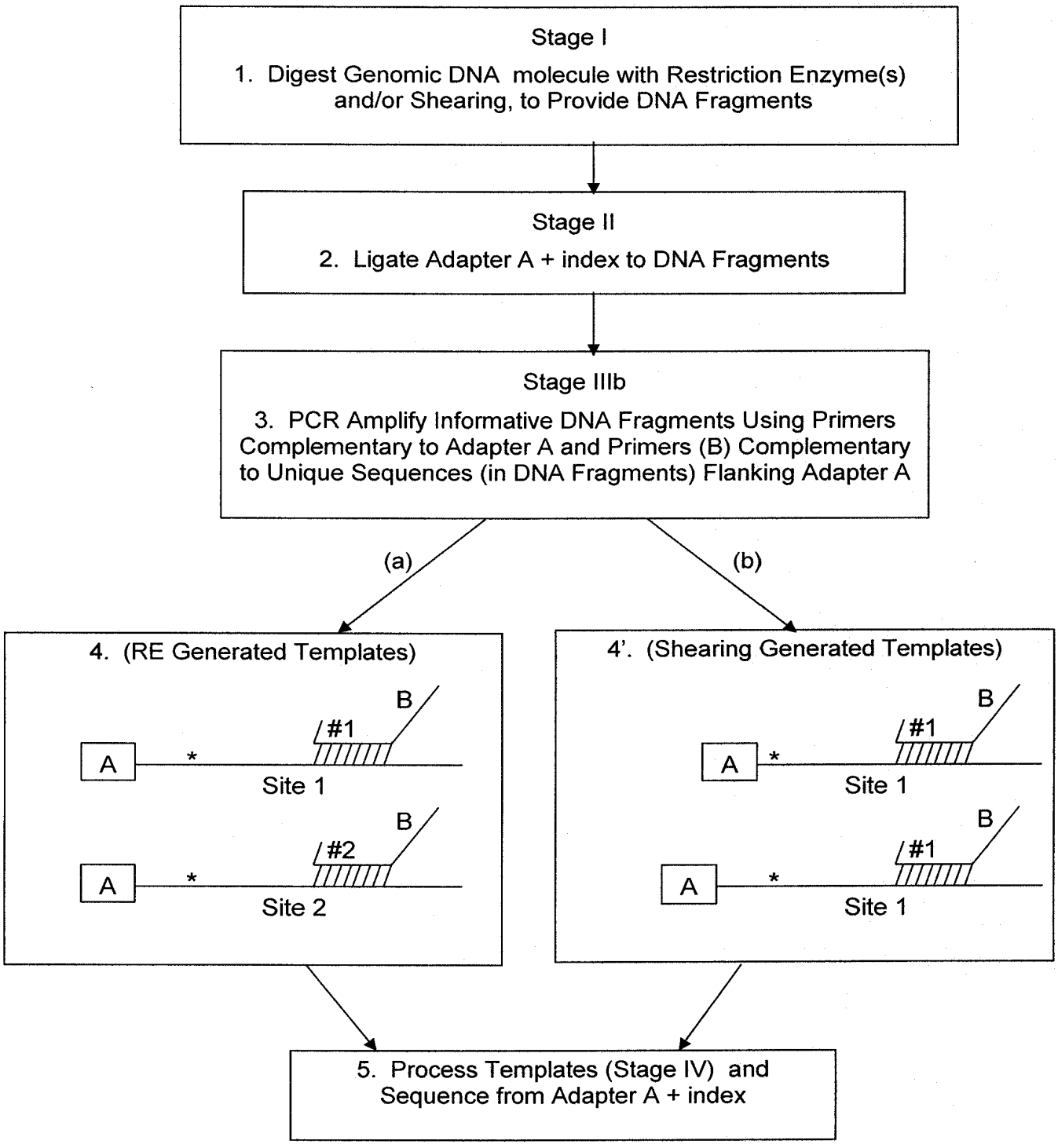

Stage I
1. Digest Genomic DNA molecule with Restriction Enzyme(s) and/or Shearing, to Provide DNA Fragments
Stage II
2. Ligate Adapter A + index to DNA Fragments
Stage IIIb
3. PCR Amplify Informative DNA Fragments Using Primers Complementary to Adapter A and Primers (B) Complementary to Unique Sequences (in DNA Fragments) Flanking Adapter A
(a)
(b)
4. (RE Generated Templates)
A
*
B
#1
Site 1
A
*
B
#2
Site 2
4'. (Shearing Generated Templates)
A
*
B
#1
Site 1
A
*
B
#1
Site 1
5. Process Templates (Stage IV) and Sequence from Adapter A + index

FIG. 5

A) 5948239-5948265 - (2) BTx_13830_40 IS3620c_12408_37

B) BTx_13830_40 - ATCCCGGCCGGCCCTTTCT**CTCT**GTCCTAG (SEQ ID NO: 7)

||||

IS3620c_12408_37 - GTCCCGGCCGGCCCTTTCT----GTCCTAGGTTA (SEQ ID NO: 8)

4 bp indel at bp 20

C) >chromosome_1 chromosome_1:5948239,5948465 (SEQ ID NO: 9)

CCGGCCGGCCCTTTCT**CTCT**GTCCTAGGTTAAACCAAAAACTCAACATGCAGTCAGAGAG

TTA*ACTATACAACGGGCCCATTTGCTG*CCTTTATATGTTCTTGGATGGATGACATATATC

TTTGGCTCCAATAATATAATTAATAGCTCATGGATAATGGTTATGGGTACACTGGAAACT

AGATACATTTAGATGCATGATTACATTTTGTTATATTTAGAACACAC

D)

---Forward PCR Primer→ → →

(----FseI Adapter----)CCGGCCGGCCCTTTCT**CTCT**GTCCTAGGT.....TGCAGTCAGAGAGTTA*ACTATACAACGGGCCCATTTGCTG*

TGATATGTTGCCCGGGTAAACGAC

← ← ←Reverse PCR Primer

(SEQ ID NO: 10)

FIG. 6

A) **6177338-6177564 (2) BTx_1133_128 IS3620c_186_136**

B) BTx_1133_128 - ATCCCGGCCTGCTGCACC**TTA**TTAGTTCAT (SEQ ID NO: 11)
IS3620c_186_136 - GTCCCGGCCTGCTGCACC---TTAGTTCATGCT (SEQ ID NO: 12)

3 bp indel at bp 19

C) >chromosome_1 chromosome_1:6177338,6177564 (reverse complemented)
(SEQ ID NO: 13)
CCGGCCTGCTGCACC**TTA**TTAGTTCATGCTGTCATGTGCCTTTGCTCTGAACCTCTAATC
AGGCTGCCTACCAGCAGTTATATTTCGCTTTTCGATGATACCAACAATGGTTAA*GTTAGT*
*GGACGATAAGCACTCCCA*TTTTCCAAAGATTAAAGCAGGTTCAGGTCATCAGGTGGCGCA
CGCAGCTCGCAGCACTACTAGTAGCTGTAGCCTGCGTATCCTGTAGA

FIG. 7

A) **7355570-7355596 (2) BTx_12360_47 IS3620c_13710_32**

B) BTx_12360_47 - ATCCCGGCC**G**GCAAACGGCAG**T**AGCAGGCG (SEQ ID NO: 15)
IS3620c_13710_32 - GTCCCGGCC**A**GCAAACGGCAG**C**AGCAGGCG (SEQ ID NO: 16)

diff bp 10 G-A
diff bp 22 T-C

C) >chromosome_1 chromosome_1:7355570,7355796
(SEQ ID NO: 17)
CCGGCC**G**GCAAACGGCAG**T**AGCAGGCGTTTAAGTAGCCGGTGGAAGATCTGACGACGCGC
CGATGGCTCCGGCTGCCTGCCTGCCTGCCGGGA*CAAGAAACATTGCACACGGAGAGG*CTC
ATGGCTGCCGCCTCGCTCTACGCAATCTACCAGCTAGCGCCCGGCTCGTTTCGGCAAAGT
CGCCAAGGCTCCATGTGCCCATGACCTGATTCAGTTCGTCGTCGCGC

FIG. 8

A) **68573364-68573590 (2) BTx_6881_78 IS3620c_3711_79**

B) BTx_6881_78 - ATCCCGGCCGCTCCCCGCGC**TAGC**TTGTCG (SEQ ID NO: 19)
IS3620c_3711_79 - GTCCCGGCCGCTCCCCGCGC----TTGTCGTCCA (SEQ ID NO: 20)

4 bp indel at bp 21

C) >chromosome_2 chromosome_2:68573364,68573590
(SEQ ID NO: 21)
CCGGCCGCTCCCCGCGC**TAGC**TTGTCGTCCAGCCTCTGCATAGTACTCTTTGCCCTCCGA
TCCCCAAGGAGATCGATCTCCCAAGCTAGCTAGCTCCGTCCAGTTGTTGGCCAAGGCGGA
TCGAGCAAAAAAAAAGCTTGTGTCTTGCAAGCTGTCTCCGGGTTT*TCATTCATCTCCCGTT*
*GTTGACGC*CGGGGTTGTCGTACGTCCTCGGTCAGGCCCTCCCTCCCT

FIG. 9

A) **5123239-5123265 (2) BTx_11756_50 IS3620c_6003_65**

B) BTx_11756_50 - ATCCCGGCCGGCT**G**CC**A**TTTCCTGAATCTT (SEQ ID NO: 23)
IS3620c_6003_65 - GTCCCGGCCGGCT**C**CC**G**TTTCCTGAATCTT (SEQ ID NO: 24)

diff bp 14 G-C
diff bp 17 A-G

C) >chromosome_3 chromosome_3:5123039,5123265 (reverse complemented)
(SEQ ID NO: 25)
CCGGCCGGCT**G**CC**A**TTTCCTGAATCTTTCCGTCGTCGCGCCAACGGGTTAACGACTGACG
AGTGTTCTTTGCTGGTTT*GTCAGATTGATCCCAAGTCTGTGG*CCGACGGCGACACCGTCA
ACGTGTACGTCGACACGGCGGACCCCCGGGAGGCCGGCAGCGTGCCCCGCGAGGTGCAGA
AGGCTGCGGCGGAGCGGGCCAAGGCGCGGGCCGCCAAGAACTACCAG

FIG. 10

A) **6565424-6565450 (2) BTx_7145_76 IS3620c_3291_82**

B) BTx_7145_76 - ATCCCGGCCCAAGAAGCGTCAGCGGACCGA (SEQ ID NO: 27)

IS3620c_3291_82 - GTCCCGGCCCGAGAAGCGTTAGCGGACCGA (SEQ ID NO: 28)

diff bp 11 A-G
diff bp 20 C-T

C) >chromosome_4 chromosome_4:6565224,6565450 (reverse complemented)
(SEQ ID NO: 29)
CCGGCCCAAGAAGCGTCAGCGGACCGAATCCCTCTGCAACCGGCTGCCTACCATCACCAT
CTGCTGCCCTATCGCTTTCACAT*GGCAATACGTGTCGATGCATTAGC*TTATCTCAACGCC
TCTTGCATAAGTGCCCTCTTATTATCTAGAAACTACTCCCTCCGTCTCAGCATAAAACAT
CTCGTTCAAAGAGTAGTTAATTTTTTAAAAAGTTTTTATAAATATAT

FIG. 11

A) **6565447-6565473 (2) BTx_3311_103 IS3620c_1860_93**

B) BTx_3311_103 - ATCCCGGCCTACGCGTGCGTCTGGCTCCTC (SEQ ID NO: 31)

IS3620c_1860_93 - GTCCCGGCCTATGCGTACGTCTGCCTCCTC (SEQ ID NO: 32)

diff bp 12 C-T
diff bp 17 G-A
diff bp 24 G-C

C) >chromosome_4 chromosome_4:6565447,6565673
(SEQ ID NO: 33)
CCGGCCTACGCGTGCGTCTGGCTCCTCGCGATATTTTGAAGAAGTCACAATTTATCGCGT
GGCCGCTAAAACGGTTCTAGAATGTCTAATTTTTATTTAAAAATAGTTTTTAGAAAAATA
TTTAGCAGAGCTCCTCTAGAGTCGGAAAAAAACTCCGAGCTAGAGTTGGAGACGACCCTT
GAGCCA*GAGCTAGAGACGAACCCTGTCAAA*CAAATCCATAGACAGGT

FIG. 12

A) **40467298-40467324 (2) BTx_4341_95 IS3620c_2685_86**

B) BTx_4341_95 - ATCCCGGCCAGG**C**TAAT**T**CT**C**CACAAAAGC (SEQ ID NO: 35)
IS3620c_2685_86 - GTCCCGGCCAGG**G**TAAT**G**CT**T**CACAAAAGC (SEQ ID NO: 36)

diff bp 13 C-G
diff bp 18 T-G
diff bp 21 C-T

C) >chromosome_6 chromosome_6:40467298,40467524
(SEQ ID NO: 37)
CCGGCCAGG**C**TAAT**T**CT**C**CACAAAAGCTCTCCAATACAACTCTCCCAATAAAATAATATT
AGGATAACCTAATACCATTTTCTAAATTTATT*GGGACAGTTGCTACAGCAGATCAT*CAAT
AATCTCCTTCAATATATAGGACATACTTGTCACTAACAATTTATTCATTTATTTATTTAT
TTATTTATTTATTTATTTTTCTTTCATCTTCTCCAATCGGGACTGCA

FIG. 13

A) **12029252-12029278 (2) BTx_8584_67 IS3620c_14289_30**

B) BTx_8584_67 - ATCCCGGCCGCCTATAAAA**A**C**G**CCCCCGCA (SEQ ID NO: 39)
IS3620c_14289_30 - GTCCCGGCCGCCTATAAAA-C**C**CCCCCGCAC (SEQ ID NO: 40)

1 bp indel at bp 20
diff bp 22 G-C

C) >chromosome_7 chromosome_7:12029252,12029478
(SEQ ID NO: 41)
CCGGCCGCCTATAAAA**A**C**G**CCCCCGCACCTCCCACGCCTCGCCGGCAGCAGCCAGCGTTC
CAGCAGCTGTTGTAATTACACGTACACGTACAGTACCCATTGCCGCTGCC*GCGTCCACCA*
*GCATTCATTCATTG*ATTTCACAGCCCATCGCCTCACCTACTGCCTCCAAACCAACCCCCT
CCGATCCCTCGGTCGCCGATCGAGTGACGTCGTGACGACGACCATGG

FIG. 14

A) **57701767-57701793 (2) BTx_14040_39 IS3620c_16226_23**

B) BTx_14040_39 - ATCCCGGCCTGCCGGGGGG-TGAAACCACCG (SEQ ID NO: 43)

IS3620c_16226_23 - GTCCCGGCCGGGCCGGGGGGTGAAACCACC (SEQ ID NO: 44)

diff bp 10 T-G
diff bp 12 C-G
diff bp 14 G-C
1 bp indel at bp 20

C) >chromosome_7 chromosome_7:57701567,57701793 (reverse complemented) (SEQ ID NO: 45)

CCGGCCTGCCGGGGGGG-TGAAACCACCGCCACGGGGGAGGGTAACGTGACAACGTCGCGTC
GGGAACGTGCCCGGCCGGGCGGCGCCATCCTCGTGCCACAGGGCCGGCCGGCAAGTAGTG
ACACCTGAGCGAGCGAG*AGATACGTGGGTTGGTATGGAGAG*GAAGACACCAAAGCTTCCT
TTCTCCCTCGCGTTATCGCTTCGAGCGGCGCCGGTCGTTCCCCCCTC

FIG. 15

Espacenet

Bibliographic data: WO2011003630 (A1) — 2011-01-13

METHOD FOR DIFFERENTIATION OF POLYNUCLEOTIDE STRANDS

Inventor(s): SOLDATOV ALEKSEY [DE]; BORODINA TATIANA [DE]; LEHRACH HANS [DE] ± (SOLDATOV, ALEKSEY, ; BORODINA, TATIANA, ; LEHRACH, HANS)

Applicant(s): MAX PLANCK GESELLSCHAFT [DE]; SOLDATOV ALEKSEY [DE]; BORODINA TATIANA [DE]; LEHRACH HANS [DE] ± (MAX-PLANCK-GESELLSCHAFT ZUR FOERDERUNG DER WISSENSCHAFTEN E. V, ; SOLDATOV, ALEKSEY, ; BORODINA, TATIANA, ; LEHRACH, HANS)

Classification: **- international:** ***C12Q1/68***
- cooperative: **C12Q1/6848**; **C12Q1/6853**; **C12Q1/6855**; **C12Q1/6869**; **C12Q1/6876** more

Application number: WO2010EP04399 20100706 Global Dossier

Priority number(s): EP20090008808 20090706 ; US20090213730P 20090708

Also published as: EP2272976 (A1) EP2451973 (A1) EP2451973 (B1) US2012237943 (A1) US2015105299 (A1) more

Abstract of WO2011003630 (A1)

Objective of the present invention is to provide a method for keeping of directional information in double-stranded DNA. We suggest to convert polynucleotide into a hybrid double-stranded DNA. One particular strand of this hybrid double-stranded DNA should be synthesised using at least one modified nucleotide. Thus, this particular strand would contain modified nucleotides along the whole length. Density of directional markers would not depend on the length of polynucleotides. Any internal fragments of the hybrid double-stranded DNA would have directional information. When it is necessary the modified strand may be easily degraded or separated from the other strand. It was found that such hybrid double-stranded DNA may be easily generated in a number of molecular biology tasks and may be used for molecular cloning, library preparation and strand separation.

IDT Exhibit 1020.1587

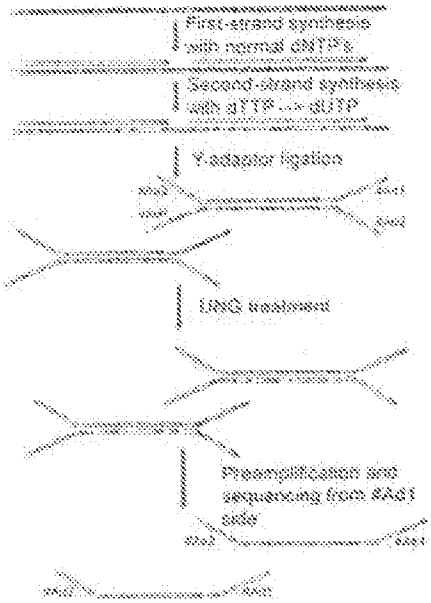


Figure 4

IDT Exhibit 1020.1588

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 January 2011 (13.01.2011)

PCT

(10) International Publication Number
WO 2011/003630 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01)

(21) International Application Number:
PCT/EP2010/004399

(22) International Filing Date:
6 July 2010 (06.07.2010)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

09008808.9	6 July 2009 (06.07.2009)	EP
61/213,730	8 July 2009 (08.07.2009)	US

(71) Applicant *(for all designated States except US)*: **MAX-PLANCK-GESELLSCHAFT ZUR FÖRDERUNG DER WISSENSCHAFTEN E. V.** [DE/DE]; Hofgartenstraße 8, 80539 München (DE).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **SOLDATOV, Aleksey** [RU/DE]; Taylorstrasse 12b, 14195 Berlin (DE). **BORODINA, Tatiana** [RU/DE]; Taylorstrasse 12b, 14195 Berlin (DE). **LEHRACH, Hans** [AT/DE]; Terrassenstr. 31, 14129 Berlin (DE).

(74) Agent: **ARTH, Hans-Lothar**; ABK Patent Attorneys, Jasminweg 9, 14052 Berlin (DE).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

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

Published:

— *with international search report (Art. 21(3))*

(54) **Title**: METHOD FOR DIFFERENTIATION OF POLYNUCLEOTIDE STRANDS

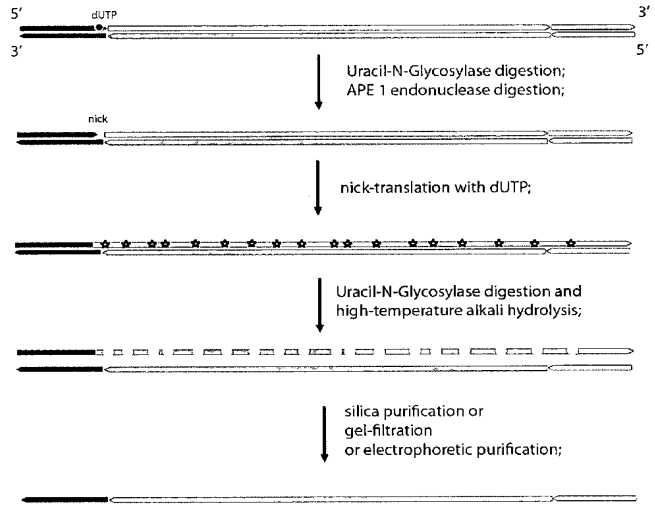


Figure 3A

(57) **Abstract**: Objective of the present invention is to provide a method for keeping of directional information in double-stranded DNA. We suggest to convert polynucleotide into a hybrid double-stranded DNA. One particular strand of this hybrid double-stranded DNA should be synthesised using at least one modified nucleotide. Thus, this particular strand would contain modified nucleotides along the whole length. Density of directional markers would not depend on the length of polynucleotides. Any internal fragments of the hybrid double-stranded DNA would have directional information. When it is necessary the modified strand may be easily degraded or separated from the other strand. It was found that such hybrid double-stranded DNA may be easily generated in a number of molecular biology tasks and may be used for molecular cloning, library preparation and strand separation.

WO 2011/003630 A1

Method for differentiation of polynucleotide strands

Description

Cloning of single-stranded RNA and DNA molecules is much easier if they are transformed into double-stranded DNA. The invention relates to methods of preserving information about the direction of the original single-stranded molecules after being transformed into a double-stranded form.

Single-stranded RNA and DNA molecules play an important role in molecular biology. A number of viruses have single-stranded genomes. Most RNAs are present in single-stranded form. Recent studies have demonstrated an unexpected complexity of eukaryotic transcriptomes. A transcriptome is the set of all RNA transcripts produced in one or a population of cells. In addition to classical mRNAs (messenger ribonucleic acids) which cover approximately 1.5 percent of the genome in most higher eukaryotes, a large number of non-coding RNAs (ribonucleic acids) with widely varying expression levels have been identified. The biological functions of these novel transcripts are largely unknown and represent a new research area requiring high throughput transcriptome studies.

The polarity of the transcript is important for correct annotation of novel genes because it provides essential information about the possible function of a gene both on the RNA (structure and hybridization to other nucleic acid molecules) and the protein level. In addition, many genomic regions give rise to transcripts from both strands simultaneously. Antisense transcription is characteristic for eukaryotic genes and is thought to play an important regulatory role. Overlapping genes are common for compact genomes of prokaryotes and lower eukaryotes. Knowledge of transcript orientation helps to resolve the problem of colliding transcripts and to determine correctly gene expression levels in the presence of antisense transcription.

Background of the invention

From the art several methods are known that preserve directional information using distinguishable adapters for different ends of double-stranded DNA.

Adapters may be ligated directly to single-stranded RNA molecules.

Special composite primers may be used for first-strand cDNA synthesis: Herein the 5'- part of the primer provides a directional information, the 3'-part (oligo(dT)) or random nucleotides) is able to form a duplex with the RNA. Recently, first-strand cDNA synthesis was performed from a tagged random hexamer primer and second strand synthesis from a DNA-RNA template-switching-primer.

Special homopolymeric adapters may be synthesised at the 3'-end of DNA molecules by Terminal Nucleotide Transferase.

DNA-RNA template-switching is used to attach a specific adapter to 3'-end of the DNA copy of the original single-stranded RNA molecule.

The main disadvantage of the adaptor-based approach is that only the ends of the clones preserve directional information. The problems are:

(i) strand-specific operations (directional immobilization on a solid phase; production of a single-stranded DNA for hybridization, etc.) should act on the adaptor sequence. For long clones the efficiency of such operations is normally low. For example, strand separation where one of the strands contains a biotin molecule at the 5'-end is getting complicated if the length of the duplex is higher than 1kb.

(ii) after fragmentation internal parts of the clones lose the directional information.

A special approach which is able to keep directional information in any part of the sequence was suggested recently: All cytidine residues in the RNA were converted to undines by bisulfite treatment prior to cDNA synthesis. The problem is that this approach is laborious and also leads to the loss of about 30% of uniquely matched sequencing reads because a part of the genome complexity is lost during the transformation from the four-bases into the three-bases code.

A lot of biologically important nucleic acids function in a single-stranded form where the polarity is extremely important. But for basic molecular cloning procedures single-stranded polynucleotides should be converted into double-stranded DNA molecules, because

* most of restriction enzymes work only on double-stranded DNA;
* ligation of double-stranded molecules is much simpler in comparison with the ligation of single-stranded molecules;
* the result of PCR is a double-stranded DNA.

At the moment most of the protocols keep directional information in the form of distinguishable adaptors on the ends of double-stranded molecules (Fig. 1B).

US 2007/0281863 discloses a method for characterizing the amounts of nucleic acids, including plus/minus determinations. To this end the mRNA is transcribed into double stranded cDNA and two different RNA promoters are integrated in both cDNA strands respectively. By using the different RNA promoters in the presence of allylamine-UTP nucleotides plus or minus RNA with modified allylamine-UTP nucleotides is generated. The modified allylamine-UTP nucleotides can be labeled with NHS-Cy3/5 and are then hybridized in different nucleic acid arrays.

The main disadvantage of this approach is that the RNA cannot be used for high-throughput complementary DNA sequencing (RNA-Seq). Direct cDNA sequencing (RNA-Seq) is a new tool for whole-transcriptome analysis. Second generation sequencing machines have increased sequencing throughput by about two orders of magnitude, making global transcriptome sequencing feasible. Since sequencing costs are constantly decreasing (contrary to those of microarrays) cDNA sequencing will capture a considerable portion of transcriptome analyses in the future.

The RNA-Seq procedure is simple, has a large dynamic range and high sensitivity, and can unequivocally identify splicing and RNA editing products as well as allele-specific transcripts. RNA-Seq provides a number of further advantages over other high throughput approaches like microarray hybridization, gene-specific and tiling arrays or SAGE-analyses. The depth of RNA-Seq analyses is flexible, providing a dynamic range typically an order of magnitude greater than one can achieve with hybridization arrays. The digital character of the RNA-Seq data permits to compare and pool results from different laboratories. No prior information about transcripts sequences is required, allowing detection of novel transcripts. It is possible to estimate the absolute level of gene expression and to study structure of transcripts.

However, the weakness of RNA-Seq was the inability to determine the polarity of RNA transcripts without laborious modification of the protocol. The method according to the invention solves this problem allowing the strand specific differentiation with an easy, robust and cheap procedure. The method is less susceptible and more robust compared to the method described in US 2007/0281863 because there is no need to transcribe the mRNA first in cDNA then inserting two different promoters in the cDNA and then using two different polymerases to transcribe the cDNA again in RNA including an additional step to label the marked RNA with NHS-Cy3/5.

Thus, the modified allylamine-UTP RNA according to US 2007/0281863 can only be used in hybridization arrays that are very expensive, less sensitive and cannot detect

novel transcripts. Additionally, the method is laborious including the integration of promoters and different polymerases and works with RNA as an end product, which is very susceptible to degradation because of abundant RNases and can only be used in expensive hybridization arrays that are less sensitive then e.g. RNA-Seq.

Further disadvantages include that many basic molecular cloning procedures are performed on double stranded cDNA, which are necessary for many subsequent analyzing methods. The method according to US 2007/0281863 generates a double stranded cDNA with two promoters as an intermediate product. However, these intermediate cDNAs are not preferred for cloning procedures,

As we already explained above the main disadvantage of adaptor-based approaches is that only the ends of the clones preserve directional information. The integration of promoters in the cDNA is basically the same procedure. During cloning procedures the promoters can be changed, damaged etc., especially because most cloning procedures modify the ends of the cDNA. The subsequent strand specific transcription of the cDNA with promoter specific polymerases is thus not very robust and the strand specific information can be lost during molecular cloning procedures. Taking into account the susceptibility of RNA to degradation, the method according to the instant application is more functional, i.e. the RNA is transcribed directly into cDNA and the strand specific information is already integrated in this step. Furthermore, the inventive method is much more robust, e.g. the cDNA already contains the strand specific information along the whole length of the transcript and no molecular cloning procedures can compromise this information. Consequently, the inventive method is convenient, reliable and highly reproducible.

WO 97/12061 discloses a method for characterizing nucleic acid molecules, comprising synthesizing DNA with a nucleic acid matrix, primers, polymerase, four canonical deoxynucleotides and at least one non-canonical deoxynucleotide. The DNA is then contacted with N-glycosylase and the abasic sites are treated in a way to leading to breakage of the phosphodiester backbone. The resulting DNA fragments are separated according to their size.

This method characterizes DNA according to their size by using non-canonical deoxynucleotide to label one synthesized DNA strand, but the method is not able to preserve the information about the direction of the original single-stranded molecules after being transformed into a double-stranded form.

Description of the invention

The objective of the present invention is to provide a method for keeping the directional information in double-stranded DNA. Surprisingly, it was found that this objective can be reached by the inventive method as disclosed by the description, the figures and the examples. A polynucleotide strand is converted into a hybrid double-stranded DNA. One particular strand of this hybrid double-stranded DNA should be synthesised using at least one modified nucleotide. Thus, this particular strand would contain modified nucleotides along the whole length of the strand. The density of directional markers would not depend on the length of hybrid double-stranded DNA. Any internal fragment of the hybrid would have directional information. It was found that such hybrid double-stranded DNA may be easily generated in a number of molecular biology tasks and may be used for molecular cloning, library preparation and strand separation. If necessary, one of the strands (modified or non-modified) may be easily degraded or separated from the other strand. Also some modifications might prevent the strand containing modified nucleotides from participation in downstream applications.

The directional library construction method and the method of synthesis of single-stranded DNA according to the invention are easy to perform, inexpensive and highly reliable and thus overcome the shortcomings of the state of the art approaches.

The inventive method for preserving information about the direction of single-stranded nucleic acid molecules during molecular cloning operations with double-stranded DNA is provided. One of the strands of the double-stranded DNA is synthesized by using the other strand as a matrix and contains at least one modified dNTP. Thus, one strand of double-stranded DNA would contain unmodified nucleotides and the other strand at least one modified nucleotide along the whole length (Fig. 1C).

The new method of keeping of direction of polynucleotides is fully compatible with methods known from the art. Double-stranded DNA molecules may have simultaneously distinguishable primers and distinguishable strands (Fig. 1D). Fig. 2A shows how molecules with distinguishable strands may be converted into molecules with distinguishable primers. In an opposite operation the conversion of molecules with distinguishable primers into molecules with distinguishable strands is shown in Fig. 2B.

The main advantages compared to the adaptor-based method are:

(i) the density of modified nucleotides (directional markers) does not depend on the length of the double-stranded molecule. This is important for strand separation or strand-specific digestion of long DNA molecules.

(ii) any internal fragment of the clone carries directional information. After fragmentation internal parts of the clones do not lose the directional information.

Two examples demonstrate the convenience of the inventive method for differentiation of polynucleotide strands.

The method of synthesis of a single-stranded DNA is described in Fig. 3.

There is a number of methods of synthesis of single-stranded DNA relying on the difference between flanked adaptors:

* asymmetric PCR may be performed with the different amounts of PCR primers;

* the 5' end of one of the strands may contain thionucleotides which protect this strand from 5' to 3' exonucleases;

* protective thionucleotides may be introduced into the 3'-end part of only one strand to protect it from 3' to 5' exonucleases;

* the 5'-part of one of the strands may bear biotin for attachment of this strand to streptavidine-coated solid particles.

The problem is that all these methods proved to be unstable and show low efficiency for molecules larger than 1 kb.

Fig. 3 shows two possible ways for generating single-stranded DNA using the inventive method for differentiation of polynucleotide strands.

In a first embodiment of the present invention (Fig. 3A) one strand of PCR product is substituted by a strand where dTTPs are substituted by dUTPs. After UNG treatment and APE 1 digestion (or alkalic hydrolisis) this strand is converted into a number of short single-stranded fragments which may be easily separated from the long unmodified strand by electrophoresis, anion-exchange chromatography, silica-based purification or gel-filtration.

In a second embodiment of the present invention (Fig. 3B) one strand of the PCR product is substituted by a strand where a significant percentage of the dTTPs is substituted by biotin-associated dUTPs, respectively biotin-associated dTTPs. The

concentration of biotin in a modified strand does not depend on the length of the double-stranded DNA, so the efficiency of the purification on streptavidine-coated particles should not depend on the length of the fragment.

The key steps of the method according to the invention include the generation of a cDNA strand including at least one modified nucleotide that in the course of the method is recognized by a selection factor which allows for selection of the strand containing the at least one modified nucleotide from the strand containing the corresponding unmodified nucleotide or nucleotides.

The key steps of the inventive method are to be read in the following order:

I) Providing polynucleotide templates, primers, buffer solution, at least one enzyme with polymerase activity, sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof, wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection;

II) Performing first-strand synthesis on said polynucleotides; and

III) Performing strand-specific selection by using a selection factor for selecting strands containing the at least one modified nucleotide from strands containing the corresponding unmodified nucleotide.

The key steps of the inventive method can also be read in the following order:

I) Providing polynucleotide templates, primers, buffer solution, at least one enzyme with polymerase activity, sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof, wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific differentiation;

II) Performing specific strand synthesis on said polynucleotides; and

III) Performing strand-specific differentiation by using a selection factor for differentiating strands containing the at least one modified nucleotide from strands containing the corresponding unmodified nucleotide.

The term *strand specific differentiation* refers to the ability to differentiate in the double stranded form between the template and the synthesized matching strand.

The term *specific strand synthesis* refers to synthesis of the matching strand containing the at least one modified nucleotide from the template.

Another embodiment of the present invention relates to single-stranded RNA. Thus, if the polynucleotide template is a single-stranded RNA a second cDNA strand has to be synthesized. For this purpose the at least one modified dNTP has to be removed from the solution and has to be replaced by the corresponding unmodified dNTP or dNTPs or analogues thereof. By the use of all four unmodified dNTPs a second cDNA strand is generated.

In this case the key steps according to the inventive method read as follows:

I) Providing polynucleotide templates, primers, buffer solution, at least one enzyme with polymerase activity, sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof, wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection;

II) Performing first-strand synthesis on said polynucleotides;

III) Performing second-strand synthesis with all four unmodified dNTPs or analogues thereof after removing the at least one modified dNTP; and

IV) Performing strand-specific selection by using a selection factor for selecting strands containing the at least one modified nucleotide from strands containing the corresponding unmodified nucleotide.

As said before, if the polynucleotide templates are single-stranded RNA a second-strand cDNA has to be synthesized after the exchange of the at least one modified dNTP by the corresponding unmodified dNTP or dNTPs. If this double-stranded cDNA shall undergo cloning it will be advantageous to perform one or more of the modifications for the molecular cloning operation as described before. By combining these conditions the key steps of the invention read as follows:

I) Providing polynucleotide templates, primers, buffer solution, at least one enzyme with polymerase activity, sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof, wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection;

II) Performing first-strand synthesis on said polynucleotides;

III) Performing second-strand synthesis with all four unmodified dNTPs or analogues thereof after removing the at least one modified dNTP;

III) Performing one or more molecular cloning operations on the double-stranded DNA, wherein the molecular cloning operations are selected

from shearing, restriction, digestion, end-polishing, A-tailing, G-tailing and adaptor ligation; and

V) Performing strand-specific selection by using a selection factor for selecting strands containing the at least one modified nucleotide from strands containing the corresponding unmodified nucleotide.

The initially provided polynucleotide templates may be a RNA such as mRNA, tRNA, rRNA, hnRNA, snRNA, snoRNA, scaRNA, asRNA, micro RNA, siRNA, short transcripts, single-strand oligonucleotides, single-strand polynucleotides, synthetic polynucleotides and semisynthetic polynucleotides as well as physiological or externally induced modifications thereof, wherein mRNA is preferred.

mRNAs are protein-coding RNA. tRNAs are transfer RNAs responsible for providing correct amino acids in peptide (protein) synthesis in ribosomes. rRNAs are ribosomal RNAs, being a constitutive element of ribosomes. hnRNAs (heterogeneous nuclear RNA) is a precursor mRNA in the nucleus of eukaryotic cells. It is often named also pre-mRNA or precursor mRNA. snRNA (small nuclear RNA) is responsible in the eukaryotic nucleus for the processing of hnRNAs (spliceosomes). snoRNA (small nucleolar RNA) occurs in the nucleolus of eukaryotic cells. scaRNA (small Cajal body-specific RNA) is strongly related to snoRNA and occurs in Cajal bodies. asRNA (antisense RNA) plays an important role in regulating gene expression. microRNAs are involved in the regulation of cellular processes such as apoptosis and cell death. siRNA (small interfering RNA) is processed from cellular RNA into fragments with a length of ca. 22 base pairs by the enzyme Dicer and integrated into the enzyme complex RISC (RNA-induced silencing complex). RISC can thus regulate the expression of DNA segments or RNAs by switching off the respective coding region through analogous base pairing (silencing). This method by also used in gene therapeutic approaches. Retroviruses use RNA as the carrier of their genetic information.

Short transcripts refer to small RNA fragments with up to 20-40 base pairs.

Alternatively, the polynucleotide templates used within the inventive methods may be a DNA such as siDNA, DNA copies from retroviral or hepadnaviral RNA, single-stranded DNA, synthetic polynucleotides and semisynthetic polynucleotides as well as physiological or externally induced modifications thereof, wherein single-stranded DNA is preferred.

It shall be pointed out that in this invention the term *single-stranded DNA* can also refer to the first-strand cDNA derived from a RNA as templates.

siDNA (small interfering DNA) is the corresponding DNA of siRNAs and mainly fulfils the same tasks in the cell. Single-stranded DNA is a single-stranded type of satellite DNA, occurring among others in certain soil-dwelling bacteria species.

According to the invention also double-stranded DNA can be the polynucleotide templates. Specific primers and optionally adjuvants are needed to allow the generation of a cDNA strand against a particular strand of the double-stranded DNA (see below and Fig. 2B). Herein, the double-stranded DNA is destabilized in the region of hybridization of said specific primer for allowing annealing of the specific primer. Such a destabilization can be achieved for example by Uracil-N-Glycosylase digestion of deoxyuridines.

In a specific embodiment the double-stranded DNA templates are PCR products.

Sources for RNAs and DNAs can be eukaryotes (animals including humans and other mammals, plants, fungi), prokaryotes (bacteria, archaea), viruses, synthetic or semisynthetic constructs wherein these constructs or semisynthetic constructs may mirror at least partially a naturally occurring nucleic acids or may have a widely artificial sequence.

Said RNAs and DNAs can be isolated by well-known methods (Molecular Cloning: A Laboratory Manual, Third Edition. J. Sambrook, D. Russell).

The term *classic dNTPs*, respectively *unmodified dNTPS* refers to dATP (deoxyadenosine triphosphate), dCTP (deoxycytidine triphosphate), dGTP (deoxyguanosine triphosphate) and dTTP (deoxythymidine triphosphate). These four dNTPs (deoxynucleoside triphosphates) are usually used in nature for the generation of DNA strands.

It shall be pointed out that the plural form *classic dNTPs* refers to the total of dATP, dCTP, dGTP and dTTP. A referral in a quantifying sense to one or more of these classic dNTPs is not intended.

The term *analogues of dNTPs*, respectively *dNTPs and analogues thereof* refers to any molecule suitable for replacing a dNTP in presence and behavioural characteristics in a respective polynucleotide-containing solution and particularly in

its matching characteristics to be able to undergo a base pair matching identical or similar to the classic dNTP it replaces. In particular, such analogues of dNTPs should be able to build the same or similar selective hydrogen bonds that allow for a selective pairing of base pairs (2 hydrogen bonds between A/T, T/A, A/U or U/A, and 3 hydrogen bonds between G/C and C/G), thus allowing for a correct translation of a polynucleotide sequence into its antisense sequence. Furthermore, it should be sterically similar to its corresponding classic dNTP so that it allows integration into the base pair sequence. Not only the nucleotide can be varied in such an analogue, but also the saccharide moiety (usually a ribose or deoxyribose) or the phosphate backbone. For example, the phosphate backbone may be partially or entirely substituted by a sulphonated backbone and/or by a phosphonate backbone. Also combinations of such variations are subsumed under this term.
Moreover the analogues of the dNTPs cannot be used for stand specific selection or strand differentiation. Thus the analogues of the dNTPs do not react under the conditions which are used for the modified dNTPs such as degradation (e.g. dUTP, heat), differentiation by e.g. electrophoretic methods, adsorption or binding to suitable counterparts such as biotin – avidin. Thus the analogues of the dNTPs are used as substitutes for the classic or canonical dNTPs and not as substitutes for the modified dNTPs.

The quantitative term *sufficient amount* herein refers to any amount, respectively concentration of an agent which allows for a qualitative and quantitative implementation of the method according to the invention in respect of the expected amounts which shall be needed for this purpose. Usually, the respective agents, in particular dNTPs, are added in exceeding amounts in order to ensure that all possible copies can be effected qualitatively and quantitatively.

The term *template* refers in molecular biology to a DNA or RNA strand which sets the base pair sequence to which a corresponding strand with an antisense sequence shall by polymerized. In a broader sense it may also refer to a RNA encoding a sequence for translation of this sequence into an amino acid sequence such as a peptide or a protein in ribosomes. The term *templates* in plural refers to the total sum of polynucleotides provided, each single template representing a single polynucleotide. Templates also refers to the originally provided sum of polynucloetid strands. to A referral in a quantifying sense to one or more of these templates is not intended.
The term *template* refers also to the original single stranded polynucleotide strand, where the preservation of information about the direction is desired after being

transformed into a double-stranded form. If the inventive method is applied to double stranded polynucleotides (e.g. DNA/RNA, DNA/DNA with distinguishable adapters), in this case the template can be freely chosen from both strands, depending for which strand the preservation of information about the direction is desired.

The term *matching strand* refers to the complementary strand to the template.

The term *first-strand synthesis* refers to the synthesis of the first strands using the templates as matrix for synthesizing the complementary strands. The nucleotide sequence of the first strand corresponds to the sequence of the matching strand.

The term *second-strand synthesis* refers to the synthesis of the second strands that uses the first strands as matrix for synthesizing the complementary strands, i.e. the second strands. The nucleotide sequence of the second strand corresponds to the sequence of the template.

As used herein the term "matrix" or "matrices" refers to any polynucleotides which can serve as starting material for polymerase-reactions.

As used herein the term "at least one nucleotide" or "at least one modified nucleotide" is not meant as a referral in a quantifying sense to one or more nucleotides but refers to a species of modified dNTPs. Thus, this term refers to a plurality of dNTP molecules of the same kind. Thus using "one" modified nucleotide does not mean that a single modified nucleotide is used. Using "one" modified nucleotide means that one kind or one sort of nucleotides are used which are modified such as modified dATP or modified dCTP or modified dGTP or modified dTTP.

The term "replaced by a corresponding modified dNTP" refers to a substitution of one of the classic or canonical dNTP species in the reaction mixture used for strand synthesis by a modified dNTP species able to fulfil the same biological function. This means that the substitution is not performed on the completed strand after strand synthesis. The term "same biological function" refers to modified dNTPs able to preferably build the same or similar selective hydrogen bonds that allow for a selective pairing of base pairs (2 hydrogen bonds between A/T, T/A, A/U or U/A, and 3 hydrogen bonds between G/C and C/G). Furthermore, it should preferably be sterically similar to its corresponding classic dNTP so that it allows integration into the base pair sequence. Not only the nucleotide can be varied in such a modified dNTP, but also the saccharide moiety (usually a ribose or deoxyribose) or the phosphate backbone.

The term *cDNA* stands for complementary deoxyribonucleic acid. It refers to a DNA which is generated as a copy of a RNA (usually by means of Reverse Transcriptase) or from a single DNA strand (usually by means of a DNA polymerase). Subsequently, a second cDNA strand may be generated by using the first cDNA strand as a matrix (usually by means of a DNA polymerase).

Enzymes with reverse transcriptase (RNA-dependent DNA polymerase) activity are able to transcribe a RNA sequence into a corresponding cDNA sequence. This group of enzymes includes among others HIV-1 reverse transcriptase from human immunodeficiency virus type 1 (PDB 1HMV), M-MLV reverse transcriptase from the Moloney murine leukemia virus, AMV reverse transcriptase from avian myeloblastosis virus and telomerase reverse transcriptase fulfilling a role in the maintenance of telomeres of eukaryotic chromosomes as well as commercially available forms such as "M-MLV reverse transcriptase RNase H Minus", "SuperScript® II", "SuperScript® III" or Tth polymerase. Among this group M-MLV reverse transcriptase RNase H Minus, SuperScript® II and SuperScript® III polymerases are preferred.

The terms *primers* and *random primers* refer to oligonucleotides which shall serve as starting point for DNA-replicating enzymes such as DNA polymerases. They can consist of a few base pairs, but in selected cases can reach up to 2.000 base pairs. Preferred are 5 to 50 base pairs. DNA polymerases need a hydroxy group for starting transcription. This hydroxy group is provided by the primer with its 3' hydroxy end. Often, the DNA sequence to be investigated is known. Hence suitable primers can be synthesized precisely according to the antisense sequence of aforesaid polynucleotide sequence. If this polynucleotide sequence is not or only incompletely known so-called random primers can be used. In such a method a large number of primers with a random base pair sequence are provided. One or more of these random primers statistically anneal to the polynucleotide sequence that shall be transcribed. Thus corresponding transcripts of different length can be obtained, depending where the random primer anneals to the polynucleotide sequence of interest. Alternatively, also composite primers with a random 3' part can be used. A further variation corresponding to the inventive method is the use of primers which are complementary to a homo-oligomeric tail synthesized by terminal transferase on the polynucleotide which serves as a matrix.

If the polynucleotide templates are RNA with a polyA tail it is preferred that said primers are oligo(dT), oligo(dU) or oligo(U) primers or composite primers with oligo(dT), oligo(dU) or oligo(U) region on 3' part. Alternatively, specific primers complementary to said single-stranded RNA templates can be used.

If double-stranded DNA templates are used it is likewise possible to mark one particular strand of the double-stranded DNA with at least one modified nucleotide. Herein specific primers complementary to a particular strand of said double-stranded DNA can be used. In this case RecA protein may facilitate the hybridization of said primers to the double-stranded DNA templates. Specific primers may appear as a result of introducing a nick in a particular strand of said double-stranded DNA templates. Such nicks can for example result by the use of nicking endonuclease. Alternatively, such nicks may result from the digestion of the at least one modified nucleotide in a particular strand of said double-stranded DNA templates.

In a specific embodiment the double-stranded DNA templates are PCR products. In this case, the recognition site of a nicking exonuclease or at least one modified nucleotide is located in one of the PCR primers. Herein it is preferred that said modified nucleotide is deoxyuridine.

If double-stranded DNA templates are used it is preferred that said polymerase is a DNA-dependent DNA polymerase with 5' to 3' exonuclease activity and said strand-synthesis on such double-stranded DNA templates are effected by means of a nick-translation.

In analogy, the polynucleotide templates may be also a DNA::RNA hybrid. Hybrid herein means that a DNA strand is matching with a RNA strand having a corresponding antisense sequence. For example, but not limiting, it is possible that such a DNA::RNA hybrid is the product of a first-strand synthesis with unmodified nucleotides generated on single-strand RNA as templates. For generating a cDNA of the DNA strand specific primers complementary to a particular DNA region of said double-stranded DNA::RNA hybrid have to be used. It is preferred that said primers are complementary to a homo-oligomeric tail synthesized by terminal transferase. In another preferred embodiment this specific primer is a template-switch-primer. Such template-switch primers bind to the extra random nucleotides attached to the 3' end of the newly synthesized cDNA strand by using SuperScript® II enzyme (Invitrogen). This allows for full-length reverse transcription of the RNA strand. If the cDNA strand shall be generated against the RNA strand of the DNA::RNA hybrid said polymerase is DNA-dependent DNA polymerase with 5' to 3' exonuclease activity.

If the cDNA strand shall be generated against the DNA strand of the DNA::RNA hybrid the specific primers may appear as a result of introducing a nick into the RNA strand of said DNA::RNA hybrid. Such a nick may be set by using the RNA-specific enzyme RNase H. In a preferred embodiment RNase H, DNA polymerase I and

E.coli ligase are used simultaneously. DNA ligases are enzymes able to ligate DNA strands by building an ester bond between the phosphate residue and the saccharide deoxyribose in the backbone of the DNA strands.

For second-strand cDNA synthesis using the first strand cDNA as a matrix an enzyme with DNA polymerase activity is needed. Also if the original polynucleotide is a single-stranded DNA and only one cDNA strand is generated such an enzyme with DNA polymerase activity is needed. DNA-dependent DNA polymerases are usually enzymes which generate a DNA copy strand from a DNA matrix. This group includes DNA polymerase I, II, III holoenzyme and IV, Taq polymerase, Tth polymerase, Pfu polymerase, wherein DNA polymerase I and Taq polymerase are preferred. This enzyme with polymerase activity does not have to be the same as used optionally for amplifying the remaining intact cDNA.

If the cDNA strand shall be generated against the RNA strand of the DNA::RNA hybrid the procedure is similar to that for generating a cDNA strand against one strand of a single-stranded RNA.

The term *modified dNTP* refers to any molecule suitable for substituting exactly one corresponding classic dNTP in presence and behavioural characteristics in a respective polynucleotide-containing solution and particularly in its matching characteristics to be able to undergo a base pair matching identical or similar to the classic dNTP it replaces. Additionally, this modified dNTP must be suitable either for a specific degradation in which selectively this modified dNTP is degraded by a suitable degrading agent, thus rendering the DNA strand containing at least one of this degraded dNTP essentially unfit for amplification and/or hybridization, or it must render the DNA strand containing this modified dNTP eligible for specific removal of this DNA strand from the solution, or at least allowing a local and thus functional separation of polynucleotide strands. Such a removal respectively separation can be achieved by molecules or particles binding selectively to such exchange dNTPs, thus selectively only to one polynucleotide strand. Removal can occur by purification, adsorption to suitable structures or surfaces, binding to suitable selective structures such as antibodies, by selective degradation through certain chemicals, heat or radiation, or by centrifugation, magnetic or electrophoretic methods, isoelectric focusing. Also any kind of immobilization can be perceived.

The term *canonical* is used to refer to the four common nucleic acid bases adenine, cytosine, guanine and thymine that are commonly found in DNA or to the respective deoxyribonucleosides, deoxyribonucleotides or 2′-deoxyribonucleoside- 5′-

triphosphates that contain a canonical base, The tern "non-canonical" is used to refer to nucleic acid bases in DNA other than the four canonical bases, or to the respective deoxyrinonucleosides, deoxyribonucleotides, or 2´ -deoxyribonucleoside- 5´-triphosphates that contain a non-canonical base. Although uracil is a common nucleic acid base in RNA, uracil is a non-canonical base in DNA. "Non canonical bases" are found in nucleic acids as a result of incorporation of non-canonical nucleotides or as a result of modification of existing bases (canonical or non-canonical).

Different purification methods may be used for removal of modified and unmodified dNTPs from the solution: electrophoretic separation, salt-alcohol precipitation, *anion*-exchange chromatography, silica purification, gel filtration. Purification on silica and gel-filtration are preferred.

Degradation respectively degradable herein means that a deoxynucleotide is specifically modified by addition of at least one enzymatic, chemical, heat, radiation (X-ray, UV, microwave) or optical methods such as lasers or combinations thereof in such a way that the resulting deoxynucleotide does not serve anymore for hybridization, cloning or as a suitable matrix in an amplification system such as PCR.

Analogously, the plural form *modified dNTPs* refers to the total of the respective modified dNTPs. A referral in a quantifying sense to one or more of these modified dNTPs is not intended.

The term *at least one modified dNTP* refers to the option that not only one dNTP is exchanged. According to the invention either one dNTP such as dATP, dCTP, dGTP or dTTP, or two dNTPs can be exchanged in the combinations dATP + dCTP, dATP + dGTP, dATP + dTTP, dCTP + dGTP, dCTP + dTTP or dGTP + dTTP, or three dNTPs can be exchanged in the combinations dATP + dCTP + dGTP, dATP + dCTP + dTTP, dATP + dGTP + dTTP or dCTP + dGTP + dTTP, or all four dNTPs are exchanged in the combination dATP + dCTP + dGTP + dTTP. It is understood that also analogues of these dNTPs as defined before shall be subsumed under the respective dNTPs. Also possible quantitative combinations thereof shall be subsumed under this combinational scheme.

The term *selection*, *selecting* or *strand specific selection* refers to selecting in ds (double stranded) polynucleotides the strand with at least one modified nucleotide to distinguish the template from the matching strand. The selected strand with the at least one modified nucleotide in ds (double stranded) polynucleotides can be

selectively removed or specifically destroyed, while the strand without modified nucleotide is not impaired preferably throughout its whole length. The removed or destroyed or destroyable strand with the modified nucleotide can be used separately. If the strand with the modified nucleotide is not removed and/or destroyed, it can remain in the same solution with the matching strand. The modified nucleotide in the strand prevents full-length amplification of the strand.

There are several principles by which a selection factor can be used to select between a strand marked with at least one modified nucleotide and a strand with the corresponding unmodified nucleotide or nucleotides. For example, the at least one modified nucleotide has a chemically active group that can be modified after strand-specific synthesis. Another possibility is that the at least one nucleotide carries an affinity group and the selection factor binds selectively to this affinity group. Such am affinity group may for example be recognized by an antibody. A further possibility is that the at least one modified nucleotide carries a group which allows for selective recognition and binding or attachment by another agent which then enables selective removal of the such recognized strand. A further selective factor may be a different sensitivity to specific nucleases. Also the incorporation of ribonucleotides (NTPs) instead of deoxyribonucleotides (dNTPs) is suitable for enabling selection between two strands.

The at least one modified nucleotide may display a different sensitivity to a nuclease, in comparison with the corresponding unmodified nucleotide or nucleotides. Thus such nuclease degrades specifically the at least one modified nucleotide while the unmodified nucleotides are not or only insignificantly affected by such nuclease action. Alternatively, such nuclease degrades specifically the unmodified nucleotides. In both cases the strand containing the such degraded nucleotide or nucleotides has become unfit for amplification or hybridization. Thus both strands can be differentiated, and only one particular strand may be further used.

In a preferred embodiment the selection of strands of polynucleotides is carried out with the modified dNTP being dUTP (deoxyuridine triphosphate). Herein the modified dNTP is dUTP (deoxyuridine triphosphate) as a substitute for dTTP. dUTP displays the same matching characteristics as dTTP, i.e. it is able to match with an opposed adenosine, but not with opposed guanosines, cytosines, thymidines or uridines. dUTP can be specifically degraded by the nuclease Uracil-N-Glycosylase (UNG). UNG is neutral towards dTTP and the other dNTPs or analogues thereof. This enzyme degrades uracil specifically by cleaving the N-glycosylic bond resulting in an abasic site. The such degraded DNA strand is unfit for undergoing appropriate

amplification. At temperatures of over 50°C in slightly alkali conditions nicks occur in the DNA strand at the abasic positions. Additionally, Phusion DNA polymerase preferentially used for amplification (see below) is strongly suppressed by deoxyuridine-containing matrices (Hogrefe, H.H., Hansen, C.J., Scott, B.R. and Nielsen K.B. 2002. Archaeal dUTPase enhances PCR amplifications with archaeal DNA polymerases by preventing dUTP incorporation. Proc Natl Acad Sci USA 99: 596-601). Consequently, only the DNA strand that does not contain dUTP can be amplified properly.

The combination of dUTP as exchange dNTP and UNG as the degrading agent of the exchange dNTP is a preferred embodiment. It is also preferred to replace only dTTP.

In another preferred embodiment the at least one modified nucleotide is thio modified and said nuclease is an exonuclease that degrades thio bonds with low efficiency. Such exonucleases sensitive to thio bonds can be lambda-exonuclease, Exo III nuclease, T4 polymerase and T7 exonuclease.

Other preferred dNTP-specific degradation systems include the use of specific biotinylated dNTPs. Herein NHS esters of biotin are reacted with free amino groups of the respective nucleotide. These biotinylated nucleotides can be recognized by polymerase enzymes as eligible for integration into the actually formed strand. The matching properties to their counter nucleotide on the other strand are comparable. The biotinylated strand might be removed using streptavidin- or avidin-coated magnetic particles. Such specifically attached magnetic particles can be removed by applying a suitable magnetic field.

In another preferred embodiment the at least one modified polynucleotide is an aminoallyl nucleotide (5-(3-aminoallyl)-nucleotide, aa-dNTP). Herein the nucleotide has an aminoallyl group attached to carbon 5 of the pyrimidine ring of uracil and cytosine. This aminoallyl group reacts with dyes from the cyanine series such as HiLyte Fluor® or Alexa Fluor® dyes that also been modified to be amine-reactive. As these dyes use to react also with standard buffer mediums a carbonate buffer is typically used. Herein it is preferred that the modified nucleotide is aminoallyl-dUTP.

In a further preferred embodiment the modified nucleotide can be selectively recognized by an antibody. The antibody binds specifically to this modified nucleotide and the strands are selected by binding or immobilizing said antibody. It is further preferred that such antibody recognizes specifically the modified nucleotide 5-

bromodeoxyuridine. It is also further preferred that such antibody or antibodies recognize specifically at least one modified nucleotide that has been modified by biotinylation.

If the strand carrying at least one modified nucleotide which carries a group which interacts which an affinity factor it is preferred that such strand is immobilized on a solid phase.

In a further preferred embodiment the at least one modified nucleotide is a ribonucleotide. Then it is preferred that said polymerase is a DNA polymerase displaying enhanced ability to incorporate modified substrates such as ribonucleotides (for example, Therminator™ II DNA Polymerase). The such incorporated at least one ribonucleotide is then specifically degraded by a specific nuclease, in this case a RNase.

If at least one classic dNTP is replaced by the corresponding modified dNTP(s) to generate a modified DNA strand the ratio of the at least one modified dNTP shall amount to at least 5% of the corresponding classic dNTP in the solution. Preferred are 20% ratio, more preferred 30%, more preferred 40%, more preferred 50%, more preferred 60%, more preferred 70%, more preferred 80% and particularly preferred more then 90%. Theoretically, only one degraded nucleotide is enough to render the degraded strand unfit for appropriate amplification. This requires that the DNA is long enough that at least one modified dNTP is integrated into the DNA strand. This is a matter of probability (when an even distribution of all four nucleotides is ideally assumed for the original polynucleotides) according to the distribution of the corresponding classic dNTPs and exchange dNTPs. With a length of at least 60 base pairs the probability is close to nil that not at least one modified dNTP has to be integrated into the second-strand cDNA. In order to ensure that also cDNAs with a length less than 60 base pairs always include at least one degradable nucleotide it is preferred that a sufficiently large percentage of the respective nucleoside triphosphate is modified.

If an amplification step is carried out in any of the embodiments of this invention an enzyme with polymerase activity is needed which is able to use the selected cDNA strand as a matrix. In theory, any such enzyme with polymerase activity is suitable for this step. However, those enzymes that are used in conventional PCR protocols are preferred. They should grant a high reliability and a high throughput under conventional PCR conditions.

To preserve information about the direction of the strand selected from a polynucleotide duplex for the further amplification., it is essential to mark differently the 5' and 3' ends of the said selected strand. This way the polarity of the original strand is preserved even after amplification. This may be achieved by for example the following ways:

1) by ligation of Y-shaped adapters to ds polynucleotide with differentiating strands, where Y-shaped adapters comprise a complementary double stranded sequence (e.g. ~13nt long) which ligates to the ds DNA, and further comprises at least one single stranded tail (e.g. 20nt long) of a different sequence. After ligation of such adapters 5'ends of each strand in ds cDNA fragment are specifically marked with the same sequence, different from that attached to the 3' ends. The adapter sequences correspond to primer sequences used later in the procedure. After selection of the strand for amplification, where the other one is either removed, or destroyed or incapable to be amplified due to sensitivity of the polymerase used to the at least one modified nucleotide, the amplification is carried out with the selected strand, wherein 5` and 3' ends bear different sequences for PCR primers (Fig. 2A-1).
2) By ligation of directional adapters (Fig.2A-2)
3) by ligation of two types of at least partly ds adapters to the ds polynucleotide; wherein one of the strands of one of adapters bears at least one biotinylated nucleotide. Through the biotin group, ds DNA with at least one biotinylated adapter is attached to streptavidine coated beads, and the non-biotinylated strand is washed away, leaving on the beads only biotinylated strand which may be with and without modified nucleotides. Then the strand for amplification is selected, while the other one is either removed, or destroyed or incapable to be amplified due to sensitivity of the polymerase used to the at least one modified nucleotide, and amplified with primers corresponding to the adapters sequences. Since the biotinylated adapter is known (whether it is attached to 5' or 3' end of the selected strand), the polarity of the selected strand is preserved after amplification (Figure 2A-3). Instead of biotin, it is possible to use any chemical group capable to bind to some surface and not interfering with amplification.

The term *PCR* refers to polymerase chain reaction. This is a general term comprising all protocols, conditions, automatized and half-automatized devices in which at least one DNA template is amplified by default by means of a enzyme with polymerase activity. In general, such molecules with polymerase activity are enzymes such as

Taq DNA polymerase, Pfu DNA polymerase, Phusion DNA polymerase, wherein Taq and Phusion DNA Polymerases are preferred.

It is particularly preferred that the at least one modified nucleotide is dUTP and said amplification is a Pfu DNA polymerase-dependent amplification.

If an amplification step is carried out in any of the embodiments of this invention a useful PCR (polymerase chain reaction) protocol shall be used by means of which the remaining intact DNA strand shall be amplified. Suitable PCR protocols of the include the known protocols for Standard PCR, Allele-Specific PCR, Alu PCR, Assembly PCR, Asymmetric PCR, Helicase Dependent Amplification HDA, Hot Start PCR, Inverse PCR, Intersequence-Specific PCR (ISSR PCR), Late-PCR, Ligation-mediated PCR, Long PCR, Methylation-Specific PCR, Miniprimer PCR, Multiplex PCR, Nested PCR, Overlap-Extension PCR, Quantitative PCR, Real Time PCR, Solid Phase PCR, Thermal Assymetric Interlaced PCR (TAIL-PCR), Touchdown PCR or combinations thereof. Preferred is the use of Standard PCR.

Herein the term "at least one intact DNA strand" means that the number of the intact DNA strands to be amplified can vary from one to many billions of different polynucleotides.

In a further optional step the resulting sequences can be aligned with already known sequence patterns of the corresponding DNA, if applicable. Such results can be plotted in a way such as shown in Fig. 6.

Optionally, the method according to the invention comprises the step of contig (contiguous) construction from the obtained sequences. Herein the sequences of different overlapping (contiguous) DNA fragments derived from the same source can be used for (re-)constructing the original respectively complete DNA sequence of this source.

Alternatively to the amplification step the at least one intact DNA strand can be hybridized to a nucleic acid structure displaying an antisense sequence of the at least one intact DNA strand. Suitable hybridization methods include *in situ* hybridization (hybridization of a DNA to a complementary sequence in isolated cells and/or tissues), Southern Blot (a method for verifying the presence of a DNA sequence in a DNA sample by separating the DNAs in the sample by size using gel electrophoresis and bringing said sample to a filter membrane for probe hybridization), Northern Blot (the equivalent of Southern Blot with RNA samples) and Southwestern Blot

(identifying the presence of a DNA sequence in a DNA sample prepared according to the Southern blot method by hybridizing the sample on a nitrocellulose gel with characterizing DNA-binding proteins) microarray hybridization and correspondent "reverse" hybridizations were intact DNA's are loaded on a membrane and hybridised with particular probes.

When an amplification is intended the key steps of the inventive method are to be read in the following order:

I) Providing polynucleotide templates, primers, buffer solution(s), at least one enzyme with polymerase activity, sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof, wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection;

II) Performing first-strand synthesis on the polynucleotide templates, so that double-stranded polynucleotides each consisting of a template and a matching strand are generated, wherein the matching strand contains the at least one modified nucleotide;

III) If said polynucleotide templates on step I) are single-stranded RNA, second-strand synthesis with unmodified dNTPs or corresponding analogues thereof is performed;

IV) Mark differently the 5' and 3' ends of the ds cDNA to preserve the polarity of the original strand after amplification;

V) Optionally, performing one or more molecular cloning operations on the double- stranded DNA, wherein the molecular cloning operations are selected from shearing, restriction, digestion, end-polishing, A-tailing, G-tailing and adaptor ligation;

VI) Amplifying the at least one remaining intact cDNA strand by performing PCR, whereby the other strand is removed, or destroyed or incapable to be amplified due to the at least one modified nucleotide.

The afore-mentioned method further comprises preferably the step of:

VII) Specific selection of at least one of the strands of the said polynucleotide duplex using different interaction of modified and unmodified nucleotides with at least one selection factor.

In the particularly preferred embodiment 1 of the method according to the invention a strand specific cloning of RNAs such as a whole transcriptome of a cell or a tissue shall be carried out. Therefore, at least one single strand RNA is provided, for example by isolation from the respective cell or tissue. The four dNTPs dATP, dCTP,

dGTP and dTTP are added in sufficient amounts, i.e. amounts which allow performance of the inventive method, to the solution containing the at least one RNA strand. Then, a first cDNA strand is generated by using random primers (or oligo(dT), oligo (dU), oligo (U) primers) for annealing to said at least one RNA strand and an enzyme with Reverse Transcriptase activity. In the following at least one of the aforesaid dNTPs is replaced by at least one corresponding modified dNTP. For example, the modified dNTP may be a dNTP which can be specifically degraded by adding a degrading agent in order that the strand containing the such degraded at least one modified dNTP is not suitable anymore for amplification, or the modified dNTP may expose the such marked cDNA strand for being removed by interaction with specific agents interacting with the additional functional group or groups on the at least one modified dNTP. With the four dNTPs including the at least one modified dNTP second strand cDNA synthesis is performed by using the first cDNA strand as a matrix and a suitable enzyme with polymerase activity. In the next step polarity specific adaptors are ligated to the ends of cDNA strands, allowing a polymerase to dock to the adaptor structure for PCR amplification. After the ligation the second cDNA strand containing the at least one modified dNTP is for example degraded by degrading the at least one modified dNTP with a suitable degrading agent, or the at least one second cDNA strand containing the at least one modified dNTP is separated from the at least one first cDNA strand by special molecules or particles attaching to characteristic structures on the modified dNTP. Finally, the remaining intact first cDNA strand is amplified by PCR using a suitable enzyme with polymerase activity.

Thus the steps of this embodiment of the inventive method can be summarized in the following schematic order:

1) Providing at least one single-stranded RNA molecule;
2) Adding sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof;
3) Generating a first cDNA strand;
4) Replacing at least one dNTP with at least one corresponding modified dNTP;
5) Performing second-strand synthesis on the first cDNA strand;
6) Ligating polarity-specific adaptors to one or both ends of the cDNA double strand;
7) Specifically modifying or removing the second cDNA strand containing the at least one modified dNTP; and
8) Amplifying the at least one remaining intact cDNA strand by performing PCR.

In embodiment 2 strand specific cloning of RNA can be achieved also by varying the steps of the inventive method. In this embodiment the first cDNA strand shall contain the at least one modified dNTP. In the replacement step this at least one modified dNTP is substituted by the corresponding classic dNTP. Thus in this embodiment the at least one first cDNA strand will be either degraded or removed. The amplification step is then performed on the remaining intact second cDNA strand. All other steps are adjusted correspondingly.

The steps of this embodiment of the inventive method can be summarized in the following schematic order:

1) Providing at least one single-stranded RNA molecule;
2) Adding sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof, wherein at least one of the four dNTPs is a modified dNTP;
3) Generating a first cDNA strand;
4) Replacing the at least one modified dNTP with the corresponding dNTPs or analogues of the corresponding dNTPs;
5) Performing second-strand synthesis on the first cDNA strand;
6) Ligating polarity-specific adaptors to one or both ends of the cDNA double strand;
7) Specifically modifying or removing the first cDNA strand containing the at least one modified dNTP; and
8) Amplifying the at least one remaining intact cDNA strand by performing PCR.

In embodiment 3 strand specific selection according to the invention can also be achieved using single-stranded DNA as a template. Here at least one DNA single strand is provided. A cDNA strand is synthesized wherein at least one modified dNTP is integrated. In this embodiment only one cDNA strand is synthesized since a second cDNA strand would be identical to the original DNA single strand. Finally, the original single-stranded DNA shall be amplified. Thus all transcripts resulting from the PCR will bear the same reading direction of the sequence as the original single-stranded DNA. The other steps are adjusted accordingly.

This leads to the following schematic order:

1) Providing at least one single-stranded DNA molecule;
2) Adding sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof, wherein at least one of the

four dNTPs is a modified dNTP used for the strand specific differentiation;

3) Generating a cDNA strand;

4) Ligating polarity-specific adaptors to one or both ends of the resulting DNA double strand;

5) Specifically modifying or removing the cDNA strand containing the at least one modified dNTP; and

6) Amplifying the at least one intact DNA strand by performing PCR.

In embodiment 4 the purpose of the inventive method is not an amplification but a hybridization. This hybridization can occur against any structure carrying a corresponding antisense sequence to the at least one intact first cDNA strand (see below). For a hybridization without amplification no adaptors have to be ligated. Consequently, this step is omitted in this embodiment and the amplification step is replaced by a hybridization step. Thus the order of the steps of the inventive method reads as follows:

1) Providing at least one single-stranded RNA molecule;

2) Adding sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof;

3) Generating a first cDNA strand;

4) Replacing at least one dNTP with at least one corresponding modified dNTP used for the strand specific differentiation;

5) Performing second-strand synthesis on the first cDNA strand;

6) Specifically modifying or removing the second cDNA strand containing the at least one modified dNTP; and

7) Hybridizing the at least one remaining intact cDNA strand to a suitable antisense DNA sequence.

Embodiment 5 varies embodiment 4 in so far that herein the first cDNA strand contains the at least one modified dNTP. Thus the remaining intact second cDNA strand is hybridized. Consequently, the sequence against which the remaining intact second cDNA strand is hybridized must display an antisense sequence of the second strand. The steps of the inventive method are adjusted according to embodiments 2 and 4 and read as follows:

1) Providing at least one single-stranded RNA molecule;

2) Adding sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof, wherein at least one of the four dNTPs is a modified dNTP used for the strand specific differentiation;

3) Generating a first cDNA strand;
4) Replacing the at least one modified dNTP with the corresponding dNTPs or analogues of the corresponding dNTPs;
5) Performing second-strand synthesis on the first cDNA strand;
6) Specifically modifying or removing the first cDNA strand containing the at least one modified dNTP; and
7) Hybridizing the at least one remaining intact cDNA strand to a suitable antisense DNA sequence.

In embodiment 6 the hybridization step follows after an amplification step. In various hybridization modes it may be advantageous to hybridize not only one or a few copies of the respective intact cDNA strand, but a large number as resulting from an amplification. Thus the steps of ligating a polarity-specific adaptor and of an amplification have to be integrated again. The order of the steps reads like embodiment 1 and a subsequent hybridization step:

1) Providing at least one single-stranded RNA molecule;
2) Adding sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof;
3) Generating a first cDNA strand;
4) Replacing at least one dNTP with at least one corresponding modified dNTP;
5) Performing second-strand synthesis on the first cDNA strand;
6) Ligating polarity-specific adaptors to one or both ends of the cDNA double strand;
7) Specifically modifying or removing the second cDNA strand containing the at least one modified dNTP; and
8) Amplifying the at least one remaining intact cDNA strand by performing PCR.
9) Hybridizing the amplified cDNA to a suitable antisense DNA sequence.

Correspondingly, in embodiment 7 the hybridization step can ensue after performing an amplification of the intact second-strand cDNA, as in embodiment 2:

1) Providing at least one single-stranded RNA molecule;
2) Adding sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof, wherein at least one of the four dNTPs is a modified dNTP;
3) Generating a first cDNA strand;
4) Replacing the at least one modified dNTP with the corresponding dNTPs or analogues of the corresponding dNTPs;

5) Performing second-strand synthesis on the first cDNA strand;
6) Ligating polarity-specific adaptors to one or both ends of the cDNA double strand;
7) Specifically modifying or removing the first cDNA strand containing the at least one modified dNTP; and
8) Amplifying the at least one remaining intact cDNA strand by performing PCR.
9) Hybridizing the amplified cDNA to a suitable antisense DNA sequence.

Also in the case of single-stranded DNA it may be advantageous to combine an amplification and a hybridization step. Thus in embodiment 8 the order of the steps reads as follows:

1) Providing at least one single-stranded DNA molecule;
2) Adding sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof, wherein at least one of the four dNTPs is a modified dNTP used for the strand specific differentiation;
3) Generating a cDNA strand;
4) Ligating polarity-specific adaptors to one or both ends of the resulting DNA double strand;
5) Specifically modifying or removing the cDNA strand containing the at least one modified dNTP; and
6) Amplifying the at least one intact DNA strand by performing PCR.
7) Hybridizing the amplified DNA to a suitable antisense DNA sequence.

Embodiment 9: If the sequence of the original RNA is not known it may be desirable to perform sequencing. Thus such a step is added to the step order of embodiment 1:

1) Providing at least one single-stranded RNA molecule;
2) Adding sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof;
3) Generating a first cDNA strand;
4) Replacing at least one dNTP with at least one corresponding modified dNTP;
5) Performing second-strand synthesis on the first cDNA strand;
6) Ligating polarity-specific adaptors to one or both ends of the cDNA double strand;
7) Specifically modifying or removing the second cDNA strand containing the at least one modified dNTP; and

8) Optionally, amplifying the at least one remaining intact cDNA strand by performing PCR.
9) Sequencing the resulting copies of the at least one intact cDNA strand.

Embodiment 10: Consequently, it is also possible to include a sequencing step if the first-stranded included the at least one modified dNTP and the intact second-strand cDNA is amplified, as in embodiment 2:

1) Providing at least one single-stranded RNA molecule;
2) Adding sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof, wherein at least one of the four dNTPs is a modified dNTP;
3) Generating a first cDNA strand;
4) Replacing the at least one modified dNTP with the corresponding dNTPs or analogues of the corresponding dNTPs;
5) Performing second-strand synthesis on the first cDNA strand;
6) Ligating polarity-specific adaptors to one or both ends of the cDNA double strand;
7) Specifically modifying or removing the first cDNA strand containing the at least one modified dNTP; and
8) Optionally, amplifying the at least one remaining intact cDNA strand by performing PCR.
9) Sequencing the resulting copies of the at least one intact cDNA strand.

Embodiment 11: After sequencing it may be desirable to determine the polarity of the resulting sequence(s). This may serve for several analytic or diagnostic purposes (see below) or for aligning RNA transcripts to gene sequences known from DNA sequencing projects. If the remaining intact first strand cDNA was sequenced the order of the steps of the inventive method read as follows:

1) Providing at least one single-stranded RNA molecule;
2) Adding sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof;
3) Generating a first cDNA strand;
4) Replacing at least one dNTP with at least one corresponding modified dNTP;
5) Performing second-strand synthesis on the first cDNA strand;
6) Ligating polarity-specific adaptors to one or both ends of the cDNA double strand;
7) Specifically modifying or removing the second cDNA strand containing the at least one modified dNTP; and

8) Optionally, amplifying the at least one remaining intact cDNA strand by performing PCR.
9) Sequencing the resulting copies of the at least one intact cDNA strand.
10) Aligning of sequencing reads to the reference genome or reconstructing of the RNA sequence from individual sequencing reads.
11) Determining the polarity of the at least one RNA transcript.

Embodiment 12: Consequently, such a determination of the polarity and optionally an alignment to a gene structure known from a sequencing project can also be performed if the remaining intact second strand cDNA was amplified and sequenced. As in this case the antisense sequence is resulting the sequence has to be inverted, preferably by means of a computer programme, into the correct reading sequence.

Thus the order of the steps of the inventive method is:
1) Providing at least one single-stranded RNA molecule;
2) Adding sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof, wherein at least one of the four dNTPs is a modified dNTP;
3) Generating a first cDNA strand;
4) Replacing the at least one modified dNTP with the corresponding dNTPs or analogues of the corresponding dNTPs;
5) Performing second-strand synthesis on the first cDNA strand;
6) Ligating polarity-specific adaptors to one or both ends of the cDNA double strand;
7) Specifically modifying or removing the first cDNA strand containing the at least one modified dNTP; and
8) Optionally, amplifying the at least one remaining intact cDNA strand by performing PCR;
9) Sequencing the resulting copies of the at least one intact cDNA strand;
10) Aligning of sequencing reads to the reference genome or reconstructing of the RNA sequence from individual sequencing reads;
11) Determining the polarity of the at least one RNA transcript

Embodiment 13: In analogy, it may be also desirable to sequence the amplified copies of a single-stranded DNA if the sequence isn't known yet. Thus a sequencing step ensues to the order of the steps of the inventive method of embodiment 3:
1) Providing at least one single-stranded DNA molecule;
2) Adding sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof, wherein at least one of the

four dNTPs is a modified dNTP used for the strand specific differentiation;

3) Generating a cDNA strand;
4) Ligating polarity-specific adaptors to one or both ends of the resulting DNA double strand;
5) Specifically modifying or removing the cDNA strand containing the at least one modified dNTP; and
6) Optionally, amplifying the at least one intact DNA strand by performing PCR.
7) Sequencing the resulting copies of the at least one intact DNA strand.

Embodiment 14: Also for single-stranded DNA it may be desirable to determine the polarity of the resulting copies after amplification and sequencing. As for single-stranded DNA by needs only one cDNA strand is generated including the at least one modified dNTP the resulting sequence must be inverted, as in embodiment 12.

1) Providing at least one single-stranded DNA molecule;
2) Adding sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof, wherein at least one of the four dNTPs is a modified dNTP used for the strand specific differentiation;
3) Generating a cDNA strand;
4) Ligating polarity-specific adaptors to one or both ends of the resulting DNA double strand;
5) Specifically modifying or removing the cDNA strand containing the at least one modified dNTP; and
6) Optionally, amplifying the at least one intact DNA strand by performing PCR;
7) Sequencing the resulting copies of the at least one intact DNA strand;
8) Aligning of sequencing reads to the reference genome or reconstructing of the DNA sequence from individual sequencing reads;
9) Determining the polarity of the at least one DNA molecule.

In another preferred embodiment 15 the original polynucleotide strand is a double-stranded DNA. Herein it may be necessary to set a nick into the particular strand of the double-stranded DNA which shall be substituted. A possible sequence of the steps according to this embodiment is:

1) Providing at least one double-stranded DNA;
2) Setting a nick in one particular strand of the double-stranded DNA by using a nicking endonuclease;

3) Adding sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof, wherein at least one of the four dNTPs is a modified dNTP used for the strand specific differentiation;
4) Performing nick-translation of the nicked DNA strand by means of a DNA polymerase with 5' to 3' exonuclease activity;
5) Specifically modifying or removing the cDNA strand containing the at least one modified dNTP.

In a further preferred embodiment 16 the original polynucleotide strand is a double-stranded DNA obtained as a result of PCR reaction. A special modified nucleotide is introduced into one PCR primer to set a nick into the particular strand of the double-stranded DNA which shall be substituted. A possible sequence of the steps according to this embodiment is:

1) Providing at least one double-stranded DNA obtained in PCR reaction where one of PCR-primers contains dUTP;
2) Setting a nick in one particular strand of the double-stranded DNA by UNG treatment and APE 1 digestion or UNG treatment and alkalic hydrolisis;
3) Adding sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof, wherein at least one of the four dNTPs is a modified dNTP used for the strand specific differentiation;
4) Performing nick-translation of the nicked DNA strand by means of a DNA polymerase with 5' to 3' exonuclease activity;
5) Specifically modifying or removing the cDNA strand containing the at least one modified dNTP.

Embodiment 17: In another preferred embodiment the method according to the invention is performed in the following way:

1) Providing at least one single-stranded RNA molecule;
2) Adding sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof, wherein at least one of the four dNTPs is a modified dNTP;
3) Generating a first cDNA strand, wherein at least one nucleotide is a modified dNTP;
4) Replacing the at least one modified dNTP with the corresponding dNTPs or analogues of the corresponding dNTPs;
5) Performing second-strand synthesis on the first cDNA strand;

6) Specifically modifying or removing the first cDNA strand containing the at least one modified dNTP; and

Embodiment 18: In another preferred embodiment the method according to the invention is performed in the following way:

1) Providing at least one single-stranded RNA molecule;
2) Adding sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof;
3) Generating a first cDNA strand,
4) Replacing at least one dNTP with at least one corresponding modified dNTP;
5) Performing second-strand synthesis on the first cDNA strand, generating a second cDNA strand, wherein at least one nucleotide is a modified dNTP;
6) Specifically modifying or removing the second cDNA strand containing the at least one modified dNTP.

Embodiment 19: In another preferred embodiment the method according to the invention is performed in the following way:

1) Performing first-strand synthesis on a polynucleotide template by providing polynucleotide template, primers, correspondent buffer, polymerase and sufficient amounts of the four dNTPs: dATP, dCTP, dGTP and dTTP or corresponding analogues thereof, wherein at least one of the four dNTPs is a modified nucleotide used for strand specific selection, so that a double-stranded polynucleotide comprising a template a matching strand containing the at least one modified nucleotide is generated;
2) Optionally, second-strand synthesis with unmodified dNTPs or corresponding analogues thereof, if said polynucleotide template on step I) is a single-stranded RNA;
3) Optionally, performing molecular cloning operations with double-stranded DNA; and
4) Strand specific selection using different interaction of modified and unmodified nucleotides with some selection factor.

The inventive method for selection of a particular strand where the preservation of information about the direction is desired after being transformed into a double-stranded form comprises also the following steps:

I) Providing polynucleotide templates, primers, buffer solution, at least one enzyme with polymerase activity, sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof, wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection;

II) Performing first-strand synthesis on the polynucleotide templates, so that double-stranded polynucleotides each consisting of a template and a matching strand are generated, wherein the matching strand contains the at least one modified nucleotide;

III) If said polynucleotide templates on step I) are single-stranded RNA, second-strand synthesis with unmodified dNTPs or corresponding analogues thereof is performed;

IV) Optionally, performing molecular cloning operations with double-stranded DNA; and

V) Specific selection of at least one of the strands of the said polynucleotide duplex using different interaction of modified and unmodified nucleotides with at least one selection factor.

The inventive method for selection of a single stranded RNA from a double stranded polynucleotide comprises the following steps:

I) Providing RNA templates, primers, buffer solution, at least one enzyme with polymerase activity, sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof, wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection;

II) Performing first-strand synthesis on the RNA templates, so that double-stranded polynucleotides each consisting of a template and a matching strand are generated, wherein the matching strand contains the at least one modified nucleotide;

III) Performing second-strand synthesis with unmodified dNTPs or corresponding analogues thereof;

IV) Optionally, performing molecular cloning operations with double-stranded DNA; and

V) Specific selection of at least one of the strands of the said polynucleotide duplex using different interaction of modified and unmodified nucleotides with at least one selection factor.

The inventive method for selection of a single stranded RNA from a double stranded polynucleotide can also comprise the following steps:

I) Providing RNA templates, primers, buffer solution, at least one enzyme with polymerase activity, sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof, wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection;

II) Performing first-strand synthesis on the RNA templates with unmodified dNTPs or corresponding analogues thereof;

III) Performing second-strand synthesis wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection, so that double-stranded polynucleotides each consisting of a template and a matching strand are generated, wherein the template contains the at least one modified nucleotide;

IV) Optionally, performing molecular cloning operations with double-stranded DNA; and

V) Specific selection of at least one of the strands of the said polynucleotide duplex using different interaction of modified and unmodified nucleotides with at least one selection factor.

The inventive method for selection of a single stranded DNA from a double stranded polynucleotide comprises the following steps:

I) Providing DNA templates, primers, buffer solution, at least one enzyme with polymerase activity, sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof, wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection;

II) Performing first-strand synthesis on the DNA templates, so that double-stranded polynucleotides each consisting of a template and a matching strand are generated, wherein the matching strand contains the at least one modified nucleotide;

III) Optionally, performing molecular cloning operations with double-stranded DNA; and

VI) Specific selection of at least one of the strands of the said polynucleotide duplex using different interaction of modified and unmodified nucleotides with at least one selection factor.

The inventive method for selection of a single stranded RNA from a double stranded DNA/RNA hybrid comprises the following steps:

I) Providing DNA/RNA templates, primers, buffer solution, at least one enzyme with polymerase activity, sufficient amounts of the four

dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof, wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection;

II) Performing first-strand synthesis on the RNA templates, so that double-stranded polynucleotides each consisting of a template and a matching strand are generated, wherein the matching strand contains the at least one modified nucleotide;

III) Performing second-strand synthesis with unmodified dNTPs or corresponding analogues thereof;

IV) Optionally, performing molecular cloning operations with double-stranded DNA; and

V) Specific selection of at least one of the strands of the said polynucleotide duplex using different interaction of modified and unmodified nucleotides with at least one selection factor.

The inventive method for selection of a single stranded RNA from a double stranded DNA/RNA hybrid comprises the following steps:

I) Providing DNA/RNA templates, primers, buffer solution, at least one enzyme with polymerase activity, sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof, wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection;

II) Performing first-strand synthesis on the RNA templates with unmodified dNTPs or corresponding analogues thereof;

III) Performing second-strand synthesis wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection, so that double-stranded polynucleotides each consisting of a template and a matching strand are generated, wherein the template contains the at least one modified nucleotide;

IV) Optionally, performing molecular cloning operations with double-stranded DNA; and

V) Specific selection of at least one of the strands of the said polynucleotide duplex using different interaction of modified and unmodified nucleotides with at least one selection factor.

The inventive method for selection of a single stranded DNA from a double stranded DNA/RNA hybrid comprises the following steps:

I) Providing DNA/RNA templates, primers, buffer solution, at least one enzyme with polymerase activity, sufficient amounts of the four

dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof, wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection;

II) Performing first-strand synthesis on the DNA templates, so that double-stranded polynucleotides each consisting of a template and a matching strand are generated, wherein the matching strand contains the at least one modified nucleotide;

III) Optionally, performing molecular cloning operations with double-stranded DNA; and

VI) Specific selection of at least one of the strands of the said polynucleotide duplex using different interaction of modified and unmodified nucleotides with at least one selection factor.

The inventive method for selection of a single stranded DNA from a double stranded DNA/RNA hybrid comprises the following steps:

I) Providing DNA/RNA templates, primers, buffer solution, at least one enzyme with polymerase activity, sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof, wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection;

II) Performing first-strand synthesis on the DNA templates with unmodified dNTPs or corresponding analogues thereof;

III) Performing second-strand synthesis wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection, so that double-stranded polynucleotides each consisting of a template and a matching strand are generated, wherein the template contains the at least one modified nucleotide;

III) Optionally, performing molecular cloning operations with double-stranded DNA; and

VI) Specific selection of at least one of the strands of the said polynucleotide duplex using different interaction of modified and unmodified nucleotides with at least one selection factor.

The inventive method for selection of a single stranded DNA from a double stranded polynucleotide can also comprise the following steps:

I) Providing DNA templates, primers, buffer solution, at least one enzyme with polymerase activity, sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof,

wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection;

II) Performing first-strand synthesis on the DNA templates with unmodified dNTPs or corresponding analogues thereof;

III) Performing second-strand synthesis wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection, so that double-stranded polynucleotides each consisting of a template and a matching strand are generated, wherein the template contains the at least one modified nucleotide;

IV) Optionally, performing molecular cloning operations with double-stranded DNA; and

V) Specific selection of at least one of the strands of the said polynucleotide duplex using different interaction of modified and unmodified nucleotides with at least one selection factor.

The inventive method for selection of a particular strand from a double stranded polynucleotide can further also comprise or consist of the following steps:

I) Providing polynucleotide templates, primers, buffer solution, at least one enzyme with polymerase activity, sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof, wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection;

II) Performing first-strand synthesis on the polynucleotide templates with unmodified dNTPs or corresponding analogues thereof;

III) Performing second-strand synthesis wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection, so that double-stranded polynucleotides each consisting of a template and a matching strand are generated, wherein the template contains the at least one modified nucleotide;

IV) Optionally, performing molecular cloning operations with double-stranded DNA; and

V) Specific selection of at least one of the strands of the said polynucleotide duplex using different interaction of modified and unmodified nucleotides with at least one selection factor.

The inventive method for selection of a particular strand from a double stranded polynucleotide can further also comprise or consist of the following steps:

I′) Providing polynucleotide templates, primers, buffer solution, at least one enzyme with polymerase activity, sufficient amounts of the four

dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof, wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection;

IIa´) Performing first-strand synthesis on the polynucleotide templates with unmodified dNTPs or corresponding analogues thereof;

or

IIb´) Performing first-strand synthesis on the polynucleotide templates, so that double-stranded polynucleotides each consisting of a template and a matching strand are generated, wherein the matching strand contains the at least one modified nucleotide;

IIIa`) If step IIa´ was performed, second-strand synthesis is performed wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection, so that double-stranded polynucleotides each consisting of a template and a matching strand are generated, wherein the template contains the at least one modified nucleotide;

IIIb´) If step IIb´ was performed and said polynucleotide templates on step I) are single-stranded RNA, second-strand synthesis with unmodified dNTPs or corresponding analogues thereof is performed

and

IV´) Optionally, performing molecular cloning operations with double-stranded DNA; and

V´) Specific selection of at least one of the strands of the said polynucleotide duplex using different interaction of modified and unmodified nucleotides with at least one selection factor.

In a further embodiment the inventive method comprises selecting in ds (double stranded) polynucleotides the strand with at least one modified nucleotide to distinguish the template from the matching strand, wherein the selected strand with the at least one modified nucleotide in ds (double stranded) polynucleotides can be removed or destroyed and the removed strand with the modified nucleotide can be

used separately, whereby if the strand with the modified nucleotide is not removed and destroyed, it can remain in the same solution with the matching strand, wherein the modified nucleotide in the strand prevents the amplification of the strand comprising the following steps:

I) Providing polynucleotide templates, primers, buffer solution, at least one enzyme with polymerase activity, sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof, wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection;

II) Performing first-strand synthesis on the polynucleotide templates, so that double-stranded polynucleotides each consisting of a template and a matching strand are generated, wherein the matching strand contains the at least one modified nucleotide;

III) If said polynucleotide templates on step I) are single-stranded RNA, second-strand synthesis with unmodified dNTPs or corresponding analogues thereof is performed;,;

IV) Optionally, performing molecular cloning operations with double-stranded DNA; and

V) Specific selection of at least one of the strands of the said polynucleotide duplex using different interaction of modified and unmodified nucleotides with at least one selection factor.

It is obvious to a person skilled in the art that the different steps mentioned before in the respective embodiments can be combined for solving similar or different tasks provided the order and the selection of the different steps results in a meaningful order of the steps. All these possible combinations shall fall under the scope and the spirit of the invention.

By way of example, the method according to the invention was applied on the yeast transcriptome (Example 1). In a similar fashion, it was applied on the transcriptome of the mouse brain (Example 2). These two examples were chosen because comparative data from the Applicants' laboratory and other laboratories are available. Also complete genomes of these organisms are published. By way of comparing the results from conventional RNA-Seq and the inventive method according to the invention the advantages of the inventive method become apparent. As shown e.g. in Example 1 and Fig. 6, the inventive method allows a differentiation in sense and antisense transcripts. In this plot sense transcripts are marked in blue and antisense transcripts in red. Thus the method according to the invention can use the maintained

information of the polarity of a transcript to generate more precise gene expression maps. Practical applications are discussed further below.

The reliability of the method according to the invention was validated in several control experiments. First, it was tested to which degree the inventive method can reproduce the results of conventional RNA sequencing. Herein the amplified sense and antisense transcripts according to the invention were pooled for allowing a correlation analysis. The results such as shown in Example 3 and Fig. 7A show a high correlation between the two methods. A high correlation corresponds to a distribution close to unity. A low correlation would be represented by a widely scattered distribution pattern. Thus the inventive method does not disturb the overall expression and transcriptome pattern in a sample. So the results of both methods can be compared to one another.

Next, two different strand-specific expression profilings were performed according to the invention on the same sample. Herein, only sense transcripts were included. The results were plotted again as described before. An excellent reproducibility of the inventive method is given, as can be seen in Example 4 and Fig. 7B.

In a further control experiment, the combination of sense and antisense transcripts obtained were compared to the results from another laboratory. There is still a very high correlation between the results, although with a somewhat broader scatter pattern (see Example 4 and Fig. 7C). Such a variation nearly always occurs when comparing the results from different laboratories. This experiment shows that this method does not work only under the specific conditions of the Applicants' laboratory, but can be universally used.

In order to test whether sense and antisense transcripts obtained according to the inventive method can be appropriately differentiated sense and antisense transcripts were compared and depicted in such an aforesaid scatter plot. As can be seen in Fig. 7D there is nearly no correlation. Thus sense and antisense transcripts obtained according to the inventive method are unequivocally different. This is a proof that the inventive method is highly suitable for differentiating accurately between sense and antisense transcripts.

The stability of the inventive method in respect of a complete exchange of at least one dNTP by at least one corresponding degradable analogue was tested. For this purpose in one reaction dTTP was exchanged by 100% dUTP. In another reaction dTTP was exchanged only by 75% dUTP, with 25% dTTP remaining in the reaction

batch. The results from both experiments were compared for the sense transcripts and antisense transcripts (see Example 5 and Fig. 8). There is no significant deterioration in correlation when the exchange of the at least one dNTP was incomplete. When extrapolating these results it appears that only a part of the at least one dNTP must be replaced by the at least one corresponding degradable analogue to keep the method going with acceptable results, as shown in the correlation coefficient.

Actinomycin D is a substance which selectively inhibits DNA synthesis from DNA templates, but not from RNA templates. In a further experiment it could be shown that the presence of actinomycin D does not deteriorate significantly the results from the inventive protocol, neither for sense nor for antisense transcripts (see Example 6 and Fig. 9).

Practical applications of the inventive method include but are not limited to the provision of suitable kits for performing the inventive method according to the different protocols (embodiments) disclosed above.

Fig. 10 presents examples where information about transcription orientation helps to resolve overlaps of annotated and new genes in both yeast and mouse transcriptomes.

A kit in molecular biology is a package which includes all necessary ingredients for performing a certain method or singular step. Standard chemicals as present in any standard molecular biology laboratory are normally not included. Nevertheless some of these standard chemicals may be indispensable to carry out the inventive method properly. It is understood that all ingredients are provided in quantities that allow for a proper execution of the inventive method for the majority of scientific, diagnostic and industrial applications.

Often, but not always, these ingredients are provided in already prepared solutions ready- or close to ready-for-use. There may be also combinations of different ingredients already added together. A further advantage is that such kits use to be verified. Therefore the operator doesn't have to prove again the viability of the method and can save on at least some control experiments. Therefore kits are a very popular tool in molecular biology laboratories in research, diagnostics and industry.

For carrying out the key steps of the method according to the invention such a kit shall include the following components:

A) All four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof, additionally at least one modified dNTP used for the strand specific selection;
B) Primers
C) One or more DNA polymerases;
D) One or more degrading agents corresponding to the modified dNTPs and/or one or more removing agents apt for removing the strand containing the modified dNTPs.

If the protocol of the inventive method that shall be carried out includes an amplification step another embodiment of the kit should include:

A) All four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof, additionally at least one modified dNTP used for the strand specific selection;
B) Primers
C) One or more DNA polymerases;
D) One or more degrading agents corresponding to the modified dNTPs and/or one or more removing agents apt for removing the strand containing the modified dNTPs.
E) One or more polarity-specific adaptor.

If the protocol of the inventive method that shall be carried out includes an amplification step and the original polynucleotide is a RNA another embodiment of the kit should include:

A) All four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof, additionally at least one modified dNTP used for the strand specific selection;
B) Primers
C) One or more DNA polymerases;
D) One or more degrading agents corresponding to the modified dNTPs and/or one or more removing agents apt for removing the strand containing the modified dNTPs.
E) One or more polarity-specific adaptor;
F) One or more Reverse transcriptases.

In any of the aforegoing kits according to the invention it may be preferable to include additionally components for removal of unmodified nucleotides by gel-filtration or purification on silica-column. Further kits are also preferred:

KIT A:

A) All four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof, additionally at least one modified dNTP used for the strand specific selection;
B) Primers
C) One or more DNA polymerases;
D) One or more degrading agents corresponding to the modified dNTPs and/or one or more removing agents apt for removing the strand containing the modified dNTPs;
E) Components for specific removal of at least one modified or at least one unmodified nucleotide by gel-filtration or purification on silica-column.

KIT B:

A) All four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof, additionally at least one modified dNTP used for the strand specific selection;
B) Primers
C) One or more DNA polymerases;
D) One or more degrading agents corresponding to the modified dNTPs and/or one or more removing agents apt for removing the strand containing the modified dNTPs.
E) One or more polarity-specific adaptor.
F) Components for specific removal of at least one modified or at least one unmodified nucleotide by gel-filtration or purification on silica-column.

KIT C:

A) All four dNTPs dATP, dCTP, dGTP and dTTP or corresponding analogues thereof, additionally at least one modified dNTP used for the strand specific selection;
B) Primers
C) One or more DNA polymerases;
D) One or more degrading agents corresponding to the modified dNTPs and/or one or more removing agents apt for removing the strand containing the modified dNTPs.
E) One or more polarity-specific adaptor;
F) One or more Reverse transcriptases.
G) Components for specific removal of at least one modified or at least one unmodified nucleotide by gel-filtration or purification on silica-column.

KIT D:

A kit for performing differentiation of polynucleotide strands comprising:

A) All four dNTPs dATP, dCTP, dGTP and dTTP and corresponding modified deoxynucleotide;

B) Primers;

C) One or more DNA polymerases;

D) One or more selection factors corresponding to the modified deoxynucleotide and/or one or more selective media for removing the strand containing the modified nucleotide;

E) One or more DNA polymerases with reverse transcriptase activity;

Optionally comprising:

F) Components for specific removal of at least one modified or at least one unmodified nucleotide by gel-filtration or purification on silica-column.

KIT E:

A kit for ds cDNA synthesis from RNA preserving information about direction of original RNA strand comprising:

A) All four dNTPs dATP, dCTP, dGTP and dTTP/dUTP and at least one modified nucleotide corresponding to at least one from dATP,ATP, dCTP, CTP, dGTP, GTP, dTTP, TTP, dUTP and UTP;

B) Primers;

C) One or more DNA polymerases with reverse transcriptase activity;

D) One or more DNA polymerases; and

E) One or more selection factors suitable to specifically interact with said at least one modified nucleotide and/or one or more selective media for removing the strand containing the modified nucleotide;

Or additionally comprising

F) Components for specific removal of at least one modified or at least one unmodified nucleotide by gel-filtration or purification on silica-column.

To prevent spurious second-strand cDNA synthesis, which was shown to be the major source of artifactual antisense transcripts. Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. Actinomycin D was included in the reverse transcription reaction. Actinomycin D specifically inhibits DNA-dependent, but not RNA-dependent, DNA synthesis. Actinomycin D is included in the kit to make the reaction more reproducible.

Thus in a preferred embodiment the kit also includes an inhibitor of DNA-dependent synthesis characterized in that said inhibitor does inhibit DNA-dependent, but not RNA-dependent, DNA synthesis.

Another application is the use of the results of this method for improved quantitative and qualitative diagnosis tests in selected diseases, as a part of the treatment or for prophylactic purposes. In general every disease is eligible which includes a temporary or permanent change in RNA expression levels. It is also possible to obtain a better monitoring of the effects of pharmaceutical agents used for therapy and/or prophylaxis of diseases provided these pharmaceutical agents produce a characteristic expression pattern different from that in untreated patients (humans or animals). Also improved gene expression patterns of plant, fungi, bacteria and viruses can be obtained by this method when the polarity of transcripts is available.

A further application is the production of improved and more selective microarrays. When the polarity of transcripts is known only the transcripts (respectively the antisense sequence) with a correct reading direction are meaningfully included in the microarray so that the limited space on the microarray is more economically used. Respectively, microarrays used for the generation of characteristic patterns instead of individual results may obtain a higher resolution by omitting non-characteristic transcripts of the incorrect reading direction.

As shown before, the advantages of the procedure according to the invention are providing a simple, inexpensive and reliable method for strand-specific cDNA cloning. This method is a novel and easy approach for determining the polarity of transcripts and using this information for purposes as outlined above. In particular, there is no need for the use of promoters or primer constructs containing promoter regions. The essence of the invention is the preservation of information about the direction of the original single-stranded molecules after being transformed into a double-stranded form or of particular strands in the original double stranded molecule and selection of the original strand according to the method of the invention. . In a preferred embodiment the single-stranded molecule is a RNA and the double-stranded form a DNA.

Moreover, the method according to the invention has a higher accuracy in measuring expression levels of genes compared to RNA-Seq. Ignoring orientation of transcripts leads for example to overestimation of the expression level for about 10% of the mouse and about 15% of the yeast genes with noticeable expression level.

A further advantage of the method according to the invention is that the RNA is transcribed in cDNA and all the following steps are carried out with the cDNA which is much more resistant to degradation compared to RNA. In particular, strand specific differentiation is achieved by modifying a nucleotide in the cDNA and not in the RNA.

Figures

Fig. 1: Directional double-stranded DNA

(A) original single-stranded polynucleotide; (B) directional double-stranded DNA with distinguishable adaptors (state of the art); (C) directional double-stranded DNA with distinguishable strands; (D) directional double-stranded DNA with distinguishable strands and distinguishable adapters.

Fig. 2A: Conversion of distinguishable strands into distinguishable adaptors

(A1) Y-shaped adaptors are ligated to double-stranded DNA with distinguishable strands; modified strands are separated from unmodified ones; after synthesis of complementary strands or PCR-amplification double-stranded DNA with distinguishable adapters is generated.

(A2) directional adaptors are ligated to double-stranded DNA with distinguishable strands; modified strands are separated from unmodified ones; after PCR-amplification double-stranded DNA with distinguishable adapters is generated.

(A3) adaptors are ligated to double-stranded DNA with distinguishable strands, with one of adaptors bearing biotin in one of the strands. Streptavidine coated beads are used to fish out fragments, containing biotin. Uridine-containing strand is destroyed, remaining strand is amplified.

Fig. 2B: Conversion of distinguishable adaptors into distinguishable strands

(B1) nick introduced into one of the adaptors using nicking endonuclease;

(B2) dUTP in one of the adaptors is converted into a nick: Uracil-N-Glycosylase converts dUTP into an abasic site and APE 1 nuclease converts the abasic site into a nick;

(B3) double-stranded DNA is destabilized in one of the Uracil-N-Glycosylase; new primer is annealed in the destabilized region;

(B4) position for annealing of the primer is organized by action 5' to 3' exonuclease; another adapter is protected from action of 5' to 3' exonuclease by thio-nucleotides; nick-translation is performed by DNA polymerase I (or Taq polymerase) in presence of modified nucleotides.

Fig. 3A: Preparation of single-stranded DNA according to the inventive method using dUTP as a modified nucleotide

* double-stranded DNA is a product of PCR reaction where one of the PCR-primers has dUTP at 3'-part;
* dUTP in the PCR primer is converted into a nick: Uracil-N-Glycosylase converts dUTP into an abasic site and APE 1 nuclease converts the abasic site into a nick;
* nick-translation is performed by DNA polymerase I (or Taq polymerase) in presence of dUTP;
* dUTP-containing strand is destroyed by Uracil-N-Glycosylase digestion and alkali hydrolysis;
* single-stranded DNA is purified by silica-based purification or gel-filtration or electrophoretic purification.

Fig. 3B: Preparation of single-stranded DNA according to the inventive method using biotinylated nucleotides

* the double-stranded DNA is a product of PCR reaction where one of the PCR-primers has dUTP at 3'-part;
* dUTP in the PCR primer is converted into a nick: Uracil-N-Glycosylase converts dUTP into an abasic site and APE 1 nuclease converts the abasic site into a nick;
* nick-translation is performed by DNA polymerase I (or Taq polymerase) in presence of biotin-associated dTTPs and/or biotin-associated dUTPs;
* strands are separated using streptavidine-coated particles;

Fig. 4: Flowchart of the inventive method

RNA is shown in red colour, DNA in green. Arrows are in 5' to 3' direction.

Fig. 5: Schematic view of the inventive method on the level of base pairs

Fig. 6: Expression profile of a yeast gene

(A) A conventional transcriptional profile of the yeast YGR203W gene without orientation is shown. Vertical lines mark the boundaries of the YGR203W gene as they were determined previously

(B) A transcriptional profile of the yeast YGR203W gene with orientation according to the inventive method is depicted. Reads mapped in the forward direction are shown in blue, in the reverse direction in red.

Fig. 7: Scatter plots comparing mouse mRNA expression data (number of reads in annotated genes)

(A) The same mouse liver sample, strand-specific (according to the method, X-axis) and strand-unspecific (RNA-Seq, Y-axis) protocols (Pearson correlation coefficient (cc) = 0.999). A distribution close to unity would corresponds to a high correlation in these scatter plots.

(B) results according to the method for two biological replicas (mouse whole brain mRNA); cc=0.990.

(C) Mouse whole brain expression data according to the method (X-axis) and data from Mortazavi et al. 2008 (Y-axis); cc=0.817.

(D) Sense (X-axis) and antisense (Y-axis) expression in mouse brain. Thus the inventive method allows for a good differentiation of sense and antisense transcripts.

Fig. 8: Comparison of ssRNA-Seq results for protocols using 75%/25% dUTP/dTTP mixture and 100% dUTP

(A) Scatter plot of levels of sense transcription.

(B) Scatter plot of levels of antisense transcription.

Still a good correlation can be reached. Thus the inventive method does not depend on a complete exchange of at least one dNTP by a degradable analogue.

Fig. 9: Comparison of ssRNA-Seq results for protocols with and without actinomycin D added to the reverse transcription reaction

(A) Scatter plot of levels of sense transcription.

(B) Scatter plot of levels of antisense transcription.

X-Axe: with actinomycin D; Y-Axe: without actinomycin D. The use of actinomycin D does not bias significantly the correlation when using sense strands.

Fig. 10: Examples of overlapping genes and novel gene-candidates

A. *S. cerevisiae* transcriptome. Overlapping tails of two genes:

B. *S. cerevisiae* transcriptome. Overlapping of three genes:

C. *S. cerevisiae* transcriptome. Novel gene-candidate between

D. Mouse whole brain transcriptome. Overlapping genes: Ncaph2 and Ecgf1.

E. Mouse whole brain transcriptome. genes: Mrpl24 and BC023814.

F. Mouse whole brain transcriptome. One unannotated and one incorrectly annotated exon of Cdc42bpa mouse gene.
G. Mouse whole brain transcriptome. Novel exon of Chd3 gene.
H. Mouse whole brain transcriptome. Novel gene-candidate.

ssRNA-Seq stands for single-strand RNA sequencing carried out according to the method.

Examples

Example 1: Transcriptome of Saccharomyces cerevisiae

In this example an application of the method according to the invention in a preferred embodiment is presented. Strand-specific sequencing of the yeast transcriptome is performed. Subsequently, the sequence of each clone is determined and matched with already existing DNA sequence of the yeast genome.

RNA isolation:
Yeast strain BY4741 (MATa; *his3Δ1; leu2Δ0; met15Δ0; ura3Δ0*) was grown in rich medium (YPD; BD Company) at 30°C overnight, diluted to an OD_{600} of 0.15 and grown until reaching an OD_{600} of 0.87. The cells were harvested by centrifugation at room temperature, washed once with 1xPBS, and frozen in liquid nitrogen. Total RNA was extracted using the RiboPure™-Yeast kit (Ambion) and analyzed by an Agilent 2100 bioanalyzer (Agilent Technologies).

polyA$^+$ RNA purification:
polyA$^+$ RNA was purified with the Dynabeads mRNA purification kit (Invitrogen) following manufacturer's instructions and treated for 30 min at 37°C with 0.2 u TURBO™ DNase (Ambion) per 1 µg of RNA.

First strand synthesis (FSS):
A typical FSS reaction was prepared by mixing 0.5 µg of polyA$^+$ RNA, 40 ng of $(dN)_6$ primers (Invitrogen) and 25 pmol of oligo(dT) primer (Invitrogen) in 8.5 µl of 1x reverse transcription buffer (Invitrogen), 0.5 mM dNTP's, 5 mM $MgCl_2$ and 10 mM DTT. The mixture was incubated at 98°C for 1 min to melt RNA secondary structures, then at 70°C for 5 min and was ramped to 15°C at 0.1°C/sec. Slow temperature ramp was used to make secondary structures of RNA and primers annealing as

reproducible as possible. At 15°C 0.5 µl of actinomycin D solution (120 ng/µl), 0.5 µl of RNase OUT (40 u/µl, Invitrogen) and 0.5 µl of Superscript III polymerase (200 u/µl, Invitrogen) were added to the reaction. Reverse transcription reaction temperature was increased gradually as a compromise between survival of the enzyme, stability of the primers and denaturation of RNA secondary structures: ramp from 15 to 25°C at 0.1°C/sec; incubation at 25°C for 10 min; ramp from 25 to 42°C at 0.1°C/sec; incubation at 42°C for 45 min; ramp from 42 to 50°C at 0.1°C/sec; incubation at 50°C for 25 min. Superscript III polymerase was finally inactivated at 75°C for 15 min.

Removal of dNTPs:

20 µl of EB (10 mM TrisCl, pH 8.5, Qiagen) was added to the reaction. dNTPs were removed by purification of the first strand mixture on a self-made 200 µl G-50 gel-filtration spin-column equilibrated with 1 mM TrisCl, pH 7.0.

Second strand synthesis (SSS):

The Invitrogen kit for second strand synthesis was used. Therefore the first strand synthesis buffer was restored after gel-filtration column. It is possible to prepare self-made buffer and reduce the number of buffers. Water was added to the purified FSS reaction to bring the final volume to 52.5 µl. The mixture was cooled on ice. Then 22.5 µl of the "second strand mixture" (1 µl of 1O x reverse transcription buffer (Invitrogen); 0.5 µl of 100 mM $MgCl_2$; 1 µl of 0.1 M DTT; 2 µl of 10 mM mixture of each: dATP, dGTP, dCTP, dUTP; 15 µl of 5x SSS buffer (Invitrogen); 0.5 µl of E.coli ligase (10 u/µl, NEB); 2 µl of DNA polymerase I (10 u/µl, NEB) and 0.5 µl RNase H (2u/µl, Invitrogen)) were added. SSS reaction was incubated at 16°C for 2 hours. ds cDNA was purified on QIAquick columns (Qiagen) according to the manufacturer's instructions.

DNA fragmentation:

About 250 ng of double-stranded cDNA was fragmented by sonication on a UTR200 (Hielscher Ultrasonics GmbH, Germany) under the following conditions: 1 hour, 50% pulse, 100% power, and continuous cooling by 0°C water flow-through.

Preparation of libraries for Illumina sequencing platform:

Libraries were prepared using the DNA sample kit (#FC-102-1002, Illumina), as described in (Sultan et al. 2008), but with the following modifications: just before library amplification uridine digestion was performed at 37°C for 15 min in 5 µl of 1 x TE buffer, pH 7.5 with 1 u UNG (Applied Biosystems).

Procedure of paired-end sequencing library preparation was the same except that other ligation adapters and PCR primers were used (#PE-102-1002, Illumina).

Sequencing:

Amplified material was loaded onto flowcell at 4 pM concentration. Sequencing was carried out on the Illumina 1G Genome Analyser by running 36 cycles according to the manufacturer's instructions.

Data analysis:

Images deconvolution, quality values calculation, mapping of exonic reads and exon junctions were performed as described in the art. Sequencing reads were aligned to *S. cerevisiae* (UCSC sacCerl) genomes using modification of Eland software (Gerald module v.1.27, Illumina). The mapping criteria of Eland are the following: sequencing read should be uniquely match to the genome allowing up to two mismatches, without insertions or deletions. The following recursive modification of Eland procedure was applied: first 32 bp of reads (trimming the last 4 bp of 36 bp reads due to Eland limitations) were aligned, then reads which do not match on reference by Eland criteria were trimmed to 31 bp, and aligned again. This 3'-end trimming of unmatched reads was done recursively down to 25 bp length. The modified procedure increases gain of uniquely aligned fragments typically by 20-50%, because the sequencing errors, which prevent successful alignment by Eland criteria, are located mostly on reads ends, which are gradually trimmed. Under these conditions, ~60% of the reads obtained here are matched to unique locations of the reference genome, whereas ~25% of the reads map to more than one genomic position and about 15% do not map.

Mapping end tags:

Unmapped sequencing reads with 1-11 nt long leading oligo(dT) stretches were used for mapping of the 3'-gene boundaries. Leading oligo(dT) stretches were removed, and if the remaining fragment length was longer or equal to 25 bp, it was aligned on reference genome.

Search for novel genes:

The search for novel genes in yeast using RNA-Seq was previously performed. Such an algorithm was not used herein, since it was not accurately described in the art. To estimate roughly how the knowledge of RNA orientation influences the number of novel gene-candidates the following procedure was used. A search algorithm presented below gives 377 novel gene candidates without taking into account RNA

orientation. It is about the same number as in the art. To use information about RNA orientation the search was performed twice (for reads mapped in the forward direction and then for reads mapped in the reverse direction), yielding in 549 novel gene candidates in forward orientation and 512 - in reverse. The whole genome was split into 50 bp windows (non-sliding). A "new gene" was defined as a joined group of more than 2 consecutive windows, with at least 2 sequence reads (of the same direction) mapped per window. The gap between "new genes" should be at least 50 bp, the gap between "new gene" and an annotated gene (with the same transcription direction as the "new gene") - at least 100 bp.

With the RNA-Seq procedure it is possible to determine the 3'-boundaries of genes using those sequencing reads, which overlap with 3'-borders of genes. These reads may be mapped to the reference genome only after removal of oligo(dT) tail. The inventive method has the advantage of reduced noise in comparison with the RNA-Seq protocol because only one orientation of the homopolymeric stretch is allowed.

Example 2: Transcriptome of the brain of *Mus musculus*

In this example a further application of the method according to the invention is presented. Strand-specific sequencing of the whole mouse brain transcriptome is performed. Subsequently, the sequence of each clone is determined and matched with already existing DNA sequence of the mouse genome. Methods are similar to those used in Example 1. Therefore only steps which differ from Example 1 are indicated.

RNA isolation:

Two eleven weeks old female mice (C57B1/6J) were dissected and the whole brain was taken for RNA preparation. Total RNA was extracted using the Trizol method.

Data analysis:

The resulting sequencing reads were aligned to the *M. musculus* (UCSC mm9) genomes using modification of Eland software (Gerald module v.1.27, Illumina).

Example 3: Comparison of strand-specific and non-strand-specific amplification of polynucleotides

To demonstrate that second-strand synthesis with deoxyuridine does not disturb the transcriptional landscape both strand-specific (according to the inventive method) and non-strand-specific transcriptome (RNA-Seq) sequencing for the same RNA

sample was performed (mouse liver transcriptome). The resulting scatter plot (Fig. 7A) shows that both RNA-Seq and the inventive method produce identical transcription patterns.

Example 4: Reproducibiliity of the inventive method and comparison with previously published results from another laboratory

In order to test the reproducibility of the inventive method independent analyses of whole brain mouse (Fig. 7B) transcriptomes were carried out under identical conditions. A high reproducibility can be seen in the respective scatter plots. A comparison of the data with previously published results also demonstrates good correlation (Fig. 7C).

Example 5: Stability of the inventive method

The inventive method results in a high degree of certainty in identifying transcript polarity. In theory, one uridine base in a molecule of a sequencing library is enough to prevent the appearance of a sequencing read from the false strand. Even if UNG occasionally does not remove the uridine base, the molecule still would not be amplified, since uridine-containing template strongly suppresses Phusion DNA polymerase used for library amplification. Apparently, most second cDNA strands in the library contain more than one uridine base, because a 75%/25% dUTP/dTTP mixture in the second strand synthesis reaction gives essentially the same results as 100% dUTP (Fig. 8A and 8B).

Example 6: Influence of actinomycin D

To prevent spurious second-strand cDNA synthesis, which was shown to be the major source of artifactual antisense transcripts. Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. Actinomycin D was included in the reverse transcription reaction. Actinomycin D specifically inhibits DNA-dependent, but not RNA-dependent, DNA synthesis. The presence of actinomycin D in the reaction practically did not influence the level of antisense transcription (Fig. 9A and 9B). Actinomycin D is included in the protocol to make the reaction more reproducible.

Example 7: Transcriptome of the liver of *Mus musculus*

In this example a further application of the method according to the invention is presented. Strand-specific sequencing of the whole mouse liver transcriptome is performed. Subsequently, the sequence of each clone is determined and matched with already existing DNA sequence of the mouse genome. Methods are similar to those used in Example 1. Therefore only steps which differ from Example 1 are indicated.

ds cDNA synthesis:

dTTP was substituted by dUTP during first strand cDNA synthesis. For the second strand cDNA synthesis, the mixture of standard dNTPs was used.

Data analysis:

The resulting sequencing reads were aligned to the *M. musculus* (UCSC mm9) genomes using modification of Eland software (Gerald module v.1.27, Illumina).

Example 8: Transcriptome of the human samples from blood and tumor (melanoma)

In this example a further application of the method according to the invention is presented. Strand-specific sequencing of the human transcriptome is performed. Subsequently, the sequence of each clone is determined and matched with already existing DNA sequence of the human genome (hg18, NCBI build 36.1). Methods are similar to those used in Example 1. Therefore only steps which differ from Example 1 are indicated.

ds cDNA preparation:

Poly A RNA was chemically fragmented prior to ds cDNA synthesis. Compared to ultrasound shearing after ds cDNA synthesis this approach resulted in a more even distribution of reads along the transcript (no 3' bias).

Example 9: Transcriptome of E.coli

In this example a further application of the method according to the invention is presented. Strand-specific sequencing of the E.coli transcriptome is performed. Subsequently, the sequence of each clone is determined and matched with already existing DNA sequence of the E.coli (K12 reference genome (NC_000913).

Methods are similar to those used in Example 1. Therefore only steps which differ from Example 1 are indicated.

ds cDNA preparation:
Instead of dUTP, biotinylated dTTP was used for the second strand cDNA synthesis.

Sequencing library preparation.
After ligation of polarity-specific Y-shaped adapters, biotinylated strand was separated from non-biotinylated using streptavidine coated magnetic beads. Biotinylated strand was further used for amplification.

Data analysis:
The resulting sequencing reads were aligned to K12 reference genome (NC_000913) using the Eland program (Gerald module v.1.27 of the Illumina pipeline)).

Example 10:Preparation of single-stranded hybridisation target

The task was to prepare an array with spotted single-stranded molecules, corresponding to different genomic regions (300-500bp long). First, selected regions were amplified using genomic DNA as a template and PCR primers corresponding to the genomic DNA, where one primer in each pair was selected so, that it had at least one T within last 3 nucleotides on the 3' end of the primer. This dTTP was substituted by dUTP during the primer synthesis. Usage of such uridine-containing primer in PCR resulted in obtaining PCR product with an uridine-marked strand.
PCR products were purified. After UNG treatment and APE 1 digestion a nick appeared in the uridine-containing strand in the place of uridine. From that nick, a nick-translation reaction was performed using dNTP mix with dUTP instead of dTTP and DNA polymerase I in a corresponding buffer. Nick-translation resulted in substitution of one of the strands in a PCR product to a strand containing multiple UTPs along its whole length (Fig. 3A). After UNG treatment and heating, the uridine strand was destroyed. Resulting small fragments were washed off by purification on a silica column. Single stranded DNA molecules were spotted on Nexterion slides (Schott) according to the manufacturers instructions.

Example 11:Preparation of single-stranded hybridisation target

The task and methods were similar to those used in Example 10. The difference was that biotin-associated dTTP were used instead of dUTP in the nick-translation reaction. Biotinylated strand was further separated from non-biotinylated by immobilisation on streptavidine-coated magnetic beads. The non-biotinylated strand was further used for hybridization

Claims

1. A method for selection of a particular strand from a double stranded polynucleotide comprising the following steps:
   I) Providing polynucleotide templates, primers, buffer solution, at least one enzyme with polymerase activity, sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof, wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection;
   II) Performing first-strand synthesis on the polynucleotide templates with unmodified dNTPs or corresponding analogues thereof;
   III) Performing second-strand synthesis wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection, so that double-stranded polynucleotides each consisting of a second strand and a matching strand are generated, wherein the secondstrand contains the at least one modified nucleotide;
   IV) Optionally, performing molecular cloning operations with double-stranded DNA; and
   V) Specific selection of at least one of the strands of the said polynucleotide duplex using different interaction of modified and unmodified nucleotides with at least one selection factor.

2. A method for selection of a particular strand from a double stranded polynucleotide comprising the following steps:
   I) Providing polynucleotide templates, primers, buffer solution, at least one enzyme with polymerase activity, sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof, wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection;
   II) Performing first-strand synthesis on the polynucleotide templates, so that double-stranded polynucleotides each consisting of a template and a matching strand are generated, wherein the matching strand contains the at least one modified nucleotide;
   III) If said polynucleotide templates on step I) are single-stranded RNA, second-strand synthesis with unmodified dNTPs or corresponding analogues thereof is performed;
   IV) Optionally, performing molecular cloning operations with double-stranded DNA; and

V) Specific selection of at least one of the strands of the said polynucleotide duplex using different interaction of modified and unmodified nucleotides with at least one selection factor.

3. The method for selection of a particular strand from a double stranded polynucleotide according to claim 1 or 2, wherein said polynucleotide templates are selected from DNA/DNA hybrid with distinguishable adapters, single-stranded RNA, single-stranded DNA and DNA/RNA hybrid.

4. A method for selection of a particular strand from a double stranded polynucleotide according to claim 3 comprising the following steps:

I) Providing polynucleotide templates, primers, buffer solution, at least one enzyme with polymerase activity, sufficient amounts of the four dNTPs dATP, dCTP, dGTP and dTTP or analogues thereof, wherein at least one of the four dNTPs is replaced by a corresponding modified dNTP used for strand-specific selection;

If said polynucleotide templates are single-stranded DNA

IIa) Performing first-strand synthesis by primer extension reaction from either random or sequence-specific primer, where at least one of dNTP´s is a modified nucleotide further used for selection of a specific strand;

or if said polynucleotide templates are single-stranded RNA,

IIb) Performing first-strand synthesis by reverse transcription reaction using either random or oligo(dT)/oligo(dU) or sequence specific primer, where at least one of dNTPs is a modified nucleotide further used for selection of a specific strand; and

III) then performing second-strand synthesis with all four unmodified dNTPs or analogues thereof

or if said polynucleotide template are DNA/DNA hybrid with distinguishable adapters,

IIc) Performing first-strand synthesis by nick-translation or strand displacement reactions initiated either by setting a nick in one of the adapters or annealing a new primer to the adapter region where at least one of dNTPs is a modified nucleotide further used for selection of a specific strand;

or if said polynucleotide template are DNA/RNA hybrid,

IId) Performing first-strand synthesis with either the DNA or the RNA as template, where at least one of dNTPs is a modified nucleotide further used for selection of a specific strand;

IV) Optionally, performing one or more molecular cloning operations on the double-stranded DNA, wherein the molecular cloning operations are selected from shearing, restriction, digestion, end-polishing, A-tailing, G-tailing and adaptor ligation; and

V) Performing strand-specific selection by using a selection factor for selecting strands containing the at least one modified nucleotide from strands containing the corresponding unmodified nucleotide.

5. The method for selection of a particular strand from a double stranded polynucleotide according to any of claims 1 - 4, wherein said selection factor for selecting strands is at least one chemical agent modifying selectively the at least one modified nucleotide or one or more chemical groups of the at least one modified nucleotide.

6. The method for selection of a particular strand from a double stranded polynucleotide according to any of claims 1 - 4, wherein said selection factor for selecting strands is at least one factor displaying selective affinity to an affinity group of the at least one modified nucleotide.

7. The method for selection of a particular strand from a double stranded polynucleotide according to any of claims 1 - 4, wherein said selection factor for selecting strands is one or more antibodies binding selectively to the at least one modified nucleotide.

8. The method for selection of a particular strand from a double stranded polynucleotide according to any of claims 1 - 4, wherein said selection factor for selecting strands is a solid phase at which the strand containing the at least one modified nucleotide is immobilized.

9. The method for selection of a particular strand from a double stranded polynucleotide according to any of claims 1 - 4, wherein said selection factor for selecting strands is a different sensitivity to nucleases of the modified and the unmodified nucleotide.

10. The method for selection of a particular strand from a double stranded polynucleotide according to claim 9, wherein said nuclease is Uracil-N-Glycosylase and said modified nucleotide is dUTP.

11. The method for selection of a particular strand from a double stranded polynucleotide according to claim 6, wherein said factor displaying selective affinity to an affinity group of the at least one modified nucleotide is streptavidin or avidin and the at least one modified nucleotide is a biotinylated nucleotide.

12. The method for selection of a particular strand from a double stranded polynucleotide according to claim 5, wherein the at least one modified nucleotide with said chemical group is Aminoallyl-dUTP.

13. The method for selection of a particular strand from a double stranded polynucleotide according to claim 7, wherein the at least one modified nucleotide recognized selectively by an antibody is 5-bromodeoxyuridine.

14. The method for selection of a particular strand from a double stranded polynucleotide according to any of claims 1 - 4, wherein the ratio of the at least one modified dNTP towards the corresponding unmodified dNTP or analogues thereof is at least 1:20.

15. The method for selection of a particular strand from a double stranded polynucleotide according to any of claims 1 - 4, wherein if the polynucleotide strand is a single RNA strand the RNA is selected from mRNA, tRNA, rRNA, hnRNA, snRNA, snoRNA, scaRNA, asRNA, micro RNA, siRNA, short transcripts, synthetic RNA polynucleotides and semisynthetic RNA polynucleotides as well as physiological or externally induced modifications thereof, or if the polynucleotide strand is a DNA the DNA is selected from siDNA, single-stranded DNA copies from retroviral or hepadnaviral RNA and single-stranded DNA, single-stranded synthetic DNA polynucleotides and single-stranded semisynthetic DNA polynucleotides as well as physiological or externally induced modifications thereof.

16. A kit for ds cDNA synthesis from RNA preserving information about direction of original RNA strand comprising:

A) All four dNTPs dATP, dCTP, dGTP and dTTP and at least one modified nucleotide corresponding to at least one from dATP, ATP, dCTP, CTP, dGTP, GTP, dTTP, TTP, dUTP and UTP;

B) Primers;

C) One or more DNA polymerases with reverse transcriptase activity;

D) One or more DNA polymerases; and

E) One or more selection factors suitable to specifically interact with said at least one modified nucleotide and/or one or more selective media for removing the strand containing the modified nucleotide;

Or additionally comprising

F) Components for specific removal of at least one modified or at least one unmodified nucleotide by gel-filtration or purification on silica-column.

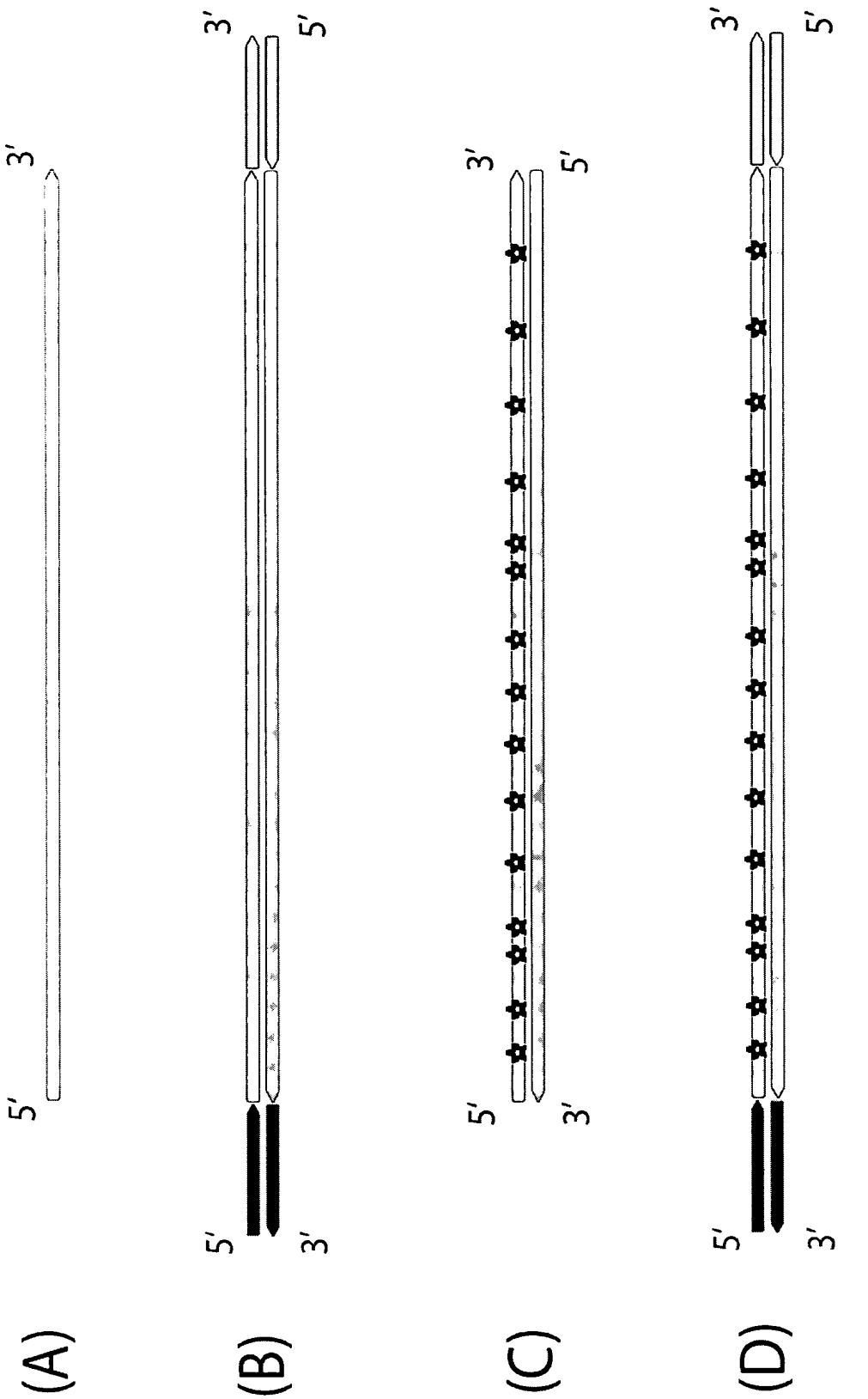


Figure 1

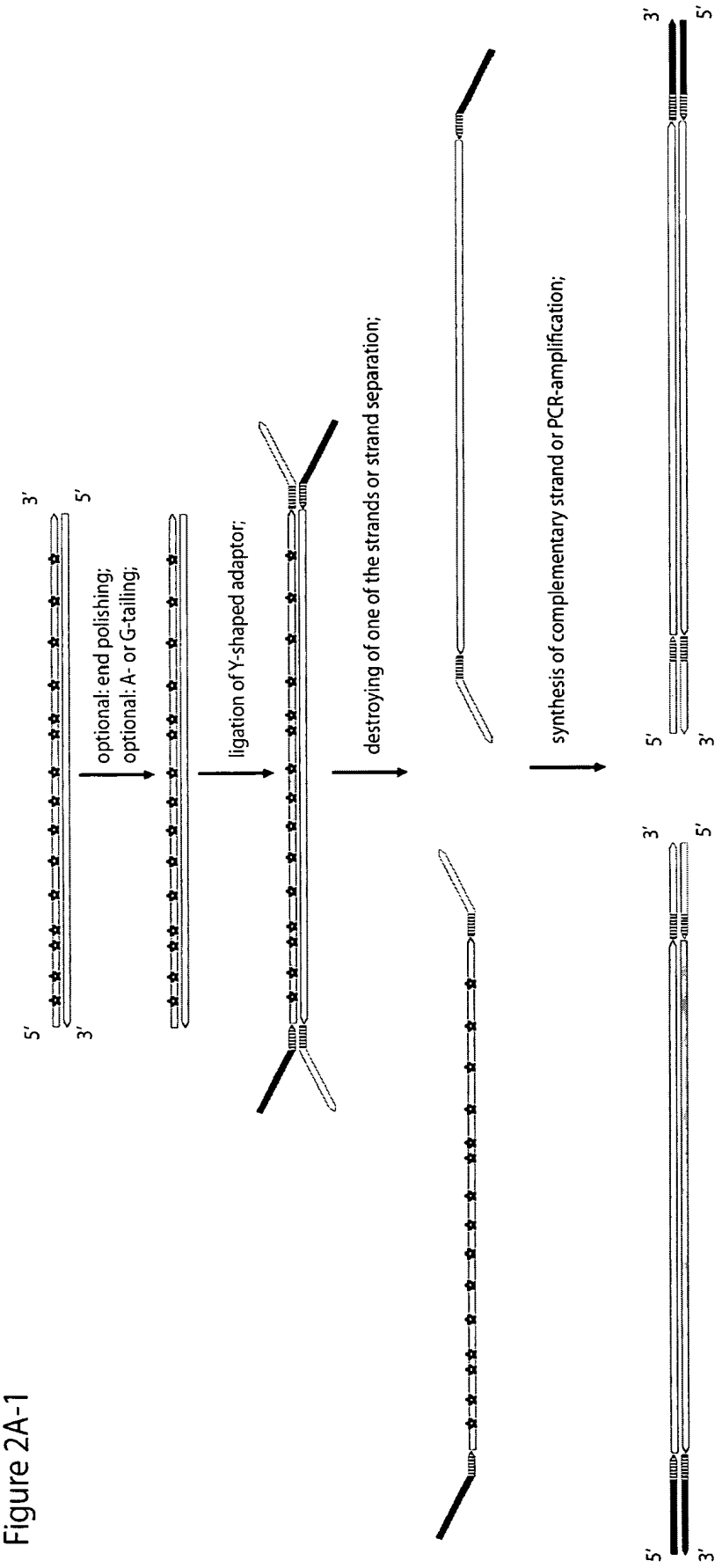

5'
3'
optional: end polishing;
optional: A- or G-tailing;
ligation of Y-shaped adaptor;
destroying of one of the strands or strand separation;
synthesis of complementary strand or PCR-amplification;
3'
5'
5'
3'
3'
5'
5'
3'

Figure 2A-1

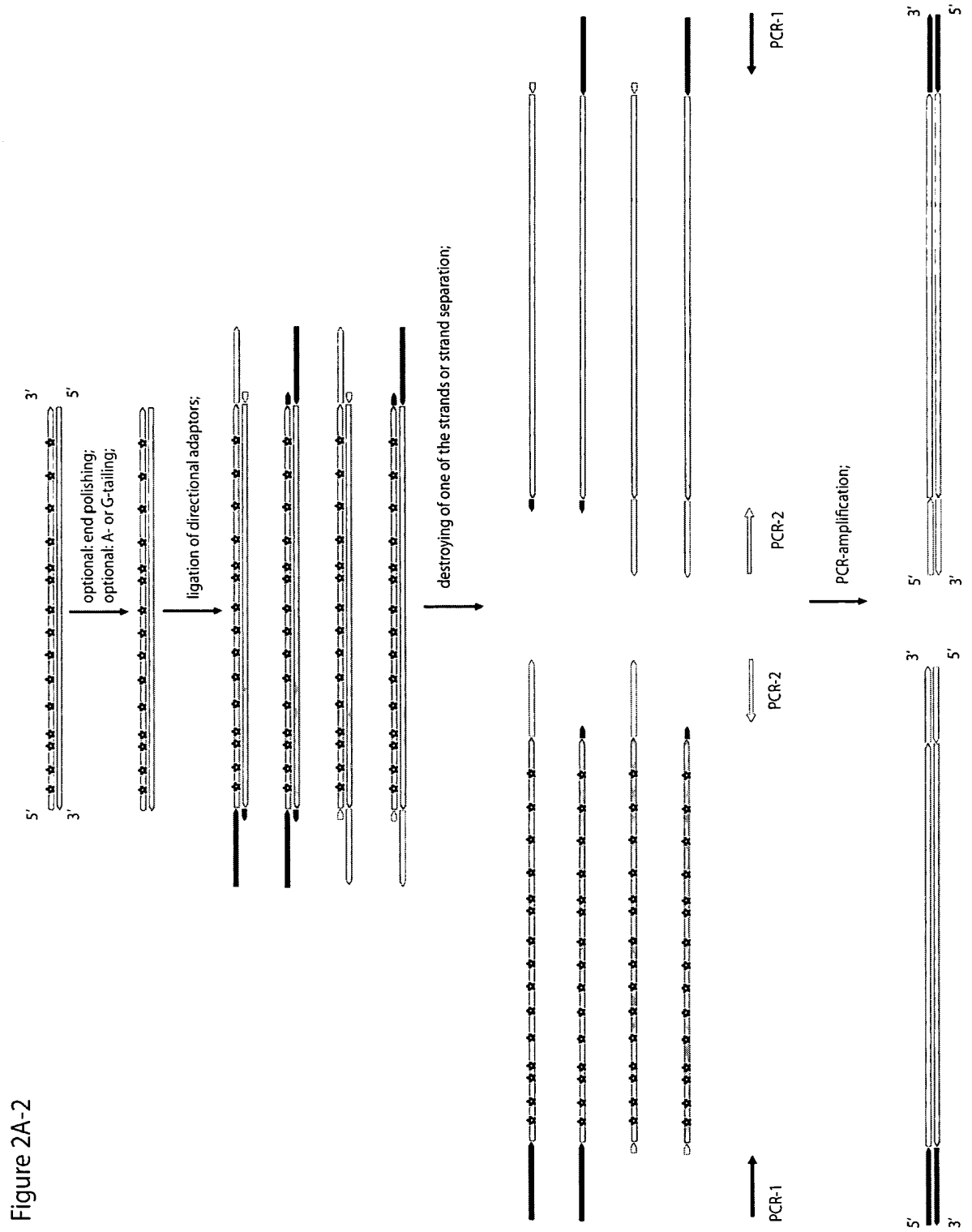

5'
3'
3'
5'
optional: end polishing;
optional: A- or G-tailing;
ligation of directional adaptors;
destroying of one of the strands or strand separation;
PCR-1
PCR-2
PCR-2
PCR-1
PCR-amplification;
3'
5'
5'
3'
5'
3'

Figure 2A-2

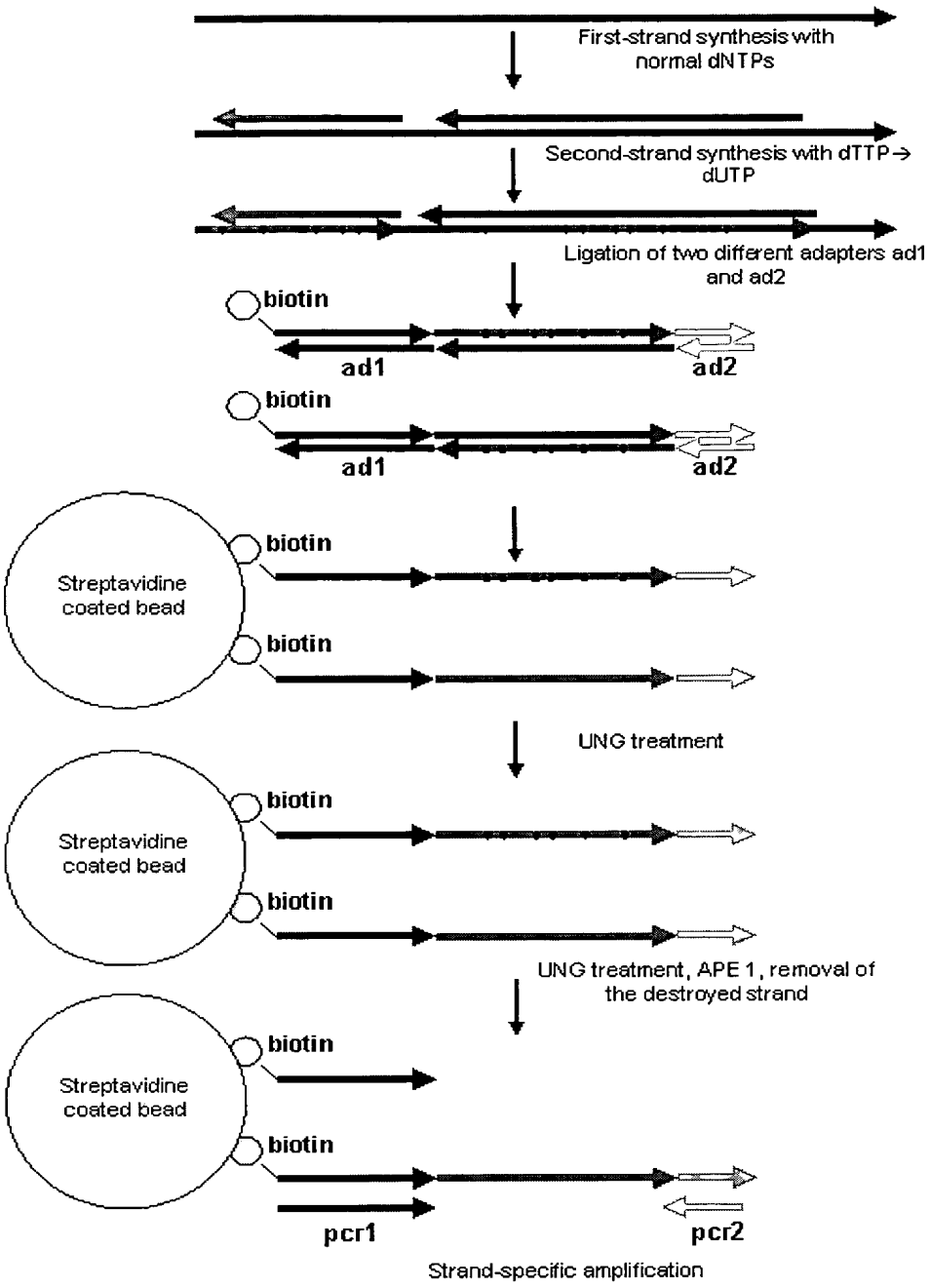

First-strand synthesis with normal dNTPs
Second-strand synthesis with dTTP→ dUTP
Ligation of two different adapters ad1 and ad2
biotin
ad1
ad2
biotin
ad1
ad2
Streptavidine coated bead
biotin
biotin
UNG treatment
Streptavidine coated bead
biotin
biotin
UNG treatment, APE 1, removal of the destroyed strand
Streptavidine coated bead
biotin
biotin
pcr1
pcr2
Strand-specific amplification

Figure 2A-3

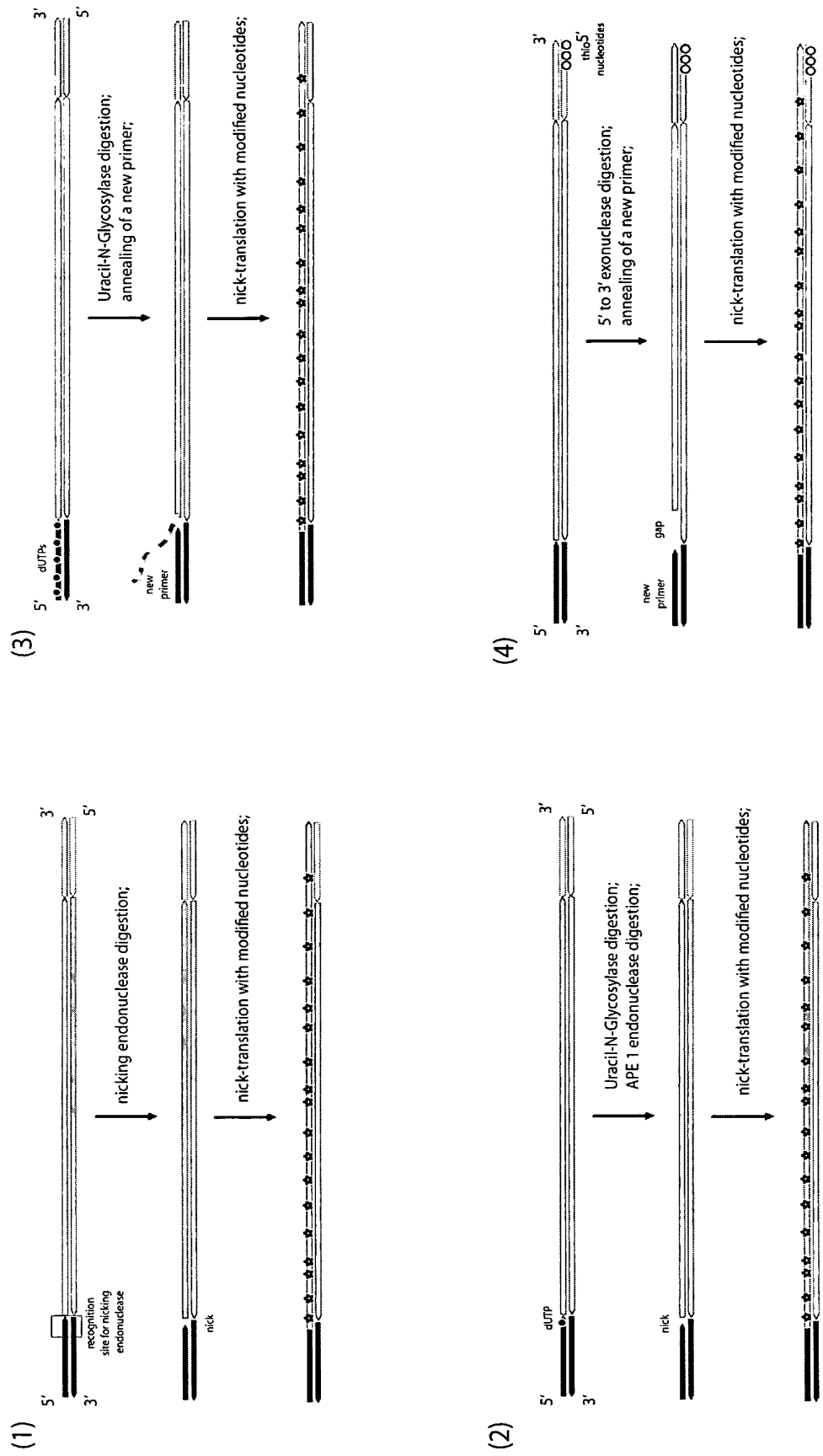

(1)
5'
3'
3'
5'
recognition site for nicking endonuclease
nicking endonuclease digestion;
nick
nick-translation with modified nucleotides;
(2)
5'
3'
dUTP
3'
5'
Uracil-N-Glycosylase digestion;
APE 1 endonuclease digestion;
nick
nick-translation with modified nucleotides;
(3)
5'
dUTPs
3'
3'
5'
Uracil-N-Glycosylase digestion;
annealing of a new primer;
new primer
nick-translation with modified nucleotides;
(4)
3'
thio
5'
nucleotides
5'
3'
5' to 3' exonuclease digestion;
annealing of a new primer;
new primer
gap
nick-translation with modified nucleotides;

Figure 2B

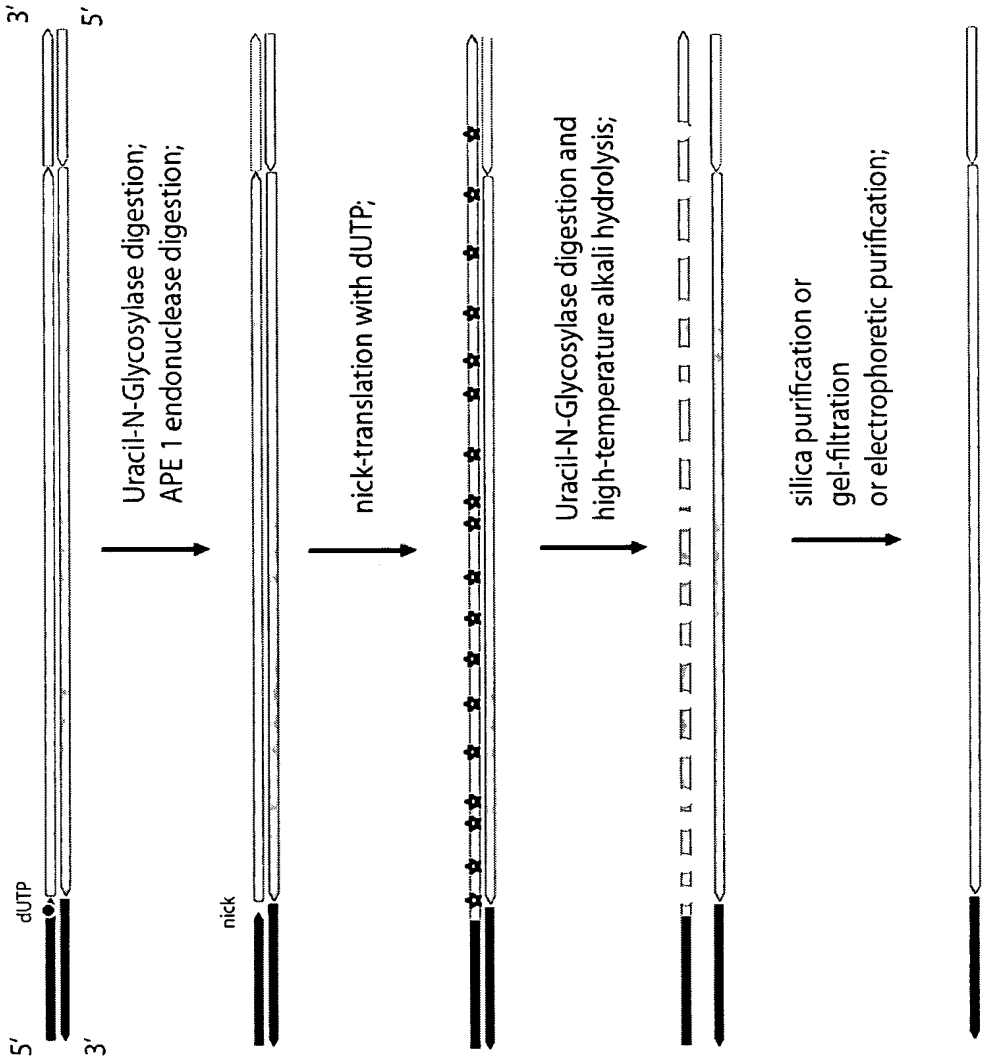

3′
5′
dUTP
5′
3′
Uracil-N-Glycosylase digestion;
APE 1 endonuclease digestion;
nick
nick-translation with dUTP;
Uracil-N-Glycosylase digestion and
high-temperature alkali hydrolysis;
silica purification or
gel-filtration
or electrophoretic purification;

Figure 3A

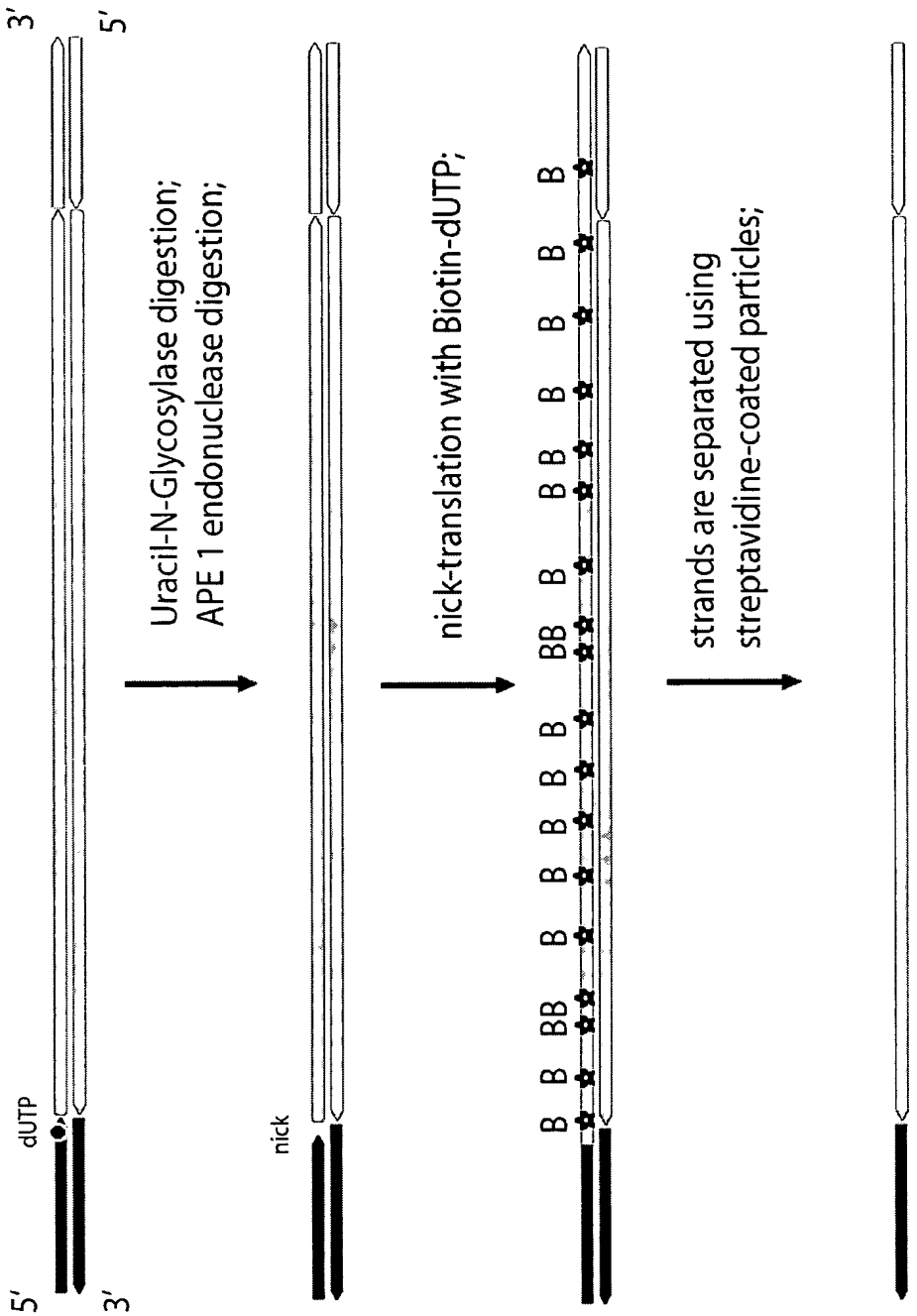

3'
5'
dUTP
5'
3'
Uracil-N-Glycosylase digestion;
APE 1 endonuclease digestion;
nick
nick-translation with Biotin-dUTP;
B B BB B B B B B BB B B B B B B B
strands are separated using
streptavidine-coated particles;

Figure 3B

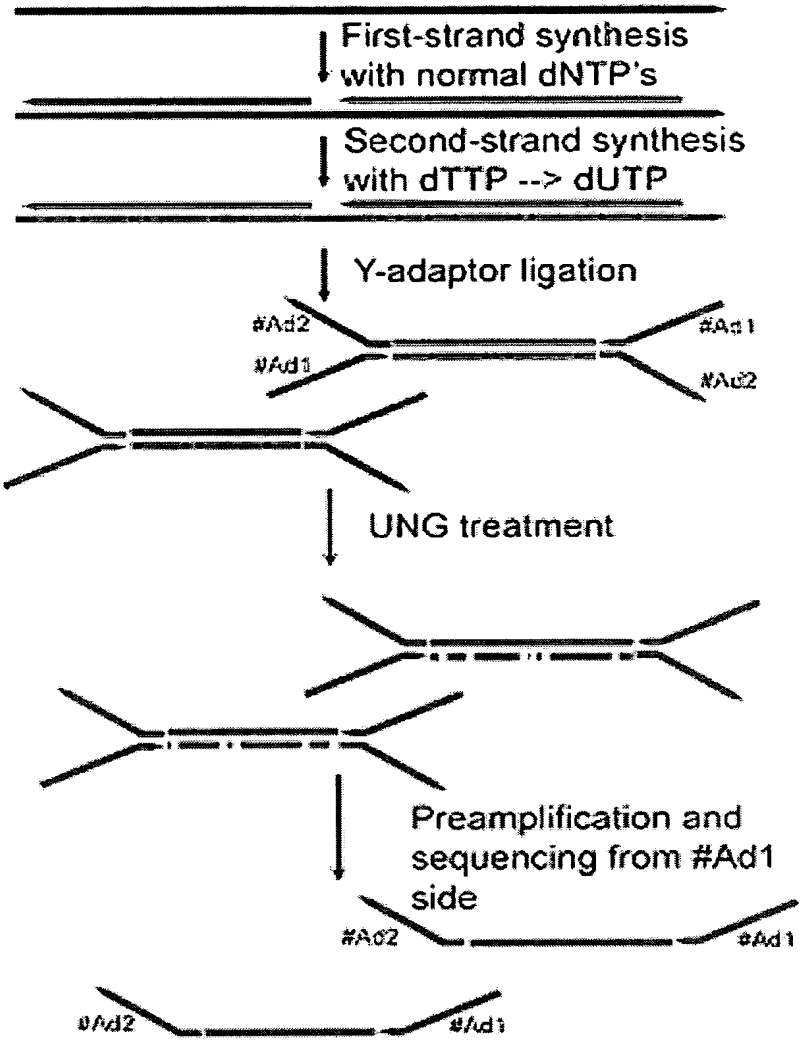
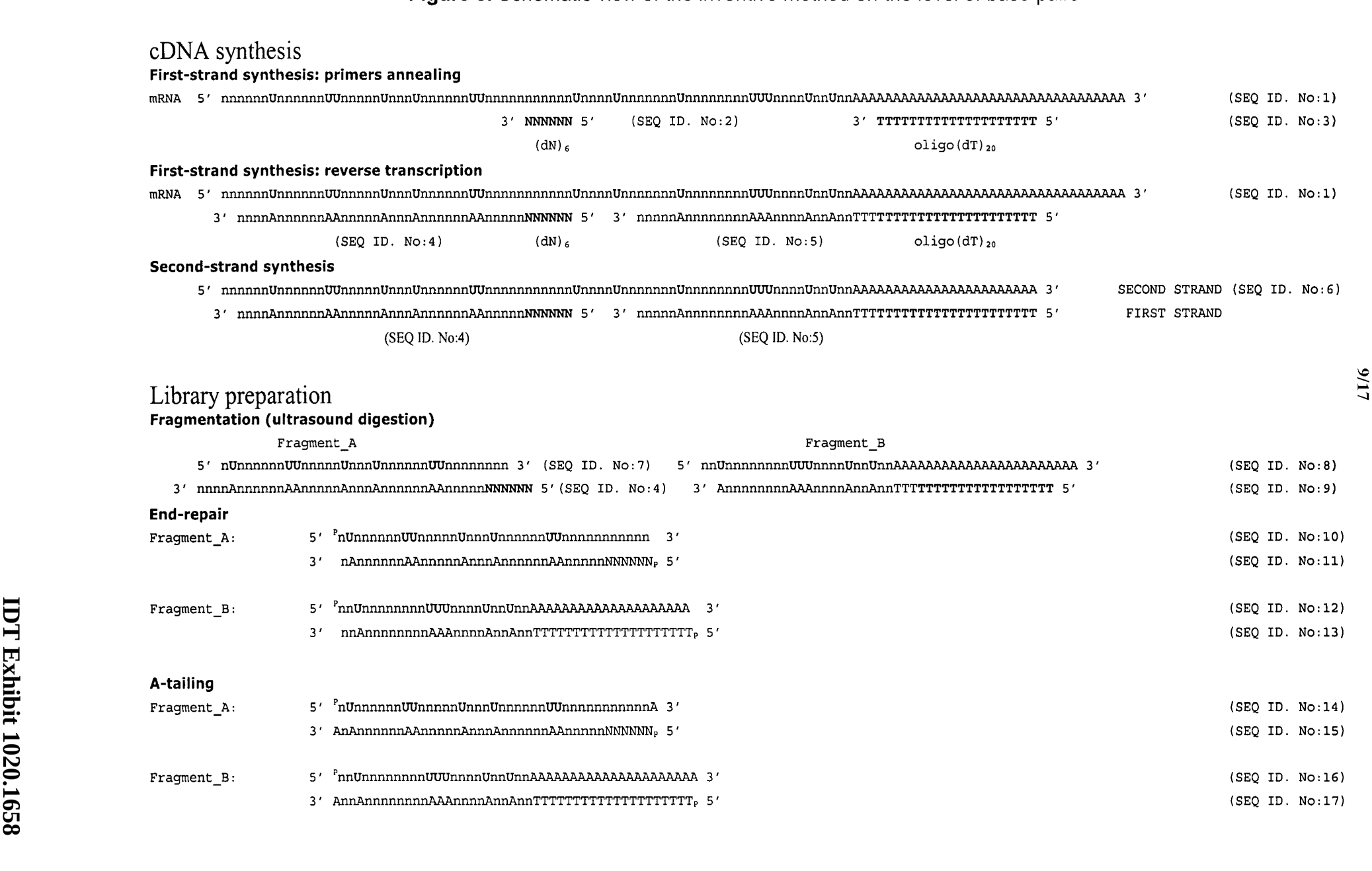

First-strand synthesis
with normal dNTP's
Second-strand synthesis
with dTTP --> dUTP
Y-adaptor ligation
#Ad2
#Ad1
#Ad1
#Ad2
UNG treatment
Preamplification and
sequencing from #Ad1
side
#Ad2
#Ad1
#Ad2
#Ad1

Figure 4

Figure 5. Schematic view of the inventive method on the level of base pairs



IDT Exhibit 1020.1658

Adaptor ligation

Fragment_A:

#Ad_1 #Ad_2

5' TACACTCTTTCCCTACACGAC TCGTATGCCGTCTTCTGCTTG 3' (SEQ ID. No:18)

GCTCTTCCGATCT nUnnnnnnUUnnnnnnUnnnUnnnnnnUUnnnnnnUnnnnA GATCGGAAGAGC

CGAGAAGGCTAG AnAnnnnnnAAnnnnnnAnnnAnnnnnnAAnnnnnnNNNNNN TCTAGCCTTCTCG

3' GTTCGTCTTCTGCCGTATGCT CAGCACATCCCTTTCTCACAT 5' (SEQ ID. No:19)

#Ad_2 #Ad_1

Fragment_B:

#Ad_1 #Ad_2

5' TACACTCTTTCCCTACACGAC TCGTATGCCGTCTTCTGCTTG 3' (SEQ ID. No:20)

GCTCTTCCGATCT nnUnnnnnnnnnUUUnnnnnUnnnUnnAAAAAAAAAAAAAAAAAAAA GATCGGAAGAGC

CGAGAAGGCTAG AnnAnnnnnnnnnAAAnnnnnAnnAnnTTTTTTTTTTTTTTTTTTTT TCTAGCCTTCTCG

3' GTTCGTCTTCTGCCGTATGCT CAGCACATCCCTTTCTCACAT 5' (SEQ ID. No:21)

#Ad_2 #Ad_1

UDGase treatment

Fragment_A:

#Ad_1 #Ad_2

SEQ ID. No:22) (SEQ ID. No:24) (SEQ ID. No:26)

5' TACACTCTTTCCCTACACGAC (SEQ ID. No:23) (SEQ ID. No:25) (SEQ ID. No:27) TCGTATGCCGTCTTCTGCTTG 3'

GCTCTTCCGATCT n·nnnnnn···nnnnn·nnn·nnnnnn···nnnnnnnnnnnnA GATCGGAAGAGC

CGAGAAGGCTAG AnAnnnnnnAAnnnnnnAnnnAnnnnnnAAnnnnnnNNNNNN TCTAGCCTTCTCG

3' GTTCGTCTTCTGCCGTATGCT CAGCACATCCCTTTCTCACAT 5' (SEQ ID. No:19)

#Ad_2 #Ad_1

Fragment_B:

#Ad_1 #Ad_2

(SEQ ID. No:28) (SEQ ID. No:30) (SEQ ID. No:32)

5' TACACTCTTTCCCTACACGAC (SEQ ID. No:29) (SEQ ID. No:31) TCGTATGCCGTCTTCTGCTTG 3'

GCTCTTCCGATCT nn·nnnnnnnn···nnnn·nn·nnAAAAAAAAAAAAAAAAAAAA GATCGGAAGAGC

CGAGAAGGCTAG AnnAnnnnnnnnAAAnnnnnAnnAnnTTTTTTTTTTTTTTTTTTTT TCTAGCCTTCTCG

3' GTTCGTCTTCTGCCGTATGCT CAGCACATCCCTTTCTCACAT 5' (SEQ ID. No:21)

#Ad_2 #Ad_1

IDT Exhibit 1020.1659

Preamplification

Fragment_A:

```
                          #PCR_2                                                                                    #PCR 1
5' CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT 3'      (SEQ ID. No:33)          3' TCTAGCCTTCTCGCAGCACATCCCTTTCTCACATCTAGAGCCACCAGCGGCATAGTAA 5' (SEQ ID. No:34)
  |<------------ 34bp ------------>|<              50-150bp            >|<------------------------ 58bp ------------------------>|
Fragment_A:
3' GTTCGTCTTCTGCCGTATGCTCGAGAAGGCTAGAnAnnnnnnnAAnnnnnAnnnAnnnnnnnAAnnnnnNNNNNNTCTAGCCTTCTCGCAGCACATCCCTTTCTCACAT 5'                    (SEQ ID. No:19
             #Ad_2                                                                                  #Ad_1
Fragment_B:
3' GTTCGTCTTCTGCCGTATGCTCGAGAAGGCTAGAnnAnnnnnnnnnAAAnnnnnAnnAnnTTTTTTTTTTTTTTTTTTTTTTCTAGCCTTCTCGCAGCACATCCCTTTCTCACAT 5'                    (SEQ ID. No:21
             #Ad_2                                                                                  #Ad_1
```

FlowCell preparation

Amplification

Fragment_A:

```
5' CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTnTnnnnnnnTTnnnnnnTnnnTnnnnnnnTTnnnnnnnnnnnnAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT 3' (SEQ ID. No:35)
3' GTTCGTCTTCTGCCGTATGCTCGAGAAGGCTAGAnAnnnnnnnAAnnnnnAnnnAnnnnnnnAAnnnnnNNNNNNTCTAGCCTTCTCGCAGCACATCCCTTTCTCACATCTAGAGCCACCAGCGGCATAGTAA 5' (SEQ ID. No:36)
Fragment_B:
                                                                                                                                                (SEQ ID. No:37
5' CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTnnTnnnnnnnnnTTTnnnnTnnTnnAAAAAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT 3'
3' GTTCGTCTTCTGCCGTATGCTCGAGAAGGCTAGAnnAnnnnnnnnnAAAnnnnnAnnAnnTTTTTTTTTTTTTTTTTTTTTTCTAGCCTTCTCGCAGCACATCCCTTTCTCACATCTAGAGCCACCAGCGGCATAGTAA 5'
                                                                                                                                                (SEQ ID. No:38
```

Linearisation

Fragment_A:

```
5' CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTnTnnnnnnnTTnnnnnnTnnnTnnnnnnnTTnnnnnnnnnnnnAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT 3' (SEQ ID. No:35)
Fragment_B:
5' CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTnnTnnnnnnnnnTTTnnnnTnnTnnAAAAAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT 3'
                                                                                                                                                (SEQ ID. No:37
```

IDT Exhibit 1020.1660

Sequencing primer annealing

Fragment_A:

sequencing primer (SEQ ID. No:39)

<---- 3' TCTAGCCTTCTCGCAGCACATCCCCTTTCTCACA 5'

||||||||||||||||||||||||||||||||||

5' CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTnTnnnnnnnTTnnnnnnTnnnTnnnnnnnTTnnnnnnnnnnnnnnAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT 3' (SEQ ID. No:35)

Fragment_B:

<---- 3' TCTAGCCTTCTCGCAGCACATCCCTTTCTCACA 5' (SEQ ID. No:39)

||||||||||||||||||||||||||||||||||

5' CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTnnTnnnnnnnnTTTnnnnTnnTnnTnnAAAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCCGTATCATT (SEQ ID. No:37)

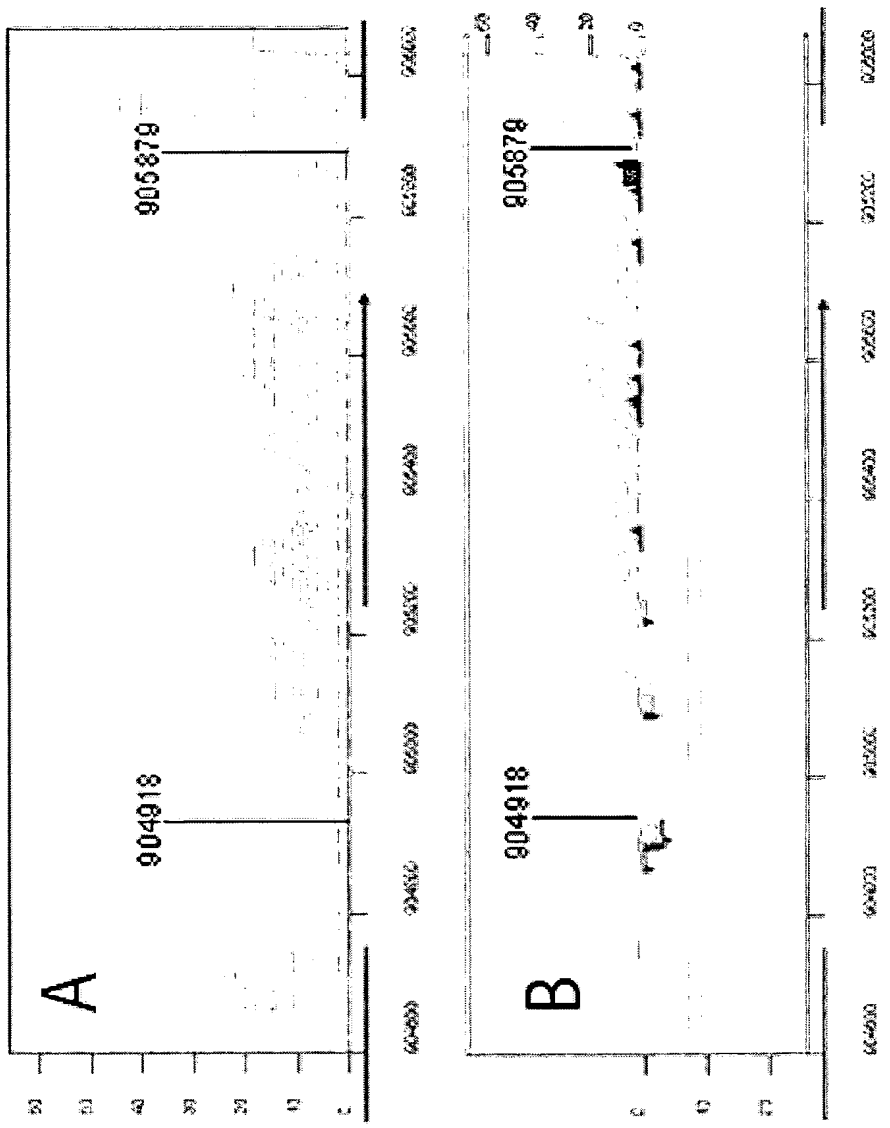


Figure 6

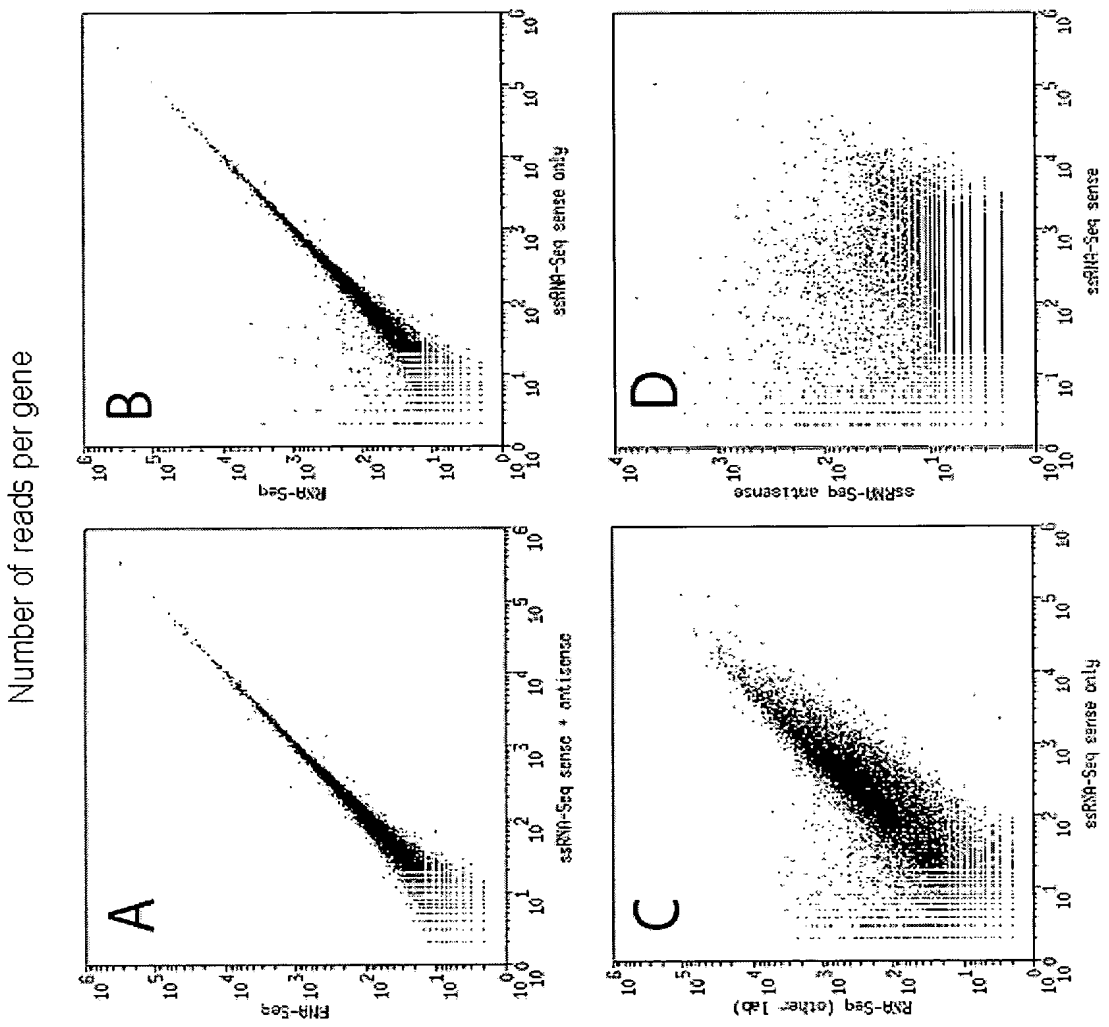

Number of reads per gene
A
B
C
D
RNA-Seq
ssRNA-Seq sense + antisense
RNA-Seq
ssRNA-Seq sense only
RNA-Seq (other lab)
ssRNA-Seq sense only
ssRNA-Seq antisense
ssRNA-Seq sense

Figure 7

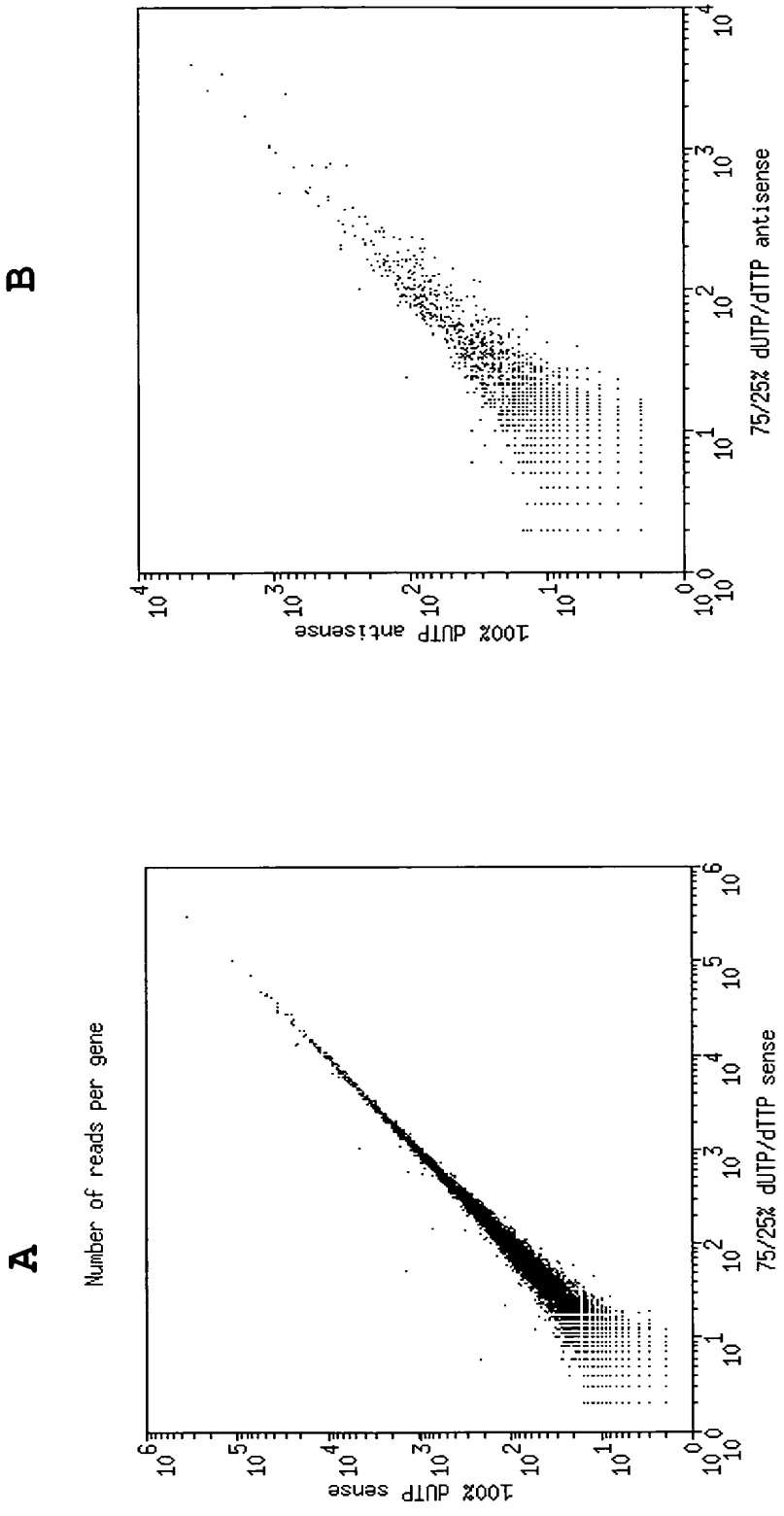

A
Number of reads per gene
100% dUTP sense
75/25% dUTP/dTTP sense
B
100% dUTP antisense
75/25% dUTP/dTTP antisense

Figure 8

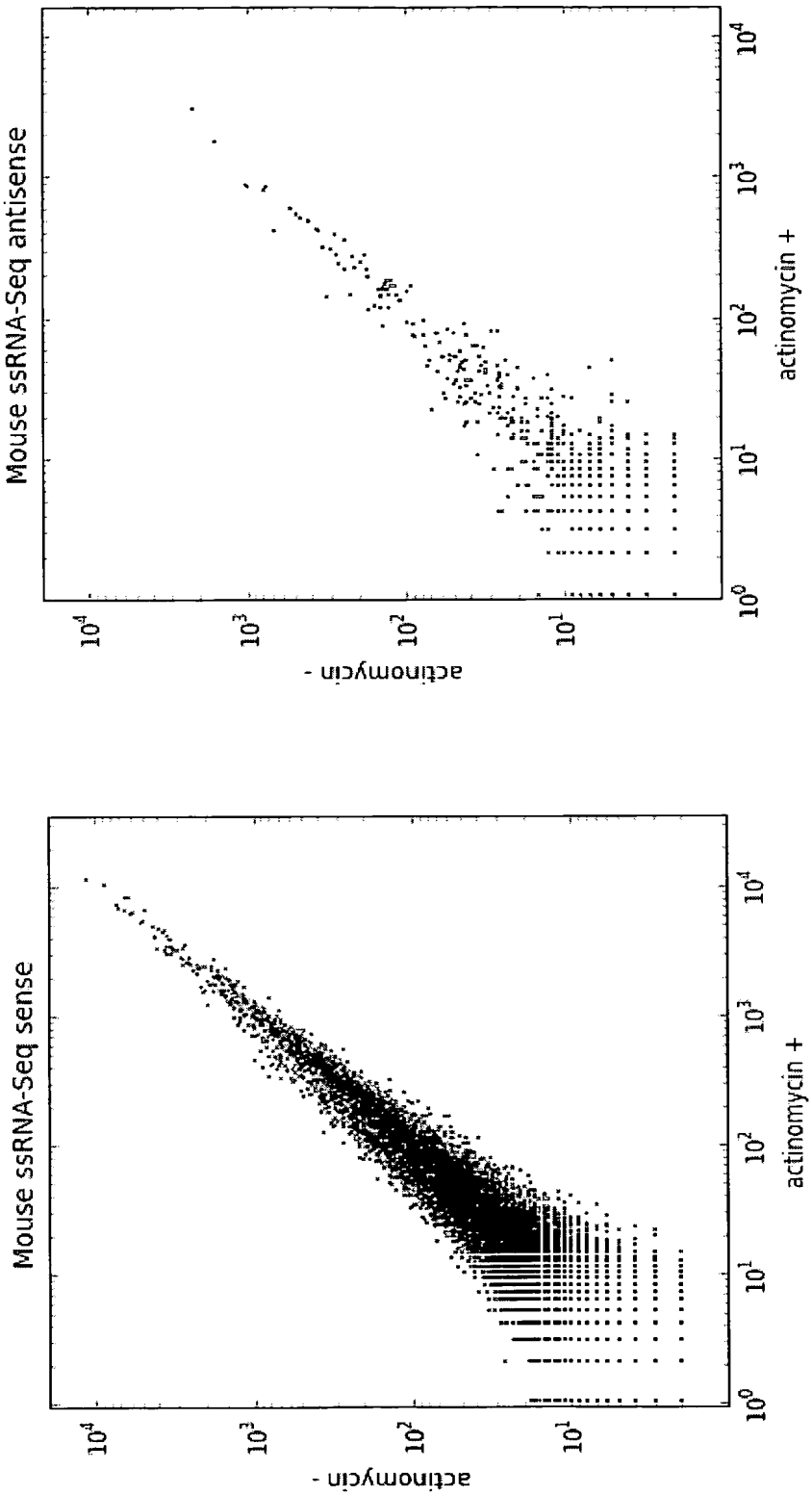

Mouse ssRNA-Seq antisense
actinomycin -
actinomycin +
Mouse ssRNA-Seq sense
actinomycin -
actinomycin +

Figure 9

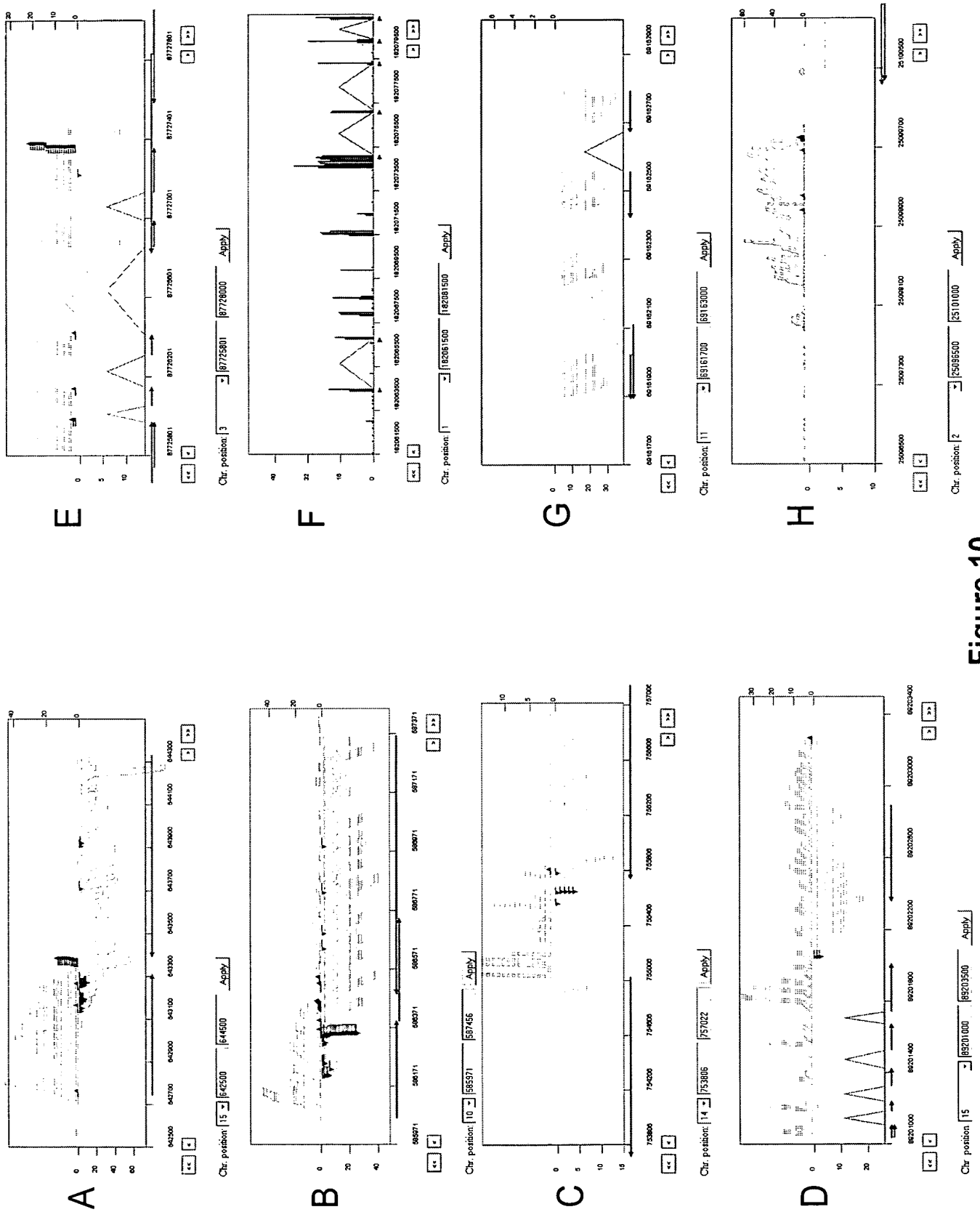


Figure 10

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2010/004399

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12Q1/68
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, BIOSIS, EMBASE, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X A	US 2007/281863 A1 (RABBANI ELAZAR [US] ET AL) 6 December 2007 (2007-12-06) * abstract; claims 100-114; figure 2; examples 1,6 page 3, column 1, paragraph 4 - column 2, paragraph 1 page 6, column 2, paragraph 3; figure 2	16 1-15
A	WO 2009/006438 A2 (EPICT TECHNOLOGIES CORP [US]; DAHL GARY [US]; SOOKNANAN ROY RABINDRANA) 8 January 2009 (2009-01-08) * abstract page 60, paragraph 3	1-16
A	WO 03/008538 A2 (MOLECULAR STAGING INC [US]) 30 January 2003 (2003-01-30) * abstract; claims 26-29; figure 5; example 9	7,13

-/--

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
25 November 2010	08/12/2010

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Fax: (+31–70) 340–3016	Tilkorn, A

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2010/004399

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X,P	PARKHOMCHUK DMITRI ET AL: "Transcriptome analysis by strand-specific sequencing of complementary DNA.", NUCLEIC ACIDS RESEARCH OCT 2009 LNKD-PUBMED:19620212, vol. 37, no. 18, 20 July 2009 (2009-07-20) , page E123, XP002608410, ISSN: 1362-4962 the whole document	1-16

Form PCT/ISA/210 (continuation of second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/EP2010/004399

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item 1.b of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application and necessary to the claimed invention, the international search was carried out on the basis of:

   a. (means)

   [X] on paper

   [X] in electronic form

   b. (time)

   [ ] in the international application as filed

   [ ] together with the international application in electronic form

   [ ] subsequently to this Authority for the purpose of search

2. [X] In addition, in the case that more than one version or copy of a sequence listing and/or table relating thereto has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that in the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments:

IDT Exhibit 1020.1669

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2010/004399

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2007281863 A1	06-12-2007	NONE	
WO 2009006438 A2	08-01-2009	CA 2707436 A1	08-01-2009
		EP 2173872 A2	14-04-2010
		US 2009053775 A1	26-02-2009
WO 03008538 A2	30-01-2003	AU 2002259196 A1	03-03-2003
		US 2004091857 A1	13-05-2004

IDT Exhibit 1020.1670

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
27 December 2013 (27.12.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/190441 A2

(51) International Patent Classification:
C12Q 1/68 (2006.01)

(21) International Application Number:
PCT/IB2013/054898

(22) International Filing Date:
14 June 2013 (14.06.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

61/662,878	21 June 2012 (21.06.2012)	US
61/682,725	13 August 2012 (13.08.2012)	US
61/695,795	31 August 2012 (31.08.2012)	US
61/711,172	8 October 2012 (08.10.2012)	US
13/801,748	13 March 2013 (13.03.2013)	US

(71) Applicant: **THE CHINESE UNIVERSITY OF HONG KONG** [CN/CN]; Knowledge Transfer Office, Shatin, New Territories, Hong Kong, SAR (CN).

(72) Inventor; and
(71) Applicant : **CHIU, Wai, Kwun, Rossa** [AU/CN]; House 31, Double Haven, 52 Ma Lok Path, Shatin, New Territories, Hong Kong (CN).

(72) Inventors: **LO, Yuk-Ming, Dennis**; 4th Floor, 7King Tak Street, Homantin Kowloon, Hong Kong (SAR) (CN). **CHAN, Kwan, Chee**; Flat A, 13/F, Block 34, Broadway Street, Mei Foo Sun Chuen, Kowloon, Hong Kong SAR (CN). **JIANG, Peiyong**; Flat 7. 1st floor of Block B, Kwong Lam Court, Nos 62-66 Siu Lek Yuen Road, Shatin New Territories, Hong Kong SAR (CN).

(74) Agent: **GRIFFITH HACK & CO**; Level 10, 161 Collins Street, Melbourne, Victoria 3000 (AU).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK,

[Continued on next page]

(54) Title: MUTATIONAL ANALYSIS OF PLASMA DNA FOR CANCER DETECTION

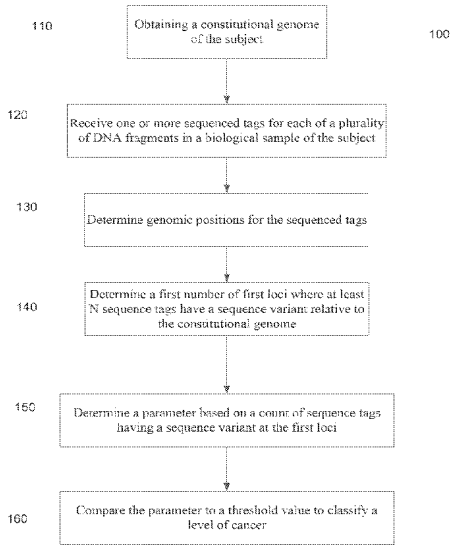


FIG. 1

(57) Abstract: A frequency of somatic mutations in a biological sample (e.g., plasma or serum) of a subject undergoing screening or monitoring for cancer, can be compared with that in the constitutional DNA of the same subject. A parameter can derived from these frequencies and used to determine a classification of a level of cancer. False positives can be filtered out by requiring any variant locus to have at least a specified number of variant sequence reads (tags), thereby providing a more accurate parameter. The relative frequencies for different variant loci can be analyzed to determine a level of heterogeneity of tumors in a patient.

WO 2013/190441 A2

IDT Exhibit 1020.1671

EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

MUTATIONAL ANALYSIS OF PLASMA DNA FOR CANCER DETECTION

CROSS-REFERENCES TO RELATED APPLICATION

[0001] This application is a non-provisional of and claims the benefit of U.S. Provisional Patent Application No. 61/662,878, entitled "MUTATIONAL ANALYSIS OF PLASMA DNA FOR CANCER DETECTION," filed on June 21, 2012; U.S. Provisional Patent Application No. 61/682,725, entitled "MUTATIONAL ANALYSIS OF PLASMA DNA FOR CANCER DETECTION," filed on August 13, 2012; U.S. Provisional Patent Application No. 61/695,795, entitled "MUTATIONAL ANALYSIS OF PLASMA DNA FOR CANCER DETECTION," filed on August 31, 2012; and U.S. Provisional Patent Application No. 61/711,172, entitled "MUTATIONAL ANALYSIS OF PLASMA DNA FOR CANCER DETECTION," filed on October 8, 2012, which are herein incorporated by reference in its entirety for all purposes.

BACKGROUND

[0002] It has been shown that tumor-derived DNA is present in the cell-free plasma/serum of cancer patients (Chen XQ et al. Nat Med 1996; 2: 1033-1035). Most current methods are based on the direct analysis of mutations known to be associated with cancer (Diehl F et al. Proc Natl Acad Sci 2005; 102: 16368-16373; Forshew T et al. Sci Transl Med 2012; 4: 136ra68). Another method has investigated cancer-associated copy number variations detected by random sequencing of plasma DNA (U.S. Patent Publication 2013/0040824 by Lo et al.).

[0003] It is known that with time, more than one cancer cell would acquire growth advantage and produce multiple clones of daughter cells. Ultimately, the tumorous growth and/or its metastatic foci would contain a conglomerate of groups of clonal cancer cells. This phenomenon is typically referred as tumor heterogeneity (Gerlinger M et al. N Engl J Med 2012; 366: 883-892; Yap TA et al. Sci Transl Med 2012; 4: 127ps10).

[0004] Cancers are known to be highly heterogeneous, i.e. mutation profile of cancers of the same tissue type can vary widely. Therefore, the direct analysis of specific mutations can typically detect only a subset of the cases within a particular cancer type known to be associated with those specific mutations. Additionally, tumor-derived DNA is usually the minor species of DNA in human plasma; the absolute concentration of DNA in plasma is low.

IDT Exhibit 1020.1673

Therefore, the direct detection of one or a small group of cancer-associated mutations in plasma or serum may achieve low analytical sensitivity even among patients with cancers known to be harboring the targeted mutations. Furthermore, it has been shown that there is significant intratumoral heterogeneity in terms of mutations even within a single tumor. The mutations can be found in only a subpopulation of the tumor cells. The difference in the mutational profiles between the primary tumor and the metastatic lesions is even bigger. One example of intratumoral and primary-metastasis heterogeneity involves the *KRAS*, *BRAF* and *PIK3CA* genes in patients suffering from colorectal cancers (Baldus et al. Clin Cancer Research 2010. 16:790-9.).

[0005] In a scenario in which a patient has a primary tumor (carrying a *KRAS* mutation but not a *PIK3CA* mutation) and a concealed metastatic lesion (carrying a *PIK3CA* mutation but not a *KRAS* mutation), if one focused on the detection of the *KRAS* mutation in the primary tumor, the concealed metastatic lesion cannot be detected. However, if one included both mutations in the analysis, both the primary tumor and the concealed metastatic lesion can be detected. Hence, the test involving both mutations would have a higher sensitivity in the detection of residual tumor tissues. Such a simple example becomes more complex when one is screening for cancer, and as one has little or no clue of the types of mutations that might occur.

[0006] It is therefore desirable to provide new techniques to perform a broad screening, detection, or assessment for cancer

SUMMARY

[0007] Embodiments can observe a frequency of somatic mutations in a biological sample (e.g., plasma or serum) of a subject undergoing screening or monitoring for cancer, when compared with that in the constitutional DNA of the same subject. Random sequencing can be used to determine these frequencies. A parameter can derived from these frequencies and used to determine a classification of a level of cancer. False positives can be filtered out by requiring any variant locus to have at least a specified number of variant sequence reads (tags), thereby providing a more accurate parameter. The relative frequencies for different variant loci can be analyzed to determine a level of heterogeneity of tumors in a patient.

[0008] In one embodiment, the parameter can be compared with the same parameter derived from a group of subjects without cancer, or with a low risk of cancer. A significant

IDT Exhibit 1020.1674

difference in the parameter obtained from the test subject and that from the group of subjects without cancer, or with a low risk of cancer, can indicate an increased risk that the test subject has cancer or a premalignant condition or would develop cancer in the future. Thus, in one embodiment, plasma DNA analysis can be conducted without prior genomic information of the tumor. Such an embodiment is thus especially useful for the screening of cancer.

[0009] In another embodiment, embodiments can also be used for monitoring a cancer patient following treatment and to see if there is residual tumor or if the tumor has relapsed. For example, a patient with residual tumor or in whom the tumor has relapsed would have a higher frequency of somatic mutations than one in whom there is no residual tumor or in whom no tumor relapse is observed. The monitoring can involve obtaining samples from a cancer patient at multiple time points following treatment for ascertaining the temporal variations of tumor-associated genetic aberrations in bodily fluids or other samples with cell-free nucleic acids, e.g. plasma or serum.

[0010] According to one embodiment, a method detects cancer or premalignant change in a subject. A constitutional genome of the subject is obtained. One or more sequence tags are received for each of a plurality of DNA fragments in a biological sample of the subject, where the biological sample includes cell-free DNA. Genomic positions are determined for the sequence tags. The sequence tags are compared to the constitutional genome to determine a first number of first loci. At each first loci, a number of the sequence tags having a sequence variant relative to the constitutional genome is above a cutoff value, where the cutoff value is greater than one. A parameter is determined based on a count of sequence tags having a sequence variant at the first loci. The parameter is compared to a threshold value to determine a classification of a level of cancer in the subject.

[0011] According to another embodiment, a method analyzes a heterogeneity of one or more tumors of a subject. A constitutional genome of the subject is obtained. One or more sequence tags are received for each of a plurality of DNA fragments in a biological sample of the subject, where the biological sample includes cell-free DNA. Genomic positions are determined for the sequence tags. The sequence tags are compared to the constitutional genome to determine a first number of first loci. At each first loci, a number of the sequence tags having a sequence variant relative to the constitutional genome is above a cutoff value,

IDT Exhibit 1020.1675

where the cutoff value is greater than one. A measure of heterogeneity of the one or more tumors is calculated based on the respective first numbers of the set of first genomic locations.

[0012] According to another embodiment, a method determines a fractional concentration of tumor DNA in a biological sample including cell-free DNA. One or more sequence tags are received for each of a plurality of DNA fragments in the biological sample. Genomic positions are determined for the sequence tags. For each of a plurality of genomic regions, a respective amount of DNA fragments within the genomic region is determined from sequence tags having a genomic position within the genomic region. The respective amount is normalized to obtain a respective density. The respective density is compared to a reference density to identify whether the genomic region exhibits a 1-copy loss or a 1-copy gain. A first density is calculated from respective densities identified as exhibiting a 1-copy loss or from respective densities identified as exhibiting a 1-copy gain. The fractional concentration is calculated by comparing the first density to another density to obtain a differential, wherein the differential is normalized with the reference density.

[0013] Other embodiments are directed to systems and computer readable media associated with methods described herein.

[0014] A better understanding of the nature and advantages of the present invention may be gained with reference to the following detailed description and the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] FIG. 1 is a flowchart of a method 100 for detecting cancer or premalignant change in a subject according to embodiments of the present invention.

[0016] FIG. 2 shows a flowchart of a method comparing the sample genome (SG) directly to the constitutional genome (CG) according to embodiments of the present invention.

[0017] FIG. 3 shows a flowchart of a method 300 comparing the sample genome (SG) to the constitutional genome (CG) using the reference genome (RG) according to embodiments of the present invention.

[0018] FIG. 4 is a table 400 showing the number of cancer-associated single nucleotide mutations correctly identified using different number of occurrences as the criterion for classifying a mutation as being present in the sample according to embodiments of the

IDT Exhibit 1020.1676

present invention when the fractional concentration of tumor-derived DNA in the sample is assumed to be 10%.

[0019] FIG. 5 is a table showing the expected number of false-positive loci and the expected number of mutations identified when the fractional concentration of tumor-derived DNA in the sample is assumed to be 5%.

[0020] FIG. 6A is a graph 600 showing the detection rate of cancer-associated mutations in plasma with 10% and 20% plasma fractional concentrations of tumor-derived DNA and using four and six occurrences (r) as criteria for calling potential cancer-associated mutations. FIG. 6B is a graph 650 showing the expected number of nucleotide positions falsely classified as having a nucleotide change using criteria of occurrence (r) of 4, 5, 6 and 7 vs. sequencing depth.

[0021] FIG. 7A is a graph 700 showing the number of true cancer-associated mutation sites and false-positive sites with difference sequencing depths when the fractional concentration of tumor-derived DNA in the sample is assumed to be 5%. FIG. 7B is a graph 750 showing the predicted number of false-positive sites involving the analysis of the whole genome (WG) and all exons.

[0022] FIG. 8 is a table 800 showing results for 4 HCC patients before and after treatment, including fractional concentrations of tumor-derived DNA in plasma according to embodiments of the present invention.

[0023] FIG. 9 is a table 900 showing detection of the HCC-associated SNVs in 16 healthy control subjects according to embodiments of the present invention.

[0024] FIG. 10A shows a distribution plot of the sequence read densities of the tumor sample of an HCC patient according to embodiments of the present invention. FIG. 10B shows a distribution plot 1050 of z-scores for all the bins in the plasma of a HCC patient according to embodiments of the present invention.

[0025] FIG. 11 shows a distribution plot 1100 of z-scores for the plasma of an HCC patient according to embodiments of the present invention.

[0026] FIG. 12 is a flowchart of a method 1200 of determining a fractional concentration of tumor DNA in a biological sample including cell-free DNA according to embodiments of the present invention.

IDT Exhibit 1020.1677

[0027] FIG. 13A shows a table 1300 of the analysis of mutations in the plasma of the patient with ovarian cancers and a breast cancer at the time of diagnosis according to embodiments of the present invention.

[0028] FIG. 13B shows a table 1350 of the analysis of mutations in the plasma of the patient with bilateral ovarian cancers and a breast cancer after tumor resection according to embodiments of the present invention.

[0029] FIG. 14A is a table 1400 showing detection of single nucleotide variations in plasma DNA for HCC1. FIG. 14B is a table 1450 showing detection of single nucleotide variations in plasma DNA for HCC2.

[0030] FIG. 15A is a table 1500 showing detection of single nucleotide variations in plasma DNA for HCC3. FIG. 15B is a table 1550 showing detection of single nucleotide variations in plasma DNA for HCC4.

[0031] FIG. 16 is a table 1600 showing detection of single nucleotide variations in plasma DNA for the patient with ovarian (and breast) cancer.

[0032] FIG. 17 is a table 1700 showing the predicted sensitivities of different requirements of occurrence and sequencing depths.

[0033] FIG. 18 is a table 1800 showing the predicted numbers of false positive loci for different cutoffs and different sequencing depths.

[0034] FIG. 19 shows a tree diagram illustrating the number of mutations detected in the different tumor sites.

[0035] FIG. 20 is a table 2000 showing the number of fragments carrying the tumor-derived mutations in the pre-treatment and post-treatment plasma sample.

[0036] FIG. 21 is a graph 2100 showing distributions of occurrence in plasma for the mutations detected in a single tumor site and mutations detected in all four tumor sites.

[0037] FIG. 22 is a graph 2200 showing predicted distribution of occurrence in plasma for mutations coming from a heterogeneous tumor

[0038] FIG. 23 demonstrates the specificity of embodiments for 16 healthy control subjects were recruited.

IDT Exhibit 1020.1678

[0039] FIG. 24 is a flowchart of a method 2400 for analyzing a heterogeneity of one or more tumors of a subject according to embodiments of the present invention.

[0040] FIG. 25 shows a block diagram of an example computer system 2500 usable with system and methods according to embodiments of the present invention.

DEFINITIONS

[0041] As used herein, the term "*locus*" or its plural form "*loci*" is a location or address of any length of nucleotides (or base pairs) which may have a variation across genomes. A "*bin*" is a region of predetermined length in a genome. A plurality of bins may have a same first length (resolution), while a different plurality can have a same second length. In one embodiment, the bins do not overlap each other.

[0042] The term "*random sequencing*" as used herein refers to sequencing whereby the nucleic acid fragments sequenced have not been specifically identified or predetermined before the sequencing procedure. Sequence-specific primers to target specific gene loci are not required. The term "*universal sequencing*" refers to sequencing where sequencing can start on any fragment. In one embodiment, adapters are added to the end of a fragment, and the primers for sequencing attached to the adapters. Thus, any fragment can be sequenced with the same primer, and thus the sequencing can be random.

[0043] The term "*sequence tag*" (also referred to as sequence read) as used herein refers to string of nucleotides sequenced from any part or all of a nucleic acid molecule. For example, a sequenced tag may be a short string of nucleotides (e.g., ~ 30) sequenced from a nucleic acid fragment, a short string of nucleotides at both ends of a nucleic acid fragment, or the sequencing of the entire nucleic acid fragment that exists in the biological sample. A nucleic acid fragment is any part of a larger nucleic acid molecule. A fragment (e.g. a gene) may exist separately (i.e. not connected) to the other parts of the larger nucleic acid molecule.

[0044] The term "*constitutional genome*" (also referred to a CG) is composed of the consensus nucleotides at loci within the genome, and thus can be considered a consensus sequence. The CG can cover the entire genome of the subject (e.g., the human genome), or just parts of the genome. The constitutional genome (CG) can be obtained from DNA of cells as well as cell-free DNA (e.g., as can be found in plasma). Ideally, the consensus nucleotides should indicate that a locus is homozygous for one allele or heterozygous for two alleles. A heterozygous locus typically contains two alleles which are members of a genetic

IDT Exhibit 1020.1679

polymorphism. As an example, the criteria for determining whether a locus is heterozygous can be a threshold of two alleles each appearing in at least a predetermined percentage (e.g., 30% or 40%) of reads aligned to the locus. If one nucleotide appears at a sufficient percentage (e.g., 70% or greater) then the locus can be determined to be homozygous in the CG. Although the genome of one healthy cell can differ from the genome of another healthy cell due to random mutations spontaneously occurring during cell division, the CG should not vary when such a consensus is used. Some cells can have genomes with genomic rearrangements, e.g., B and T lymphocytes, such as involving antibody and T cell receptor genes. Such large scale differences would still be a relatively small population of the total nucleated cell population in blood, and thus such rearrangements would not affect the determination of the constitutional genome with sufficient sampling (e.g., sequencing depth) of blood cells. Other cell types, including buccal cells, skin cells, hair follicles, or biopsies of various normal body tissues, can also serve as sources of CG.

[0045] The term "*constitutional DNA*" refers to any source of DNA that is reflective of the genetic makeup with which a subject is born. For a subject, examples of "constitutional samples", from which constitutional DNA can be obtained, include healthy blood cell DNA, buccal cell DNA and hair root DNA. The DNA from these healthy cells defines the CG of the subject. The cells can be identified as healthy in a variety of ways, e.g., when a person is known to not have cancer or the sample can be obtained from tissue that is not likely to contain cancerous or premalignant cells (e.g., hair root DNA when liver cancer is suspected). As another example, a plasma sample may be obtained when a patient is cancer-free, and the determined constitutional DNA compared against results from a subsequent plasma sample (e.g., a year or more later). In another embodiment, a single biologic sample containing <50% of tumor DNA can be used for deducing the constitutional genome and the tumor-associated genetic alterations. In such a sample, the concentrations of tumor-associated single nucleotide mutations would be lower than those of each allele of heterozygous SNPs in the CG. Such a sample can be the same as the biological sample used to determine a sample genome, described below.

[0046] The term "*biological sample*" as used herein refers to any sample that is taken from a subject (*e.g.*, a human, a person with cancer, a person suspected of having cancer, or other organisms) and contains one or more cell-free nucleic acid molecule(s) of interest. A biological sample can include cell-free DNA, some of which can have originated from

IDT Exhibit 1020.1680

healthy cells and some from tumor cells. For example, tumor DNA can be found in blood or other fluids, e.g., urine, pleural fluid, ascitic fluid, peritoneal fluid, saliva, tears or cerebrospinal fluid. A non-fluid example is a stool sample, which may be mixed with diarrheal fluid. For some of such samples, the biological sample can be obtained non-invasively. In some embodiments, the biological sample can be used as a constitutional sample.

[0047] The term "sample genome" (also referred to as SG) is a collection of sequence reads that have been aligned to locations of a genome (e.g., a human genome). The sample genome (SG) is not a consensus sequence, but includes nucleotides that may appear in only a sufficient number of reads (e.g., at least 2 or 3, or higher cutoff values). If an allele appears a sufficient number of times and it is not part of the CG (i.e., not part of the consensus sequence), then that allele can indicate a "single nucleotide mutation" (also referred to as an SNM). Other types of mutations can also be detected using the current invention, e.g. mutations involving two or more nucleotides, (such as affect the number of tandem repeat units in a microsatellite or simple tandem repeat polymorphism), chromosomal translocation (which can be intrachromosomal or interchromosomal) and sequence inversion.

[0048] The term "*reference genome*" (also referred to as RG) refers to a haploid or diploid genome to which sequence reads from the biological sample and the constitutional sample can be aligned and compared. For a haploid genome, there is only one nucleotide at each locus. For a diploid genome, heterozygous loci can be identified, with such a locus having two alleles, where either allele can allow a match for alignment to the locus.

[0049] The term "*level of cancer*" can refer to whether cancer exists, a stage of a cancer, a size of tumor, and/or other measure of a severity of a cancer. The level of cancer could be a number or other characters. The level could be zero. The level of cancer also includes premalignant or precancerous conditions (states) associated with mutations or a number of mutations. The level of cancer can be used in various ways. For example, screening can check if cancer is present in someone who is not known previously to have cancer. Assessment can investigate someone who has been diagnosed with cancer. Detection can mean 'screening' or can mean checking if someone, with suggestive features of cancer (e.g. symptoms or other positive tests), has cancer.

IDT Exhibit 1020.1681

DETAILED DESCRIPTION

[0050] Embodiments are provided for the detection of cancer by the analysis of a biological sample (e.g., a blood plasma/serum sample) that is not taken directly from a tumor and includes cell-free nucleic acids. The cell-free nucleic acids can originate for various types of tissue throughout the body. In this manner, a broad analysis for the detection of various cancers can be performed.

[0051] Genetic aberrations (including single nucleotide mutations, deletions, amplifications, and rearrangements) accumulate in the tumor cells during the development of cancers. In embodiments, massively parallel sequencing can be used to detect and quantify the single nucleotide mutations (SNMs), also called single nucleotide variations (SNVs), in body fluids (e.g. plasma, serum, saliva, ascitic fluid, pleural fluid and cerebrospinal fluid) so as to detect and monitor cancers. A quantification of the number of SNMs (or other types of mutations) can provide a mechanism for identifying early stages of cancer as part of screening tests. In various implementations, care is taken to distinguish sequencing errors and to distinguish spontaneous mutations occurring in healthy cells (e.g., by requiring multiple SNMs to be identified at a particular locus, e.g., at least 3, 4, or 5).

[0052] Some embodiments also provide noninvasive methods for the analysis of tumor heterogeneity, which can involve cells within the same tumor (i.e. intratumoral heterogeneity) or cells from different tumors (from either the same site or from different sites) within a body. For example, one can noninvasively analyze the clonal structure of such tumor heterogeneity, including an estimation of the relative tumor cell mass containing each mutation. Mutations that are present in higher relative concentrations are present in a larger number of malignant cells in the body, e.g., cells that have occurred earlier on during the tumorigenic process relative to other malignant cells still in the body (Welch JS et al. Cell 2012; 150: 264-278). Such mutations, due to their higher relative abundance, are expected to exhibit a higher diagnostic sensitivity for detecting cancer DNA than those with lower relative abundance. A serial monitoring of the change of the relative abundance of mutations would allow one to noninvasively monitor the change in the clonal architecture of tumors, either spontaneously as the disease progresses, or in response to treatment. Such information would be of use in assessing prognosis or in the early detection of tumor resistance to treatment.

IDT Exhibit 1020.1682

I. INTRODUCTION

[0053] Mutations can occur during cell division because of errors in DNA replication and/or DNA repair. One type of such mutations involve the alteration of single nucleotides, which can involve multiple sequences from different parts of the genome. Cancers are generally believed to be due to the clonal expansion of a single cancer cell which has acquired growth advantage. This clonal expansion would lead to the accumulation of mutations (e.g. single nucleotide mutations) in all the cancer cells originating from the ancestral cancer cell. These progeny tumor cells would share a set of mutations (e.g. single nucleotide mutations). As described herein, cancer-associated single nucleotide mutations are detectable in the plasma/serum of cancer patients.

[0054] Some embodiments can effectively screen for all mutations in a biological sample (e.g., the plasma or serum). As the number of mutations are not fixed (hundreds, thousands, or millions of cancer-associated mutations from different subpopulations of tumor cells can be detected), embodiments can provide a better sensitivity than techniques that detect specific mutations. The number of mutations can be used to detect cancer.

[0055] To provide such a screening of many or all mutations, embodiments can perform a search (e.g., a random search) for genetic variations in a biological sample (e.g., bodily fluids, including plasma and serum), which could contain tumor-derived DNA. The use of a sample, such as plasma, obviates the need to perform an invasive biopsy of the tumor or cancer. Also, as the screening can cover all or large regions of the genome, the screening is not limited to any enumerable and known mutations, but can use the existence of any mutation. Moreover, since the number of mutations is summed across all or large regions of the genome, a higher sensitivity can be obtained.

[0056] However, there are polymorphic sites, including single nucleotide polymorphisms (SNPs), in the human genome, which should not be counted in the mutations. Embodiments can ascertain whether genetic variations that have been detected are likely to be cancer-associated mutations or are polymorphisms in the genome. For example, as part of determining between cancer-associated mutations and polymorphisms in the genome, embodiments can determine a constitutional genome, which can include polymorphisms. The polymorphisms of the constitutional genome (CG) can be confined to polymorphisms that are exhibited with a sufficiently high percentage (e.g., 30-40%) in the sequencing data.

IDT Exhibit 1020.1683

[0057] The sequences obtained from the biological sample can then be aligned to the constitutional genome and variations that are single nucleotide mutations (SNMs), or other types of mutations, identified. These SNMs would be variations that are not included in the known polymorphisms, and thus can be labeled as cancer-associated, and not part of the constitutional genome. A healthy person may have a certain number of SNMs due to random mutations among healthy cells, e.g., created during cell division, but a person with cancer would have more.

[0058] For example, for a person with cancer, the number of SNMs detectable in a bodily fluid would be higher than the polymorphisms present in the constitutional genome of the same person. A comparison can be made between the amounts of variations detected in a bodily fluid sample containing tumor-derived DNA and a DNA sample containing mostly constitutional DNA. In one embodiment, the term 'mostly' would mean more than 90%. In another preferred embodiment, the term 'mostly' would mean more than 95, 97%, 98%, or 99%. When the amount of variations in the bodily fluid exceeds that of the sample with mostly constitutional DNA, there is an increased likelihood that the bodily fluid might contain tumor-derived DNA.

[0059] One method that could be used to randomly search for variations in DNA samples is random or shotgun sequencing (e.g., using massively parallel sequencing). Any massively parallel sequencing platform may be used, including a sequencing-by-ligation platform (e.g. the Life Technologies SOLiD platform), the Ion Torrent/Ion Proton, semiconductor sequencing, Roche 454, single molecular sequencing platforms (e.g. Helicos, Pacific Biosciences and nanopore). Yet, it is known that sequencing errors can occur and may be misinterpreted as a variation in the constitutional DNA or as mutations derived from tumor DNA. Thus, to improve the specificity of our proposed approach, the probability of the sequencing error or other components of analytical errors can be accounted for, e.g., by using an appropriate sequencing depth along with requiring at least a specified number (e.g., 2 or 3) of detected alleles at a locus for it to be counted as an SNM.

[0060] As described herein, embodiments can provide evidence for the presence of tumor-derived DNA in a biological sample (e.g., a bodily fluid) when the amount of randomly detected genetic variations present in the sample exceeds that expected for constitutional DNA and variations that may be inadvertently detected due to analytical errors (e.g., sequencing errors). The information could be used for the screening, diagnosis,

IDT Exhibit 1020.1684

prognostication and monitoring of cancers. In the following sections, we describe analytical steps that can be used for the detection of single nucleotide mutations in plasma/serum or other samples (e.g., bodily fluids). Bodily fluids could include plasma, serum, cerebrospinal fluid, pleural fluid, ascitic fluid, nipple discharge, saliva, bronchoalveolar lavage fluid, sputum, tears, sweat and urine. In addition to bodily fluids, the technology can also be applied to stools sample, as the latter has been shown to contain tumor DNA from colorectal cancer (Berger BM, Ahlquist DA. Pathology 2012; 44: 80-88).

II. GENERAL SCREENING METHOD

[0061] FIG. 1 is a flowchart of a method 100 for detecting cancer or premalignant change in a subject according to embodiments of the present invention. Embodiments can analyze cell-free DNA in a biological sample from the subject to detect variations in the cell-free DNA likely resulting from a tumor. The analysis can use a constitutional genome of the subject to account for polymorphisms that are part of healthy cells, and can account for sequencing errors. Method 100 and any of the methods described herein may be totally or partially performed with a computer system including one or more processors.

[0062] In step 110, a constitutional genome of the subject is obtained. The constitutional genome (CG) can be determined from the constitutional DNA of the tested subject. In various embodiments, the CG can be read from memory or actively determined, e.g., by analyzing sequence reads of constitutional DNA, which may be in cells from the sample that includes the cell-free DNA. For example, when a non-hematological malignancy is suspected, blood cells can be analyzed to determine the constitutional DNA of the subject.

[0063] In various implementations, the analysis of the constitutional DNA could be performed using massively parallel sequencing, array-based hybridization, probe-based in-solution hybridization, ligation-based assays, primer extension reaction assays, and mass spectrometry. In one embodiment, the CG can be determined at one time point in a subject's life, e.g., at birth or even in the prenatal period (which could be done using fetal cells or via cell-free DNA fragment, see U.S. Publication 2011/0105353), and then be referred to when bodily fluids or other samples are obtained at other times of the subject's life. Thus, the CG may simply be read from computer memory. The constitutional genome may be read out as a list of loci where the constitutional genome differs from a reference genome.

IDT Exhibit 1020.1685

[0064] In step 120, one or more sequence tags are received for each of a plurality of DNA fragments in a biological sample of the subject, where the biological sample includes cell-free DNA. In one embodiment, the one or more sequence tags are generated from a random sequencing of DNA fragments in the biological sample. More than one sequence tag may be obtained when paired-end sequencing is performed. One tag would correspond to each end of the DNA fragment.

[0065] The cell-free DNA in the sample (e.g., plasma, serum or other body fluid) can be analyzed to search for genetic variations. The cell-free DNA can be analyzed using the same analytical platform as that has been used to analyze the constitutional DNA. Alternatively, a different analytical platform could be used. For example, the cell-free DNA sample can be sequenced using massively parallel sequencing or parts of the genome could be captured or enriched before massively parallel sequencing. If enrichment is used, one could, for example, use solution-phase or solid-phase capture of selected parts of the genome. Then, massively parallel sequencing can be carried out on the captured DNA.

[0066] In step 130, genomic positions for the sequence tags are determined. In one embodiment, the sequence tags are aligned to a reference genome, which is obtained from one or more other subjects. In another embodiment, the genomic sequence tags are aligned to the constitutional genome of the tested subject. The alignment can be performed using techniques known to one skilled in the art, e.g., using Basic Local Alignment Search Tool (BLAST).

[0067] In step 140, a first number of loci are determined where at least N sequence tags have a sequence variant relative to the constitutional genome (CG). N is equal to or greater than two. As discussed in more detail below, sequencing errors as well as somatic mutations occurring randomly in cells (e.g., due to cell division) can be removed by having N equal 2, 3, 4, 5, or higher. The loci that satisfy one or more specified criteria can be identified as a mutation (variant) or mutation loci (variant loci), whereas a locus having a variant but not satisfying the one or more criteria (e.g., as just one variant sequence tag) is referred to as a potential or putative mutation. The sequence variant could be for just one nucleotide or multiple nucleotides.

[0068] N may be determined as percentage of total tags for a locus, as opposed to an absolute value. For example, a variant locus can be identified when the fractional

IDT Exhibit 1020.1686

concentration of tumor DNA inferred from the variant reads is determined to be equal to or greater than 10% (or some other percentage). In other words, when the locus is covered by 200 sequence reads, a criterion of at least 10 sequence reads showing the variant allele can be required to define the variant as a mutation. The 10 sequence reads of the variant allele and 190 reads of the wildtype allele would give an fractional concentration of tumor DNA of 10% (2x10/(10+190)).

[0069] In one embodiment, the sequence tags (collectively referred to as the sample genome) can be compared directly to the CG to determine the variants. In another embodiment, the sample genome (SG) is compared to the CG via a reference genome (RG) to determine the variants. For example, both the CG and SG can be compared to the RG to determine respective numbers (e.g., sets) of loci exhibiting variants, and then a difference can be taken to obtain the first number of loci. The first number can simply be obtained as a number or may correspond to a specific set of loci, which may then be analyzed further to determine a parameter from the sequence tags at the first loci.

[0070] In one implementation, sequencing results of constitutional DNA and plasma DNA are compared to determine if a single nucleotide mutation is present in the plasma DNA. The regions at which the constitutional DNA is homozygous can be analyzed. For illustration purposes, assume the genotype of a particular locus is homozygous in the constitutional DNA and is AA. Then in the plasma, the presence of an allele other than A would indicate the potential presence of a single nucleotide mutation (SNM) at the particular locus. The loci indicating the potential presence of an SNM can form the first number of loci in step 140.

[0071] In one embodiment, it could be useful to target parts of the genome that are known to be particularly prone to mutation in a particular cancer type or in a particular subset of the population. Of relevance to the latter aspect, embodiments can look for types of mutations that are particularly prevalent in a specific population group, e.g. mutations that are especially common in subjects who are carriers of hepatitis B virus (for liver cancer) or human papillomavirus (for cervical cancer) or who have genetic predisposition to somatic mutations or subjects with germline mutations in a DNA mismatch repair gene. The technology would also be useful to screen for mutations in ovarian and breast cancers in subjects with *BRCA1* or *BRCA2* mutations. The technology would similarly be useful to screen for mutations in colorectal cancer in subjects with *APC* mutations.

IDT Exhibit 1020.1687

[0072] In step 150, a parameter is determined based on a count of sequence tags having a sequence variant at the first loci. In one example, the parameter is the first number of loci where at least N DNA fragments have a sequence variant at a locus relative to the constitutional genome. Thus, the count can be used simply to ensure that a locus has more than N copies of a particular variant identified before being included in the first number. In another embodiment, the parameter can be or include the total number of sequence tags having a sequence variant relative to the constitutional genome at the first loci.

[0073] In step 160, the parameter for the subject is compared to a threshold value (e.g., derived from one or more other subjects) to determine a classification of a level of cancer in the subject. Examples of a level of cancer includes whether the subject has cancer or a premalignant condition, or an increased likelihood of developing cancer. In one embodiment, the threshold value may be determined from a previously obtained sample from the subject.

[0074] In another embodiment, the one or more other subjects may be determined to not have cancer or a low risk of cancer. Thus, threshold value may be a normal value, a normal range, or indicate a statistically significant deviation from a normal value or range. For example, the number of mutations relative to the CG of a specific subject, detectable in the plasma of subjects without a cancer or with a low risk of cancer, can be used as the normal range to determine if the number of mutations detected in the tested subject is normal. In another embodiment, the other subjects could be known to have cancer, and thus a similar number of mutations can indicate cancer.

[0075] In one implementation, the other subjects can be selected to have clinical characteristics that are matched to those of the test subject, e.g. sex, age, diet, smoking habit, drug history, prior disease, family history, genotypes of selected genomic loci, status for viral infections (e.g. hepatitis B or C virus or human papillomavirus or human immunodeficiency virus or Epstein-Barr virus infection) or infections with other infectious agents (such as bacteria (e.g. *Helicobacter pylori*) and parasites (e.g. *Clonorchis sinensis*), etc. For example, subjects who are carriers of hepatitis B or C virus have an increased risk of developing hepatocellular carcinoma. Thus, test subjects who have a similar number or pattern of mutations as a carrier of hepatitis B or C can be considered to have an increased risk of developing hepatocellular carcinoma. On the other hand, a hepatitis B or C patient who exhibits more mutations than another hepatitis patient can properly be identified as having a higher classification of a level of cancer, since the proper baseline (i.e. relative to another

IDT Exhibit 1020.1688

hepatitis patient) is used. Similarly, subjects who are carriers of human papillomavirus infection have increased risk for cervical cancer, and head and neck cancer. Infection with the Epstein-Barr virus has been associated with nasopharyngeal carcinoma, gastric cancer, Hodgkin's lymphoma and non-Hodgkin's lymphoma. Infection with *Helicobacter pylori* has been associated with gastric cancer. Infection with *Clonorchis sinensis* has been associated with cholangiocarcinoma.

[0076] The monitoring of the changes of the number of mutations at different time points can be used for monitoring of the progress of the cancer and the treatment response. Such monitoring can also be used to document the progress of a premalignant condition or change in the risk that a subject would develop cancer.

[0077] The amount of sequence tags showing variations can also be used to monitor. For example, a fractional concentration of variant reads at a locus can be used. In one embodiment, an increase in the fractional concentrations of tumor-associated genetic aberrations in the samples during serial monitoring can signify the progression of the disease or imminent relapse. Similarly, a decrease in the fractional concentrations of tumor-associated genetic aberrations in the samples during serial monitoring can signify response to treatment and/or remission and/or good prognosis.

III. DETERMINING GENOMES

[0078] The various genomes discussed above are explained in more detail below. For example, the reference genome, constitutional genome, and the sample genome are discussed.

A. Reference Genome

[0079] The reference genome (RG) refers to a haploid or diploid genome of a subject or consensus of a population. The reference genome is known and thus can be used to compare sequencing reads from new patients. The sequence reads from a sample of a patient can be aligned and compared to identify variations in the reads from the RG. For a haploid genome, there is only one nucleotide at each locus, and thus each locus can be considered hemizygous. For a diploid genome, heterozygous loci can be identified, with such a locus having two alleles, where either allele can allow a match for alignment to the locus.

[0080] A reference genome can be the same among a population of subjects. This same reference genome can be used for healthy subjects to determine the appropriate threshold to

IDT Exhibit 1020.1689

be used for classifying the patient (e.g., having cancer or not). However, different reference genomes can be used for different populations, e.g., for different ethnicities or even for different families.

B. Constitutional Genome

[0081] The constitutional genome (CG) for a subject (e.g., a human or other diploid organism) refers to a diploid genome of the subject. The CG can specify heterozygous loci where a first allele is from a first haplotype and a different second allele is from a second haplotype. Note that the structures of two haplotypes that cover two heterozygous loci need not be known, i.e., which allele on one heterozygous locus is on the same haplotype as an allele of another heterozygous locus need not be known. Just the existence of the two alleles at each heterozygous locus can be sufficient.

[0082] The CG can differ from the RG due to polymorphisms. For example, a locus on the RG can be homozygous for T, but the CG is heterozygous for T/A. Thus, the CG would exhibit a variation at this locus. The CG can also be different from the RG due to inherited mutations (e.g., that run in families) or de novo mutations (that occur in a fetus, but which are not present in its parents). The inherited mutation is typically called 'germline mutation'. Some of such mutations are associated with predisposition to cancer, such as a *BRCA1* mutation that runs in a family. Such mutations are different from 'somatic mutations' that can occur due to cell division during one's lifetime and can push a cell and its progeny on the way to become a cancer.

[0083] A goal of determining the CG is to remove such germline mutations and de novo mutations from the mutations of the sample genome (SG) in order to identify the somatic mutations. The amount of somatic mutations in the SG can then be used to assess the likelihood of cancer in the subject. These somatic mutations can be further filtered to remove sequencing errors, and potentially to remove somatic mutations that occur rarely (e.g., only one read showing a variant), as such somatic mutations are not likely related to cancer.

[0084] In one embodiment, a CG can be determined using cells (buffy coat DNA). However, the CG can also be determined from cell-free DNA (e.g. plasma or serum) as well. For a sample type in which most of the cells are non-malignant, e.g. the buffy coat from a healthy subject, then the majority or consensus genome is the CG. For the CG, each genomic locus consists of the DNA sequence possessed by the majority of cells in the sampled tissue.

IDT Exhibit 1020.1690

The sequencing depth should be sufficient to elucidate heterozygous sites within the constitutional genome.

[0085] As another example, plasma can be used as the constitutional sample to determine the CG. For example, for cases in which the tumor DNA in plasma is less than 50% and an SNM is in a heterozygous state, e.g., the mutation is the addition of a new allele, then the new allele can have a concentration of less than 25%. Whereas, the concentration of the heterozygous alleles of SNPs in the CG should amount to approximately 50%. Thus, a distinction can be made between a somatic mutation and a polymorphism of the CG. In one implementation, a suitable cutoff can be between 30-40% for determining a somatic mutation from a polymorphism when using plasma, or other mixtures with significant tumor concentration. A measurement of tumor DNA concentration can be useful to ensure that the tumor DNA in plasma is less than 50%. Examples of determining a tumor DNA concentration are described herein.

C. Sample Genome

[0086] The sample genome (SG) is not simply a haploid or diploid genome as is the case for the RG and CG. The SG is a collection of reads from the sample, and can include: reads from constitutional DNA that correspond to the CG, reads from tumor DNA, reads from healthy cells that show random mutations relative to the CG (e.g., due to mutations resulting from cell division), and sequencing errors. Various parameters can be used to control exactly which reads are included in the SG. For example, requiring an allele to show up in at least 5 reads can decrease the sequencing errors present in the SG, as well as decrease the reads due to random mutations.

[0087] As an example, assume the subject is healthy, i.e., does not have cancer. For illustration purposes, the DNA from 1000 cells is in 1 ml of plasma (i.e. 1000 genome-equivalents of DNA) obtained from this subject. Plasma DNA typically consists of DNA fragments of about 150 bp. As the human genome is $3x10^9$ bp, there would be about $2x10^7$ DNA fragments per haploid genome. As the human genome is diploid, there would be about $4x10^7$ DNA fragments per ml of plasma.

[0088] As millions to billions of cells are releasing their DNA in the plasma per unit time and fragments from these cells would mix together during circulation, the $4x10^7$ DNA fragments could have come from $4x10^7$ different cells. If these cells do not bear a recent (as

IDT Exhibit 1020.1691

opposed to distant, e.g., the original zygote) clonal relationship to each other (i.e. that they do not share a recent ancestral cell), then it is statistically likely that no mutation will be seen more than once amongst these fragments.

[0089] On the other hand, if amongst the 1000 genome-equivalents per ml of plasma DNA, there is a certain percentage of cells that share a recent ancestral cell (i.e., they are related to each other clonally), then one could see the mutations from this clone to be preferentially represented in the plasma DNA (e.g. exhibiting a clonal mutational profile in plasma). Such clonally related cells could be cancer cells, or cells that are on their way to become a cancer but not yet there (i.e. pre-neoplastic). Thus, requiring a mutation to show up more than once can remove this natural variance in the "mutations" identified in the sample, which can leave more mutations related to cancer cells or pre-neoplastic cells, thereby allowing detection, especially early detection of cancer or precancerous conditions.

[0090] In one approximation, it has been stated that on average, one mutation will be accumulated in the genome following every cell division. Previous work has shown that most of the plasma DNA is from hematopoietic cells (Lui YY et al. Clin Chem 2002: 48: 421-427). It has been estimated that hematopoietic stem cells replicate once every 25-50 weeks (Catlin SN, et al. Blood 2011; 117: 4460-4466). Thus, as a simplistic approximation, a healthy 40-year-old subject would have accumulated some 40 to 80 mutations per hematopoietic stem cell.

[0091] If there are 1000 genome-equivalents per ml in this person's plasma, and if each of these cells is derived from a different hematopoietic stem cell, then 40,000 to 80,000 mutations might be expected amongst the $4x10^{10}$ DNA fragments (i.e. $4x10^{7}$ DNA fragments per genome, and 1000 genome-equivalents per ml of plasma). However, as each mutation would be seen once, each mutation can still be below a detection limit (e.g., if cutoff value N is greater than 1), and thus these mutations can be filtered out, thereby allowing the analysis to focus on mutations that are more likely to result from cancerous conditions. The cutoff value can be any value (integer or non-integer) greater than one, and may be dynamic for different loci and regions. The sequencing depth and fractional concentration of tumor DNA can also affect the sensitivity of detecting mutations (e.g., percentage of mutations detectable) from cancer cells or pre-neoplastic cells.

IDT Exhibit 1020.1692

IV. COMPARING SG DIRECTLY TO CG

[0092] Some embodiments can identify nucleotide positions that the CG is homozygous, but where a minority species (i.e. the tumor DNA) in the SG is heterozygous. When sequencing a position in high depth (e.g., over 50-fold coverage), one can detect if there are one or two alleles at that position in the DNA mixture of healthy and cancer cells. When there are two alleles detected, either (1) the CG is heterozygous or (2) the CG is homozygous but the SG is heterozygous. These two scenarios can be differentiated by looking at the relative counts of the major and the minor alleles. In the former scenario, the two alleles would have similar numbers of counts; but for the latter scenario, there would be a large difference in their numbers of counts. This comparison of the relative allele counts of the reads from the test sample is one embodiment for comparing sequence tags to the constitutional genome. The first loci of method 100 can be determined as loci where the number of alleles is below an upper threshold (threshold corresponding to a polymorphism in the CG) and above a lower threshold (threshold corresponding to errors and somatic mutations occurring at a sufficiently low rate to not be associated with a cancerous condition). Thus, the constitutional genome and the first loci can be determined at the same time.

[0093] In another embodiment, a process for identifying mutations can determine the CG first, and then determine loci having a sufficient number of mutations relative to the CG. The CG can be determined from a constitutional sample that is different from the test sample.

[0094] FIG. 2 shows a flowchart of a method 200 comparing the sample genome (SG) directly to the constitutional genome (CG) according to embodiments of the present invention. At block 210, a constitutional genome of the subject is obtained. The constitutional genome can be obtained, for example, from a sample taken previously in time or a constitutional sample that is obtained and analyzed just before method 200 is implemented.

[0095] At block 220, one or more sequence tags are received for each of a plurality of DNA fragments in a biological sample of the subject. The sequencing may be performed using various techniques, as mentioned herein. The sequence tags are a measurement of what the sequence of a fragment is believed to be. But, one or more bases of a sequence tag may be in error.

IDT Exhibit 1020.1693

[0096] At block 230, at least a portion of the sequence tags are aligned to the constitutional genome. The alignment can account for the CG being heterozygous at various loci. The alignment would not require an exact match so that variants could be detected.

[0097] At block 240, sequence tags that have a sequence variant at a locus relative to the constitutional genome are identified. It is possible that a sequence tag could have more than one variant. The variants for each locus and for each sequence tag can be tracked. A variant could be any allele that is not in the CG. For example, the CG could be heterozygous for A/T and the variant could be G or C.

[0098] At block 250, for each locus with a variant, a computer system can count a respective first number of sequence tags that align to the locus and have a sequence variant at the locus. Thus, each locus can have an associated count of the number of variants seen at the locus. Typically, fewer variants will be seen at a locus compared to sequence tags that correspond to the CG, e.g., due to the tumor DNA concentration being less than 50%. However, some samples may have a concentration of tumor DNA that is greater than 50%.

[0099] At block 260, a parameter is determined based on the respective first numbers. In one embodiment, if a respective number is greater than a cutoff value (e.g., greater than two), then the respective number can be added to a sum, which is the parameter or is used to determine the parameter. In another embodiment, the number of loci having a respective number greater than the cutoff value is used as the parameter.

[0100] At block 270, the parameter is compared to a threshold value to classify a level of cancer. As described above, the threshold value may be determined from the analysis of samples from other subjects. Depending on the healthy or cancer state of these other subjects, the classification can be determined. For example, if the other subjects had stage 4 cancer, then if the current parameter was close (e.g., within a specific range) to the value of the parameter obtained from the other subjects, then the current subject might be classified as having stage 4 cancer. However, if the parameter is exceeds the threshold (i.e., greater than or less, depending on how the parameter is defined), then the classification can be identified as being less than stage 4. A similar analysis can be made when the other subjects do not have cancer.

[0101] Multiple thresholds may be used to determine the classification, where each threshold is determined from a different set of subjects. Each set of subjects may have a

IDT Exhibit 1020.1694

common level of cancer. Thus, the current parameter may be compared to the values for each set of subjects, which can provide a match to one of the sets or provide a range. For example, the parameter might be about equal to the parameter obtained for subjects that are precancerous or at stage 2. As another example, the current parameter can fall in a range that can possibly match to several different levels of cancer. Thus, the classification can include more than one level of cancer.

V. USING REFERENCE GENOME

[0102] The genomic sequences of both the constitutional DNA and the DNA from the biological sample can be compared to the human reference genome. When there are more changes in the plasma sample than the constitutional DNA as compared with the reference genome, then there is a higher probability for cancer. In one embodiment, the homozygous loci in the reference genome are studied. The amounts of heterozygous loci in both the constitutional DNA and DNA from the biological sample are compared. When the amount of heterozygous sites detected from the DNA of the biological sample exceeds that of the constitutional DNA, there is a higher probability of cancer.

[0103] The analysis could also be limited to loci that are homozygous in the CG. SNMs can be defined for heterozygous loci as well, but this would generally require the generation of a third variant. In other words, if the heterozygous locus is A/T, a new variant would be either C or G. Identifying SNMs for homozygous loci is generally easier.

[0104] The degree to which an increase in the amount of heterozygous loci in the biological sample DNA relative to the constitutional DNA can be suggestive of cancer or a premalignant state when compared to the rate of change seen in healthy subjects. For example, if the degree of increase in such sites exceeds that observed in healthy subjects by a certain threshold, one can consider the data to be suggestive of cancer or a premalignant state. In one embodiment, the distribution of mutations in subjects without cancer is ascertained and a threshold can be taken as a certain number of standard deviations (e.g., 2 or 3 standard deviations).

[0105] One embodiment can require at least a specified number of variants at a locus before that locus is counted. Another embodiment provides a test even for the data based on seeing a change once. For example, when the total number of variations (errors + genuine

IDT Exhibit 1020.1695

mutations or polymorphisms) seen in plasma is statistically significantly higher than that in the constitutional DNA, then there is evidence for cancer.

[0106] FIG. 3 shows a flowchart of a method 300 comparing the sample genome (SG) to the constitutional genome (CG) using the reference genome (RG) according to embodiments of the present invention. Method 300 assumes that the RG is already obtained, and that the sequence tags for the biological sample have already been received.

[0107] At block 310, at least a portion of the sequence tags are aligned to the reference genome. The alignment can allow mismatches as variations are being detected. The reference genome can be from a similar population as the subject. The aligned sequence tags effectively comprise the sample genome (SG)

[0108] At block 320, a first number (A) of potential variants, e.g., single nucleotide mutations (SNMs), are identified. The potential SNMs are loci where a sequence tag of the SG shows a nucleotide that is different from the RG. Other criteria may be used, e.g., the number of sequence tags showing a variation must be greater than a cutoff value and whether a locus is homozygous in the RG. The set of potential SNMs may be represented as set A when specific loci are identified and tracked by storing the loci in memory. The specific loci may be determined or simply a number of such SNMs can be determined.

[0109] At block 330, a constitutional genome is determined by aligning sequence tags obtained by sequencing DNA fragments from a constitutional sample to a reference genome. This step could have been performed at any time previously and using a constitutional sample obtained at any time previously. The CG could simply be read from memory, where the aligning was previously done. In one embodiment, the constitutional sample could be blood cells.

[0110] At block 340, a second number (B) of loci where an aligned sequence tag of the CG has a variant (e.g., an SNM) at a locus relative to the reference genome are identified. If a set of loci is specifically tracked, then B can represent the set, as opposed to just a number.

[0111] At block 350, set B is subtracted from set A to identify variants (SNMs) that are present in the sample genome but not in CG. In one embodiment, the set of SNMs can be limited to nucleotide positions that the CG is homozygous. To achieve this filtering, specific loci where the CG is homozygous can be identified in set C. In another embodiment, a locus is not counted in the first number A or the second number B, if the CG is not homozygous at

IDT Exhibit 1020.1696

the locus. In another embodiment, any known polymorphism (e.g. by virtue of its presence in a SNP database) can be filtered out.

[0112] In one embodiment, the subtraction in block 350 can simply be a subtraction of numbers, and thus specific potential SNMs are not removed, but simply a value is subtracted. In another embodiment, the subtraction takes a difference between set A and set B (e.g., where set B is a subset of set A) to identify the specific SNMs that are not in set B. In logical values, this can be expressed as [A AND NOT(B)]. The resulting set of identified variants can be labeled C. The parameter can be determined as the number C or determined from the set C.

[0113] In some embodiments, the nature of the mutations can be taken into consideration and different weighting attributed to different classes of mutations. For example, mutations that are commonly associated with cancer can be attributed a higher weighting (also called an importance value when referring to relative weightings of loci). Such mutations can be found in databases of tumor-associated mutations, e.g., the Catalogue of Somatic Mutations in Cancer (COSMIC) (www.sanger.ac.uk/genetics/CGP/cosmic/). As another example, mutations associated with non-synonymous changes can be attributed a higher weighting.

[0114] Thus, the first number A could be determined as a weighted sum, where the count of tags showing a variant at one locus may have a different weighting than the count of tags at another locus. The first number A can reflect this weighted sum. A similar calculation can be performed B, and thus the number C and the parameter can reflect this weighting. In another embodiment, the weightings are accounted for when a set C of specific loci is determined. For example, a weighted sum can be determined for the counts for the loci of set C. Such weights can be used for other methods described herein.

[0115] Accordingly, the parameter that is compared to a threshold to determine the classification of a level of cancer can be the number of loci exhibiting a variation for the SG and the CG relative to the RG. In other embodiments, the total number of DNA fragments (as counted via the sequence tags) showing a variation can be counted. In other embodiments, such numbers can be used in another formula to obtain the parameter.

[0116] In one embodiment, the concentration of the variant at each locus can be a parameter and compared with a threshold. This threshold can be used to determine if a locus is a potential variant locus (in addition to the cutoff of a specific number of reads showing the

IDT Exhibit 1020.1697

variant), and then have the locus be counted. The concentration could also be used as a weighting factor in a sum of the SNMs.

VI. DECREASING FALSE POSITIVES USING CUTOFF VALUES

[0117] As mentioned above, the single nucleotide mutations can be surveyed in a large number of cell-free DNA fragments (e.g. circulating DNA in plasma) for a large genomic region (e.g. the entire genome) or a number of genomic regions to improve the sensitivity of the approach. However, analytical errors such as sequencing errors can affect the feasibility, accuracy and specificity of this approach. Here, we use the massively parallel sequencing platform as an example to illustrate the importance of sequencing errors. The sequencing error rate of the Illumina sequencing-by-synthesis platform is approximately 0.1% to 0.3% per sequenced nucleotide (Minoche et al. Genome Biol 2011, 12:R112). Any massively parallel sequencing platform may be used, including a sequencing-by-ligation platform (e.g. the Life Technologies SOLiD platform), the Ion Torrent/Ion Proton, semiconductor sequencing, Roche 454, single molecular sequencing platforms (e.g. Helicos, Pacific Biosciences and nanopore).

[0118] In a previous study on hepatocellular carcinoma, it was shown that there are approximately 3,000 single nucleotide mutations for the whole cancer genome (Tao Y et al. 2011 Proc Natl Acad Sci USA; 108: 12042-12047). Assuming that only 10% of the total DNA in the circulation is derived from the tumor cells and we sequence the plasma DNA with an average sequencing depth of one fold haploid genome coverage, we would encounter 9 million ($3 \times 10^9 \times 0.3\%$) single nucleotide variations (SNVs) due to sequencing errors. However, most of the single nucleotide mutations are expected to occur on only one of the two homologous chromosomes. With a sequencing depth of one-fold haploid genome coverage of a sample with 100% tumor DNA, we expect to detect only half of the 3,000 mutations, i.e. 1,500 mutations. When we sequence the plasma sample containing 10% tumor-derived DNA to one haploid genome coverage, we expect to detect only 150 (1,500 x 10%) cancer-associated single nucleotide mutations. Thus, the signal-to-noise ratio for the detection of cancer-associated mutations is 1 in 60,000. This very low signal-to-noise ratio suggests that the accuracy of using this approach for differentiating normal and cancer cases would be very low if we simply use all the single nucleotide changes in the biological sample (e.g., plasma) as a parameter.

IDT Exhibit 1020.1698

[0119] It is expected that with the progress in sequencing technologies, there would be continual reduction in the sequencing error rate. One can also analyze the same sample using more than one sequencing platform and through a comparison of the cross-platform sequencing results, pinpoint the reads likely to be affected by sequencing errors. Another approach is to analyze two samples taken at different times from the same subject. However, such approaches are time consuming.

[0120] In one embodiment, one way to enhance the signal-to-noise ratio in the detection of single nucleotide mutations in the plasma of cancer patients is to count a mutation only if there are multiple occurrences of the same mutation in the sample. In selected sequencing platforms, the sequencing errors involving particular nucleotide substitutions might be more common and would affect the sequencing results of the test sample and the constitutional DNA sample of both the test subject and the control subjects. However, in general, sequencing errors occur randomly.

[0121] The chance of having a sequencing error is exponentially lower when one observes the same change at the same nucleotide position in multiple DNA fragments. On the other hand, the chance of detecting a genuine cancer-associated mutational change in the sample is affected by the sequencing depth and the fractional concentration of the tumoral DNA in the sample. The chance of observing the mutation in multiple DNA fragments would increase with the sequencing depth and fractional concentration of tumoral DNA. In various embodiments using samples with cell-free tumoral DNA (such as in plasma), the fractional concentration can be 5%, 10%, 20%, and 30%. In one embodiment, the fractional concentration is less than 50%.

[0122] FIG. 4 is a table 400 showing the number of cancer-associated single nucleotide mutations correctly identified using different number of occurrences as the criterion for classifying a mutation as being present in the sample according to embodiments of the present invention. The numbers of nucleotide positions that are falsely identified as having mutation because of sequencing error based on the same classification criteria are also shown. The sequencing error rate is assumed to be 0.1% (Minoche et al. Genome Bio 2011, 12:R112). The fractional concentration of tumor-derived DNA in the sample is assumed to be 10%.

IDT Exhibit 1020.1699

[0123] FIG. 4 shows that the ratio between the number of cancer-associated mutations detected in the plasma and the number of false-positive calls would increase exponentially with the increasing number of times the same change is seen in the sample for defining a mutation, when the fractional concentration of tumor-derived DNA in the sample is assumed to be 10%. In other words, both the sensitivity and specificity for cancer mutation detection would improve. In addition, the sensitivity for detecting the cancer-associated mutations is affected by the sequencing depth. With 100-fold haploid genome coverage of sequencing, 2,205 (73.5%) of the 3,000 mutations can be detected even using the criterion of the occurrence of the particular mutation in at least 4 DNA fragments in the sample. Other values for the minimum number of fragments may be used, such as 3, 5, 8, 10, and greater than 10.

[0124] FIG. 5 is a table 500 showing the expected number of false-positive loci and the expected number of mutations identified when the fractional concentration of tumor-derived DNA in the sample is assumed to be 5%. With a lower fractional concentration of tumor-derived DNA in the sample, a higher sequencing depth would be required to achieve the same sensitivity of detecting the cancer-associated mutations. A more stringent criterion would also be required to maintain the specificity. For example, the criterion of the occurrence of the particular mutation in at least 5 DNA fragments, instead of the criterion of at least 4 occurrences in the situation of 10% tumor DNA fraction, in the sample would need to be used. Tables 400 and 500 provide guidance for the cutoff value to use given the fold coverage and a tumor DNA concentration, which can be assumed or measured as described herein.

[0125] Another advantage of using the criteria of detecting a single nucleotide change more than one time to define a mutation is that this is expected to minimize false positives detection because of single nucleotide changes in non-malignant tissues. As nucleotide changes can occur during mitosis of normal cells, each healthy cell in the body can harbor a number of single nucleotide changes. These changes may potentially lead to false positive results. However, the changes of a cell would be present in the plasma/serum when the cell dies. While different normal cells are expected to carry different sets of mutations, the mutations occurring in one cell are unlikely to be present in numerous copies in the plasma/serum. This is in contrast to mutations within tumor cells where multiple copies are expected to be seen in plasma/serum because tumor growth is clonal in nature. Thus,

IDT Exhibit 1020.1700

multiple cells from a clone would die and release the signature mutations representative of the clones.

[0126] In one embodiment, target enrichment for specific genomic regions can be performed before sequencing. This target enrichment step can increase the sequencing depth of the regions of interest with the same total amount of sequencing performed. In yet another embodiment, a round of sequencing with relatively low sequencing depth can first be performed. Then regions showing at least one single nucleotide change can be enriched for a second round of sequencing which has higher fold coverage. Then, the criterion of multiple occurrences can be applied to define a mutation for the sequencing results with target enrichment.

VII. DYNAMIC CUTOFFS

[0127] As described above, a cutoff value N for the number of reads supporting a variant (potential mutation) can be used to determine whether a locus qualifies as a mutation (e.g., an SNM) to be counted. Using such a cutoff can reduce false positives. The discussion below provides methods for selecting a cutoff for different loci. In the following embodiments, we assume that there is a single predominant cancer clone. Similar analysis can be carried out for scenarios involving multiple clones of cancer cells releasing different amounts of tumor DNA into the plasma.

A. Number Of Cancer-Associated Mutations Detected In Plasma

[0128] The number of cancer-associated mutations detectable in plasma can be affected by a number of parameters, for example: (1) The number of mutations in the tumor tissue (N_T) - the total number of mutations present in the tumor tissue is the maximum number of tumor-associated mutations detectable in the plasma of the patient; (2) The fractional concentration of tumor-derived DNA in the plasma (f) - the higher the fractional concentration of tumor-derived DNA in the plasma, the higher the chance of detecting the tumor-associated mutations in the plasma would be; (3) Sequencing depth (D) - Sequencing depth refers to the number of times the sequenced region is covered by the sequence reads. For example, an average sequencing depth of 10-fold means that each nucleotide within the sequenced region is covered on average by 10 sequence reads. The chance of detecting a cancer-associated mutation would increase when the sequencing depth is increased; and (4) The minimum number of times a nucleotide change that is detected in the plasma so as to define it as a

IDT Exhibit 1020.1701

potential cancer-associated mutation (r), which is a cutoff value used to discriminate sequencing errors from real cancer-associated mutations.

[0129] In one implementation, the Poisson distribution is used to predict the number of cancer-associated mutations detected in plasma. Assuming that a mutation is present in a nucleotide position on one of the two homologous chromosomes, with a sequencing depth of D, the expected number of times a mutation is present in the plasma (M_P) is calculated as: $M_P = D \times f/2$.

[0130] The probability of detecting the mutation in the plasma (Pb) at a particular mutation site is calculated as:

$$Pb = 1 - \sum_{i=0}^{r-1} Poisson(i, M_P)$$

where r (cutoff value) is the number of times that a nucleotide change is seen in the plasma so as to define it as a potential tumor-associated mutation; Poisson(i,M_P) is the Poisson distribution probability of having *i* occurrences with an average number of M_P.

[0131] The total number of cancer-associated mutations expected to be detected in the plasma (NP) can be calculated as: $N_P = N_T \times Pb$, where N_T is the number of mutations present in the tumor tissue. The following graphs show the percentages of tumor-associated mutations expected to be detected in the plasma using different criteria of occurrences (r) for calling a potential mutation and different sequencing depths.

[0132] FIG. 6A is a graph 600 showing the detection rate of cancer-associated mutations in plasma with 10% and 20% plasma fractional concentrations of tumor-derived DNA and using four and six occurrences (r) as criteria for calling potential cancer-associated mutations. With the same r, a higher fractional concentration of tumor-derived DNA in plasma would result in a higher number of cancer-associated mutations detectable in the plasma. With the same fractional concentration of tumor-derived DNA in plasma, a higher r would result in a smaller number of detected mutations.

B. Number Of False-Positive Single Detected Due To Errors

[0133] Single nucleotide changes in the plasma DNA sequencing data can occur due to sequencing and alignment errors. The number of nucleotide positions with false-positive

IDT Exhibit 1020.1702

single nucleotide changes can be predicted mathematically based on a binomial distribution. The parameters affecting the number of false-positive sites (N_{FP}) can include: (1) Sequencing error rate (E) - Sequencing error rate is defined as the proportion of sequenced nucleotide being incorrect; (2) Sequencing depth (D) - With a higher sequencing depth, the number of nucleotide positions showing a sequencing error would increase; (3) The minimum number of occurrences of the same nucleotide change for defining a potential cancer-associated mutation (r); and (4) The total number of nucleotide positions within the region-of-interest (N_I).

[0134] The occurrence of mutations can generally be regarded as a random process. Therefore, with the increase of the criteria of occurrence for defining a potential mutation, the number of false-positive nucleotide positions would exponentially decrease with r. In some of the existing sequencing platforms, certain sequence contexts are more prone to having sequencing errors. Examples of such sequencing contexts include the GGC motif, homopolymers (e.g. AAAAAAA), and simple repeats (e.g. ATATATATAT). These sequence contexts will substantially increase the single nucleotide change or insertion/deletion artifacts (Nakamura K et al. Nucleic Acids Res 2011;39,e90 and Minoche AE et al. Genome Biol 2011;12,R112). In addition, repeat sequences, such as homopolymers and simple repeats, would computationally introduce ambiguities in alignment and, hence, lead to false-positive results for single nucleotide variations.

[0135] The larger the region-of-interest, the higher the number of false-positive nucleotide positions would be observed. If one is looking for mutations in the whole genome, then the region-of-interest would be the whole genome and the number of nucleotides involved would be 3 billion. On the other hand, if one focuses on the exons, then the number of nucleotides encoding the exons, i.e. approximately 45 million, would constitute the region-of-interest.

[0136] The number of false-positive nucleotide positions associated with sequencing errors can be determined based on the following calculations. The probability (P_{Er}) of having the same nucleotide change at the same position due to sequencing errors can be calculated as:

$$P_{Er} = C(D,r)E\left(\frac{E}{3}\right)^{r-1}$$

IDT Exhibit 1020.1703

where $C(D,r)$ is the number of possible combinations for choosing r elements from a total of D elements; r is the number of occurrences for defining a potential mutation; D is the sequencing depth; and E is the sequencing error rate. $C(D,r)$ can be calculated as:

$$C(D,r) = \frac{\mathrm{D!}}{\mathrm{r!\,(D-r)!}}$$

[0137] The number of nucleotide positions (N_{FP}) being false-positives for mutations can be calculated as:

$$N_{FP} = N_I P_{Er}$$

where N_I is the total number of nucleotide positions in the region-of-interest.

[0138] FIG. 6B is a graph 650 showing the expected number of nucleotide positions falsely classified as having a nucleotide change using criteria of occurrence (r) of 4, 5, 6 and 7 vs. sequencing depth. The region-of-interest is assumed to be the whole genome (3 billion nucleotide positions) in this calculation. The sequencing error rate is assumed to be 0.3% of the sequenced nucleotides. As one can see, the value of *r* has a significant impact on the false positives. But, as can be seen from FIG. 6A, a higher value of *r* also reduces the number of mutations detected, at least until significantly higher sequencing depths are used.

C. Choosing Minimum Occurrence (r)

[0139] As discussed above, the number of true cancer-associated mutation sites and false-positive sites due to sequencing errors would increase with sequencing depth. However, their rates of increase would be different. Therefore, it is possible to make use of the choice of sequencing depth and the value of r to maximize the detection of true cancer-associated mutations while keeping the number of false-positive sites at a low value.

[0140] FIG. 7A is a graph 700 showing the number of true cancer-associated mutation sites and false-positive sites with difference sequencing depths. The total number of cancer-associated mutations in the tumor tissue is assumed to be 3,000 and the fractional concentration of tumor-derived DNA in the plasma is assumed to be 10%. The sequencing error rate is assumed to be 0.3%. In the legend, TP denotes the true-positive sites at which a corresponding mutation is present in the tumor tissue, and FP denotes false-positive sites at which no corresponding mutation is present in the tumor tissue and the nucleotide changes present in the sequencing data are due to sequencing errors.

IDT Exhibit 1020.1704

[0141] From graph 700, at a sequencing depth of 110-fold, approximately 1,410 true cancer-associated mutations would be detected if we use the minimum occurrence of 6 as the criterion (r=6) to define a potential mutation site in the plasma. Using this criterion, only approximately 20 false-positive sites would be detected. If we use the minimum of 7 occurrences (r=7) as the criterion to define a potential mutation, the number of cancer-associated mutations that could be detected would be reduced by 470 to approximately 940. Therefore, the criterion of r=6 would make the detection of cancer-associated mutations in plasma more sensitive.

[0142] On the other hand, at a sequencing depth of 200-fold, the number of true cancer-associated mutations detected would be approximately 2,800 and 2,600, if we use the criteria of minimum occurrence (r) of 6 and 7, respectively, to define potential mutations. Using these two values of *r*, the numbers of false-positive sites would be approximately 740 and 20, respectively. Therefore, at a sequencing depth of 200-fold, the use of a more stringent criterion of r=7 for defining a potential mutation can greatly reduce the number of false-positive sites without significantly adversely affecting the sensitivity for detecting the true cancer-associated mutations.

D. Dynamic cutoff for sequencing data for defining potential mutations in plasma

[0143] The sequencing depth of each nucleotide within the region-of-interest would be different. If we apply a fixed cutoff value for the occurrence of a nucleotide change to define a potential mutation in plasma, the nucleotides that are covered by more sequence reads (i.e. a higher sequencing depth) would have higher probabilities of being falsely labeled as having nucleotide variation in the absence of such a change in the tumor tissue due to sequencing errors compared with nucleotides that have lower sequencing depths. One embodiment to overcome this problem is to apply a dynamic cutoff value of *r* to different nucleotide positions according to the actual sequencing depth of the particular nucleotide position and according to the desired upper limit of the probability for calling false-positive variations.

[0144] In one embodiment, the maximum allowable false-positive rate can be fixed at 1 in 1.5×10^8 nucleotide positions. With this maximum allowable false-positive rate, the total number of false-positive sites being identified in the whole genome would be less than 20. The value of *r* for different sequencing depths can be determined according to the curves

IDT Exhibit 1020.1705

shown in FIG. 6B and these cutoffs are shown in Table 1. In other embodiments, other different maximum allowable false-positive rates, e.g. 1 in 3 x 10^8, 1 in 10^8 or 1 in 6 x 10^7, can be used. The corresponding total number of false-positive sites would be less than 10, 30 and 50, respectively.

Sequencing depth of a particular nucleotide position	Minimum number of occurrence of a nucleotide change to be present in the plasma DNA sequencing data to define a potential mutation (r)
<50	5
50 – 110	6
111 – 200	7
201 – 310	8
311 – 450	9
451 – 620	10
621 – 800	11

Table 1. The minimum number of occurrences of a nucleotide change present in plasma to define a potential mutation (r) for different sequencing depths of the particular nucleotide position. The maximum false-positive rate is fixed at 1 in 1.5x10^8 nucleotides.

E. Target-enrichment sequencing

[0145] As shown in FIG. 7A, a higher sequencing depth can result in a better sensitivity for detecting cancer-associated mutations while keeping the number false-positive sites low by allowing the use of a higher value of *r*. For example, at a sequencing depth of 110-fold, 1,410 true cancer-associated mutations can be detected in the plasma using an *r* value of 6 whereas the number of true cancer-associated mutations detected would be 2,600 when the sequencing depth increases to 200-fold and an *r* value of 7 is applied. The two sets of data would give an expected number of false-positive sites of approximately 20.

[0146] While the sequencing of the whole genome to a depth of 200-fold is relatively expensive at present, one possible way of achieving such a sequencing depth would be to focus on a smaller region-of-interest. The analysis of a target region can be achieved for example by, but not limited to, the use of DNA or RNA baits to capture genomic regions of interest by hybridization. The captured regions are then pulled down, e.g., by magnetic means and then subjected to sequencing. Such target capture can be performed, for example, using the Agilent SureSelect target enrichment system, the Roche Nimblegen target enrichment

IDT Exhibit 1020.1706

system and the Illumina targeted resequencing system. Another approach is to perform PCR amplification of the target regions and then perform sequencing. In one embodiment, the region-of-interest is the exome. In such an embodiment, target capturing of all exons can be performed on the plasma DNA, and the plasma DNA enriched for exonic regions can then be sequenced.

[0147] In addition to having higher sequencing depth, the focus on specific regions instead of analyzing the whole genome would significantly reduce the number of nucleotide positions in the search space and would lead to a reduction in the number of false-positive sites given the same sequencing error rate.

[0148] FIG. 7B is a graph 750 showing the predicted number of false-positive sites involving the analysis of the whole genome (WG) and all exons. For each type of analysis, two different values, 5 and 6, for r are used. At a sequencing depth of 200-fold, if $r = 5$ is used to define mutations in plasma, the predicted number of false-positive sites are approximately 23,000 and 230 for the whole genome and all exons, respectively. If $r = 6$ is used to define mutations in plasma, the predicted number of false-positive sites are 750 and 7, respectively. Therefore, the limit of the number of nucleotides in the region-of-interest can significantly reduce the number of false-positives in plasma mutational analysis.

[0149] In exon-capture or even exome-capture sequencing, the number of nucleotides in the search space is reduced. Therefore, even if we allow a higher false-positive rate for the detection of cancer-associated mutations, the absolute number of false-positive sites can be kept as a relatively low level. The allowance of higher false-positive rate would allow a less stringent criterion of minimum occurrences (r) for defining a single nucleotide variation in plasma to be used. This would result in a higher sensitivity for the detection of true cancer-associated mutations.

[0150] In one embodiment, we can use a maximum allowable false-positive rate of 1.5×10^{-6}. With this false-positive rate, the total number of false-positive sites within the targeted exons would only be 20. The values of r for different sequencing depths using a maximum allowable false-positive rate of 1.5×10^{-6} are shown in Table 2. In other embodiments, other different maximum allowable false-positive rates, e.g. 1 in 3×10^{-6}, 1 in 10^{-6} or 1 in 6×10^{-5}, can be used. The corresponding total number of false-positive sites would be less than 10, 30

IDT Exhibit 1020.1707

and 50, respectively. In one embodiment, different classes of mutations can be attributed different weightings, as described above.

Sequencing depth of a particular nucleotide position	Minimum number of occurrence of a nucleotide change to be present in the plasma DNA sequencing data to define a potential mutation (r)
<50	4
50 – 125	5
126 – 235	6
236 – 380	7
381 – 560	8
561 – 760	9

Table 2. The minimum number of occurrence of a nucleotide change present in plasma to define a potential mutation (r) for different sequencing depths of the particular nucleotide position. The maximum false-positive rate is fixed at 1 in 1.5×10^6 nucleotides.

VIII. CANCER DETECTION

[0151] As mentioned above, the counts of sequence tags at variant loci can be used in various ways to determine the parameter, which is compared to a threshold to classify a level of cancer. The fractional concentration of variant reads relative to all reads at a locus or many loci is another parameter that may be used. Below are some examples of calculating the parameter and the threshold.

A. Determination of Parameter

[0152] If the CG is homozygous at a particular locus for a first allele and a variant allele is seen in the biological sample (e.g., plasma), then the fractional concentration can be calculated as $2p / (p+q)$, where p is the number of sequence tags having the variant allele and q is the number of sequence tags having the first allele of the CG. This formula assumes that that only one of the haplotypes of the tumor has the variant, which would typically be the case. Thus, for each homozygous locus a fractional concentration can be calculated. The fractional concentrations can be averaged. In another embodiment, the count p can include the number of sequence tags for all of the loci, and similarly for the count q, to determine the fractional concentration. An example is now described.

IDT Exhibit 1020.1708

[0153] The genomewide detection of tumor derived single nucleotide variants (SNVs) in the plasma of the 4 HCC patients was explored. We sequenced tumor DNA and buffy coat DNA to mean depths of 29.5-fold (range, 27-fold to 33-fold) and 43-fold (range, 39-fold to 46-fold) haploid genome coverage, respectively. The MPS data from the tumor DNA and the buffy coat DNA from each of the 4 HCC patients were compared, and SNVs present in the tumor DNA but not in the buffy coat DNA were mined with a stringent bioinformatics algorithm. This algorithm required a putative SNV to be present in at least a threshold number of sequenced tumor DNA fragments (i.e. in a corresponding sequenced tag) before it would be classified as a true SNV. The threshold number was determined by taking into account the sequencing depth of a particular nucleotide and the sequencing error rate, e.g., as described herein.

[0154] FIG. 8 is a table 800 showing results for 4 HCC patients before and after treatment, including fractional concentrations of tumor-derived DNA in plasma according to embodiments of the present invention. The number of tumor-associated SNVs ranged from 1,334 to 3,171 in the 4 HCC cases. The proportions of such SNVs that were detectable in plasma are listed before and after treatment. Before treatment, 15%–94% of the tumor associated SNVs were detected in plasma. After treatment, the percentage was between 1.5%-5.5%. Thus, the number of detected SNVs does correlate to a level of cancer. This shows that the number of SNVs can be used as a parameter to classify a level of cancer.

[0155] The fractional concentrations of tumor-derived DNA in plasma were determined by the fractional counts of the mutant with respect to the total (i.e., mutant plus wild type) sequences. The formula is $2p/(p+q)$, where the 2 accounts for just one haplotype being mutated on the tumor. These fractional concentrations were well correlated with those determined with genomewide aggregated allelic loss (GAAL) analysis (Chan KC et al. Clin Chem 2013;59:211-24) and were reduced after surgery. Thus, the fractional concentration is also shown to be a usable parameter for determining a level of cancer.

[0156] The fractional concentration from the SNV analysis can convey a tumor load. A cancer patient with a higher tumor load (e.g., a higher deduced fractional concentration) will have a higher frequency of somatic mutations than one with a lower tumor load. Thus, embodiments can also be used for prognostication. In general, cancer patients with higher tumor loads have worse prognosis than those with lower tumor loads. The former group would thus have a higher chance of dying from the disease. In some embodiments, if the

IDT Exhibit 1020.1709

absolute concentration of DNA in a biological sample, e.g. plasma, can be determined (e.g. using real-time PCR or fluorometry), then the absolute concentration of tumor-associated genetic aberrations can be determined and used for clinical detection and/or monitoring and/or prognostication.

B. Determining Of Threshold

[0157] Table 800 may be used to determine a threshold. As mentioned above, the number of SNVs and a fractional concentration determined by SNV analysis correlate to a level of cancer. The threshold can be determined on an individual basis. For example, the pre-treatment value can be used to determine the threshold. In various implementations, the threshold could be a relative change from the pre-treatment of an absolute value. A suitable threshold could be a reduction in number of SNVs or fractional concentration by 50%. Such a threshold would provide a classification of a lower level of cancer for each of the cases in table 800. Note that such threshold may be dependent on the sequencing depth.

[0158] In one embodiment, a threshold could be used across samples, and may or may not account for pre-treatment values for the parameter. For example, a threshold of 100 SNVs could be used to classify the subject as having no cancer or a low level of cancer. This threshold of 100 SNVs is satisfied by each of the four cases in table 800. If the fractional concentration was used as the parameter, a threshold of 1.0% would classify HCC1-HCC3 as practically zero level of cancer, and a second threshold of 1.5 % would classify HCC4 as a low level of cancer. Thus, more than one threshold may be used to obtain more than two classifications.

[0159] To illustrate other possible thresholds, we analyzed the plasma of the healthy controls for the tumor-associated SNVs. Numerous measurements can be made of healthy subjects to determine a range of how many variations are expected from the biological sample relative to the constitutional genome.

[0160] FIG. 9 is a table 900 showing detection of the HCC-associated SNVs in 16 healthy control subjects according to embodiments of the present invention. Table 900 can be used to estimate the specificity of an SNV analysis approach. The 16 healthy controls are listed as different rows. The columns investigate the SNVs detected for the specific HCC patients, and show the number of sequence reads at variant loci having the variant allele and the number of sequence reads with the wildtype allele (i.e., the allele from the CG). For

IDT Exhibit 1020.1710

example, for HCC1, control C01 had 40 variant reads at such variant loci, but 31,261 reads of the wildtype allele. The last column shows the total fractional concentration across all of the SNVs for the HCC1 patients. As the HCC-associated SNVs were specific for the HCC patients, the presence of the HCC-associated SNVs represent false-positives. If a cutoff values, as described herein, are applied to these apparent sequence variants, all of these false-positives would be filtered away.

[0161] The presence of a small number of these putative tumor-associated mutations in the plasma of the 16 healthy controls represented the "stochastic noise" of this method and was likely due to sequencing errors. The mean fractional concentration estimated from such noise was 0.38%. These values show a range for healthy subjects. Thus, a threshold value for a classification of zero level of cancer for HCC could be about 0.5%, since the highest fractional concentration was 0.43%. Thus, if all the cancer cells are removed from an HCC patient, these low fractional concentrations would be expected.

[0162] Referring back to table 800, if 0.5 % was used as a threshold for zero level of cancer, then the post-treatment plasma data for HCC1 and HCC3 would be determined as having zero level based on the SNV analysis. HCC2 might be classified as one level up from zero. HCC4 might also be classified as one level up from zero, or some higher level, but still a relatively low level compared to the pre-treatment samples.

[0163] In one embodiment where the parameter corresponds to the number of variant loci, the threshold could be zero (i.e., one variant locus could indicate a non-zero level of cancer). However, with many settings (e.g., of depth), the threshold would be higher, e.g., an absolute value of 5 or 10. In one implementation where a person is monitored after treatment, the threshold can be a certain percentage of SNVs (identified by analyzing the tumors directly) showing up in the sample. If the cutoff value for the number of variant reads required at a locus was large enough, just having one variant loci might be indicative of a non-zero level of cancer.

[0164] Thus, quantitative analysis of variations (e.g., single nucleotide variations) in DNA from a biological sample (e.g., plasma) can be used for the diagnosis, monitoring and prognostication of cancer. For the detection of cancer, the number of single nucleotide variations detected in the plasma of a tested subject can be compared with that of a group of healthy subjects. In the healthy subjects, the apparent single nucleotide variations in plasma

IDT Exhibit 1020.1711

can be due to sequencing errors, non-clonal mutations from the blood cells and other organs. It has been shown that the cells in normal healthy subjects could carry a small number of mutations (Conrad DF et al. Nat Genet 2011;43:712-4), as shown in table 900. Thus, the overall number of apparent single nucleotide variations in the plasma of a group of apparently healthy subjects can be used as a reference range to determine if the tested patient has an abnormally high number of single nucleotide variations in plasma corresponding to a non-zero level of cancer.

[0165] The healthy subjects used for establishing the reference range can be matched to the tested subject in terms of age and sex. In a previous study, it has been shown that the number of mutations in the somatic cells would increase with age (Cheung NK et al, JAMA 2012;307:1062-71). Thus, as we grow older, then it would be 'normal' for one to accumulate clones of cells, even though they are relatively benign most of the time, or would take a very long time to become clinically significant. In one embodiment, reference levels can be generated for different subject groups, e.g. different age, sex, ethnicity, and other parameters (e.g. smoking status, hepatitis status, alcohol, drug history).

[0166] The reference range can vary based on the cutoff value used (i.e., the number of variant sequence tags required at a locus), as well as the assumed false positive rate and other variables (e.g., age). Thus, the reference range may be determined for a particular set of one or more criteria, and the same criteria would be used to determine a parameter for a sample. Then, the parameter can be compared to the reference range, since both were determined using the same criteria.

[0167] As mentioned above, embodiments may use multiple thresholds for determining a level of cancer. For example, a first level could determine no signs of cancer for parameters below the threshold, and at least a first level of cancer, which could be a pre-neoplastic level. Other levels could correspond to different stages of cancer.

C. Dependency on Experimental Variables

[0168] The depth of sequencing can be important for establishing the minimum detection threshold of the minority (e.g. tumor) genome. For example, if one uses a sequencing depth of 10 haploid genomes, then the minimum tumoral DNA concentration that one could detect even with a sequencing technology without any error is 1/5, i.e. 20%. On the other hand, if one uses a sequencing depth of 100 haploid genomes, then one could go down to 2%. This

IDT Exhibit 1020.1712

analysis is referring to the scenario that only one mutation locus is being analyzed. However, when more mutation loci are analyzed, the minimum tumoral DNA concentration can be lower and is governed by a binomial probability function. For example, if the sequencing depth is 10 folds and the fractional concentration of tumoral DNA is 20%, then the chance of detecting the mutation is 10%. However, if we have 10 mutations, then the chance of detecting at least one mutation would be $1 - (1 - 10\%)^{10} = 65\%$.

[0169] There are several effects for increasing the sequencing depth. The higher the sequencing depth, the more sequencing errors would be seen, see FIGS. 4 and 5. However, with a higher sequencing depth, one can more easily differentiate sequencing errors from mutations due to clonal expansion of a subpopulation of cells (e.g. cancer cells) because the sequencing errors would occur randomly in the genome but the mutations would occur at the same location for the given population of cells.

[0170] The higher the sequencing depth, the more mutations from the "healthy cells" would be identified. However, when there is no clonal expansion of these healthy cells and their mutational profiles are different, then the mutations in these healthy cells can be differentiated from the mutations by their frequencies of occurrence in the plasma (e.g., by using a cutoff N for a required number of reads exhibiting the mutation, such as having N equal to 2, 3, 4, 5, or larger).

[0171] As mentioned above, the threshold can depend on an amount of mutations in healthy cells that would be clonally expanded, and thus might not be filtered out through other mechanisms. This variance that one would expect can be obtained by analyzing healthy subjects. As clonal expansion occurs over time, the age of patient can affect a variance that one sees in healthy subjects, and thus the threshold can be dependent on age.

D. Combination With Targeted Approaches

[0172] In some embodiments, a random sequencing can be used in combination with targeted approaches. For example, one can perform random sequencing of a plasma sample upon presentation of a cancer patient. The sequencing data of plasma DNA can be analyzed for copy number aberrations and SNVs. The regions showing aberrations (e.g., amplification/deletion or high density of SNVs) can be targeted for serial monitoring purposes. The monitoring can be done over a period of time, or done immediately after the random sequencing, effectively as a single procedure. For the targeted analysis, solution-

IDT Exhibit 1020.1713

phase hybridization-based capture approaches have been successfully used to enrich plasma DNA for noninvasive prenatal diagnosis (Liao GJ et al. Clin Chem 2011;57:92-101). Such techniques are mentioned above. Thus, the targeted and random approaches can be used in combination for cancer detection and monitoring.

[0173] Thus, one could perform targeted sequencing of the loci that are found to be potentially mutated using the non-targeted, genomewide approach mentioned above. Such targeted sequencing could be performed using solution- or solid-phase hybridization techniques (e.g. using the Agilent SureSelect, NimbleGen Sequence Capture, or Illumina targeted resequencing system) followed by massively parallel sequencing. Another approach is to perform amplification (e.g. PCR based) system for targeted sequencing (Forshew T et al. Sci Transl Med 2012; 4: 135ra68).

IX. FRACTIONAL CONCENTRATION

[0174] The fractional concentration of tumor DNA can be used to determine the cutoff value for the required number of variations at a locus before the locus is identified as a mutation. For example, if the fractional concentration was known to be relatively high, then a high cutoff could be used to filter out more false positives, since one knows that a relatively high number of variant reads should exist for true SNVs. On the other hand, if the fractional concentration was low, then a lower cutoff might be needed so that some SNVs are not missed. In this case, the fractional concentration would be determined by a different method than the SNV analysis, where it is used as a parameter.

[0175] Various techniques may be used for determining the fractional concentration, some of which are described herein. These techniques can be used to determine the fractional concentration of tumor-derived DNA in a mixture, e.g. a biopsy sample containing a mixture of tumor cells and nonmalignant cells or a plasma sample from a cancer patient containing DNA released from tumor cells and DNA released from nonmalignant cells.

A. GAAL

[0176] Genomewide aggregated allelic loss (GAAL) analyzes loci that have lost heterozygosity (Chan KC et al. Clin Chem 2013;59:211-24). For a site of the constitutional genome CG that is heterozygous, a tumor often has a locus that has a deletion of one of the alleles. Thus, the sequence reads for such a locus will show more of one allele than another,

IDT Exhibit 1020.1714

where the difference is proportional to the fractional concentration of tumor DNA in the sample. An example of such a calculation follows.

[0177] DNA extracted from the buffy coat and the tumor tissues of the HCC patients was genotyped with the Affymetrix Genome-Wide Human SNP Array 6.0 system. The microarray data were processed with the Affymetrix Genotyping Console version 4.1. Genotyping analysis and single-nucleotide polymorphism (SNP) calling were performed with the Birdseed v2 algorithm. The genotyping data for the buffy coat and the tumor tissues were used for identifying loss-of-heterozygosity (LOH) regions and for performing copy number analysis. Copy number analysis was performed with the Genotyping Console with default parameters from Affymetrix and with a minimum genomic-segment size of 100 bp and a minimum of 5 genetic markers within the segment.

[0178] Regions with LOH were identified as regions having 1 copy in the tumor tissue and 2 copies in the buffy coat, with the SNPs within these regions being heterozygous in the buffy coat but homozygous in the tumor tissue. For a genomic region exhibiting LOH in a tumor tissue, the SNP alleles that were present in the buffy coat but were absent from or of reduced intensity in the tumor tissues were considered to be the alleles on the deleted segment of the chromosomal region. The alleles that were present in both the buffy coat and the tumor tissue were deemed as having been derived from the non-deleted segment of the chromosomal region. For all the chromosomal regions with a single copy loss in the tumor, the total number of sequence reads carrying the deleted alleles and the non-deleted alleles were counted. The difference of these two values was used to infer the fractional concentration of tumor-derived DNA (F_{GAAL}) in the sample using the following equation:

$$F_{GAAL} = \frac{N_{non-del} - N_{del}}{N_{non-del}}$$

where $N_{non\text{-}del}$ represents the total number of sequence reads carrying the non-deleted alleles and N_{del} represents the total number of sequence reads carrying the deleted alleles.

B. Estimation using Genomic Representation

[0179] A problem with the GAAL technique is that particular loci (i.e. ones exhibiting LOH) are identified and only sequence reads aligning to such loci are used. Such

IDT Exhibit 1020.1715

requirement can add additional steps, and thus costs. An embodiment is now described which uses only copy number, e.g., a sequence read density.

[0180] Chromosomal aberrations, for example, amplifications and deletions are frequently observed in cancer genomes. The chromosomal aberrations observed in cancer tissues typically involve subchromosomal regions and these aberrations can be shorter than 1 Mb. And, the cancer-associated chromosomal aberrations are heterogeneous in different patients, and thus different regions may be affected in different patients. It is also not uncommon for tens, hundreds or even thousands of copy number aberrations to be found in a cancer genome. All of these factors make determining tumor DNA concentration difficult.

[0181] Embodiments involve the analysis of quantitative changes resulted from tumor-associated chromosomal aberrations. In one embodiment, the DNA samples containing DNA derived from cancer cells and normal cells are sequenced using massively parallel sequencing, for example, by the Illumina HiSeq2000 sequencing platform. The derived DNA may be cell-free DNA in plasma or other suitable biological sample.

[0182] Chromosomal regions that are amplified in the tumor tissues would have increased probability of being sequenced and regions that are deleted in the tumor tissues would have reduced probability of being sequenced. As a result, the density of sequence reads aligning to the amplified regions would be increased and that aligning to the deleted regions would be reduced. The degree of variation is proportional to the fractional concentration of the tumor-derived DNA in the DNA mixture. The higher the proportion of DNA from the tumor tissue, the larger the change would be caused by the chromosomal aberrations.

1. Estimation in sample with high tumor concentration

[0183] DNA was extracted from the tumor tissues of four hepatocellular carcinoma patients. The DNA was fragmented using the Covaria DNA sonication system and sequenced using the Illumina HiSeq2000 platform as described (Chan KC et al. Clin Chem 2013;59:211-24). The sequence reads were aligned to the human reference genome (hg18). The genome was then divided into 1 Mb bins (regions) and the sequence read density was calculated for each bin after adjustment for GC-bias as described (Chen EZ et al. PLoS One. 2011;6:e21791).

[0184] After sequence reads are aligned to a reference genome, a sequence read density can be computed for various regions. In one embodiment, the sequence read density is a

IDT Exhibit 1020.1716

proportion determined as the number of reads mapped to a particular bin (e.g., 1 Mb region) divided by the total sequence reads that can be aligned to the reference genome (e.g., to a unique position in the reference genome). Bins that overlap with chromosomal regions amplified in the tumor tissue are expected to have higher sequence read densities than those from bins without such overlaps. On the other hand, bins that overlap with chromosomal regions that are deleted are expected to have lower sequence read densities than those without such overlaps. The magnitude of the difference in sequence read densities between regions with and without chromosomal aberrations is mainly affected by the proportion of tumor-derived DNA in the sample and the degree of amplification/deletion in the tumor cells.

[0185] Various statistical models may be used to identify the bins having sequence read densities corresponding to different types of chromosomal aberrations. In one embodiment, a normal mixture model (McLachlan G and Peel D. Multvariate normal mixtures. In Finite mixture models 2004: p81-116. John Wiley & Sons Press) can be used. Other statistical models, for example the binomial mixture model and Poisson regression model (McLachlan G and Peel D. Mixtures with non-normal components, Finite mixture models 2004: p135-174. John Wiley & Sons Press), can also be used.

[0186] The sequence read density for a bin can be normalized using the sequence read density of the same bin as determined from the sequencing of the buffy coat DNA. The sequence read densities of different bins may be affected by the sequence context of a particular chromosomal region, and thus the normalization can help to more accurately identify regions showing aberration. For example, the mappability (which refers to the probability of aligning a sequence back to its original position) of different chromosomal regions can be different. In addition, the polymorphism of copy number (i.e. copy number variations) would also affect the sequence read densities of the bins. Therefore, normalization with the buffy coat DNA can potentially minimize the variations associated with the difference in the sequence context between different chromosomal regions.

[0187] FIG. 10A shows a distribution plot 1000 of the sequence read densities of the tumor sample of an HCC patient according to embodiments of the present invention. The tumor tissue was obtained following surgical resection from the HCC patient. The x-axis represents the $\log_2$ of the ratio (R) of the sequence read density between the tumor tissue and the buffy coat of the patient. The y-axis represents the number of bins.

IDT Exhibit 1020.1717

[0188] Peaks can be fitted to the distribution curve to represent the regions with deletion, amplification, and without chromosomal aberrations using the normal mixture model. In one embodiment, the number of peaks can be determined by the Akaike's information criterion (AIC) across different plausible values. The central peak with a $\log_2 R = 0$ (i.e. $R = 1$) represents the regions without any chromosomal aberration. The left peak (relative to the central one) represents regions with one copy loss. The right peak (relative to central one) represents regions with one copy amplification.

[0189] The fractional concentration of tumor-derived DNA can be reflected by the distance between the peaks representing the amplified and deleted regions. The larger the distance, the higher the fractional concentration of the tumor-derived DNA in the sample would be. The fractional concentration of tumor-derived DNA in the sample can be determined by this genomic representation approach, denoted as F_{GR}, using the following equation: $F_{GR} = R_{right} - R_{left}$, where R_{right} is the R value of the right peak and R_{left} is the R value of the left peak. The largest difference would be 1, corresponding to 100%. The fractional concentration of tumor-derived DNA in the tumor sample obtained from the HCC patient is estimated to be 66%, where the values of R_{right} and R_{left} are 1.376 and 0.712, respectively.

[0190] To verify this result, another method using the genomewide aggregated allele loss (GAAL) analysis was also used to independently determine the fractional concentration of proportion of tumoral DNA (Chan KC et al. Clin Chem 2013;59:211-24). Table 3 shows the fractional concentrations of tumor-derived DNA in the tumor tissues of the four HCC patients using the genomic representation (F_{GR}) and the GAAL (F_{GAAL}) approaches. The values determined by these two different approaches agree well with each other.

HCC tumor	F_{GAAL}	F_{GR}
1	60.0%	66.5%
2	60.0%	61.4%
3	58.0%	58.9%
4	45.7%	42.2%

Table 3 showing fractional concentration determined by GAAL and genomic representation (GR).

IDT Exhibit 1020.1718

2. Estimation in sample with low tumor concentration

[0191] The above analysis has shown that our genomic representation method can be used to measure the fractional concentration of tumor DNA when more than 50% of the sample DNA is tumor-derived, i.e. when the tumor DNA is a majority proportion. In the previous analysis, we have shown that this method can also be applied to samples in which the tumor-derived DNA represents a minor proportion (i.e., below 50%). Samples that may contain a minor proportion of tumor DNA include, but not limited to blood, plasma, serum, urine, pleural fluid, cerebrospinal fluid, tears, saliva, ascitic fluid and feces of cancer patients. In some samples, the fractional concentration of tumor-derived DNA can be 49%, 40%, 30%, 20%, 10%, 5%, 2%, 1%, 0.5%, 0.1% or lower.

[0192] For such samples, the peaks of sequence read density representing the regions with amplification and deletion may not be as obvious as in samples containing a relatively high concentration of tumor-derived DNA as illustrated above. In one embodiment, the regions with chromosomal aberrations in the cancer cells can be identified by making comparison to reference samples which are known to not contain cancer DNA. For example, the plasma of subjects without a cancer can be used as references to determine the normative range of sequence read densities for the chromosome regions. The sequence read density of the tested subject can be compared with the value of the reference group. In one embodiment, the mean and standard deviation (SD) of sequence read density can be determined. For each bin, the sequence read density of the tested subject is compared with the mean of the reference group to determine the z-score using the following formula:

$z-score=\frac{(GR_{test}-\overline{GR}_{ref})}{SD_{ref}}$, where GR_{test} represents the sequence read density of the cancer patient; $\overline{GR}_{ref}$ represents the mean sequence read density of the reference subjects and SD_{ref} represents the SD of the sequence read densities for the reference subjects.

[0193] Regions with z-score < -3 signifies significant underpresentation of the sequence read density for a particular bin in the cancer patient suggesting the presence of a deletion in the tumor tissue. Regions with z-score > 3 signifies significant overpresentation of the sequence read density for a particular bin in the cancer patient suggesting the presence of an amplification in the tumor tissue

IDT Exhibit 1020.1719

[0194] Then, the distribution of the z-scores of all the bins can be constructed to identify regions with different numbers of copy gain and loss, for example, deletion of 1 or 2 copies of a chromosome; and amplification, resulting in of 1, 2, 3 and 4 additional copies of a chromosome. In some cases, more than one chromosome or more than one regions of a chromosome may be involved.

[0195] FIG. 10B shows a distribution plot 1050 of z-scores for all the bins in the plasma of a HCC patient according to embodiments of the present invention. The peaks (from left to right) representing 1-copy loss, no copy change, 1-copy gain and 2-copy gain are fitted to the z-score distribution. Regions with different types of chromosomal aberrations can then be identified, for example using the normal mixture model as described above.

[0196] The fractional concentration of the cancer DNA in the sample (F) can then be inferred from the sequence read densities of the bins that exhibit one-copy gain or one-copy loss. The fractional concentration determined for a particular bin can be calculated as $F = \frac{\left|GR_{test} - \overline{GR}_{ref}\right| \times 2}{GR_{ref}} \times 100\%$. This can also be expressed as: $F = \frac{\left|z - score \times SD_{ref}\right|}{\overline{GR}_{ref}} \times 2$, which can be rewritten as: $F = \left|z - score\right| \times CV \times 2$, where CV is the coefficient of variation for the measurement of the sequence read density of the reference subjects; and $CV = \frac{SD_{ref}}{\overline{GR}_{ref}}$.

[0197] In one embodiment, the results from the bins are combined. For example, the z-scores of bins showing a 1-copy gain can be averaged or the resulting F values averaged. In another implementation, the value of the z-score used for inferring F is determined by a statistical model and is represented by the peaks shown in FIG. 10B and FIG. 11. For example, the z-score of the right peak can be used to determine the fractional concentration for the regions exhibiting 1-copy gain.

[0198] In another embodiment, all bins with z-score < -3 and z-score > 3 can be attributed to regions with single copy loss and single copy gain, respectively, because these two types of chromosomal aberrations are the most common. This approximation is most useful when the number of bins with chromosomal aberrations is relatively small and fitting of normal distribution may not be accurate.

IDT Exhibit 1020.1720

[0199] FIG. 11 shows a distribution plot 1100 of z-scores for the plasma of an HCC patient according to embodiments of the present invention. While the number of bins overlapping with chromosomal aberrations is relatively small, all bins with z-score < -3 and z-score > 3 were fitted to the normal distributions of single copy loss and single copy gain, respectively.

[0200] The fractional concentrations of tumor-derived DNA in the plasma of the four HCC patients were determined using GAAL analysis and this GR-based approach. The results are shown in Table 4. As can be seen, the deduced fractional representation correlates well between the GAAL analysis and the GR analysis.

	Fractional concentration of tumor-derived DNA in plasma	
Samples	GAAL analysis	GR analysis
case11	4.3%	4.5%
case13	5%	5.5%
case23	52%	62%
case27	7.6%	6.1%

Table 4. Fractional concentration of tumor-derived DNA in plasma deduced by the analysis of chromosomal aberrations.

C. Method of Determining Fractional Concentration

[0201] FIG. 12 is a flowchart of a method 1200 of determining a fractional concentration of tumor DNA in a biological sample including cell-free DNA according to embodiments of the present invention. Method 1200 may be performed via various embodiments, including embodiments described above.

[0202] At block 1210, one or more sequence tags are received for each of a plurality of DNA fragments in the biological sample. Block 1210 may be performed as described herein for other methods. For example, one end of a DNA fragment may be sequenced from a plasma sample. In another embodiment, both ends of a DNA fragment may be sequenced, thereby allowing a length of the fragment to be estimated.

[0203] At block 1220, genomic positions are determined for the sequence tags. The genomic positions can be determined, e.g., as described herein by aligning the sequence tags to a reference genome. If both ends of a fragment are sequenced, then the paired tags may be

IDT Exhibit 1020.1721

aligned as a pair with a distance between the two tags constrained to be less than a specified distance, e.g., 500 or 1,000 basis.

[0204] At block 1230, for each of a plurality of genomic regions, a respective amount of DNA fragments within the genomic region is determined from sequence tags having a genomic position within the genomic region. The genomic regions may be non-overlapping bins of equal length in the reference genome. In one embodiment, a number of tags that align to a bin can be counted. Thus, each bin can have a corresponding number of aligned tags. A histogram can be computed illustrating a frequency that bins have a certain number of aligned tags. Method 1200 may be performed for genomic regions each having a same length (e.g., 1 Mb bins), where the regions are non-overlapping. In other embodiments, different lengths can be used, which may be accounted for, and the regions may overlap.

[0205] At block 1240, the respective amount is normalized to obtain a respective density. In one embodiment, normalizing the respective amount to obtain a respective density includes using a same total number of aligned reference tags to determine the respective density and the reference density. In another embodiment, the respective amount can be divided by a total number of aligned reference tags.

[0206] At block 1250, the respective density is compared to a reference density to identify whether the genomic region exhibits a 1-copy loss or a 1-copy gain. In one embodiment, a difference is computed between the respective density and the reference density (e.g., as part of determining a z-score) and compared to a cutoff value. In various embodiments, the reference density can be obtained from a sample of healthy cells (e.g., from the buffy coat) or from the respective amounts themselves (e.g., by taking an median or average value, under an assumption that most regions do not exhibit a loss or a gain).

[0207] At block 1260, a first density is calculated from one or more respective densities identified as exhibiting a 1-copy loss or from one or more respective densities identified as exhibiting a 1-copy gain. The first density can correspond to just one genomic region, or may be determined from densities of multiple genomic regions. For example, the first density may be computed from respective densities having a 1-copy loss. The respective densities provide a measure of the amount of the density difference resulting from the deletion of the region in a tumor, given the tumor concentration. Similarly, if the first density is from respective densities having a 1-copy gain, then a measure of the amount of density difference

IDT Exhibit 1020.1722

resulting from the duplication of the region in a tumor can be obtained. Sections above describe various examples of how the densities of multiple regions can be used to determine an average density to be used for the first density.

[0208] At block 1270, the fractional concentration is calculated by comparing the first density to another density to obtain a differential. The differential is normalized with the reference density, which may be done in block 1270. For example, the differential can be normalized with the reference density by dividing the differential by the reference density. In another embodiment, the differential can be normalized in earlier blocks.

[0209] In one implementation, the another density is the reference density, e.g., as in section 2 above. Thus, calculating the fractional concentration may include multiplying the differential by two. In another implementation, the another density is a second density calculated from respective densities identified as exhibiting a 1-copy loss (where the first density is calculated using respective densities identified as exhibiting a 1-copy gain), e.g., as described in section 1 above. In this case, the normalized differential can be determined by computing a first ratio (e.g., R_{right}) of the first density and the reference density and computing a second ratio (R_{left}) of the second density and the reference density, where the differential is between the first ratio and the second ratio. As described above, the identification of genomic region exhibiting a 1-copy loss or a 1-copy gain can be performed by fitting peaks to a distribution curve of a histogram of the respective densities.

[0210] In summary, embodiments can analyze the genomic representation of plasma DNA in different chromosomal regions to simultaneously determine if the chromosomal region is amplified or deleted in the tumor tissue and, if the region is amplified or deleted, to use its genomic representation to deduce the fractional concentration of the tumor-derived DNA. Some implementations use a normal mixture model to analyze the overall distribution of the genomic representation of different regions so as to determine the genomic representation associated with different types of aberrations, namely gains of 1, 2, 3 or 4 copies and the losses of 1 or 2 copies.

[0211] Embodiments have several advantages over other methods, for example genomewide aggregated allelic loss (GAAL) approach (US patent application 13/308,473; Chan KC et al. Clin Chem 2013;59:211-24) and the analysis of tumor-associated single nucleotide mutations (Forshew T et al. Sci Transl Med. 2012;4:136ra68). All sequence reads

IDT Exhibit 1020.1723

mapping to regions with chromosomal aberrations can be used to determine the sequence read density of the region and, hence, are informative regarding the fractional concentration of tumoral DNA. On the other hand, in GAAL analysis, only sequence reads covering single nucleotides that are heterozygous in the individual and located within a chromosomal region with chromosome gain or loss would be informative. Similarly, for the analysis of cancer-associated mutations, only sequence reads covering the mutations would be useful for the deduction of the tumoral DNA concentration. Therefore, embodiments can allow a more cost-effective use of the sequencing data as relatively fewer sequencing reads may be needed to achieve the same degree of accuracy in the estimation of fractional concentration of tumor-derived DNA when compared with other approaches.

X. ALTERNATIVE METHODOLOGIES

[0212] Apart from using the number of times that a particular mutation is seen on a sequence tag as a criteria for identifying a locus as being a true mutation (thereby adjusting the positive predictive value), one could employ other techniques instead of or in addition to using a cutoff value to provide greater predictive value in identifying a cancerous mutation. For example, one could use bioinformatics filters of different stringencies when processing the sequencing data, e.g., by taking into account the quality score of a sequenced nucleotide. In one embodiment, one could use DNA sequencers and sequencing chemistries with different sequencing error profiles. Sequencers and chemistries with lower sequencing error rates would give a higher positive predictive values. One can also use repeated sequencing of the same DNA fragment to increase the sequencing accuracy. One possible strategy is the circular consensus sequencing strategy of Pacific Biosciences.

[0213] In another embodiment, one could incorporate size information on the sequenced fragments into the interpretation of the data. As tumor-derived DNA is shorter than the non-tumor-derived DNA in plasma (see U.S. Patent Application No. 13/308,473), the positive predictive value of a shorter plasma DNA fragment containing a potential tumor-derived mutation will be higher than that of a longer plasma DNA fragment. The size data will be readily available if one performs paired-end sequencing of the plasma DNA. As an alternative, one could use DNA sequencers with long read lengths, thus yielding the complete length of a plasma DNA fragment. One could also perform size fractionation of the plasma DNA sample prior to DNA sequencing. Examples of methods that one could use for size

IDT Exhibit 1020.1724

fractionation include gel electrophoresis, the use of microfluidics approach (e.g. the Caliper LabChip XT system) and size-exclusion spin columns.

[0214] In yet another embodiment, the fractional concentration of tumor-associated mutations in plasma in a patient with non-hematologic cancer would be expected to increase if one focuses on the shorter DNA fragments in plasma. In one implementation, one can compare the fractional concentration of tumor-associated mutations in plasma in DNA fragments of two or more different size distributions. A patient with a non-hematologic cancer will have higher fractional concentrations of tumor-associated mutations in the shorter fragments when compared with the larger fragments.

[0215] In some embodiments, one could combine the sequencing results from two or more aliquots of the same blood sample, or from two or more blood samples taken on the same occasions or on different occasions. Potential mutations seen in more than one aliquot or samples would have a higher positive predictive value of tumor-associated mutations. The positive predictive value would increase with the number of samples that show such a mutation. The potential mutations that are present in plasma samples taken at different time points can be regarded as potential mutations.

XI. EXAMPLES

[0216] The following are example techniques and data, which should not be considered limiting on embodiments of the present invention.

A. Materials And Methods

[0217] Regarding sample collection, hepatocellular carcinoma (HCC) patients, carriers of chronic hepatitis B, and a patient with synchronous breast and ovarian cancers were recruited. All HCC patients had Barcelona Clinic Liver Cancer stage A1 disease. Peripheral blood samples from all participants were collected into EDTA-containing tubes. The tumor tissues of the HCC patients were obtained during their cancer resection surgeries.

[0218] Peripheral blood samples were centrifuged at 1,600 g for 10 min at 4 °C. The plasma portion was recentrifuged at 16,000 g for 10 min at 4 °C and then stored at 80 °C. Cell-free DNA molecules from 4.8 mL of plasma were extracted according to the blood and body fluid protocol of the QIAamp DSP DNABlood Mini Kit (Qiagen). The plasma DNA

IDT Exhibit 1020.1725

was concentrated with a SpeedVac Concentrator (Savant DNA120; Thermo Scientific) into a 40-μl final volume per case for subsequent preparation of the DNA-sequencing library

[0219] Genomic DNA was extracted from patients' buffy coat samples according to the blood and body fluid protocol of the QIAamp DSP DNA Blood Mini Kit. DNA was extracted from tumor tissues with the QIAamp DNA Mini Kit (Qiagen).

[0220] Sequencing libraries of the genomicDNAsamples were constructed with the Paired-End Sample Preparation Kit (Illumina) according to the manufacturer's instructions. In brief, 1–5 micrograms of genomic DNA was first sheared with a Covaris S220 Focused-ultrasonicator to 200-bp fragments. Afterward, DNA molecules were end-repaired with T4 DNA polymerase and Klenow polymerase; T4 polynucleotide kinase was then used to phosphorylate the 5' ends. A 3' overhang was created with a 3'-to-5' exonuclease– deficient Klenow fragment. Illumina adapter oligonucleotides were ligated to the sticky ends. The adapter-ligated DNA was enriched with a 12-cycle PCR. Because the plasma DNA molecules were short fragments and the amounts of total DNA in the plasma samples were relatively small, we omitted the fragmentation steps and used a 15-cycle PCR when constructing the DNA libraries from the plasma samples.

[0221] An Agilent 2100 Bioanalyzer (Agilent Technologies) was used to check the quality and size of the adapter-ligated DNA libraries. DNA libraries were then measured by a KAPA Library Quantification Kit (Kapa Biosystems) according to the manufacturer's instructions. The DNA library was diluted and hybridized to the paired-end sequencing flow cells. DNA clusters were generated on a cBot cluster generation system (Illumina) with the TruSeq PE Cluster Generation Kit v2 (Illumina), followed by 51_2 cycles or 76_2 cycles of sequencing on a HiSeq 2000 system (Illumina) with the TruSeq SBS Kit v2 (Illumina).

[0222] The paired-end sequencing data were analyzed by means of the Short Oligonucleotide Alignment Program 2 (SOAP2) in the paired-end mode. For each paired-end read, 50 bp or 75 bp from each end was aligned to the non–repeat-masked reference human genome (hg18). Up to 2 nucleotide mismatches were allowed for the alignment of each end. The genomic coordinates of these potential alignments for the 2 ends were then analyzed to determine whether any combination would allow the 2 ends to be aligned to the same chromosome with the correct orientation, spanning an insert size less than or equal to 600 bp, and mapping to a single location in the reference human genome. Duplicated reads

IDT Exhibit 1020.1726

were defined as paired-end reads in which the insert DNA molecule showed identical start and end locations in the human genome; the duplicate reads were removed as previously described (Lo et al. *Sci Transl Med* 2010; 2: 61ra91).

[0223] In some embodiments, the paired tumor and constitutional DNA samples were sequenced to identify the tumor-associated single nucleotide variants (SNVs). In some implementations, we focused on the SNVs occurring at homozygous sites in the constitutional DNA (in this example being the buffy coat DNA). In principle, any nucleotide variation detected in the sequencing data of the tumor tissues but absent in the constitutional DNA could be a potential mutation (i.e., a SNV). Because of sequencing errors (0.1%–0.3% of sequenced nucleotides), however, millions of false positives would be identified in the genome if a single occurrence of any nucleotide change in the sequencing data of the tumor tissue were to be regarded as a tumor-associated SNV. One way to reduce the number of false positives would be to institute the criterion of observing multiple occurrences of the same nucleotide change in the sequencing data in the tumor tissue before a tumor associated SNV would be called.

[0224] Because the occurrence of sequencing errors is a stochastic process, the number of false positives due to sequencing errors would decrease exponentially with the increasing number of occurrences required for an observed SNV to be qualified as a tumor-associated SNV. On the other hand, the number of false positives would increase with increasing sequencing depth. These relationships could be predicted with Poisson and binomial distribution functions. Embodiments can determine a dynamic cutoff of occurrence for qualifying an observed SNV as tumor associated. Embodiments can take into account the actual coverage of the particular nucleotide in the tumor sequencing data, the sequencing error rate, the maximum false-positive rate allowed, and the desired sensitivity for mutation detection.

[0225] In some examples, we set very stringent criteria to reduce false positives. For example, a mutation may be required to be completely absent in the constitutional DNA sequencing, and the sequencing depth for the particular nucleotide position had to be 20-fold. In some implementations, the cutoff of occurrence achieved a false-positive detection rate of less than 10^{-7}. In some examples, we also filtered out SNVs that were within centromeric, telomeric, and low-complexity regions to minimize false positives due to alignment artifacts.

IDT Exhibit 1020.1727

In addition, putative SNVs mapping to known SNPs in the dbSNP build 135 database were also removed.

B. Before and After Resection

[0226] FIG. 13A shows a table 1300 of the analysis of mutations in the plasma of the patient with ovarian cancers and a breast cancer at the time of diagnosis according to embodiments of the present invention. Here, we demonstrate an example for a patient with bilateral ovarian cancers and a breast cancer. The sequencing data of the plasma were compared to the sequencing results of the constitutional DNA of the patient (buffy coat). Single nucleotide changes that were present in the plasma but not in the constitutional DNA were regarded as potential mutations. The ovarian cancers on the right and left side of the patient were each sampled at two sites, making a total of four tumor samples. The tumor mutations were mutations detected in all the four ovarian tumor tissues at four different sites.

[0227] Over 3.6 million single nucleotide changes were detected in the plasma for at least one time by sequencing. Among these changes, only 2,064 were also detected in the tumor tissues giving a positive prediction value of 0.06%. Using the criterion of being detected at least two times in plasma, the number of potential mutations was significantly reduced by 99.5% to 18,885. The number of tumor mutations was only reduced by 3% to 2,003, and the positive prediction value increased to 11%.

[0228] Using the criteria of detecting at least five times in plasma, only 2,572 potential mutations were detected and amongst them, 1,814 were mutations detected in all the tumor tissues, thus, giving a positive predictive value of 71%. Other criteria for the number of occurrences (e.g. 2, 3, 4, 6, 7, 8, 9, 10, etc.) can be used for defining potential mutations depending on the sensitivity and positive predictive value required. The higher the number of occurrences is used as the criterion, the higher the positive predictive value would be with a reduction in the sensitivity.

[0229] FIG. 13B shows a table 1350 of the analysis of mutations in the plasma of the patient with bilateral ovarian cancers and a breast cancer after tumor resection according to embodiments of the present invention. Surgical resection of the patient was performed. A blood sample was taken one day after the resection of the ovarian tumors and the breast cancer. The plasma DNA was then sequenced. For this example, only the mutations from the ovarian cancers were analyzed. Over 3 million potential mutations were detected at least

IDT Exhibit 1020.1728

once in a plasma sample. However, using a criterion of having at least five occurrences, the number of potential mutations was reduced to 238. A significant reduction was observed when compared with the number of potential mutations for the sample taken at diagnosis and using the same criterion of five mutations.

[0230] In one embodiment, the number of single nucleotide changes detected in plasma can be used as a parameter for the detection, monitoring and prognostication of a cancer patient. Different number of occurrences can be used as the criterion to achieve the desired sensitivity and specificity. A patient with a higher tumor load and thus worse prognosis will be expected to have a higher mutational load seen in plasma.

[0231] For such analysis, one could establish the mutational load profile for different types of cancer. For monitoring purposes, one would see that the mutational load in plasma of a patient who responds to treatment would reduce. If the tumor has recurred, e.g. during a relapse, then the mutational load will be expected to increase. Such monitoring would allow one to monitor the efficacy of the selected modality of treatment for a patient and to detect the emergence of resistance to a particular treatment.

[0232] Through the analysis of the specific mutations that one could see in the plasma DNA sequencing results, one could also identify targets that would predict sensitivity (e.g. mutations in the epidermal growth factor receptor gene and response to tyrosine kinase inhibitor treatment) and resistance to particular targeted treatment (e.g. *KRAS* mutations in colorectal cancer and resistance to treatment by panitumumab and cetuximab), and could guide the planning of treatment regimes.

[0233] The example above was for the bilateral ovarian cancers. One could also perform the same analysis on the mutations of the breast cancer and then would be able to track the mutations of both of these cancer types in the plasma. One can also use a similar strategy to track the mutations of a primary cancer and its metastasis or metastases.

[0234] Embodiments would be useful to the screening of cancer in apparently healthy subjects or in subjects with particular risk factors (e.g. smoking status, viral status (such as hepatitis virus carriers, human papillomavirus infected subjects)). The mutational load that one could see in the plasma of such subjects would give a risk that the subject would develop symptomatic cancer within a particular timeframe. Thus, subjects with a higher mutational load in plasma would be expected to have a higher risk than those with a lower mutational

IDT Exhibit 1020.1729

load. Furthermore, the temporal profile of such mutational load in plasma would also be a powerful indicator of risk. For example, if a subject has one plasma mutational load performed each year and if the mutational loads are progressively increasing, then this subject should be referred for additional screening modalities for cancer, e.g. using chest X ray, ultrasound, computed tomography, magnetic resonance imaging or positron emission tomography.

C. Dynamic Cutoffs To Deduce Mutations From Sequencing Plasma

[0235] Four patients with hepatocellular carcinoma (HCC) and one patient with ovarian and breast cancer were recruited for this study. For the latter patient, we focused on the analysis of the ovarian cancer. Blood samples were collected from each patient before and after surgical resection of the tumors. The resected tumor tissues were also collected. The DNA extracted from the tumor tissue, the white blood cells of the preoperative blood sample and the pre- and post-operative plasma samples was sequenced using the HiSeq2000 sequencing system (Illumina). The sequencing data were aligned to the reference human genome sequence (hg18) using Short Oligonucleotide Analysis Package 2 (SOAP2) (Li R et al. Bioinformatics 2009; 25: 1966-1967). The DNA sequences of the white blood cells were regarded as constitutional DNA sequence for each study subject.

[0236] In this example, tumor-associated SNMs were first deduced from the plasma DNA sequencing data and the CG without reference to the tumor tissues. Then, the deduced results from plasma were compared with sequencing data generated from the tumor tissues (as a gold standard) to ascertain the accuracy of the deduced results. In this regard, the gold standard was made by comparing sequencing data from the tumor tissues and the constitutional sequence to work out the mutations in the tumor tissues. In this analysis, we focused on nucleotide positions at which the constitutional DNA of the studied subject was homozygous.

1. Non-targeted whole genome analysis

[0237] The sequencing depths for the white cells, the tumor tissues and the plasma DNA of each patient are shown in Table 5.

IDT Exhibit 1020.1730

Case	Median sequencing depth (folds)			
	White blood cells	Tumor tissue	Preoperative plasma	Postoperative plasma
HCC1	39	29	23	24
HCC2	39	29	25	28
HCC3	46	33	18	21
HCC4	46	27	20	23
Ovarian cancer patient	44	53	37	28

Table 5. Median sequencing depths of different samples for the four HCC cases.

[0238] The dynamic cutoffs for the minimum occurrences for defining plasma mutations (*r*) as shown in table 1 are used for identifying the mutations in the plasma of each patient. As the sequencing depth of each locus may vary, the cutoff may vary, which effectively provides a dependence of the cutoff on the total number of reads for a locus. For example, although the median depth is less than 50 (Table 5), the sequencing depth of individual loci can vary a lot and be covered >100 times.

[0239] In addition to sequencing errors, another source of error would be alignment errors. To minimize this type of errors, the sequence reads carrying a mutation was realigned to the reference genome using the Bowtie alignment program (Langmead B et al. Genome Biol 2009, 10:R25). Only reads that could be aligned to a unique position of the reference genome by SOAP2 and Bowtie were used for the downstream analysis for plasma mutations. Other combinations of alignment software packages based on different algorithms could also be used.

[0240] In order to further minimize the sequencing and alignment errors in the actual sequencing data, we applied two additional filtering algorithms for calling the nucleotide positions which showed single nucleotide variations in the sequence reads: (1) ≥ 70% of the sequence reads carrying the mutation could be realigned to the same genomic coordinate using Bowtie with mapping quality ≥Q20 (i.e. misalignment probability <1%); (2) ≥70% of the sequence reads carrying the mutation were not within 5 bp of both ends (i.e. 5' and 3'

IDT Exhibit 1020.1731

ends) of the sequence reads. This filtering rule was instituted because sequencing errors were more prevalent at both ends of a sequence read.

[0241] We also investigated the factors affecting the deducing of a tumor without prior knowledge of the tumor genome. One such parameter was the fractional concentration of tumor-derived DNA in plasma. This parameter could be regarded as another gold standard parameter and was deduced for reference purpose with prior knowledge of the tumor genome using GAAL.

[0242] Table 6 shows nucleotide variations detected in plasma before and over treatment. For HCC1, without prior knowledge of the tumor genome, a total of 961 single nucleotide variations were detected. Amongst these nucleotide variations detected in plasma, 828 were cancer-associated mutations. After surgical resection of the HCC, the total number of nucleotide variations was reduced to 43 and none of them was cancer-associated mutations.

[0243] For reference purposes, the fractional concentration of tumor-derived DNA in the pre-operative plasma sample was 53% and was deduced with prior knowledge of the tumor genome. For HCC2, HCC3 and HCC4, without prior knowledge of the tumor genomes, the numbers of single nucleotide variations in plasma were deduced as ranging from 27 to 32 for the pre-operative plasma samples. These results are compatible with the mathematical prediction that, with a sequencing depth of approximately 20-fold, a very low percentage of cancer-associated mutations could be detected in the plasma and most of the sequence variations detected in the plasma were due to sequencing errors. After tumor resection, there was no significant change in the number of sequence variations detected. For reference purposes, the fractional concentrations of tumor-derived DNA in plasma were deduced as ranging from 2.1% to 5% and were deduced with prior knowledge of the tumor genomes.

IDT Exhibit 1020.1732

	Pre-operative plasma			Post-operative plasma		
	Fractional concentration of tumor-derived DNA	Total no. of single nucleotide variations	No. of cancer-associated mutations identified	Fractional concentration of tumor-derived DNA	Total no. of single nucleotide variations	No. of cancer-associated mutations identified
HCC1	53%	961	828	0.4%	43	0
HCC2	5%	32	0	0.6%	49	0
HCC3	2.1%	29	0	0.2%	32	0
HCC4	2.6%	27	0	1.3%	35	1
Ovarian (and breast)cancer patient	46%	1718	1502	0.2%	2	0

Table 6. Nucleotide variations detected in plasma.

2. Target enrichment of the exons

[0244] As discussed above, increasing the sequencing depth for the region-of-interest can increase both the sensitivity and specificity for identifying cancer-associated mutations in plasma and, hence, increasing the discrimination power between the cancer patients and non-cancer subjects. While the increase of sequencing depth for the whole genome is still very costly, one alternative is to enrich for certain regions for sequencing. In one embodiment, selected exons or indeed the whole exome can be target-enriched for sequencing. This approach can significantly increase the sequencing depth of the target region without increasing the total amount of sequence reads.

[0245] The sequencing libraries of the plasma DNA of the HCC patients and the patient with ovarian (and breast) cancer were captured using the Agilent SureSelect All Exon kit for target enrichment of the exome. The exon-enriched sequencing libraries were then sequenced using the HiSeq 2000 sequencing system. The sequence reads were aligned to the human reference genome (hg18). After alignment, sequence reads uniquely mapped to the exons were analyzed for single nucleotide variations. For the identification of single nucleotide variations in plasma for the exome capture analysis, the dynamic cutoff values shown in table 2 are used.

IDT Exhibit 1020.1733

[0246] FIG. 14A is a table 1400 showing detection of single nucleotide variations in plasma DNA for HCC1. Without prior knowledge of the tumor genome, we deduced from the targeted sequencing data a total of 57 single nucleotide variations in plasma. In subsequent validation from the sequencing data obtained from the tumor tissues, 55 were found to be true tumor-associated mutations. As discussed before, fractional concentration of tumor-derived DNA in the pre-operative plasma was 53%. After tumor resection, no single nucleotide variations were detected in the targeted sequencing data obtained from the plasma. These results indicate that the quantitative analysis of the number of single nucleotide variations in plasma can be used for monitoring the disease progression of cancer patients.

[0247] FIG. 14B is a table 1450 showing detection of single nucleotide variations in plasma DNA for HCC2. Without prior knowledge of the tumor genome, we deduced from the targeted sequencing data of the plasma a total of 18 single nucleotide variations. All of these mutations were found in the tumor tissues. As discussed before, fractional concentration of tumor-derived DNA in the pre-operative plasma was 5%. After tumor resection, no single nucleotide variations were detected in the plasma. Compared with HCC1 which had a higher fractional concentration of tumor-derived DNA in plasma, fewer single nucleotide variations were detected in the plasma of the case involving HCC2. These results suggest that the number of single nucleotide variations in plasma can be used as a parameter to reflect the fractional concentration of tumor-derived DNA in plasma and, hence, the tumor load in the patient as it has been showed that the concentration of tumor-derived DNA in plasma is positively correlated with the tumor load (Chan KC et al. Clin Chem 2005;51:2192-5).

[0248] FIG. 15A is a table 1500 showing detection of single nucleotide variations in plasma DNA for HCC3. Without prior knowledge of the tumor genome, we did not observe from the targeted sequencing data any single nucleotide variations in both the pre- and post-resection plasma samples. This is likely to be due to the relatively low fractional concentration (2.1%) of tumor-derived DNA in plasma in this patient. Further increase in the sequencing depth is predicted to improve the sensitivity for detecting cancer-associated mutations in cases with low fractional concentration of tumor-derived DNA.

[0249] FIG. 15B is a table 1550 showing detection of single nucleotide variations in plasma DNA for HCC4. Without prior knowledge of the tumor genome, we deduced from the targeted sequencing data of the plasma a total of 3 single nucleotide variations. All of these mutations were found in the tumor tissues. Compared with HCC1 and HCC2 which had

IDT Exhibit 1020.1734

higher fractional concentrations of tumor-derived DNA in plasma, fewer single nucleotide variations were detected in the plasma of case HCC4 which had a fractional tumor DNA in plasma of 2.6%. These results suggest that the number of single nucleotide variations in plasma can be used as a parameter to reflect the fractional concentration of tumor-derived DNA in plasma and tumor load in a patient.

[0250] FIG. 16 is a table 1600 showing detection of single nucleotide variations in plasma DNA for the patient with ovarian (and breast) cancer. Without prior knowledge of the tumor genome, we deduced from the targeted sequencing data of the plasma a total of 64 single nucleotide variations. Amongst these 59 were found in the ovarian tumor tissues. The estimated fractional concentration of ovarian tumor-derived DNA in the plasma was 46%. A significant reduction in the total number of single nucleotide variations were detected in plasma after resection of the ovarian cancer.

[0251] In addition to the use of the SureSelect target enrichment system (Agilent), we also used the Nimblegen SeqCap EZ Exome+UTR target enrichment system (Roche) for enriching sequences from exons for sequencing. The Nimblegen SeqCap system covers the exon regions of the genome as well as the 5' and 3' untranslated region. The pre-treatment plasma samples of the four HCC patients, two healthy control subjects and two chronic hepatitis B carriers without a cancer were analyzed (Table 7). In other embodiments, other target enrichment systems, including but not limited to those using solution phase or solid phase hybridization, can be used.

IDT Exhibit 1020.1735

	Pre-treatment plasma			Post-treatment plasma	
	Fractional concentration of tumor-derived DNA in plasma by GAAL analysis	No. of sequence variation detected in plasma fulfilling the dynamic cutoffs	No. of sequence variation that overlap with mutations detected in the corresponding tumor tissue	No. of sequence variation detected in plasma fulfilling the dynamic cutoffs	No. of sequence variation that overlap with mutations detected in the corresponding tumor tissue
HCC1	53%	69	64	1	1
HCC2	5%	51	47	3	0
HCC3	2.1%	0	0	1	0
HCC4	2.6%	8	7	0	0

Table 7. Exome sequencing results for the four HCC patients (HCC1-4) using the Nimblegen SeqCap EZ Exome+UTR target enrichment system for sequence capture. The sequencing analysis of the pre-treatment plasma of HCC3 was sub-optimal due to a higher percentage of PCR-duplicated reads.

[0252] In the two chronic hepatitis B carriers and the two healthy control subjects, one or less single nucleotide variations that fulfilled the dynamic cutoff criteria were detected (Table 8). In three of the four HCC patients, the number of sequence variations detected in plasma that fulfilled the dynamic cutoff requirement was at least 8. In HCC3, no SNV that fulfilled the dynamic cutoff was detected. In this sample, there was a high proportion PCR-duplicated read in the sequenced reads leading to a lower number of non-duplicated sequenced reads. Marked reduction of SNVs detectable in plasma was observed after surgical resection of the tumor.

	No. of sequence variation detected in plasma fulfilling the dynamic cutoffs
HBV1	0
HBV2	1
Ctrl1	1
Ctrl2	1

IDT Exhibit 1020.1736

Table 8. Exome sequencing results for 2 chronic hepatitis B carriers (HBV1 and HBV2) and 2 healthy control subjects (Ctrl1 and Ctrl2) using the Nimblegen SeqCap EZ Exome+UTR target enrichment system for sequence capture.

XII. TUMOR HETEROGENEITY

[0253] The quantification of single nucleotide mutations in a biological sample (e.g., plasma/serum) is also useful for the analysis of tumor heterogeneity, both intra-tumoral and inter-tumoral heterogeneity. Intra-tumoral heterogeneity relates to the existence of multiple clones of tumor cells within the same tumor. Inter-tumoral heterogeneity relates to the existence of multiple clones of tumor cells for two or more tumors of the same histologic type, but present in different sites (either in the same organs, or in different organs). In certain types of tumors, the existence of tumoral heterogeneity is a bad prognostic indicator (Yoon HH et al. J Clin Oncol 2012; 30: 3932-3938; Merlo LMF et al. Cancer Prev Res 2010; 3: 1388-1397). In certain types of tumors, the higher the degree of tumoral heterogeneity, the higher would be the chance of tumor progression or the development of resistant clones following targeted treatment.

[0254] Although cancers are believed to arise from the clonal expansion of one tumor cell, the growth and evolution of a cancer would lead to the accumulation of new and different mutations in different parts of a cancer. For example, when a cancer patient develops metastasis, the tumor located at the original organ and the metastatic tumor would share a number of mutations. However, the cancer cells of the two sites would also carry a unique set of mutations that are absent in the other tumor site. The mutations that are shared by the two sites are expected to be present at higher concentrations than those mutations that are only observed in one tumor site.

A. Example

[0255] We analyzed the blood plasma of a patient who had bilateral ovarian cancers and a breast cancer. Both ovarian tumors were serous adenocarcinoma. The left one measured 6 cm and the right one measured 12 cm in the longest dimension. There were also multiple metastatic lesions at the colon and the omentum. The DNA extracted from the leukocytes were sequenced using the sequencing-by-synthesis platform from Illumina to an average of

IDT Exhibit 1020.1737

44-fold haploid genome coverage. The nucleotide locations showing only one allele, i.e. homozygous, were analyzed further for single nucleotide mutations in plasma.

[0256] DNA was extracted from four different sites of the left and right tumors and was sequenced using the Illumina sequencing platform. Two sites (sites A and B) were from the right tumor and the other two sites (sites C and D) were from the left tumor. Sites A and B were approximately 4 cm apart. The distance between sites C and D was also approximately 4 cm. Plasma samples were collected from the patient before and after surgical resection of the ovarian tumors. DNA was then extracted from the plasma of the patient. The sequencing depth of the tumor from sites A, B, C and D, as well as the plasma samples, are shown in the table 9.

Sample	No. of raw sequencing reads	No. of aligned reads	Folds of haploid genome coverage
Constitutional DNA from buffy coat	1,091,250,072	876,269,922	43.81
Right ovarian tumor (site A)	1,374,495,256	1,067,277,229	53.36
Right ovarian tumor (site B)	934,518,588	803,007,464	40.15
Left ovarian tumor (site C)	1,313,051,122	1,036,643,946	51.83
Left ovarian tumor (site D)	1,159,091,833	974,823,207	48.74
Plasma sample collected before	988,697,457	741,982,535	37.10
Plasma sample collected after	957,295,879	564,623,127	28.23

Table 9. Sequencing depth of the tumor from sites A, B, C and D.

[0257] In the current example, for defining a single tumor-associated single nucleotide mutation, the nucleotide location is sequenced at least 20 times in the tumor tissue and 30 times in the constitutional DNA. In other embodiments, other sequencing depths can be used, e.g. 35, 40, 45, 50, 60, 70, 80, 90, 100 and >100 folds. The reduction of sequencing costs would allow increased depths to be performed much more readily. The nucleotide position is homozygous in the constitutional DNA whereas a nucleotide change is observed in the tumor tissue. The criterion for the occurrence of the nucleotide change in the tumor tissue is dependent on the total sequencing depth of the particular nucleotide position in the tumor tissue. For nucleotide coverage from 20 to 30 folds, the occurrence of the nucleotide change

IDT Exhibit 1020.1738

(cutoff value) is at least five times. For coverage from 31 to 50 folds, the occurrence of the nucleotide change is at least six times. For coverage from 51 to 70 folds, the occurrence needs is at least seven times. These criteria are derived from the prediction of sensitivity of detecting the true mutations and the expected number of false positive loci using the Poisson distribution.

[0258] FIG. 17 is a table 1700 showing the predicted sensitivities of different requirements of occurrence and sequencing depths. The sensitivity would correspond to the number of true mutations detected at a particular fold depth using a particular cutoff. The higher sequencing depth, the more likely it is for a mutation to be detected for a given cutoff, as more mutation sequence reads will be obtained. For higher cutoff values, the less likely a mutant would be detected, since the criteria are more stringent.

[0259] FIG. 18 is a table 1800 showing the predicted numbers of false positive loci for different cutoffs and different sequencing depths. The number of false positives increases with increasing sequencing depth, as more sequence reads are obtained. However, no false positives are predicted for a cutoff of five or more, even up to a sequencing depth of 70. In other embodiments, different criteria of occurrence can be used so as to achieve the desired sensitivity and specificity.

[0260] FIG. 19 shows a tree diagram illustrating the number of mutations detected in the different tumor sites. The mutations were determined by sequencing the tumors directly. Site A has 71 mutations that are specific to that tumor, and site B has 122 site-specific mutations, even though they were only 4 cm apart. 10 mutations were seen in both sites A and B. Site C has 168 mutations that are specific to that tumor, and site D has 248 site-specific mutations, even though they were only 4 cm apart. 12 mutations were seen in both sites C and D. There is significant heterogeneity in the mutational profiles for the different tumor sites. For example, 248 mutations were only detected in the site D tumor but not detected in the other three tumor sites. A total of 2,129 mutations were seen across all sites. Thus, many mutations were shared among the different tumors. Thus, there were seven SNV groups. There were no observable differences among these four regions in terms of copy number aberrations

[0261] FIG. 20 is a table 2000 showing the number of fragments carrying the tumor-derived mutations in the pre-treatment and post-treatment plasma sample. The inferred fractional concentrations of tumor-derived DNA carrying the respective mutations were also

IDT Exhibit 1020.1739

shown. The category of mutation refers to the tumor site(s) where the mutations were detected. For example, category A mutations refer to mutations only present in site A whereas category ABCD mutations refer to mutations present in all the four tumor sites.

[0262] For the 2,129 mutations that were present in all four tumor sites, 2,105 (98.9%) were detectable in at least one plasma DNA fragment. On the other hand, for the 609 mutations that were present in only one of the four tumor sites, only 77 (12.6%) were detectable in at least one plasma DNA fragment. Therefore, the quantification of single nucleotide mutations in plasma can be used for reflecting the relative abundance of these mutations in the tumor tissues. This information would be useful for the study of the cancer heterogeneity. In this example, a potential mutation was called when it had been seen once in the sequencing data.

[0263] The fractional concentrations of circulating tumor DNA were determined with each SNV group. The fractional concentrations of tumor DNA in plasma before surgery and after surgery, as determined by SNVs shared by all 4 regions (i.e., group ABCD), were 46% and 0.18%, respectively. These latter percentages correlated well with those obtained in GAAL analyses, 46% and 0.66%. Mutations that were shared by all 4 regions (i.e., group ABCD) contributed the highest fractional contribution of tumor-derived DNA to the plasma.

[0264] The fractional concentrations of tumor-derived DNA in preoperative plasma determined with SNVs from groups AB and CD were 9.5% and 1.1%, respectively. These concentrations were consistent with the relative sizes of the right and left ovarian tumors. The fractional concentrations of tumor-derived DNA determined with the region-unique SNVs (i.e., those in groups A, B, C, and D) were generally low. These data suggest that for an accurate measurement of the total tumor load in a cancer patient, the use of a genomewide shotgun approach might provide a more representative picture, compared with the more traditional approach of targeting specific tumor-associated mutations. For the latter approach, if only a subset of the tumor cells possesses the targeted mutations, one might miss important information regarding imminent relapse or disease progression caused by tumor cells not possessing the targeted mutations, or one might miss the emergence of a treatment-resistant clone.

[0265] FIG. 21 is a graph 2100 showing distributions of occurrence in plasma for the mutations detected in a single tumor site and mutations detected in all four tumor sites. The

IDT Exhibit 1020.1740

bar graph 2100 shows data for two types of mutation: (1) mutations detected in only one site and (2) mutations detected in all four tumor sites. The horizontal axis is the number of times that a mutation is detected in the plasma. The vertical axis shows the percentage of mutations that correspond to a particular value on the horizontal axis. For example, about 88% of type (1) mutations showed up only once in the plasma. As you can see, the mutations that showed up in one site were detected mostly once, and not more than four times. The mutations present in a single tumor site were much less frequently detected in the plasma compared with the mutations present in all four tumor sites.

[0266] One application of this technology would be to allow the clinicians to estimate the load of tumor cells carrying the different classes of mutations. A proportion of these mutations would potentially be treatable with targeted agents. Agents targeting mutations carried by a higher proportion of tumor cells would be expected to have a more prominent therapeutic effects.

[0267] FIG. 22 is a graph 2200 showing predicted distribution of occurrence in plasma for mutations coming from a heterogeneous tumor. The tumor contains two groups of mutations. One group of mutations are present in all tumor cells and the other group of mutations are only present in ¼ of the tumor cells, based on an approximation that two sites are representative of each ovarian tumor. The total fractional concentration of tumor-derived DNA in plasma is assumed to be 40%. The plasma sample is assumed to be sequenced to an average depth of 50 times per nucleotide position. According to this predicted distribution of occurrence in plasma, the mutations that are present in all tumor tissues can be differentiated from the mutations only present in ¼ tumor cells by their occurrence in plasma. For example, the occurrence of 6 times can be used as a cutoff. For the mutations present in all tumor cells, 92.3% of the mutations would be present in the plasma for at least 6 times. In contrast, for the mutations that are present in ¼ tumor cells, only 12.4% of mutations would be present in the plasma for at least 6 times.

[0268] FIG. 23 is a table 2300 demonstrating the specificity of embodiments for 16 healthy control subjects. Their plasma DNA samples were sequenced to a median coverage of 30 folds. Detection of the mutations that were present in the plasma of the above ovarian cancer patient was performed in the plasma samples of these healthy subjects. The mutations present in the tumor of the ovarian cancer patient were very infrequently detected in the sequencing data of the plasma of the healthy control subjects and none of the category of

IDT Exhibit 1020.1741

mutations had an apparent fractional concentration of >1%. These results show that this detection method is highly specific.

B. Method

[0269] FIG. 24 is a flowchart of a method 2400 for analyzing a heterogeneity of one or more tumors of a subject according to embodiments of the present invention. Certain steps of method 2400 may be performed as described herein,

[0270] At block 2410, a constitutional genome of the subject is obtained. At block 2420, one or more sequence tags are received for each of a plurality of DNA fragments in a biological sample of the subject, where the biological sample includes cell-free DNA. At block 2430, genomic positions are determined for the sequence tags. At block 2440, the sequence tags are compared to the constitutional genome to determine a first number of first loci. At each first loci, a number of the sequence tags having a sequence variant relative to the constitutional genome is above a cutoff value, where the cutoff value is greater than one.

[0271] At block 2450, a measure of heterogeneity of the one or more tumors are calculated based on the respective first numbers of the set of first genomic locations. In one aspect, the measures can provide a value that represents a number of mutations that are shared by tumors relative to a number of mutations that are not shared by tumors. Here, various tumors can exist as a single object, with different tumors within the object, which may represent what is normally called intra-tumor heterogeneity. The measure can also relate to whether some mutations are in one or a few tumors compared to mutations that are in many or most tumors. More than one measure of heterogeneity can be calculated.

[0272] At block 2460, the heterogeneity measure can be compared to a threshold value to determine a classification of a level of heterogeneity. The one or more measured can be used in various ways. For example, one or more heterogeneity measure measures can be used to predict the chance of tumor progression. In some tumors, the more heterogeneity the higher is the chance of progression and the higher is the chance of emergence of a resistant clone following treatment (e.g. targeted treatment).

C. Tumor Heterogeneity Measures

[0273] One example of a heterogeneity measure is number of 'concentration bands' of different groups of mutations in plasma. For example, if there are two predominant tumor

IDT Exhibit 1020.1742

clones within a patient, and if these clones are present in different concentrations, then we would expect to see two different mutations with different concentrations in plasma. These different values can be computed by determining the fractional concentration for different sets of mutations, where each set corresponds to one of the tumors.

[0274] Each of these concentrations can be called a 'concentration band' or 'concentration class'. If a patient has more clones, then more concentration bands/classes will be seen. Thus, the more bands, the more heterogeneous. The number of concentration bands can be seen by plotting the fractional concentrations for various mutations. A histogram can be made for the various concentrations, where different peaks correspond to different tumors (or different clones of one tumor). A large peak will likely be for mutations that are shared by all or some tumors (or clones of a tumor). These peaks may be analyzed to determine which smaller peaks are combined to determine a larger peak. A fitting procedure may be used, e.g., similar to the fitting procedure for FIGS. 10B and 11.

[0275] In one implementation, the histogram is a plot with Y-axis being the amount (e.g., number or proportion) of loci and x-axis being the fractional concentration. Mutations that are shared by all or some tumors would result in a higher fractional concentration. The peak size would represent the amount of loci that give rise to a particular fractional concentration. The relative size of the peaks at low and high concentration would reflect the degree of heterogeneity of the tumors (or clones of a tumor). A larger peak at the high concentration reflects that most mutations are shared by most or all tumors (or clones of a tumor) and indicate a lower degree of tumor heterogeneity. If the peak at the low concentration is larger, then most mutations are shared by a few tumors (or a few clones of a tumor). This would indicate a higher degree of tumor heterogeneity

[0276] The more peaks that exist, the more site-specific mutations there are. Each peak can correspond to a different set of mutations, where the set of mutations are from a subset of the tumors (e.g., just one or two tumors – as illustrated above). For the example of FIG. 19, there might be a total of 7 peaks, with the 4 site-only peaks likely having the smallest concentration (depending on the relative size of the tumors), two peaks for AB sites and CD sites, and a peak for mutations shared by all sites.

[0277] The location of the peaks can also provide a relative size of the tumors. A larger concentration would correlate to a larger tumor, as a larger tumor would release more tumor

IDT Exhibit 1020.1743

DNA into the sample, e.g., into plasma. Thus, one could estimate the load of tumor cells carrying the different classes of mutations.

[0278] Another example of a heterogeneity measure is the proportion of mutation sites having relatively few variant reads (e.g., 4, 5, or 6) compared to the proportion of mutation reads having relatively high variant reads (e.g., 9-13). Referring back to FIG. 22, one can see that the sire-specific mutations had fewer variant reads (which also results in a smaller fractional concentration). The share mutations have more variant reads (which also results in a larger fractional concentration). A ratio of a first proportion at 6 (smaller count) divided by a second proportion at 10 (larger count) conveys a heterogeneity measure. If the ratio is small, then there are few mutations that are site-specific, and thus the level of heterogeneity is low. If the ratio is large (or at least larger than values calibrated from known specimens), then the level of heterogeneity is larger.

D. Determining Thresholds

[0279] The threshold values can be determined from subjects whose tumors are biopsied (e.g., as described above) to directly determine a level of heterogeneity. The level may be defined in various ways, such as ratios of site-specific mutations to shared mutations. Biological samples (e.g., plasma samples) can then be analyzed to determine heterogeneity measures, where a heterogeneity measure from the biological samples can be associated with the level of heterogeneity determined by analyzing the cells of the tumors directly.

[0280] Such a procedure can provide a calibration of thresholds relative to heterogeneity levels. If the test heterogeneity measure falls between two thresholds, then the level of heterogeneity can be estimated as being between the levels corresponding to the thresholds.

[0281] In one embodiment, a calibration curve can be calculated between the heterogeneity levels determined from the biopsies and the corresponding heterogeneity measure determined from the plasma sample (or other sample). In such an example, the heterogeneity levels are numeric, where these numeric levels can correspond to different classifications. Different ranges of numeric levels can correspond to different diagnoses, e.g., different stages of cancer.

IDT Exhibit 1020.1744

E. Method Using Fractional Concentration from Genomic Representation

[0282] Tumor heterogeneity can also be analyzed using the fractional concentration, e.g., as determined using embodiments of method 1200. The genomic regions that exhibit one copy loss might come from different tumors. Thus, the fractional concentration determined for various genomic regions might differ depending on whether the amplification (or deletion for 1-copy loss) exists in just one tumor or multiple tumors. Thus, the same heterogeneity measures may be used for fractional concentrations determined via embodiments of method 1200.

[0283] For example, one genomic region can be identified as corresponding to a 1-copy loss, and a fractional concentration can be determined just from a respective density at that genomic region (the respective density could be used as a fractional concentration). A histogram can be determined from the various respective densities by counting the number of regions having various densities. If only one tumor or one tumor clone or one tumor deposit had a gain in a particular region, then the density of that region would be less than the density in a region that had a gain in multiple tumors or multiple tumor clones or multiple tumor deposits (i.e., the fractional concentration of tumor DNA in the shared region would be larger than the site-specific region). The heterogeneity measures described above can thus be applied to peaks identified using the copy number gain or loss in various regions, just as the fractional concentration of different sites showed a distribution of fractional concentrations.

[0284] In one implementation, if the respective densities are used for the histogram, one would have gains and losses separated. The regions showing a gain could be analyzed separately by creating a histogram just for gains, and a separate histogram can be created just for losses. If the fractional concentration is used, then the peaks of losses and gains can be analyzed together. For example, the fractional concentrations use a difference (e.g., as an absolute value) to the reference density, and thus the fractional concentrations for gains and losses can contribute to the same peak.

XIII. COMPUTER SYSTEM

[0285] Any of the computer systems mentioned herein may utilize any suitable number of subsystems. Examples of such subsystems are shown in FIG. 25 in computer apparatus 2500. In some embodiments, a computer system includes a single computer apparatus, where the subsystems can be the components of the computer apparatus. In other embodiments, a

IDT Exhibit 1020.1745

computer system can include multiple computer apparatuses, each being a subsystem, with internal components.

[0286] The subsystems shown in FIG. 25 are interconnected via a system bus 2575. Additional subsystems such as a printer 2574, keyboard 2578, fixed disk 2579, monitor 2576, which is coupled to display adapter 2582, and others are shown. Peripherals and input/output (I/O) devices, which couple to I/O controller 2571, can be connected to the computer system by any number of means known in the art, such as serial port 2577. For example, serial port 2577 or external interface 2581 (e.g. Ethernet, Wi-Fi, etc.) can be used to connect computer system 2500 to a wide area network such as the Internet, a mouse input device, or a scanner. The interconnection via system bus 2575 allows the central processor 2573 to communicate with each subsystem and to control the execution of instructions from system memory 2572 or the fixed disk 2579, as well as the exchange of information between subsystems. The system memory 2572 and/or the fixed disk 2579 may embody a computer readable medium. Any of the values mentioned herein can be output from one component to another component and can be output to the user.

[0287] A computer system can include a plurality of the same components or subsystems, e.g., connected together by external interface 2581 or by an internal interface. In some embodiments, computer systems, subsystem, or apparatuses can communicate over a network. In such instances, one computer can be considered a client and another computer a server, where each can be part of a same computer system. A client and a server can each include multiple systems, subsystems, or components.

[0288] It should be understood that any of the embodiments of the present invention can be implemented in the form of control logic using hardware (e.g. an application specific integrated circuit or field programmable gate array) and/or using computer software with a generally programmable processor in a modular or integrated manner. As used herein, a processor includes a multi-core processor on a same integrated chip, or multiple processing units on a single circuit board or networked. Based on the disclosure and teachings provided herein, a person of ordinary skill in the art will know and appreciate other ways and/or methods to implement embodiments of the present invention using hardware and a combination of hardware and software.

IDT Exhibit 1020.1746

[0289] Any of the software components or functions described in this application may be implemented as software code to be executed by a processor using any suitable computer language such as, for example, Java, C++ or Perl using, for example, conventional or object-oriented techniques. The software code may be stored as a series of instructions or commands on a computer readable medium for storage and/or transmission, suitable media include random access memory (RAM), a read only memory (ROM), a magnetic medium such as a hard-drive or a floppy disk, or an optical medium such as a compact disk (CD) or DVD (digital versatile disk), flash memory, and the like. The computer readable medium may be any combination of such storage or transmission devices.

[0290] Such programs may also be encoded and transmitted using carrier signals adapted for transmission via wired, optical, and/or wireless networks conforming to a variety of protocols, including the Internet. As such, a computer readable medium according to an embodiment of the present invention may be created using a data signal encoded with such programs. Computer readable media encoded with the program code may be packaged with a compatible device or provided separately from other devices (e.g., via Internet download). Any such computer readable medium may reside on or within a single computer program product (e.g. a hard drive, a CD, or an entire computer system), and may be present on or within different computer program products within a system or network. A computer system may include a monitor, printer, or other suitable display for providing any of the results mentioned herein to a user.

[0291] Any of the methods described herein may be totally or partially performed with a computer system including one or more processors, which can be configured to perform the steps. Thus, embodiments can be directed to computer systems configured to perform the steps of any of the methods described herein, potentially with different components performing a respective steps or a respective group of steps. Although presented as numbered steps, steps of methods herein can be performed at a same time or in a different order. Additionally, portions of these steps may be used with portions of other steps from other methods. Also, all or portions of a step may be optional. Additionally, any of the steps of any of the methods can be performed with modules, circuits, or other means for performing these steps.

[0292] The specific details of particular embodiments may be combined in any suitable manner without departing from the spirit and scope of embodiments of the invention.

IDT Exhibit 1020.1747

However, other embodiments of the invention may be directed to specific embodiments relating to each individual aspect, or specific combinations of these individual aspects.

[0293] The above description of exemplary embodiments of the invention has been presented for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form described, and many modifications and variations are possible in light of the teaching above. The embodiments were chosen and described in order to best explain the principles of the invention and its practical applications to thereby enable others skilled in the art to best utilize the invention in various embodiments and with various modifications as are suited to the particular use contemplated.

[0294] A recitation of "a", "an" or "the" is intended to mean "one or more" unless specifically indicated to the contrary.

[0295] All patents, patent applications, publications, and descriptions mentioned here are incorporated by reference in their entirety for all purposes. None is admitted to be prior art.

IDT Exhibit 1020.1748

WHAT IS CLAIMED IS:

1. A method for detecting cancer or premalignant change in a subject, the method comprising:

obtaining a constitutional genome of the subject;

receiving one or more sequence tags for each of a plurality of DNA fragments in a biological sample of the subject, the biological sample including cell-free DNA;

determining genomic positions for the sequence tags;

comparing the sequence tags to the constitutional genome to determine a first number of first loci, wherein:

at each first loci, a number of the sequence tags having a sequence variant relative to the constitutional genome is above a cutoff value, the cutoff value being greater than one;

determining a parameter based on a count of sequence tags having a sequence variant at the first loci; and

comparing the parameter to a threshold value to determine a classification of a level of cancer in the subject.

2. The method of claim 1, wherein the threshold value is determined from one or more samples from one or more other subjects.

3. The method of claim 1, wherein the cutoff value for a locus is dependent on a total number of sequence tags that have a genomic position at the locus.

4. The method of claim 1, wherein different cutoff values are used for at least two of the first loci.

5. The method of claim 4, further comprising:

dynamically determining a first cutoff value for one of the first loci, the one loci residing within a first region.

6. The method of claim 5, wherein the first cutoff value is determined based on a sequencing depth of the one of the first loci.

IDT Exhibit 1020.1749

7. The method of claim 5, wherein the first cutoff value is determined based on a false positive rate that depends on a sequencing error rate, a sequencing depth of the first region, and a number of nucleotide positions in the first region.

8. The method of claim 7, wherein the first cutoff value is determined based on a number of true positives in the first region.

9. The method of claim 8, further comprising:

calculating a number of true positives for the first cutoff value based on sequencing depth D and fractional concentration f of tumor-derived DNA in the biological sample.

10. The method of claim 9, wherein calculating the number of true positives uses the Poisson distribution probability according to the formula: $Pb = 1 - \sum_{i=0}^{r-1} Poisson(i, M_P)$, where Pb is the probability for detecting true positives, and r is the first cutoff value, and Mp=D x f/2.

11. The method of claim 5, wherein the first cutoff value is determined using any one of the following criteria:

if the sequencing depth is less than 50 then the first cutoff value is 5,

if the sequencing depth is 50 – 110 then the first cutoff value is 6,

if the sequencing depth is 111 – 200 then the first cutoff value is 7,

if the sequencing depth is 201 – 310 then the first cutoff value is 8,

if the sequencing depth is 311 – 450 then the first cutoff value is 9,

if the sequencing depth is 451 – 620 then the first cutoff value is 10, and

if the sequencing depth is 621 – 800 then the first cutoff value is 11.

12. The method of claim 1, wherein the parameter is a weighted sum of the first number of first loci, wherein a contribution of each first loci is weighted based on an importance value assigned to the respective first loci.

13. The method of claim 1, wherein the parameter includes a sum of the sequence tags indicating a sequence variant at the first number of first loci.

IDT Exhibit 1020.1750

14. The method of claim 13, wherein the sum is a weighted sum, and wherein one of the first loci has a first weight that is different than a second weight of a second of the first loci.

15. The method of claim 14, wherein the first weight is greater than the second weight, and wherein the one of the first loci is associated with cancer, and the second of the first loci is not associated with cancer.

16. The method of claim 1, wherein the parameter is the first number of first loci.

17. The method of claim 1, wherein determining a genomic position for a sequence tag includes:

aligning at least a portion of the sequence tags to a reference genome, wherein the alignment of a sequence tag allows for one or more mismatches between the sequence tag and the constitutional genome.

18. The method of claim 17, wherein comparing the sequence tags to the constitutional genome includes:

comparing the constitutional genome to the reference genome to determine a second number of second loci having a variant relative to the reference genome;

based on the aligning, determining a third number of third loci, wherein:

at each third loci, a number of the sequence tags having a sequence variant relative to the reference genome is above a cutoff value; and

taking a difference of the third number and the second number to obtain the first number of first loci.

19. The method of claim 18, wherein taking the difference of the third number and the second number identifies the specific loci of the first loci.

20. The method of claim 19, wherein determining the parameter includes:

for each locus of the first number of first loci:

counting sequence tags that align to the locus and have a sequence variant at the locus; and

determining the parameter based on the respective counts.

IDT Exhibit 1020.1751

21. The method of claim 1, wherein the constitutional genome is derived from a constitutional sample from the subject that contains more than 50% constitutional DNA.

22. The method of claim 1, wherein determining a genomic position for a sequence tag includes:

aligning at least a portion of the sequence tags to the constitutional genome, wherein the alignment of a sequence tag allows for one or more mismatches between the sequence tag and the constitutional genome.

23. The method of claim 22, wherein comparing the sequence tags to the constitutional genome includes:

based on the aligning, identifying sequence tags that have a sequence variant at a genomic location relative to the constitutional genome of the subject;

for each genomic location exhibiting a sequence variant:

counting a respective first number of sequence tags that align to the genomic location and have a sequence variant at the genomic location;

determining a parameter based on the respective first numbers.

24. The method of claim 23, wherein determining the parameter based on the respective first numbers includes:

summing the respective first numbers to obtain a first sum; and

using the first sum to determine the parameter.

25. The method of claim 24, wherein using the first sum to determine the parameter includes:

subtracting the number of genomic locations exhibiting a sequence variant from the first sum.

26. The method of claim 24, wherein using the first sum to determine the parameter includes:

normalizing the first sum based on an amount of sequence tags aligned.

27. The method of claim 1, further comprising:

IDT Exhibit 1020.1752

obtaining a constitutional sample of the subject that contains more than 90% constitutional DNA;

performing random sequencing of DNA fragments in the constitutional sample to obtain one or more second sequence tags for each of a plurality of DNA fragments in the constitutional sample;

aligning at least a portion of the second sequence tags to a reference genome, wherein the alignment of a second sequence tag allows for a mismatch between the second sequence tag and the constitutional genome at M or less genomic locations, wherein M is an integer equal to or greater than one; and

constructing the constitutional genome based on the second sequence tags and the aligning.

28. The method of claim 27, wherein the constitutional sample is the biological sample, and wherein constructing the constitutional genome includes:

determining a consensus sequence that includes a determination of a homozygous locus or a heterozygous locus having two alleles; and

using the consensus sequence at the constitutional genome.

29. The method of claim 1, further comprising:

receiving the biological sample of the subject; and

performing the random sequencing of DNA fragments in the biological sample to generate the one or more sequence tags for each of a plurality of DNA fragments in the biological sample.

30. The method of claim 1, wherein the one or more sequence tags are generated from a random sequencing of DNA fragments in the biological sample.

31 The method of claim 1, wherein the biological sample is urine, pleural fluid, ascitic fluid, peritoneal fluid, saliva, cerebrospinal fluid, or a stool sample.

32. The method of claim 1, wherein the parameter is a fractional concentration of tumor-derived DNA.

IDT Exhibit 1020.1753

33. A computer product comprising a non-transitory computer readable medium storing a plurality of instructions that when executed control a computer system to perform method 1.

34. A method for analyzing a heterogeneity of one or more tumors of a subject, the method comprising:

obtaining a constitutional genome of the subject;

receiving one or more sequence tags for each of a plurality of DNA fragments in a biological sample of the subject, the biological sample including cell-free DNA;

determining genomic positions for the sequence tags;

comparing the sequence tags to the constitutional genome to determine a first number of first loci, wherein:

at each first loci, a number of the sequence tags having a sequence variant relative to the constitutional genome is above a cutoff value, the cutoff value being greater than one; and

calculating a measure of heterogeneity of the one or more tumors based on the respective first numbers of the set of first genomic locations.

35. The method of claim 34, further comprising:

comparing the heterogeneity measure to one or more threshold values to determine a classification of a level of heterogeneity.

36. The method of claim 35, wherein the one or more threshold values are determined from one or more other subjects whose tumors have been biopsied and analyzed to determine mutations in the biopsied tumors to determine a level of heterogeneity, and wherein a measure of heterogeneity from a biological sample including cell-free DNA of one or more other subjects is used determine a threshold value.

37. The method of claim 36, wherein comparing the heterogeneity measure to one or more threshold values includes:

inputting the heterogeneity measure to a calibration function that outputs a heterogeneity level based on the heterogeneity measure.

IDT Exhibit 1020.1754

38. The method of claim 34, wherein the one or more measures include a total number of first genomic locations at which more than one DNA fragment aligned to the constitutional DNA while having a sequence variant at that genomic location.

39. The method of claim 34, wherein a plurality of heterogeneity measures are calculated, wherein calculating the heterogeneity measures includes:

for each first loci, calculating a proportion of sequence tags having a sequence variation;

creating a histogram of a number of first loci for values of the proportions; and

identifying a plurality of peaks in the histogram.

40. The method of claim 39, wherein a heterogeneity measure corresponds to the number of identified peaks.

41. The method of claim 39, wherein a heterogeneity measure includes a ratio of heights of two peaks.

42. The method of claim 39, wherein the proportions each represent a fractional concentration of tumor DNA as measured at a particular first locus.

43. The method of claim 34, wherein the heterogeneity measure corresponds to a ratio of a first proportion of first loci having a first specified amount of sequence tags with sequence variations and a second proportion of first loci having a second specified amount of sequence tags with sequence variations.

44. The method of claim 43, wherein the first specified amount is less than the second specified amount.

45. The method of claim 44, wherein the first specified amount is a first range and the second specified amount is a second range, the first range being below the second range.

46. The method of claim 43, wherein the first specified amount and the second specified amount correspond to a fractional concentration or to an absolute number of sequence tags with a sequence variation.

IDT Exhibit 1020.1755

47. The method of claim 34, wherein the heterogeneity measure is determined from a histogram of a number of first loci corresponding to each of the respective first numbers.

48. The method of claim 47, wherein the set of first identified genomic locations include a first subset and a second subset, wherein the one or more measures includes a first histogram of the respective first numbers corresponding to the first subset and a second histogram of the respective second numbers corresponding to the second subset.

49. The method of claim 34, wherein the heterogeneity measure includes a proportion of first loci having corresponding first numbers above a specified value.

50. A method of determining a fractional concentration of tumor DNA in a biological sample including cell-free DNA, the method comprising:

receiving one or more sequence tags for each of a plurality of DNA fragments in the biological sample;

determining genomic positions for the sequence tags;

for each of a plurality of genomic regions:

determining a respective amount of DNA fragments within the genomic region from sequence tags having a genomic position within the genomic region;

normalizing the respective amount to obtain a respective density; and

comparing the respective density to a reference density to identify whether the genomic region exhibits a 1-copy loss or a 1-copy gain;

calculating a first density from one or more respective densities identified as exhibiting a 1-copy loss or from one or more respective densities identified as exhibiting a 1-copy gain; and

calculating the fractional concentration by:

comparing the first density to another density to obtain a differential, wherein the differential is normalized with the reference density.

51. The method of claim 50, further comprising:

creating a histogram of a number of genomic regions having various respective densities;

identifying a plurality of peaks in the histogram; and

IDT Exhibit 1020.1756

calculating a heterogeneity measure from a ratio of the number of genomic regions of one or more first peaks to the number of genomic regions of one or more second peaks.

52. The method of claim 51, wherein the histogram uses values of fractional concentrations determined separately for each genomic region in creating the histogram.

53. The method of claim 51, wherein the first peaks have respective densities of a first specified amount, and the second peaks have respective densities of a second specified amount.

54. The method of claim 53, wherein the first specified amount is a first range and the second specified amount is a second range, and wherein the first range is less than the second range.

55. The method of claim 50, wherein comparing the respective density to the reference density to identify whether the genomic region exhibits a 1-copy loss or a 1-copy gain includes:

computing a difference between the respective density and the reference density; and

comparing the difference to a cutoff value.

56. The method of claim 50, wherein the differential is normalized with the reference density by:

dividing the differential by the reference density.

57. The method of claim 50, wherein the another density is the reference density.

58. The method of claim 57, wherein calculating the fractional concentration further includes multiplying the differential by two.

59. The method of claim 50, wherein the first density is calculated using respective densities identified as exhibiting a 1-copy gain, and wherein the another density is a second density calculated from respective densities identified as exhibiting a 1-copy loss.

IDT Exhibit 1020.1757

60. The method of claim 59, wherein the differential is normalized with the reference density by:

computing a first ratio of the first density and the reference density;

computing a second ratio of the second density and the reference density, wherein the differential is between the first ratio and the second ratio.

61. The method of claim 59, wherein comparing the respective density to the reference density to identify whether the genomic region exhibits a 1-copy loss or a 1-copy gain includes:

fitting peaks to a distribution curve of a histogram of the respective densities, wherein the first density corresponds to a first peak and the second density corresponds to a second peak.

62. The method of claim 50, wherein all genomic regions determined to exhibit a statistically significant gain in the respective density relative to the reference density are identified as exhibiting a 1-copy gain.

63. The method of claim 50, wherein normalizing the respective amount to obtain a respective density includes using a same total number of aligned reference tags to determine the respective density and the reference density.

64. The method of claim 50, wherein normalizing the respective amount to obtain a respective density includes dividing the respective amount by a total number of aligned reference tags.

65. The method of claim 50, wherein the plurality of genomic regions each have a same length.

66. The method of claim 50, wherein the genomic regions are non-overlapping.

IDT Exhibit 1020.1758

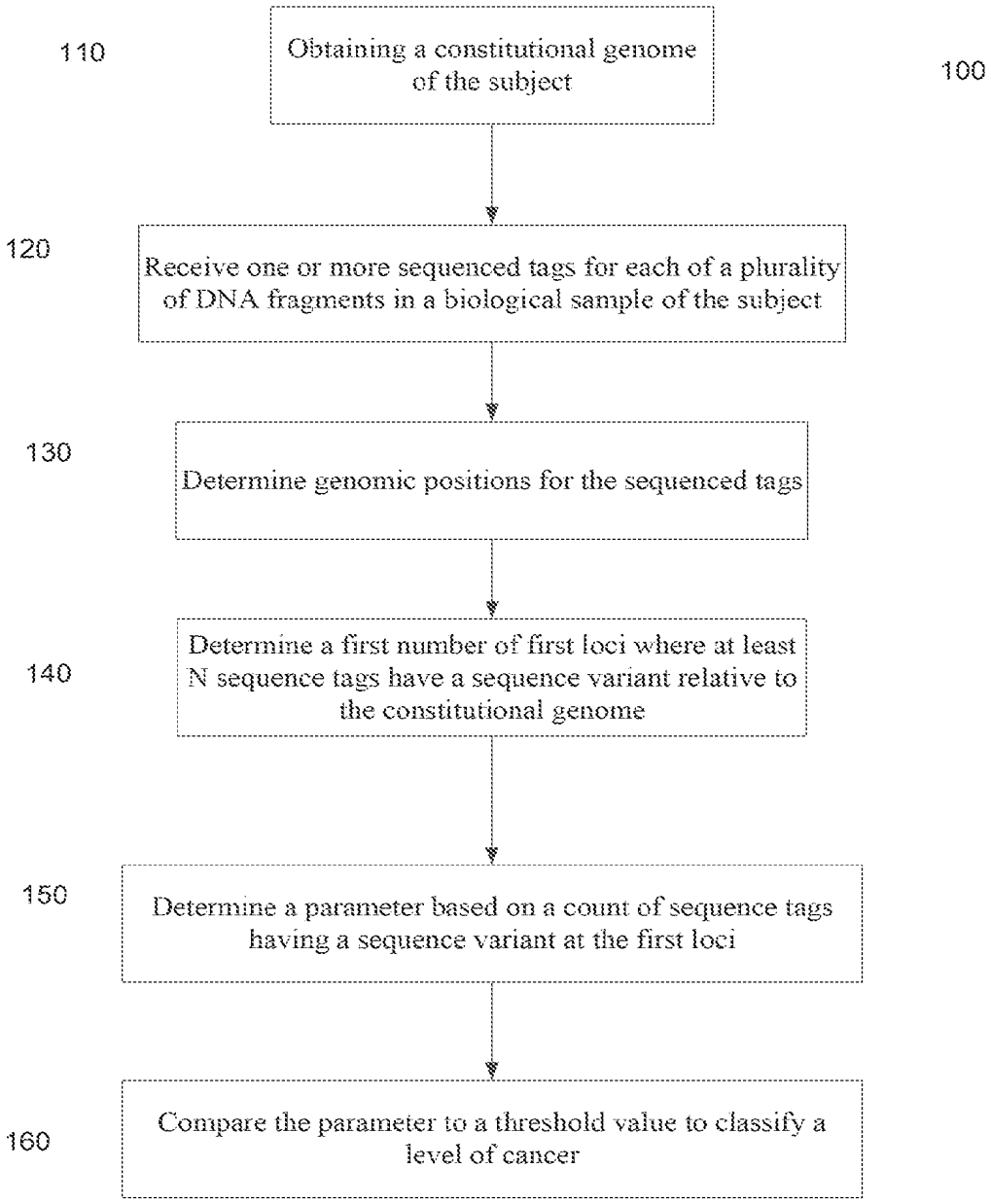

110
Obtaining a constitutional genome of the subject
100
120
Receive one or more sequenced tags for each of a plurality of DNA fragments in a biological sample of the subject
130
Determine genomic positions for the sequenced tags
140
Determine a first number of first loci where at least N sequence tags have a sequence variant relative to the constitutional genome
150
Determine a parameter based on a count of sequence tags having a sequence variant at the first loci
160
Compare the parameter to a threshold value to classify a level of cancer

FIG. 1

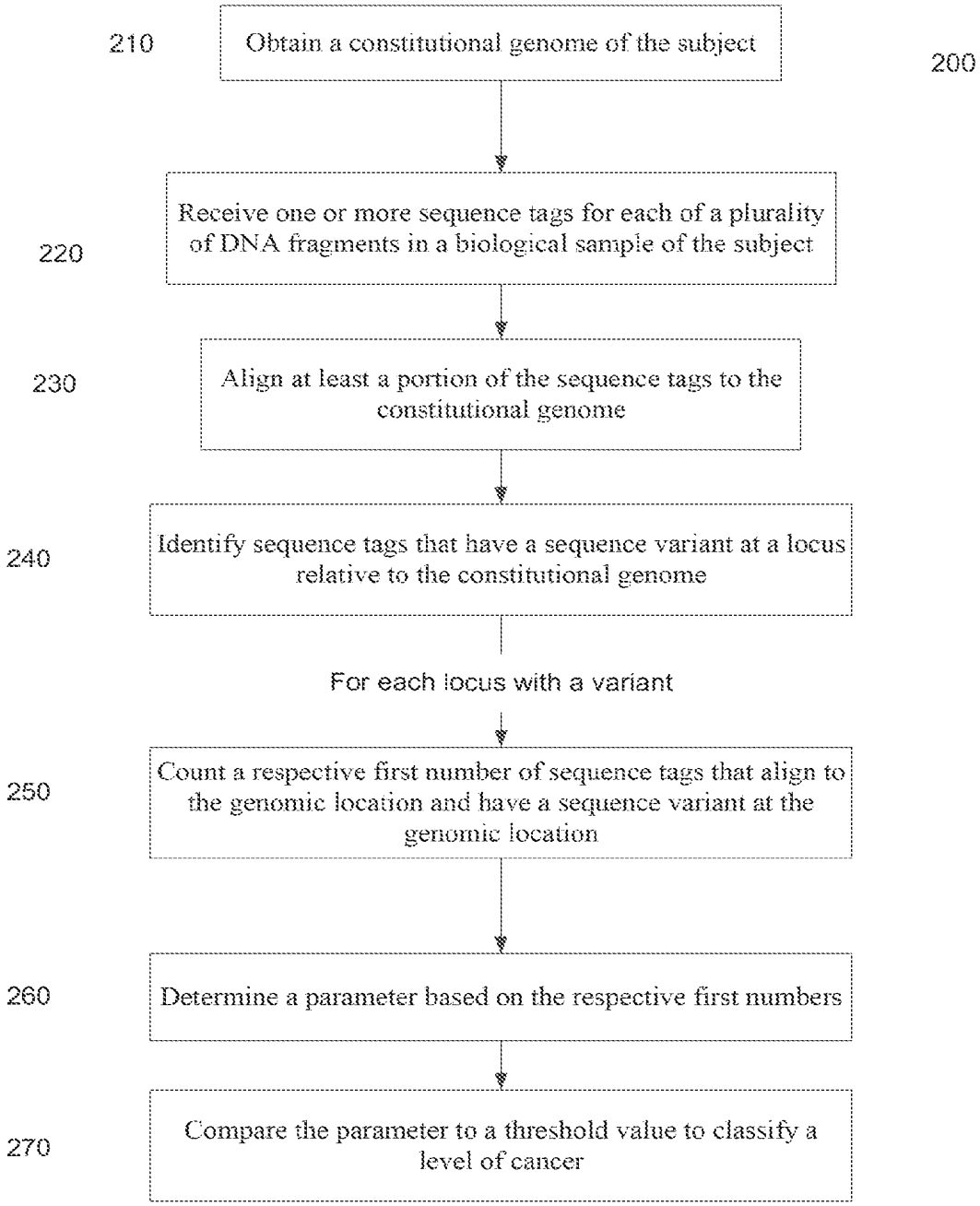

210
Obtain a constitutional genome of the subject
200
220
Receive one or more sequence tags for each of a plurality of DNA fragments in a biological sample of the subject
230
Align at least a portion of the sequence tags to the constitutional genome
240
Identify sequence tags that have a sequence variant at a locus relative to the constitutional genome
For each locus with a variant
250
Count a respective first number of sequence tags that align to the genomic location and have a sequence variant at the genomic location
260
Determine a parameter based on the respective first numbers
270
Compare the parameter to a threshold value to classify a level of cancer

FIG. 2

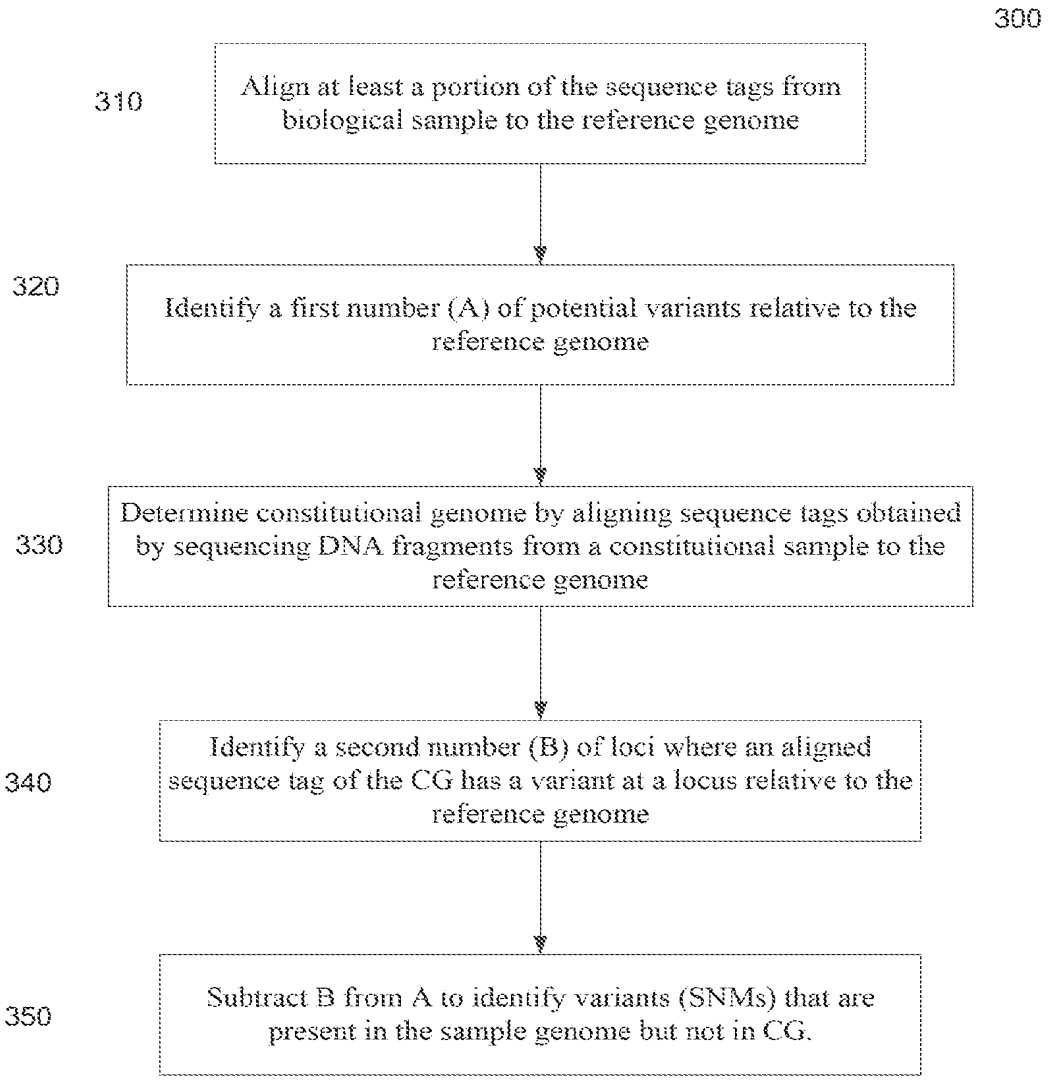

300
310
Align at least a portion of the sequence tags from biological sample to the reference genome
320
Identify a first number (A) of potential variants relative to the reference genome
330
Determine constitutional genome by aligning sequence tags obtained by sequencing DNA fragments from a constitutional sample to the reference genome
340
Identify a second number (B) of loci where an aligned sequence tag of the CG has a variant at a locus relative to the reference genome
350
Subtract B from A to identify variants (SNMs) that are present in the sample genome but not in CG.

FIG. 3

No. of times seen	1x		2x		3x		4x		5x		6x	
Sequencing Depth	Expect no. false positive loci for sequencing errors	Expect no. of mutations identified	Expect no. false positives loci for sequencing errors	Expect no. of mutations identified	Expect no. false positives loci for sequencing errors	Expect no. of mutations identified	Expect no. false positives loci for sequencing errors	Expect no. of mutations identified	Expect no. false positives loci for sequencing errors	Expect no. of mutations identified	Expect no. false positives loci for sequencing errors	Expect no. of mutations identified
1	2,968,500	146	1,499	4	0	0	0	0	0	0	0	0
10	29,850,499	1,138	45,000	271	40	43	0	5	0	1	0	0
20	59,403,980	1,896	190,000	793	380	241	1	57	0	11	0	2
40	117,631,683	2,594	780,000	1,782	3,293	970	10	429	0	158	0	50
60	174,706,399	2,851	1,770,000	2,403	11,407	1,739	54	1,033	0	554	0	252
80	230,650,961	2,945	3,160,000	2,725	27,387	2,286	176	1,709	1	1,113	0	645
100	285,487,746	2,980	4,950,000	2,879	53,900	2,626	436	2,205	3	1,679	0	1,152
120	339,238,690	2,993	7,140,000	2,948	93,613	2,814	913	2,546	7	2,145	0	1,663
140	391,925,294	2,997	9,730,000	2,978	149,193	2,911	1,703	2,755	15	2,481	0	2,098
160	443,568,633	2,999	12,720,000	2,991	223,307	2,959	2,922	2,873	30	2,701	0	2,426
180	494,189,366	3,000	16,110,000	2,996	318,620	2,981	4,700	2,936	55	2,835	1	2,653
200	543,807,741	3,000	19,900,000	2,999	437,800	2,992	7,187	2,969	94	2,912	1	2,799

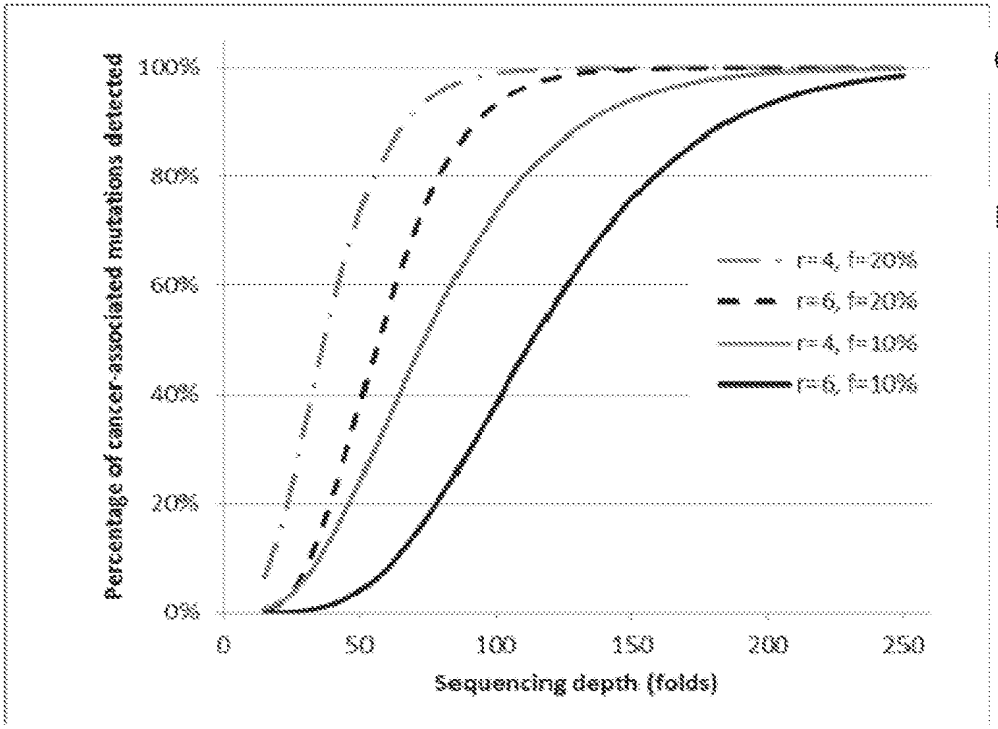
400

FIG. 4

no. of times seen	1x		2x		3x		4x		5x		6x	
Sequencing depth	Expect no. false positives loci for sequencing errors	Expect no. of mutations identified	Expect no. false positives loci for sequencing errors	Expect no. of mutations identified	Expect no. false positives loci for sequencing errors	Expect no. of mutations identified	Expect no. false positives loci for sequencing errors	Expect no. of mutations identified	Expect no. false positives loci for sequencing errors	Expect no. of mutations identified	Expect no. false positives loci for sequencing errors	Expect no. of mutations identified
1	2,998,500	74	1,499	1	0	0	0	0	0	0	0	0
10	29,850,499	664	45,000	79	40	6	0	0	0	0	0	0
20	59,403,980	1,180	190,000	271	380	43	1	5	0	1	0	0
40	117,631,683	1,896	780,000	793	3,793	241	18	57	0	11	0	2
60	174,706,399	2,331	1,770,000	1,327	11,407	573	54	197	0	56	0	13
80	230,650,961	2,594	3,160,000	1,732	27,387	976	178	429	1	158	0	50
100	285,487,746	2,754	4,950,000	2,138	53,900	1,369	436	727	3	326	0	126
120	339,238,690	2,851	7,140,000	2,403	93,613	1,730	913	1,058	7	554	0	252
140	391,925,394	2,909	9,730,000	2,592	149,193	2,037	1,765	1,390	15	824	0	427
160	443,568,633	2,945	12,720,000	2,725	223,307	2,286	2,922	1,700	30	1,113	0	645
180	494,189,366	2,967	16,110,000	2,817	318,620	2,479	4,730	1,973	55	1,404	1	891
200	543,807,741	2,980	19,900,000	2,879	437,800	2,626	7,187	2,205	94	1,679	1	1,152

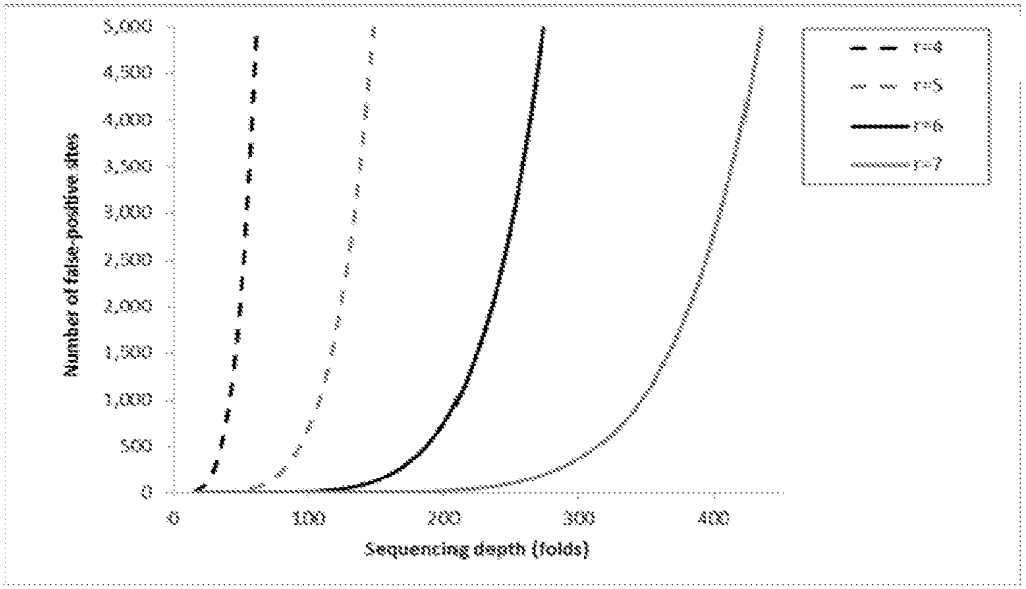
500

FIG. 5



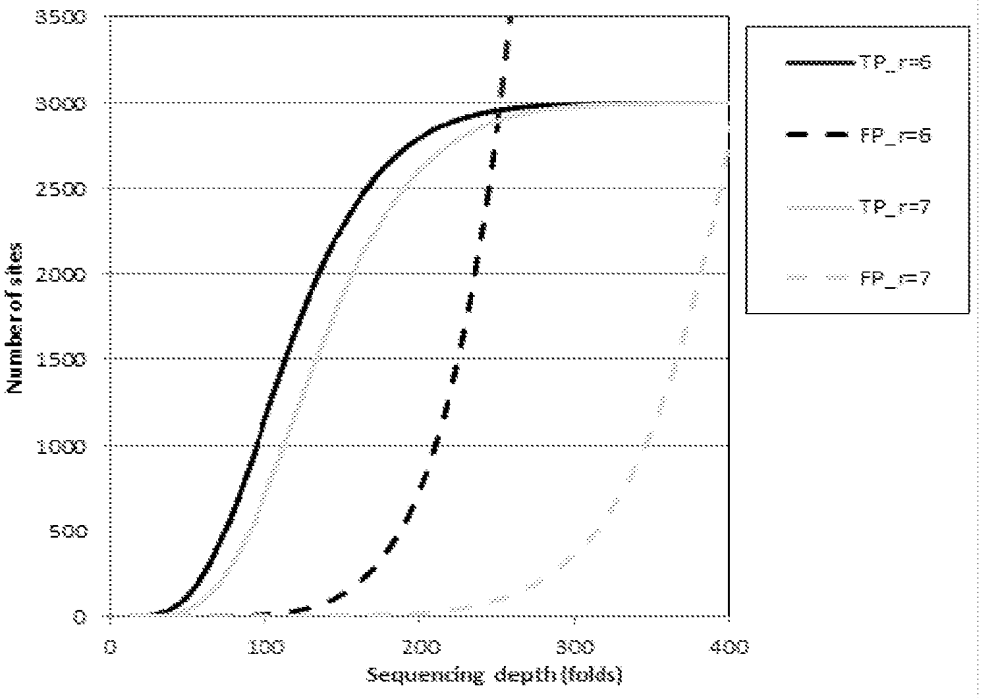
600

FIG. 6A



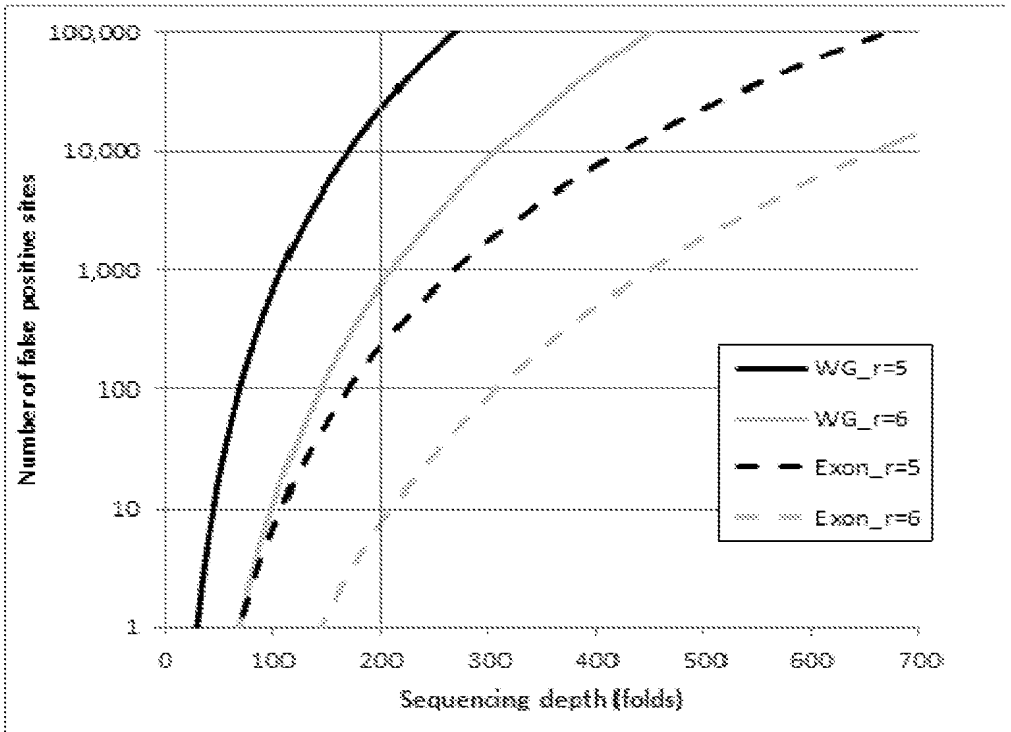
650

FIG. 6B



700

FIG. 7A



750

FIG. 7B

Case	No. of SNVs detected in the tumor tissue	Time point	No. of tumor-associated SNVs sequenced from plasma (% of SNVs seen in tumor tissue)	No. of plasma DNA sequence reads showing SNVs (p)	No. of plasma DNA sequence reads showing wildtype sequence (q)	Deduced fractional concentration of tumor-derived DNA by SNV analysis $(\frac{2p}{p+q})$	Deduced fractional concentration of tumor-derived DNA by GAAL analysis
HCC1	2,840	Pre-Tx	2,569 (94%)	11,389	31,602	53%	52%
		Post-Tx	44 (1.5%)	91	46,898	0.4%	1.4%
HCC2	3,105	Pre-Tx	1,097 (35%)	1,490	57,865	5.0%	4.3%
		Post-Tx	72 (2.3%)	206	66,692	0.6%	0.9%
HCC3	3,171	Pre-Tx	461 (15%)	525	48,886	2.1%	5.6%
		Post-Tx	31 (1%)	67	58,862	0.2%	0.9%
HCC4	1,334	Pre-Tx	201 (15%)	248	18,527	2.6%	7.6%
		Post-Tx	74 (5.5%)	149	22,144	1.3%	2.7%

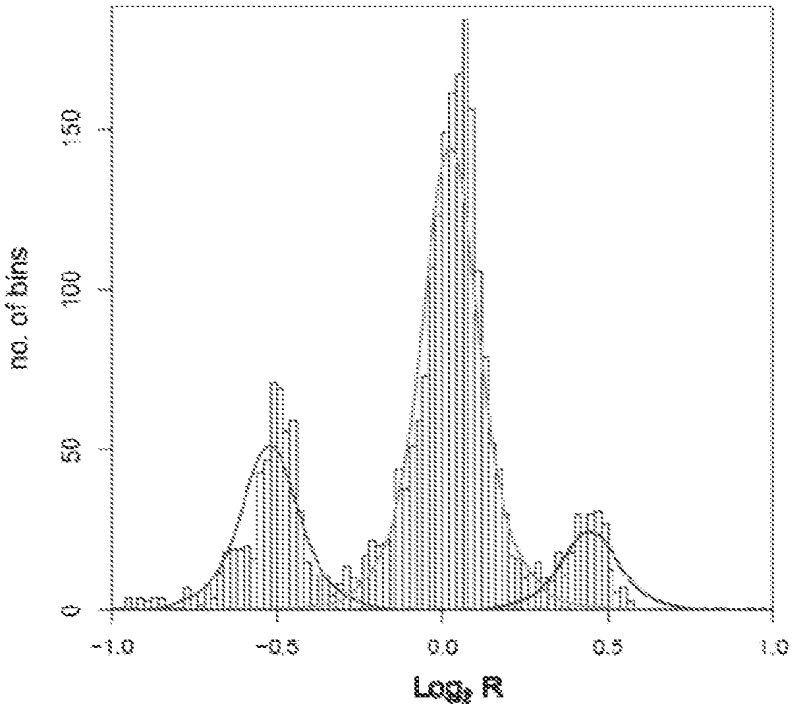
800

FIG. 8

	HCC1 SNVs		HCC2 SNVs		HCC3 SNVs		HCC4 SNVs		All SNVs		
Controls	Plasma DNA sequence reads showing tumor-associated SNVs	Plasma DNA sequence reads showing wildtype sequence	Plasma DNA sequence reads showing tumor-associated SNVs	Plasma DNA sequence reads showing wildtype sequence	Plasma DNA sequence reads showing tumor-associated SNVs	Plasma DNA sequence reads showing wildtype sequence	Plasma DNA sequence reads showing tumor-associated SNVs	Plasma DNA sequence reads showing wildtype sequence	Plasma DNA sequence reads showing tumor-associated SNVs (p)	Plasma DNA sequence reads showing wildtype sequence (q)	Fractional concentration of all tumor-associated SNVs in plasma $(\frac{2p}{p+q})$
C01	40	31,261	66	38,172	30	33,823	69	14,389	205	117,645	0.35%
C02	60	34,368	112	42,660	37	37,462	58	16,097	267	130,587	0.41%
C03	55	37,931	100	46,297	50	41,181	74	17,039	279	142,448	0.39%
C04	41	40,277	87	48,441	39	43,541	87	17,358	254	149,617	0.34%
C05	42	31,655	71	38,787	46	34,566	70	14,821	229	119,829	0.38%
C06	41	34,604	97	43,999	26	37,477	79	16,762	243	132,842	0.37%
C07	52	38,664	118	47,200	61	41,840	72	17,601	303	145,305	0.42%
C08	57	38,514	131	46,464	42	41,662	81	17,275	311	143,915	0.43%
C09	41	34,945	115	43,760	33	38,039	81	16,503	270	133,247	0.40%
C10	44	35,197	89	43,622	51	38,370	65	16,009	249	133,198	0.37%
C11	41	36,693	95	45,108	39	39,732	72	16,801	247	138,334	0.36%
C12	45	39,506	104	50,085	46	42,536	91	18,598	286	150,725	0.38%
C13	36	33,799	71	40,309	39	36,241	70	14,450	216	124,799	0.35%
C14	42	38,199	104	47,832	42	41,245	91	17,807	279	145,083	0.38%
C15	49	36,106	90	43,129	41	38,432	73	16,154	253	133,821	0.38%
C16	38	33,633	66	40,023	41	36,431	52	14,502	197	124,589	0.32%
										Mean	0.38%

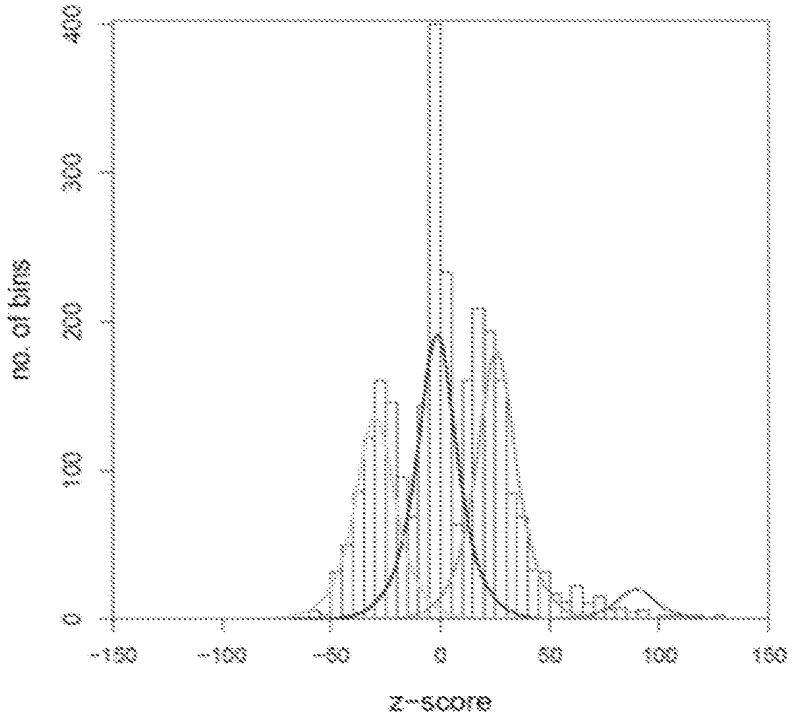
FIG. 9

900



1000

FIG. 10A



1050

FIG. 10B

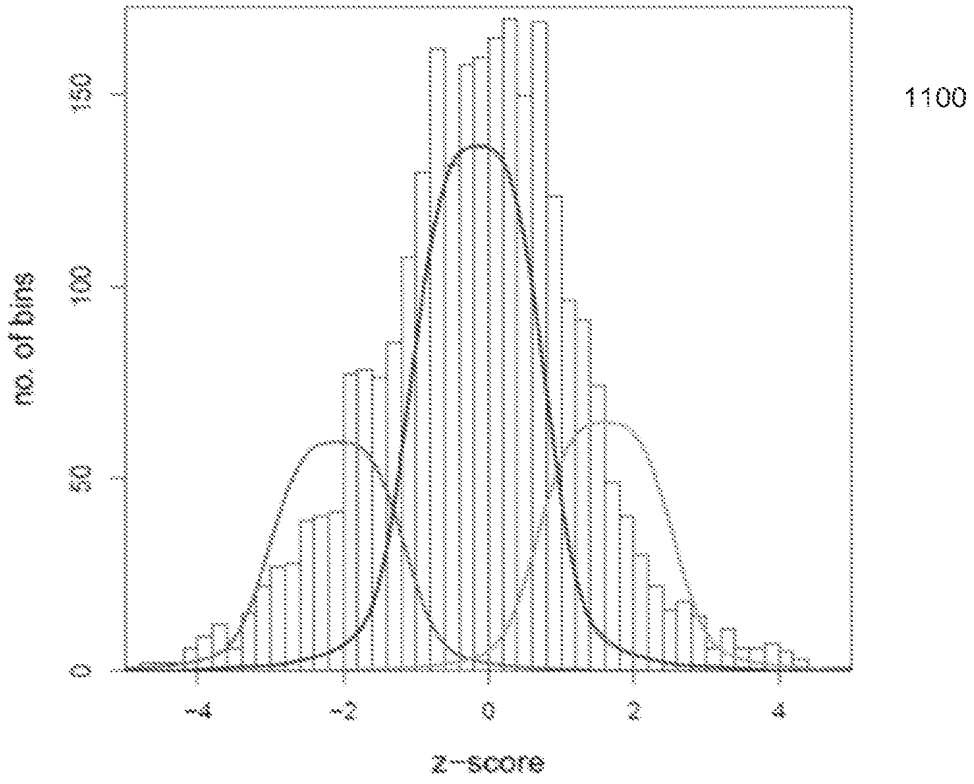


FIG. 11

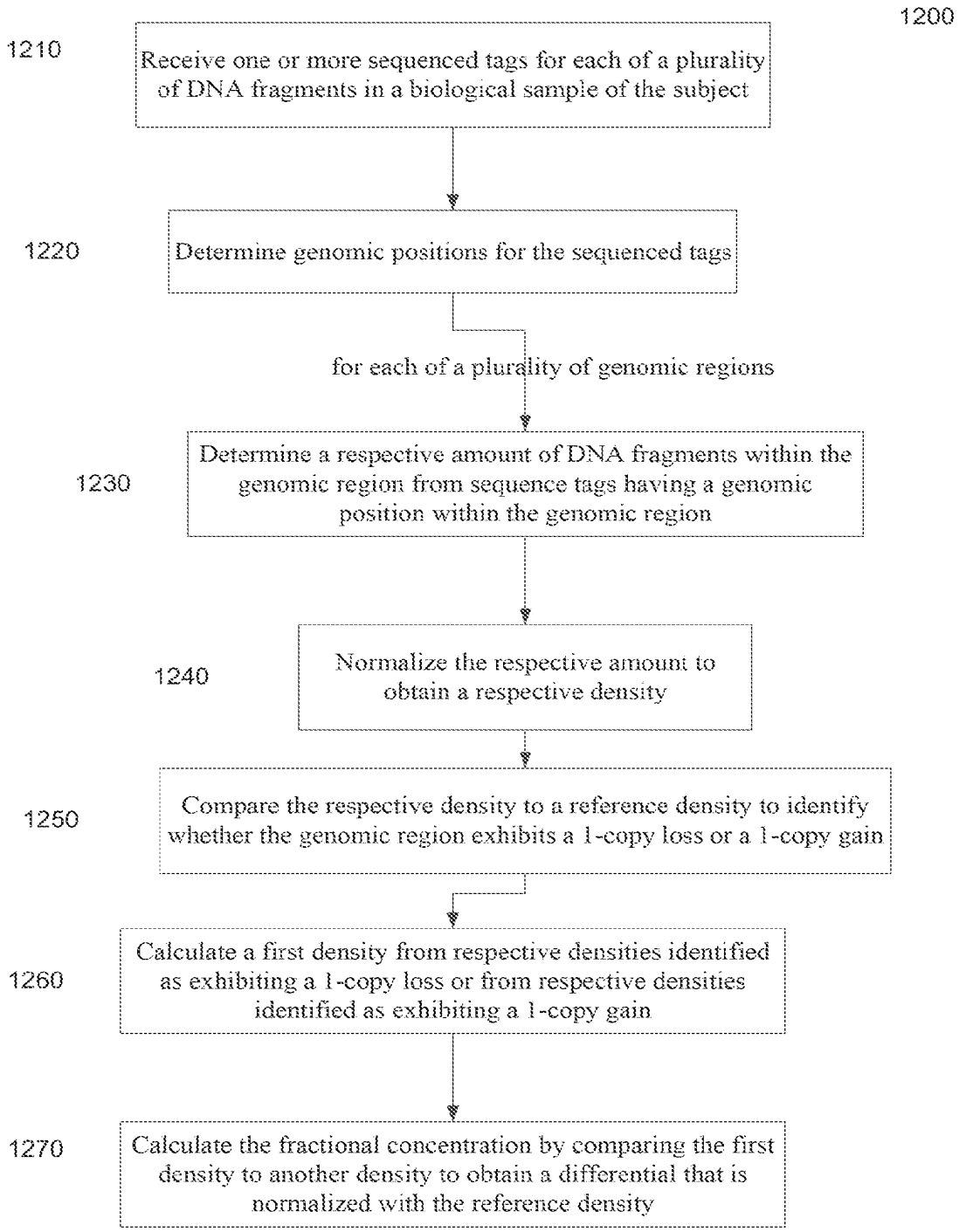

1200
1210
Receive one or more sequenced tags for each of a plurality of DNA fragments in a biological sample of the subject
1220
Determine genomic positions for the sequenced tags
for each of a plurality of genomic regions
1230
Determine a respective amount of DNA fragments within the genomic region from sequence tags having a genomic position within the genomic region
1240
Normalize the respective amount to obtain a respective density
1250
Compare the respective density to a reference density to identify whether the genomic region exhibits a 1-copy loss or a 1-copy gain
1260
Calculate a first density from respective densities identified as exhibiting a 1-copy loss or from respective densities identified as exhibiting a 1-copy gain
1270
Calculate the fractional concentration by comparing the first density to another density to obtain a differential that is normalized with the reference density

FIG. 12

No. of times detected in plasma (N)	No. of potential mutations detected N times in plasma	No. of potential mutations detected ≥ N times in plasma	Tumor mutations detected N times in plasma	Tumor mutations detected ≥N times in plasma	Percentage of mutations detected compared with N=1	Positive predictive value
1	3,605,085	3,623,970	61	2,064	100%	0.06%
2	13,518	18,885	35	2,003	97%	11%
3	2,029	5,367	71	1,968	95%	37%
4	766	3,338	83	1,897	92%	57%
5	437	2,572	140	1,814	88%	71%
6	313	2,135	161	1,674	81%	78%
7	265	1,822	165	1,513	73%	83%
8	261	1,557	192	1,348	65%	87%
9	231	1,296	185	1,156	56%	89%
10	191	1,065	160	971	47%	91%
11	179	874	165	811	39%	93%
12	148	695	137	646	31%	93%
13	122	547	109	509	25%	93%
14	99	425	92	400	19%	94%
15	74	326	71	308	15%	94%
16	54	252	49	237	11%	94%
17	39	198	39	188	9%	95%
18	35	159	33	149	7%	94%
19	24	124	23	116	6%	94%
20	20	100	20	93	5%	93%
21	18	80	17	73	4%	91%
22	11	62	10	56	3%	90%
23	9	51	9	46	2%	90%
24	8	42	5	37	2%	88%
25	6	34	6	32	2%	94%
26	7	28	5	26	1%	93%
27	8	21	8	21	1%	100%
28	1	13	1	13	1%	100%
29	4	12	4	12	1%	100%
30	8	8	8	8	0%	100%

1300

FIG. 13A

No. of times detected in plasma (N)	No. of potential mutations detected N times in plasma	No. of potential mutations detected ≥ N times in plasma	Tumor mutations detected N times in plasma	Tumor mutations detected ≥N times in plasma	Percentage of mutations detected compared with N=1	Positive predictive value
1	3,155,634	3,167,144	51	54	100%	<0.01%
2	9,835	11,510	2	3	6%	0.03%
3	1,123	1,675	1	1	2%	0.06%
4	314	552	-	-	-	-
5	132	238	-	-	-	-
6	49	106	-	-	-	-
7	22	57	-	-	-	-
8	11	35	-	-	-	-
9	6	24	-	-	-	-
10	5	18	-	-	-	-
11	4	13	-	-	-	-
12	2	9	-	-	-	-
13	2	7	-	-	-	-
14	3	5	-	-	-	-
15	-	2	-	-	-	-
16	-	2	-	-	-	-
17	1	2	-	-	-	-
18	-	1	-	-	-	-
19	1	1	-	-	-	-
20	-	-	-	-	-	-
21	-	-	-	-	-	-
22	-	-	-	-	-	-
23	-	-	-	-	-	-
24	-	-	-	-	-	-
25	-	-	-	-	-	-
26	-	-	-	-	-	-
27	-	-	-	-	-	-
28	-	-	-	-	-	-
29	-	-	-	-	-	-
30	-	-	-	-	-	-

1300

FIG. 13B

FIG. 14A

1400

Sequencing depth	Pre-treatment plasma			Post-treatment plasma		
	Percentage of nucleotides with the specified sequencing depth	Total no. nucleotide positions with single nucleotide variations detected in plasma	No. of sites with corresponding changes in tumor	Percentage of nucleotides with the specified sequencing depth	Total no. nucleotide positions with single nucleotide variations detected in plasma	No. of sites with corresponding changes in tumor
<50	36.3%	9	8	30.5%	0	0
50 – 125	29.9%	15	15	31.2%	0	0
126 – 235	18.6%	21	21	20.8%	0	0
236 – 380	9.4%	7	6	10.9%	0	0
381 – 560	4.2%	2	2	4.8%	0	0
561 – 760	1.6%	3	3	1.7%	0	0
Total		57	55		0	0

FIG. 14B

1450

Sequencing depth	Pre-treatment plasma			Post-treatment plasma		
	Percentage of nucleotides with the specified sequencing depth	Total no. nucleotide positions with single nucleotide variations detected in plasma	No. of sites with corresponding changes in tumor	Percentage of nucleotides with the specified sequencing depth	Total no. nucleotide positions with single nucleotide variations detected in plasma	No. of sites with corresponding changes in tumor
<50	30.0%	0	0	27.1%	0	0
50 – 125	33.6%	2	2	32.5%	0	0
126 – 235	21.6%	6	6	22.2%	0	0
236 – 380	10.2%	5	5	11.5%	0	0
381 – 560	3.7%	4	4	4.9%	0	0
561 – 760	0.9%	1	1	1.7%	0	0
Total		18	18		0	0

FIG. 15A

1500

Sequencing depth	Pre-treatment plasma			Post-treatment plasma		
	Percentage of nucleotides with the specified sequencing depth	Total no. nucleotide positions with single nucleotide variations detected in plasma	No. of sites with corresponding changes in tumor	Percentage of nucleotides with the specified sequencing depth	Total no. nucleotide positions with single nucleotide variations detected in plasma	No. of sites with corresponding changes in tumor
<50	38.8%	0	0	18.5%	0	0
50 – 125	45.7%	0	0	28.6%	0	0
126 – 235	15.2%	0	0	25.1%	0	0
236 – 380	0.2%	0	0	16.2%	0	0
381 – 560	<0.1%	0	0	8.3%	0	0
561 – 760	<0.1%	0	0	3.4%	0	0
Total		0	0		0	0

FIG. 15B

1550

Sequencing depth	Pre-treatment plasma			Post-treatment plasma		
	Percentage of nucleotides with the specified sequencing depth	Total no. nucleotide positions with single nucleotide variations detected in plasma	No. of sites with corresponding changes in tumor	Percentage of nucleotides with the specified sequencing depth	Total no. nucleotide positions with single nucleotide variations detected in plasma	No. of sites with corresponding changes in tumor
<50	26.7%	0	0	24.4%	0	0
50 – 125	35.9%	1	1	31.2%	1	1
126 – 235	24.7%	2	2	23.1%	0	0
236 – 380	10.4%	0	0	13.1%	0	0
381 – 560	2.2%	0	0	6.0%	0	0
561 – 760	0.1%	0	0	2.2%	0	0
Total		3	3		1	1

Sequencing depth	Pre-treatment plasma			Post-treatment plasma		
	Percentage of nucleotides with the specified sequencing depth	Total no. nucleotide positions with single nucleotide variations detected in plasma	No. of sites with corresponding changes in tumor	Percentage of nucleotides with the specified sequencing depth	Total no. nucleotide positions with single nucleotide variations detected in plasma	No. of sites with corresponding changes in tumor
<50	31.1%	7	7	30.8%	0	0
50 – 125	30.3%	14	14	33.2%	1	0
126 – 235	20.5%	19	19	21.6%	0	0
236 – 380	11.3%	13	13	10.3%	0	0
381 – 560	5.2%	7	4	3.5%	0	0
561 – 760	1.8%	4	2	0.7%	0	0
Total		64	59		1	0

FIG. 16

1600

Sequencing depth (folds)	Sensitivity when mutant is defined as being seen in the tumor tissue sequencing data for at least n times, where n=					
	4	5	6	7	8	9
20	73%	56%	38%	24%	13%	7%
21	77%	60%	43%	28%	16%	9%
22	80%	64%	47%	31%	19%	11%
23	83%	68%	51%	35%	22%	13%
24	85%	71%	55%	39%	26%	15%
25	87%	75%	59%	43%	29%	18%
26	89%	78%	63%	47%	33%	21%
27	90%	80%	67%	51%	36%	24%
28	92%	83%	70%	55%	40%	27%
29	93%	85%	73%	59%	44%	30%
30	94%	87%	76%	62%	48%	34%
31	95%	89%	78%	66%	51%	37%
32	96%	90%	81%	69%	55%	41%
33	96%	91%	83%	72%	58%	44%
34	97%	93%	85%	74%	61%	48%
35	97%	94%	87%	77%	65%	51%
36	98%	95%	88%	79%	68%	54%
37	98%	95%	90%	82%	70%	58%
38	99%	96%	91%	84%	73%	61%
39	99%	97%	92%	85%	76%	64%
40	99%	97%	93%	87%	78%	67%
41	99%	98%	94%	88%	80%	69%
42	99%	98%	95%	90%	82%	72%
43	99%	98%	96%	91%	84%	75%
44	100%	98%	96%	92%	86%	77%
45	100%	99%	97%	93%	87%	79%
46	100%	99%	97%	94%	89%	81%
47	100%	99%	98%	95%	90%	83%
48	100%	99%	98%	95%	91%	84%
49	100%	99%	98%	96%	92%	86%
50	100%	99%	99%	97%	93%	88%
51	100%	100%	99%	97%	94%	89%
52	100%	100%	99%	97%	95%	90%
53	100%	100%	99%	98%	95%	91%
54	100%	100%	99%	98%	96%	92%
55	100%	100%	99%	98%	96%	93%
56	100%	100%	99%	99%	97%	94%
57	100%	100%	100%	99%	97%	95%
58	100%	100%	100%	99%	98%	95%
59	100%	100%	100%	99%	98%	96%
60	100%	100%	100%	99%	98%	96%
61	100%	100%	100%	99%	98%	97%
62	100%	100%	100%	99%	99%	97%
63	100%	100%	100%	100%	99%	97%
64	100%	100%	100%	100%	99%	98%
65	100%	100%	100%	100%	99%	98%
66	100%	100%	100%	100%	99%	98%
67	100%	100%	100%	100%	99%	99%
68	100%	100%	100%	100%	99%	99%
69	100%	100%	100%	100%	100%	99%
70	100%	100%	100%	100%	100%	99%

1700

FIG. 17

Sequencing depth (folds)	No. of false positives in the whole genome when a mutant is defined as being seen in the tumor sequencing data for at least n times, where n=					
	4	5	6	7	8	9
20	1	0	0	0	0	0
21	1	0	0	0	0	0
22	1	0	0	0	0	0
23	1	0	0	0	0	0
24	1	0	0	0	0	0
25	1	0	0	0	0	0
26	2	0	0	0	0	0
27	2	0	0	0	0	0
28	2	0	0	0	0	0
29	3	0	0	0	0	0
30	3	0	0	0	0	0
31	3	0	0	0	0	0
32	4	0	0	0	0	0
33	5	0	0	0	0	0
34	5	0	0	0	0	0
35	6	0	0	0	0	0
36	7	0	0	0	0	0
37	7	0	0	0	0	0
38	8	0	0	0	0	0
39	9	0	0	0	0	0
40	10	0	0	0	0	0
41	11	0	0	0	0	0
42	12	0	0	0	0	0
43	14	0	0	0	0	0
44	15	0	0	0	0	0
45	17	0	0	0	0	0
46	18	0	0	0	0	0
47	20	0	0	0	0	0
48	22	0	0	0	0	0
49	24	0	0	0	0	0
50	26	0	0	0	0	0
51	28	0	0	0	0	0
52	30	0	0	0	0	0
53	33	0	0	0	0	0
54	35	0	0	0	0	0
55	38	0	0	0	0	0
56	41	0	0	0	0	0
57	44	0	0	0	0	0
58	47	0	0	0	0	0
59	51	0	0	0	0	0
60	54	0	0	0	0	0
61	58	0	0	0	0	0
62	62	0	0	0	0	0
63	66	0	0	0	0	0
64	71	0	0	0	0	0
65	75	0	0	0	0	0
66	80	0	0	0	0	0
67	85	0	0	0	0	0
68	90	0	0	0	0	0
69	96	0	0	0	0	0
70	102	0	0	0	0	0

FIG. 18

1800

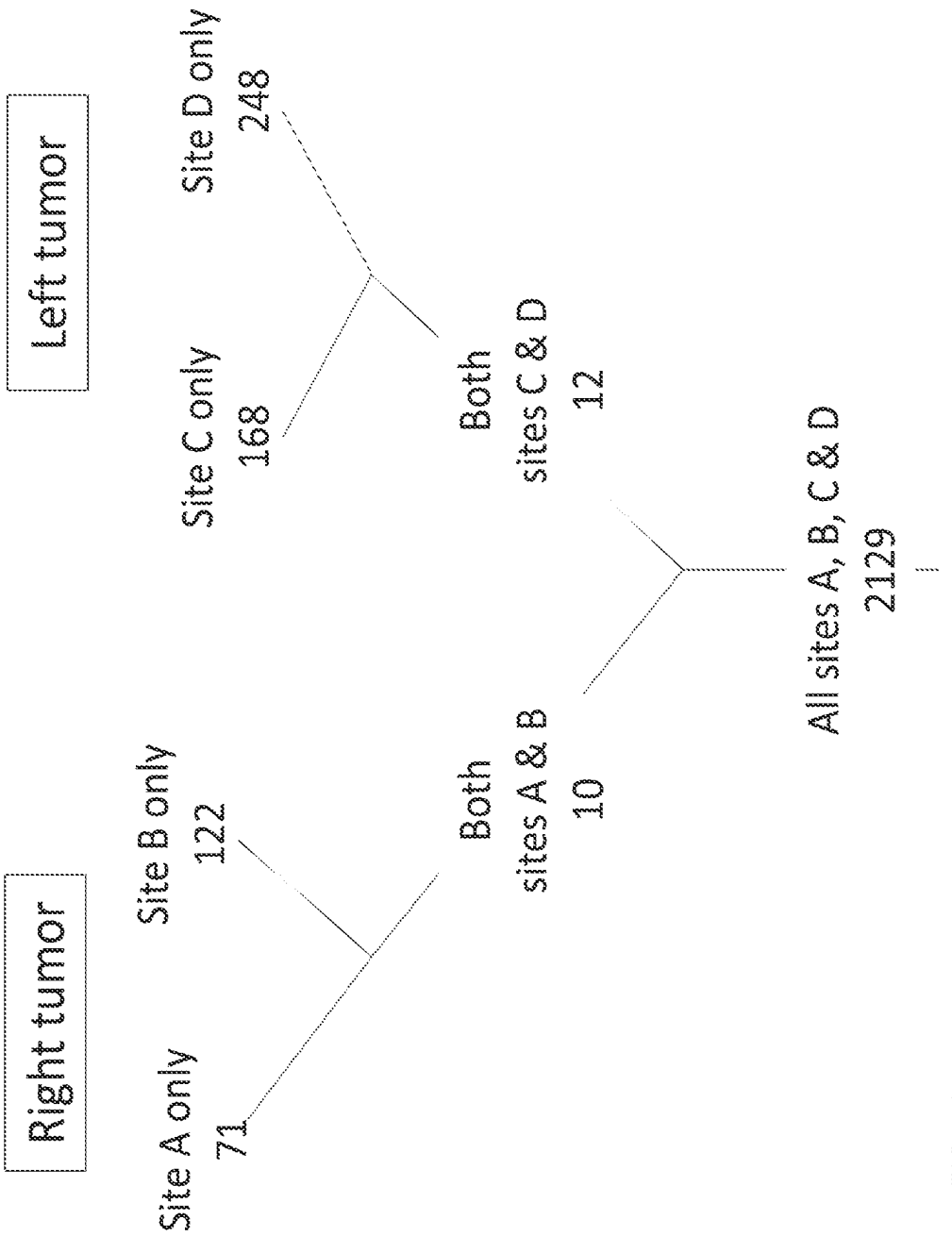

Left tumor
Site D only
248
Site C only
168
Both
sites C & D
12
Right tumor
Site B only
122
Site A only
71
Both
sites A & B
10
All sites A, B, C & D
2129

FIG. 19

Category	No. of loci with a mutation	Pre-treatment plasma			Post-treatment plasma		
		No. of DNA fragments carrying wildtype alleles	No. of DNA fragments carrying mutant alleles	Fractional concentration of tumor-derived DNA	No. of DNA fragments carrying wildtype alleles	No. of DNA fragments carrying mutant alleles	Fractional concentration of tumor-derived DNA
A	71	3,321	37	2.20%	2,149	5	0.46%
B	122	5,633	8	0.28%	3,516	1	0.06%
C	168	8,507	51	1.19%	5,438	2	0.07%
D	248	10,880	26	0.48%	7,060	2	0.06%
AB	10	423	21	9.46%	297	0	0.00%
CD	12	569	3	1.05%	333	0	0.00%
ABCD	2129	70,940	21,344	46.26%	61,417	54	0.18%

2000

FIG. 20

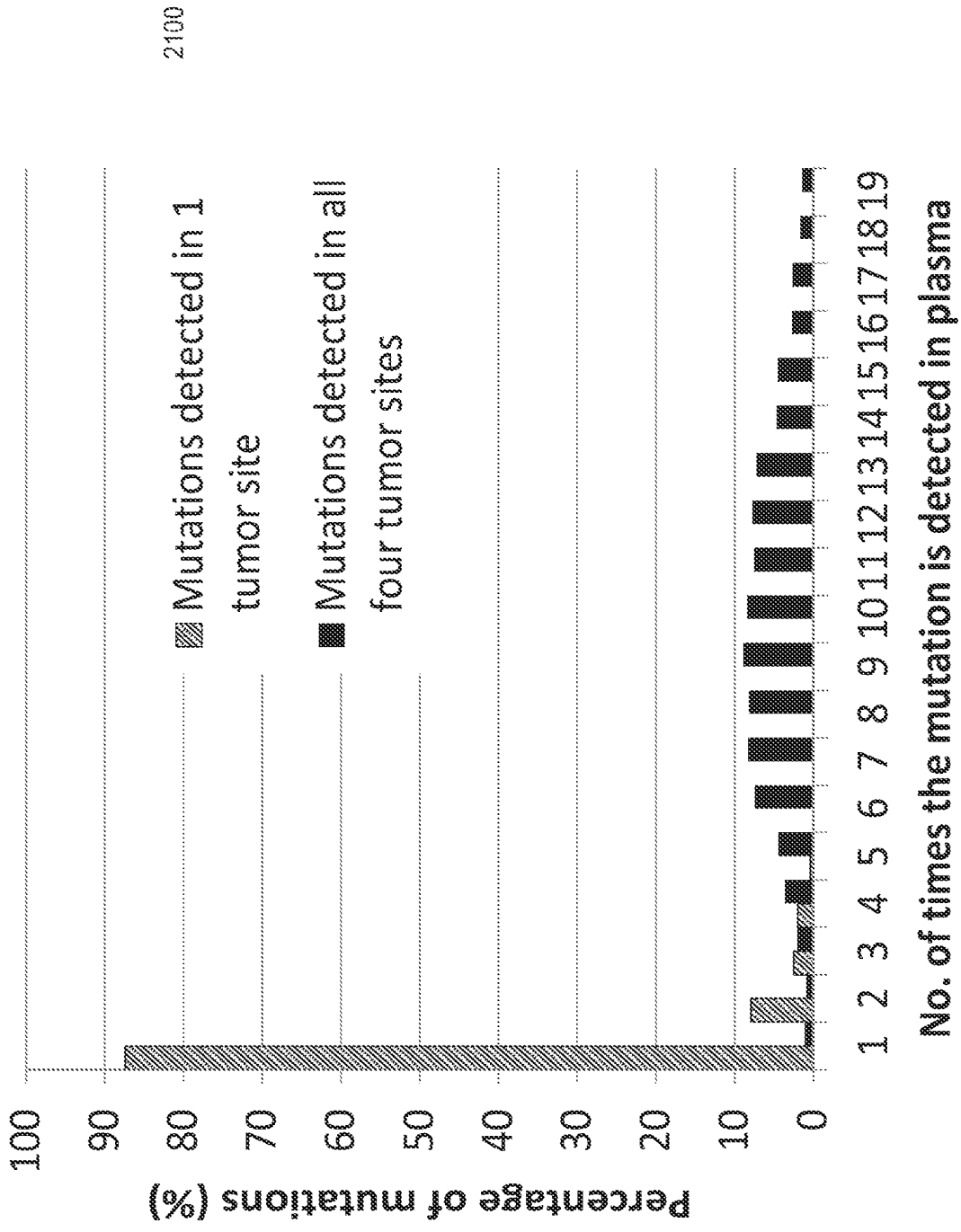

2100
Mutations detected in 1 tumor site
Mutations detected in all four tumor sites
Percentage of mutations (%)
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
No. of times the mutation is detected in plasma

FIG. 21

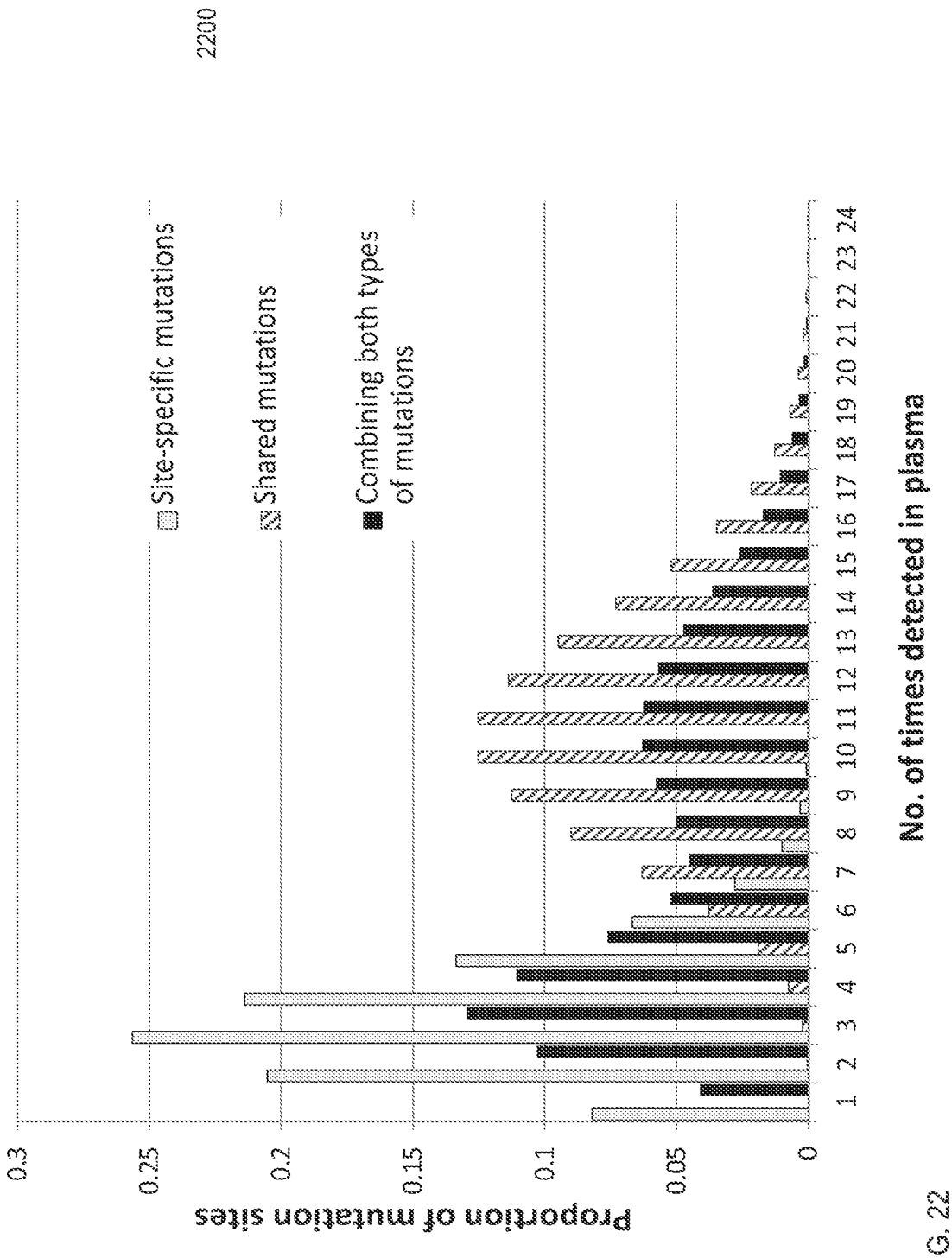

2200
Site-specific mutations
Shared mutations
Combining both types of mutations
Proportion of mutation sites
0.3
0.25
0.2
0.15
0.1
0.05
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
No. of times detected in plasma

FIG. 22

Healthy control subject																
Cat.	No. of loci	No (fractional concentration) of mutant alleles in plasma														
		TC1	TC2	TC3	TC4	TC5	TC6	TC7	TC8	TC9	TC10	TC11	TC12	TC13	TC14	TC15	TC16
A	71	0	1 (0.17%)	1 (0.18%)	0	1 (0.17%)	1 (0.18%)	1 (0.87%)	1 (0.20%)	0	6 (0.17%)	0	0	1 (0.17%)	2 (0.38%)	0	2 (0.38%)
B	122	0	0	0	0	0	0	1 (0.44%)	0	1 (0.11%)	3 (0.05%)	1 (0.10%)	1 (0.12%)	0	0	0	1 (0.11%)
C	168	0	0	0	0	0	0	2 (0.96%)	0	2 (0.16%)	0	0	0	1 (0.07%)	0	0	0
D	248	0	0	1 (0.06%)	0	0	0	0	0	0	1 (0.01%)	0	0	1 (0.05%)	0	0	0
AB	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CD	12	0	0	0	0	0	0	0	0	0	1 (0.16%)	0	0	1 (0.92%)	0	0	0
ABCD	2129	11 (0.07%)	3 (0.02%)	2 (0.01%)	3 (0.08%)	5 (0.03%)	5 (0.03%)	10 (0.82%)	4 (0.03%)	4 (0.03%)	19 (0.02%)	3 (0.02%)	0	1 (0.01%)	5 (0.03%)	2 (0.01%)	4 (0.03%)

FIG. 23

2300

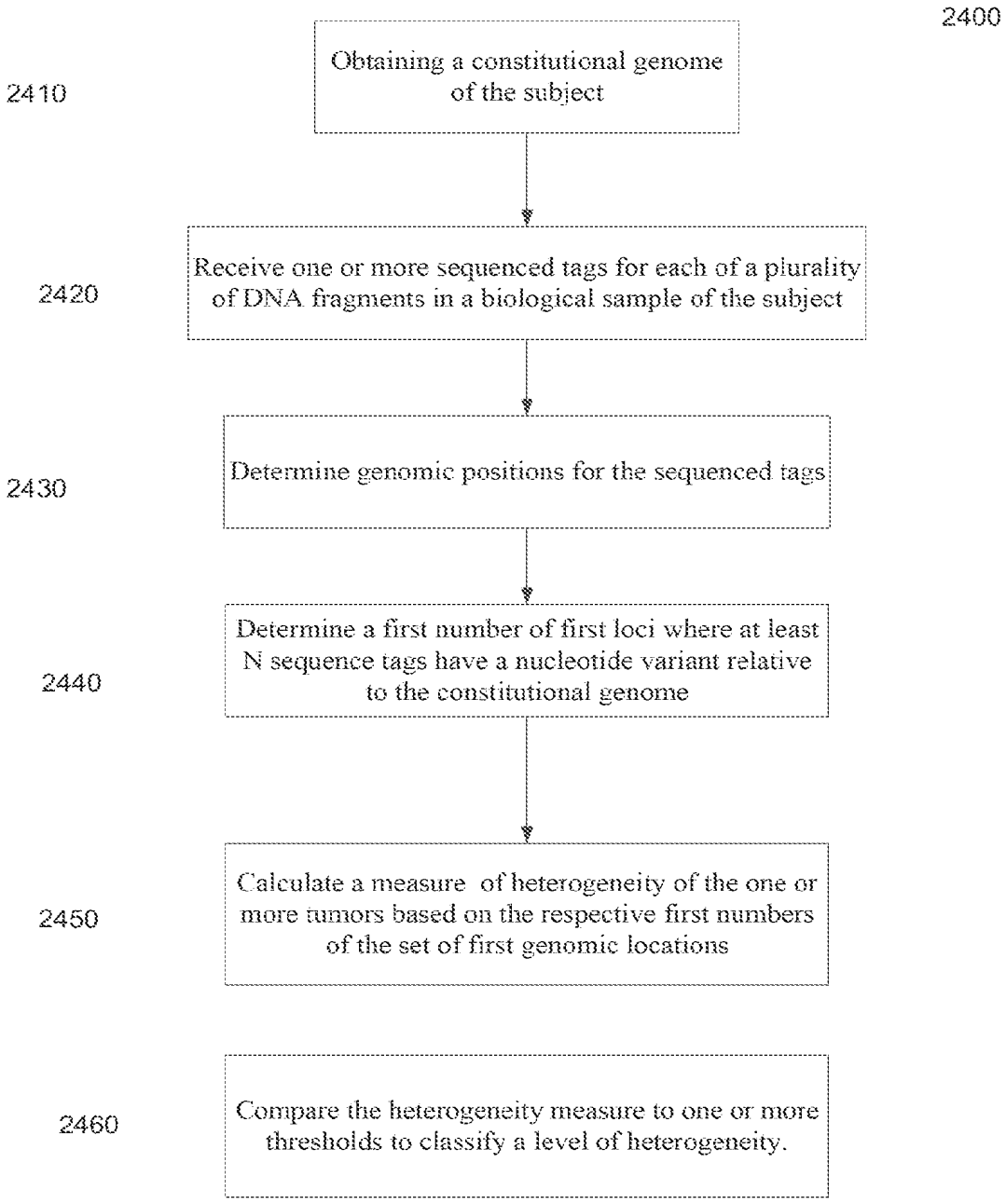

2400
2410
Obtaining a constitutional genome of the subject
2420
Receive one or more sequenced tags for each of a plurality of DNA fragments in a biological sample of the subject
2430
Determine genomic positions for the sequenced tags
2440
Determine a first number of first loci where at least N sequence tags have a nucleotide variant relative to the constitutional genome
2450
Calculate a measure of heterogeneity of the one or more tumors based on the respective first numbers of the set of first genomic locations
2460
Compare the heterogeneity measure to one or more thresholds to classify a level of heterogeneity.

FIG. 24

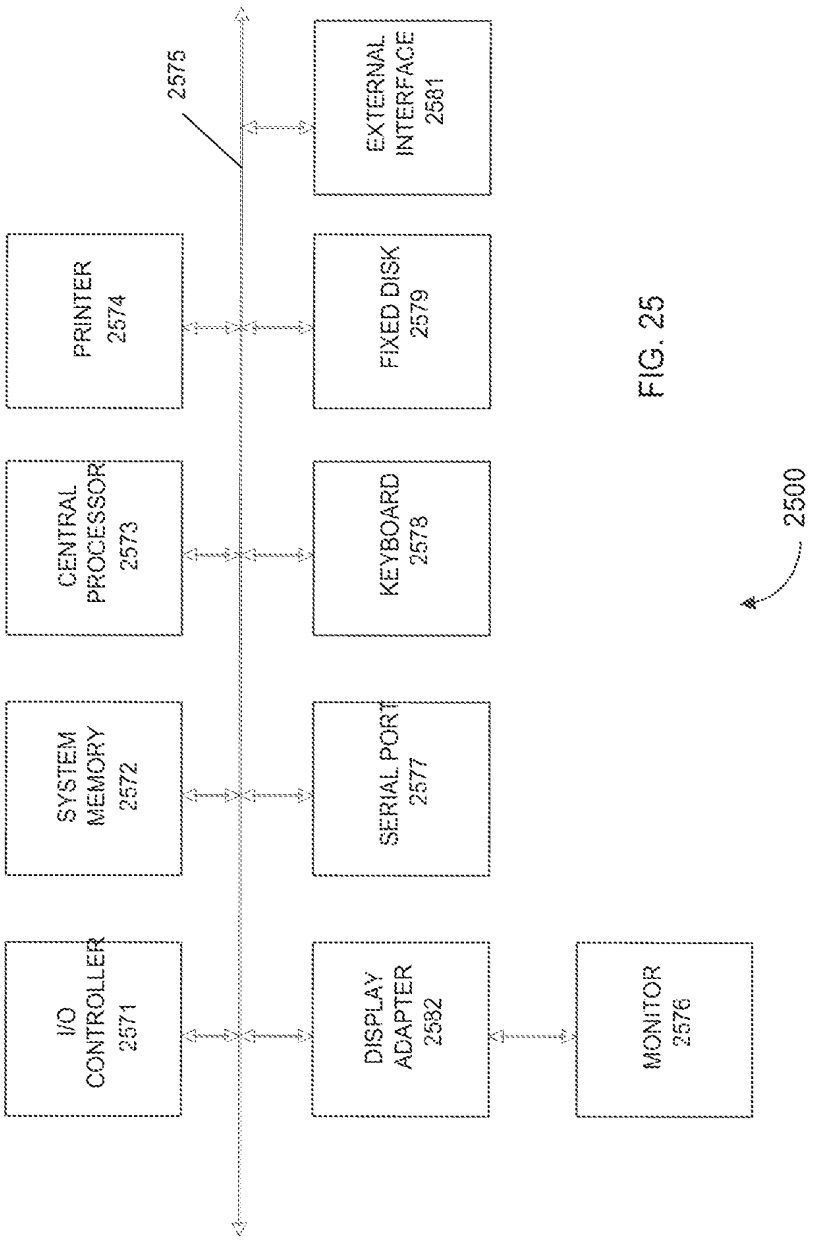

2575
EXTERNAL INTERFACE 2581
PRINTER 2574
FIXED DISK 2579
CENTRAL PROCESSOR 2573
KEYBOARD 2578
SYSTEM MEMORY 2572
SERIAL PORT 2577
I/O CONTROLLER 2571
DISPLAY ADAPTER 2582
MONITOR 2576
2500

FIG. 25

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
6 September 2013 (06.09.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/130512 A2

(51) International Patent Classification:
C12Q 1/68 (2006.01)

(21) International Application Number:
PCT/US2013/027891

(22) International Filing Date:
27 February 2013 (27.02.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/603,909 27 February 2012 (27.02.2012) US

(71) Applicant: **THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL** [US/US]; 308 Bynum Hall, Chapel Hill, NC 27599 (US).

(72) Inventors: **JABARA, Cassandra, B.**; 87 High Ridge Avenue, Ridgefield, CT 06877 (US). **ANDERSON, Jeffrey, A.**; 199 Penn View Drive, Pennington, NJ 08534 (US). **SWANSTROM, Ronald, I.**; 7021 Knotty Pine Drive, Chapel Hill, NC 27517 (US).

(74) Agents: **LETTS, Nathan, P.** et al.; The Eclipse Group LLP, 6345 Balboa Blvd., Suite 325, Encino, CA 91316 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

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

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

— *with sequence listing part of description (Rule 5.2(a))*

(54) Title: METHODS AND USES FOR MOLECULAR TAGS

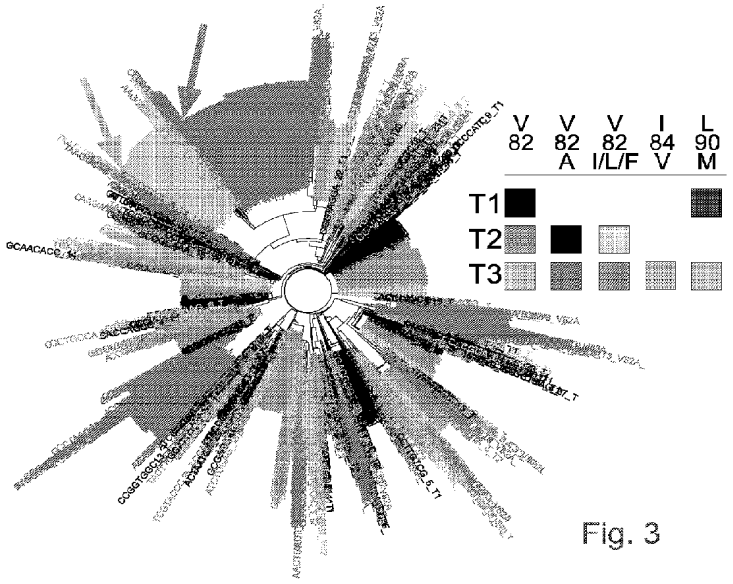


Fig. 3

(57) Abstract: Methods and uses for molecular tags are disclosed. Molecular tags may be attached to nucleic acid molecules. The attachment of the nucleic acid molecules prior to PCR amplification and sequencing improves the accuracy of genetic analysis and detection of genetic variations and diversity. Molecular tags may also be used for detection of drug-resistant variants. Methods for using molecular tags for determining and correcting PCR errors and/or sequencing error are also disclosed.

WO 2013/130512 A2

METHODS AND USES FOR MOLECULAR TAGS

Cross reference to related applications

[0001] This application claims the benefit of U.S. Provisional Appn. No. 61/603,909, filed February 27, 2012, Jabara *et al.*, entitled "Methods and Uses for Molecular Tags" having Atty. Docket No. 41977-705.101 , which is hereby incorporated by reference in its entirety.

Statement regarding federally sponsored research or development

[0002] This invention was made with government support under Grant No. AI044667 awarded by the U.S. National Institutes of Health (NIH). The government has certain rights in the invention.

FIELD OF THE INVENTION

[0003] The present invention relates generally to molecular tags and more specifically to compositions comprising molecular tags and methods for using molecular tags in genetic analysis. Molecular tags may also be used in the identification of drug-resistant variants. Also provided are methods for using molecular tags to detect and correct PCR amplification errors and sequencing errors.

BACKGROUND OF THE INVENTION

[0004] Deep sequencing technologies allow extensive sampling of genetic populations. Limitations of these technologies, however, potentially bias this sampling, particularly when a PCR step precedes the sequencing protocol. Typically, an unknown number of templates are used in initiating the PCR amplification, and this can lead to unrecognized sequence resampling creating apparent homogeneity; also, PCR-mediated recombination can disrupt linkage, and differential amplification efficiency or templates entering at different PCR cycles can skew the frequency of alleles. Finally, misincorporation of nucleotides during PCR and errors during the sequencing protocol can inflate diversity.

SUMMARY OF THE INVENTION

[0005] Limitations to current techniques can be overcome by including a unique sequence tag in the initial primer such that each template receives a unique Primer ID. After sequencing, repeated identification of a Primer ID reveals sequence resampling. These resampled sequences are then used to create an accurate consensus sequence for each template, correcting for recombination, allelic skewing, misincorporation errors, and sequencing errors. The resulting

IDT Exhibit 1020.1786

population of consensus sequences directly represents the initial sampled templates. Use of these molecular tags, can detect and correct for PCR error and/or sequencing error.

[0006] This approach can also be used in genetic analysis. Use of molecular tags, such as the Primer ID, enables analysis of the distribution of sequence variation of a gene in a complex genetic population. With this approach, major and minor polymorphisms at coding and noncoding positions have been identified. In addition, dynamic genetic changes within the population during intermittent drug exposure can be observed, including the emergence of multiple resistant alleles. The methods disclosed herein provide an unprecedented view of a complex genetic population in the absence of PCR resampling, PCR biases, and sequencing error.

[0007] Methods and uses for molecular tags are disclosed. Each copy of a nucleic acid molecule randomly chooses from a non-depleting reservoir of diverse Primer IDs. Attachment of the Primer ID to the nucleic acid molecule prior to amplification and sequencing enables direct counting of nucleic acid molecules and increased accuracy in detecting genetic variants. Primer IDs may also be used for detection of drug-resistant variants. Lastly, Primer IDs may also be used for reducing and/or correcting PCR errors and/or sequencing errors.

[0008] Disclosed herein is a method for determining genetic diversity of a sample comprising: (a) providing a sample comprising a nucleic acid template molecule; (b) attaching a primer comprising a Primer ID to each nucleic acid template molecule to be analyzed to generate a tagged nucleic acid template, wherein each tagged nucleic acid template is attached to a unique Primer ID; (c) amplifying the tagged nucleic acid template to produce tagged amplicons; and (d) detecting the tagged amplicons, thereby determining genetic diversity of a sample.

[0009] Further provided is a method for detecting genetic variants comprising: (a) providing a sample comprising a nucleic acid template molecule; (b) attaching a primer comprising a Primer ID to each nucleic acid template molecule to be analyzed to generate a tagged nucleic acid template, wherein each tagged nucleic acid template is attached to a unique Primer ID; (c) amplifying the tagged nucleic acid template to produce tagged amplicons; and (d) detecting the tagged amplicons, thereby detecting genetic variants.

[0010] Also provided herein is a method for determining or screening for drug-resistant variants comprising: (a) providing a sample comprising a nucleic acid template molecule; (b) attaching a primer comprising a Primer ID to each nucleic acid template molecule to be analyzed to generate a tagged nucleic acid template, wherein each tagged nucleic acid template is attached

IDT Exhibit 1020.1787

to a unique Primer ID; (c) amplifying the tagged nucleic acid template to produce tagged amplicons; and (d) detecting the tagged amplicons, thereby determining or screening for drug-resistant variants.

[0011] Further disclosed herein is a method for determining PCR resampling in an amplification reaction comprising: (a) providing a sample comprising a nucleic acid template molecule; (b) attaching a primer comprising a Primer ID to each nucleic acid template molecule to be analyzed to generate a tagged nucleic acid template, wherein each tagged nucleic acid template is attached to a unique Primer ID; (c) amplifying the tagged nucleic acid template to produce tagged amplicons; and (d) detecting the tagged amplicons, thereby determining PCR resampling in an amplification reaction.

[0012] Further disclosed herein is a method for determining PCR error and/or sequencing error comprising: (a) providing a sample comprising a nucleic acid template molecule; (b) attaching a primer comprising a Primer ID to each nucleic acid template molecule to be analyzed to generate a tagged nucleic acid template, wherein each tagged nucleic acid template is attached to a unique Primer ID; (c) amplifying the tagged nucleic acid template to produce tagged amplicons; and (d) detecting the tagged amplicons, thereby determining PCR error and/or sequencing error. In some embodiments, determining the PCR error and/or sequencing error comprises determining fidelity of a polymerase. In some embodiments, determining the PCR error and/or sequencing error comprises determining accuracy of oligonucleotides synthesized in vitro.

[0013] Further disclosed herein is a method for correcting PCR error and/or sequencing error comprising: (a) providing a sample comprising a nucleic acid template molecule; (b) attaching a primer comprising a Primer ID to each nucleic acid template molecule to be analyzed to generate a tagged nucleic acid template, wherein each tagged nucleic acid template is attached to a unique Primer ID; (c) amplifying the tagged nucleic acid template to produce tagged amplicons; and (d) detecting the tagged amplicons, thereby correcting PCR error and/or sequencing error.

[0014] In some embodiments, the Primer ID comprises a degenerate sequence. In some embodiments, the Primer ID comprises a semi-degenerate sequence. In some embodiments, the Primer ID comprises a mixed sequence. In some embodiments, the Primer ID comprises an ambiguous sequence. In some embodiments, the Primer ID comprises a wobble sequence. In some embodiments, the Primer ID comprises a random sequence. In some embodiments, the

IDT Exhibit 1020.1788

Primer ID comprises a predetermined sequence.

[0015] In some embodiments, the Primer ID is attached to the template by ligation. In some embodiments, the Primer ID is attached to the template by hybridization. In some embodiments, the Primer ID is attached to the template through PCR. In some embodiments, at least one template molecule is analyzed. In some embodiments, at least two different template molecules are analyzed. In some embodiments, detecting the tagged amplicons comprises sequencing the tagged amplicons. Sequencing of the tagged amplicons may occur by a variety of methods, including, but not limited to the Maxam-Gilbert sequencing method, the Sanger dideoxy sequencing method, dye-terminator sequencing method, pyrosequencing, multiple-primer DNA sequencing, shotgun sequencing, and primer walking. In some embodiments, sequencing comprises pyrosequencing.

[0016] In some embodiments, detecting the tagged amplicons further comprises counting a number of different Primer IDs associated with the tagged amplicons, wherein the number of different Primer IDs associated with the tagged amplicons reflects the number of templates sampled. In some embodiments, the method further comprises forming a consensus sequence for tagged amplicons comprising the same Primer ID.

[0017] In some embodiments, the nucleic acid template comprises a DNA template. In some embodiments, the nucleic acid template comprises an RNA template.

[0018] In some embodiments, amplifying comprises a PCR-based method. In some embodiments, the PCR-based method comprises PCR. In some embodiments, the PCR-based method comprises quantitative PCR. In some embodiments, the PCR-based method comprises emulsion PCR. In some embodiments, the PCR-based method comprises droplet PCR. In some embodiments, the PCR-based method comprises hot start PCR. In some embodiments, the PCR-based method comprises in situ PCR. In some embodiments, the PCR-based method comprises inverse PCR. In some embodiments, the PCR-based method comprises multiplex PCR. In some embodiments, the PCR-based method comprises Variable Number of Tandem Repeats (VNTR) PCR. In some embodiments, the PCR-based method comprises asymmetric PCR. In some embodiments, the PCR-based method comprises long PCR. In some embodiments, the PCR-based method comprises nested PCR. In some embodiments, the PCR-based method comprises hemi-nested PCR. In some embodiments, the PCR-based method comprises touchdown PCR. In some embodiments, the PCR-based method comprises assembly PCR. In some embodiments, the

IDT Exhibit 1020.1789

PCR-based method comprises colony PCR.

[0019] In some embodiments, amplifying comprises a non-PCR-based method. In some embodiments, the non-PCR-based method comprises multiple displacement amplification (MDA). In some embodiments, the non-PCR-based method comprises transcription-mediated amplification (TMA). In some embodiments, the non-PCR-based method comprises nucleic acid sequence-based amplification (NASBA). In some embodiments, the non-PCR-based method comprises strand displacement amplification (SDA). In some embodiments, the non-PCR-based method comprises real-time SDA. In some embodiments, the non-PCR-based method comprises rolling circle amplification. In some embodiments, the non-PCR-based method comprises circle-to-circle amplification.

BRIEF DESCRIPTION OF THE DRAWINGS

[0020] The skilled artisan will understand that the drawings described below are for illustration purposes only. The drawings are not intended to limit the scope of the present teachings in any way.

[0021] **Fig. 1A** shows an example primer comprising a Primer ID and Barcode.

[0022] **Fig. 1B** shows the use of a primer comprising a Primer ID and Barcode to detect and correct PCR biases and sequencing error (SEQ ID No. 1).

[0023] **Fig. 1C** shows the creation of a consensus sequence. In particular, Fig. 1A-1C show tagging viral RNA templates with a Primer ID before PCR amplification and sequencing allows for direct removal of artifactual errors and identifies resampling. Fig. 1A shows a primer that was designed to bind downstream of the protease coding domain. In the 5′ tail of the primer, a degenerate string of eight nucleotides created a Primer ID, allowing for 65,536 unique combinations. An *a priori* selected three nucleotide barcode was designed for the sample ID. Finally, a heterologous string of nucleotides with low affinity to the HIV-1 genome was included in the far 5′ end for use as the priming site in the PCR amplification. (Fig. 1B) PCR biases and sequencing error are introduced during amplification and sequencing of viral templates. Repetitive identification of the barcode and Primer ID allow for tracking of each templating event from a single tagged cDNA. As errors are minor components within the Primer ID population, forming a consensus sequence directly removes them, and corrects for PCR resampling. (Fig. 1C) HIV-1 RNA templates isolated from plasma samples from two pre- and one post-intermittent ritonavir drug therapy were tagged, amplified, and deep sequenced. Tagged

IDT Exhibit 1020.1790

sequences containing full-length protease were used to create a population of consensus sequences when at least three sequences contained an identical barcode and Primer ID.

[0024] **Fig. 2A-2B** shows the frequency of codon variation across all 99 positions in protease over three time points. Within a codon position, the first two bars represent untreated time points 1 and 2, respectively. Bars 3 and 4 are the third time point split based on the presence or absence of the resistance mutations to ritonavir. Bar 3 is the population of susceptible genotypes (defined as not V82A, I84V, or L90M), and bar 4 is the major resistant variant, V82A, population. Upward facing bars are nonsynonymous changes (scale in regular typeface), and downward facing bars are synonymous changes (scale in bolded typeface). Within a codon position, different shading represents different SNPs.

[0025] **Fig. 3** shows phylogenetic representation of protease population derived from deep sequencing with a Primer ID. A Neighbor-Joining tree was constructed from sequences derived from all three time points and colored based on susceptibility to ritonavir. V82 taxa represent susceptible variants (defined as not V82A/I/L/F, I84V, or L90M). V82A taxa represent variants containing the major ritonavir resistant variant, V82A. Other taxa represent the minor resistant variants V82I/L/F and the minor resistant alleles L90M and I84V, respectively. Within a group brightness is correlated with sample time. Dark arrows point to pre-RTV low-abundance sequences that clonally amplified to their respective clades.

[0026] **Fig. 4** shows a longitudinal sampling of blood plasma from a single individual infected with HIV-1 subtype B pre- and post- a failed ritonavir monotherapy regime. Two time-points ~6 mo apart were sampled before ritonavir therapy (T1 and T2). One time point was sampled after failed, intermittent ritonavir monotherapy (T3). The shaded areas represent times of therapy compliance based on self-report.

[0027] **Fig. 5** shows the logic flow of the bioinformatic pipeline that processed raw sequence reads into consensus sequences. First, when applicable, reads were converted to forward orientation. Next, reads were assessed for the cDNA synthesis tagging primer containing correctly spaced sample and primer identifying information (barcode and Primer ID, respectively). Sequences were then binned based on the barcode, and within each barcode, binned by Primer ID, then trimmed to just the protease coding domain. For full-length protease sequences, when at least 3 sequences within a barcode bin contained an identical Primer ID, a consensus sequence was made based on majority-rule and the use of ambiguous nucleotide

IDT Exhibit 1020.1791

designations for ties. Sequences were then further filtered based on background estimates of error for the in vitro RT cDNA synthesis and the first round of Taq DNA polymerase synthesis.

[0028] **Fig. 6A** shows the distribution of the number of reads per Primer ID or consensus sequence. Left gray bars represent the distribution of resampling of the filtered sequence population immediately before consensus sequence generation. Within a single Primer ID, when three or more sequences were present, a consensus sequence was formed. The right gray bars represent the distribution of the number of reads that went into each consensus sequence. The values shown represent the mean for the data from the three time points with the error bars representing the SD between the three samples. Starred bars are included to mark positions where a single sequence had high resampling occurrence.

[0029] **Fig. 6B** shows the number of consensus sequences containing an ambiguity as a function of extent of resampling. All three time points were combined. Dark gray bars represent consensus sequences without an ambiguity, and light gray bars represent consensus sequences with an ambiguity. There is a discernible pattern of an increased number of ambiguities going out to 22 reads/consensus sequence for those consensus sequences created from an even number of reads, the result of having a tie between two different sequences at one position. However, this represents only a small fraction of the total reads (5.4%). The amino acid position with the highest ambiguity total was used per Primer ID subpopulation.

[0030] **Fig. 7** shows an analysis of low abundance variants for the distribution of allelic skewing. We used discarded sequences (i.e., unique sequences represented by a single Primer ID) and transient genomes defined as having a low abundance SNP in the preconsensus population per untreated time point. Transient sequences were defined as having at least two sequences at only one of the untreated time points, or one copy at one of the untreated time points and then again at the third time point. These sequences were used to define a set of sequences that could be compared for low frequency abundance in the total data set versus the consensus sequences. The horizontal bars represent the measured frequency of a single copy sequences in the consensus population at T1 and T2. Dark points represent discarded genomes, and light points represent transient genomes with their position indicating their abundance in the total sequence population before construction of the consensus sequences. Light gray points represent sequences present at T1, darker gray points represent sequences present at T2. These data show that allelic skewing of 2-fold upward and 10-fold downward is common before the formation of

IDT Exhibit 1020.1792

the consensus sequence.

[0031] **Fig. 8A-8C** shows the major and minor allelic variants in the untreated populations. Fig. 8A shows the frequency of major and minor unique pro gene sequences. Gray colors represent pro gene sequences present between 2.5 and 0.5% in frequency. Black represents the sum of all pro gene sequences individually present at <0.5%. Fig. 8B shows SNP distribution of the most abundant pro gene sequences (>2.5%), with the shaded dots on the right indicating the corresponding sequences identified in the pie chart (Fig. 8A). Fig. 8C shows SNP distribution of variants present between 2.5 and 0.5%, the same sequences indicated in panel Fig. 8A with the gray bar. The line at the bottom indicated by the black circle represents the sum of all variants <0.5% in frequency for the sequences shown in black in the pie chart (Fig. 8A).

[0032] **Fig. 9A-9F** shows the major and minor unique pro gene sequences in the major resistant populations V82A, L90M, and I84V. (Fig. 9A) Frequency of different unique pro gene sequences carrying the V82A mutation at high frequency (colored >2.5%) and low frequency (<2.5%, black and with the abundance pooled). (Fig. 9B) Highlighter plot showing the sequence changes from the consensus sequence for the major (>2.5%) pro gene variants carrying the V82A mutation. The V82A substitution is indicated by the nucleotide change at position 245 shown in light gray. (Fig. 9C) Frequency of different unique pro gene sequences carrying the L90M mutation at high frequency (colored >2.5%) and low frequency (<2.5%, black and with the abundance pooled). (Fig. 9D) Highlighter plot showing the sequence changes from the consensus sequence for the major (>2.5%) pro gene variants carrying the L90M mutation. The L90M substitution is indicated by the nucleotide change at position 268 shown in gray. (Fig. 9E) Frequency of different unique pro gene sequences carrying the I84V mutation at high frequency (colored >2.5%) and low frequency (<2.5%, black and with the abundance pooled). (Fig. 9F) Highlighter plot showing the sequence changes from the consensus sequence for the major (>2.5%) pro gene variants carrying the I84V mutation. The I84V substitution is indicated by the nucleotide change at position 250 shown in gray.

[0033] **Fig. 10** shows the frequency of appearance of individual nucleotides at each Primer ID position (labeled 1–8) in resolved consensus sequences. Shading represents Da (dark gray), dT (black), dC (white), and dG (light gray), respectively. On the horizontal axis, each Primer ID position is subdivided by time point (T1, T2, and T3).

[0034] **Fig. 11** shows the frequency of deletions in total versus consensus sequences. The

IDT Exhibit 1020.1793

percentage and nucleotide position of single nucleotide deletions are depicted in total (upward facing bars) or consensus (downward facing bars) sequences. Shade corresponds to time point for T1, T2, and T3.

[0035] **Fig. 12A** shows a schematic of a tagging primer. Degenerate sequence tags individual vRNA templates with a unique ID. An *a priori* selected barcode serves as a sample ID. Together, individual samples and templates can be traced post-PCR amplification and sequencing. **Fig. 12B** shows patient sampling and clinical history. HIV-1 RNA was extracted from blood plasma of a single infected individual. Protease from two pre-Ritonavir therapy and one post-intermittent Ritonavir therapy was tagged and sequenced.

[0036] **Fig. 13A** shows the allele frequencies per amino acid position. Upward facing bars are coding changes, downward facing bars are silent changes. Change in color within an amino acid position correlates to change in codon. 1= T1, 2 = T2, 3S = T3 V82, T3R = T3V92A. **Fig. 13B** shows the summary of resolved sequences. Total sequences are the number of sequences containing full-length protease with the tagging primer. Within a sample, when three or more sequences contained an identical Primer ID, a consensus sequence was generated.

[0037] **Fig. 14** shows a bioinformatic pipeline. Raw sequence reads were screened for uncorrupted tagging primers and full length protease. A minimum of three identical sample and Primer IDs were used to create individual consensus sequences.

[0038] **Fig. 15** shows consensus sequence generation. Polymerase misincorporation, artificial recombination, differential amplification, and sequencing errors introduce sequence diversity and skew allelic frequencies. Creating a consensus sequence directly corrects for this (SEQ ID No. 1).

[0039] **Fig. 16** shows phylogenetic tree and highlighter plot demonstrate emergence of V82A strain. A Neighbor-Joining tree was constructed from consensus sequences for all three time points. Clades were colored based on V82 (lighter gray) or V82A (darker gray) amino acid. The highlighter plot represents SNPs (defined from the T1 and T2 consensus sequence) in the V82A population. Denoted next to each sequence is frequency in the population and the number of sequences in the construct. A; gray, T; black, G; darker gray, C, lighter gray.

[0040] **Table 1.** Frequency of nonconsensus codons per position

[0041] **Table 2.** Summary of nucleotide variation in sampled time points

IDT Exhibit 1020.1794

DETAILED DESCRIPTION OF THE INVENTION

[0042] Disclosed herein are methods, kits, and systems for analyzing one or more nucleic acid molecules in a sample. Generally, the method comprises (a) attaching a Primer ID to a nucleic acid molecule or fragment thereof to produce a tagged nucleic acid; and (b) detecting the tagged nucleic acid molecule or a derivative or product thereof.

[0043] Attachment of the Primer ID to the nucleic acid molecule may occur by any method known in the art. For example, attachment of the Primer ID may comprise ligation. Ligation may comprise blunt end ligation. Alternatively, ligation comprises sticky end ligation. Alternatively, or additionally, attachment of the Primer ID may comprise primer extension. Attachment of the Primer ID to the nucleic acid molecule may comprise transcription or reverse transcription. Attachment of the Primer ID to the nucleic acid molecule may comprise one or more end-repair techniques of the ends of the nucleic acid molecule.

[0044] The method may further comprise, prior to the detecting step, amplifying the tagged nucleic acid molecule to produce one or more tagged amplicons, wherein detecting comprises detecting the tagged amplicons. Amplification of the tagged amplicons may comprise any method known in the art. For example, amplification may comprise a PCR based amplification method. Alternatively, or additionally, amplification may comprise a non-PCR based amplification method.

[0045] The method may further comprise, prior to the attaching step, fragmenting a nucleic acid molecule to produce nucleic acid fragments, wherein the Primer IDs are attached to the nucleic acid fragments. Fragmenting the nucleic acid molecule may occur by any method known in the art. For example, fragmenting the nucleic acid molecule may comprise shearing. Shearing may comprise mechanical shearing. Fragmenting may comprise sonicating the sample. Alternatively, fragmenting may comprise one or more restriction enzymes. The one or more restriction enzymes may be a restriction endonuclease.

[0046] Detection of the tagged nucleic acid molecule may comprise any method known in the art. Detection of the tagged nucleic acid molecule may comprise hybridization, sequencing, capture of the tagged nucleic acid molecule, electrophoresis, luminescence, chemiluminescence, or any combination thereof. Detection may comprise detection of the Primer ID portion of the tagged nucleic acid molecule. The Primer ID may comprise a detectable label (e.g., fluorophore, dye, bead, antigen, antibody, peptide, etc). Detection of the tagged nucleic acid molecule may

IDT Exhibit 1020.1795

comprise hybridization of the tagged nucleic acid molecule to a solid support (e.g., array, bead, plate).

[0047] The method may further comprise detecting one or more genetic variants based on the detection of the tagged nucleic acid molecule. For example, genetic variants may be detected by sequencing the tagged nucleic acid molecule. Sequences with the same Primer ID can be grouped together to form a Primer ID family. A genetic variant can be detected when at least 50% of the nucleic acid molecules in the Primer ID family contain the same nucleotide sequence variation. When less than 105 of the nucleic acid molecules in the Primer ID family contain the same nucleotide sequence variation, then the nucleotide sequence variation can be due to sequencing and/or amplification error.

[0048] The method may further comprise determining the amplification bias of an amplification reaction based on the detection of the tagged nucleic acid molecules. Amplification bias or PCR resampling can be used interchangeable and may refer to the unequal amplification of nucleic acid templates. Amplification bias may result in a skewing of the distribution of PCR products (e.g., amplicons). Amplification bias may be due to differences in the amplification efficiency of two or more nucleic acid templates. Alternatively, or additionally, amplification bias may be due to inhibition of amplification of a nucleic acid template. Determining the amplification bias may be based on comparison of two or more ratios, wherein comparison of the two or more ratios comprises comparing a first ratio of the quantification of different Primer IDs associated with two or more types of nucleic acid molecules to a second ratio of the quantification of the total number of amplicons of two or more types of nucleic acid molecules. The first ratio may be based on the amount of different Primer IDs associated with a first type of nucleic acid molecule and the amount of different Primer IDs associated with a second type of nucleic acid molecule. The second ratio may be based on the number of total amplicons associated with the first type of nucleic acid molecule and the number of total amplicons associated with the second type of nucleic acid molecule. In some instances, the difference in the first ratio and the second ratio can reveal amplification bias.

[0049] The method may further comprise determining the amplification efficiency of a nucleic acid molecule based on detection of the tagged nucleic acid molecule. Determining the amplification efficiency may comprise quantifying the number of different Primer IDs associated with the nucleic acid molecule. The method may further comprise comparing the number

IDT Exhibit 1020.1796

different Primer IDs associated with the nucleic acid molecule to the number of different Primer IDs associated with a nucleic acid control.

[0050] Reference will now be made in detail to exemplary embodiments of the invention. While the invention will be described in conjunction with the exemplary embodiments, it will be understood that they are not intended to limit the invention to these embodiments. On the contrary, the invention is intended to cover alternatives, modifications and equivalents, which may be included within the spirit and scope of the invention.

[0051] The invention has many preferred embodiments and relies on many patents, applications and other references for details known to those of the art. Therefore, when a patent, application, or other reference, such as a printed publication, is cited or repeated below, it should be understood that it is incorporated by reference in its entirety for all purposes and particularly for the proposition that is recited.

[0052] As used in this application, the singular form "a," "an," and "the" include plural references unless the context clearly dictates otherwise. For example, the term "an agent" includes a plurality of agents, including mixtures thereof.

[0053] An individual is not limited to a human being, but may also be other organisms including, but not limited to, mammals, plants, bacteria, cells derived from any of the above, viruses or virally-infected cells.

[0054] Throughout this disclosure, various aspects of this invention can be presented in a range format. It should be understood that the description in range format is merely for convenience and brevity and should not be construed as an inflexible limitation on the scope of the invention. Accordingly, the description of a range should be considered to have specifically disclosed all the possible subranges as well as individual numerical values within that range. For example, description of a range such as from 1 to 6 should be considered to have specifically disclosed subranges such as from 1 to 3, from 1 to 4, from 1 to 5, from 2 to 4, from 2 to 6, from 3 to 6 etc., as well as individual numbers within that range, for example, 1, 2, 3, 4, 5, and 6. This applies regardless of the breadth of the range.

[0055] Suitable samples for analysis may be derived from a variety of sources. Biological samples may be of any biological tissue or fluid or cells from any organism. Frequently the sample will be a "clinical sample" which is a sample derived from a patient. Clinical samples provide a rich source of information regarding the various states of gene expression and copy

IDT Exhibit 1020.1797

number. Typical clinical samples include, but are not limited to, sputum, blood, tissue or fine needle biopsy samples, urine, peritoneal fluid, and pleural fluid, or cells therefrom. Biological samples may also include sections of tissues, such as frozen sections or formalin-fixed sections taken for histological purposes, which may include formalin-fixed, paraffin embedded (FFPE) samples and samples derived therefrom. FFPE samples are a particularly important source for study of archived tissue as they can nucleic acids can be recovered from these samples even after long term storage of the samples at room temperature. See, for example, Specht *et al. Am J. Path.* (2001), 158(2):419-429. Nucleic acids isolated from fresh-frozen samples may also be analyzed using the disclosed methods.

[0056] The practice of the present invention may employ, unless otherwise indicated, conventional techniques and descriptions of organic chemistry, polymer technology, molecular biology (including recombinant techniques), cell biology, biochemistry, and immunology, which are within the skill of the art. Such conventional techniques include polymer array synthesis, hybridization, ligation, and detection of hybridization using a label. Specific illustrations of suitable techniques can be had by reference to the example herein below. However, other equivalent conventional procedures can, of course, also be used. Such conventional techniques and descriptions can be found in standard laboratory manuals such as *Genome Analysis: A Laboratory Manual Series (Vols. I-IV), Using Antibodies: A Laboratory Manual, Cells: A Laboratory Manual, PCR Primer: A Laboratory Manual*, and Molecular *Cloning: A Laboratory Manual* (all from Cold Spring Harbor Laboratory Press), *Gait, "Oligonucleotide Synthesis: A Practical Approach" 1984, IRL Press, London*, Nelson and Cox (2000), Lehninger *et al.*,(2008) *Principles of Biochemistry* 5th Ed., W.H. Freeman Pub., New York, NY and Berg *et al.* (2006) *Biochemistry*, 6th Ed., W.H. Freeman Pub., New York, NY, all of which are herein incorporated in their entirety by reference for all purposes.

[0057] The present invention can employ solid substrates, including arrays in some preferred embodiments. Methods and techniques applicable to polymer (including protein) array synthesis have been described in U.S. Patent Pub. No. 20050074787, WO 00/58516, U.S. Patent Nos. 5,143,854, 5,242,974, 5,252,743, 5,324,633, 5,384,261, 5,405,783, 5,424,186, 5,451,683, 5,482,867, 5,491,074, 5,527,681, 5,550,215, 5,571,639, 5,578,832, 5,593,839, 5,599,695, 5,624,711, 5,631,734, 5,795,716, 5,831,070, 5,837,832, 5,856,101, 5,858,659, 5,936,324, 5,968,740, 5,974,164, 5,981,185, 5,981,956, 6,025,601, 6,033,860, 6,040,193, 6,090,555,

IDT Exhibit 1020.1798

6,136,269, 6,269,846 and 6,428,752, in PCT Publication No. WO 99/36760 and WO 01/58593, which are all incorporated herein by reference in their entirety for all purposes. Patents that describe synthesis techniques in specific embodiments include U.S. Patent Nos. 5,412,087, 6,147,205, 6,262,216, 6,310,189, 5,889,165, and 5,959,098. Nucleic acid arrays are described in many of the above patents, but many of the same techniques may be applied to polypeptide arrays.

[0058] The present invention also contemplates many uses for polymers attached to solid substrates. These uses include gene expression monitoring, transcript profiling, library screening, genotyping, epigenetic analysis, methylation pattern analysis, tumor typing, pharmacogenomics, agrigenetics, pathogen profiling and detection and diagnostics. Gene expression monitoring and profiling methods have been shown in U.S. Patent Nos. 5,800,992, 6,013,449, 6,020,135, 6,033,860, 6,040,138, 6,177,248 and 6,309,822. Genotyping and uses therefore are shown in U.S. Patent Publication Nos. 20030036069 and 20070065816 and U.S. Patent Nos. 5,856,092, 6,300,063, 5,858,659, 6,284,460, 6,361,947, 6,368,799 and 6,333,179. Other uses are embodied in U.S. Patent Nos. 5,871,928, 5,902,723, 6,045,996, 5,541,061, and 6,197,506.

[0059] The present invention also contemplates sample preparation methods in certain embodiments. Prior to or concurrent with analysis, the sample may be amplified by a variety of mechanisms. In some aspects nucleic acid amplification methods such as PCR may be combined with the disclosed methods and systems. *See*, for example, *PCR Technology: Principles and Applications for DNA Amplification* (Ed. H.A. Erlich, Freeman Press, NY, NY, 1992); *PCR Protocols: A Guide to Methods and Applications* (Eds. Innis, et al., Academic Press, San Diego, CA, 1990); Mattila et al., *Nucleic Acids Res.* 19, 4967 (1991); Eckert et al., *PCR Methods and Applications* 1, 17 (1991); *PCR* (Eds. McPherson et al., IRL Press, Oxford); and U.S. Patent Nos. 4,683,202, 4,683,195, 4,800,159, 4,965,188, and 5,333,675, each of which is incorporated herein by reference in their entireties for all purposes. Additional methods of sample preparation and techniques for reducing the complexity of a nucleic sample are described in Dong et al., *Genome Research* 11, 1418 (2001), in U.S. Patent Nos. 6,300,070 (amplification on an array), 6,361,947, 6,391,592, 6,872,529 and 6,458,530 and U.S. Patent Pub. Nos. 20030096235, 20030082543, 20030039069, 20050079536, 20040072217, 20050142577, 20050233354, 20050227244, 20050208555, 20050074799, 20050042654, and 20040067493, which are each incorporated herein by reference in their entireties.

IDT Exhibit 1020.1799

[0060] Many of the methods and systems disclosed herein utilize enzyme activities. Enzymes and related methods of use in molecular biology that may be used in combination with the disclosed methods and systems are reviewed, for example, in Rittie and Perbal, *J. Cell Commun. Signal.* (2008) 2:25-45, incorporated herein by reference in its entirety. A variety of enzymes are well known, have been characterized and many are commercially available from one or more supplier. Exemplary enzymes include DNA dependent DNA polymerases (such as those shown in Table 1 of Rittie and Perbal), RNA dependent DNA polymerase (see Table 2 of Rittie and Perbal), RNA polymerases (such as T7 and SP6), ligases (see Table 3 of Rittie and Perbal), enzymes for phosphate transfer and removal (see Table 4 of Rittie and Perbal), nucleases (see Table 5 of Rittie and Perbal), and methylases.

[0061] Other methods of genome analysis and complexity reduction include, for example, AFLP, *see* US Patent 6,045,994, which is incorporated herein by reference, and arbitrarily primed-PCR (AP-PCR) *see* McClelland and Welsh, in *PCR Primer: A laboratory Manual*, (1995) eds. C. Dieffenbach and G. Dveksler, Cold Spring Harbor Lab Press, for example, at p 203, which is incorporated herein by reference in its entirety.

[0062] Other suitable amplification methods include the ligase chain reaction (LCR) *(for example*, Wu and Wallace, *Genomics* 4, 560 (1989), Landegren et al., *Science* 241, 1077 (1988) and Barringer et al. *Gene* 89:117 (1990)), transcription amplification (Kwoh et al., *Proc. Natl. Acad. Sci. USA* 86, 1173 (1989) and W088/10315), self-sustained sequence replication (Guatelli et al., *Proc. Nat. Acad. Sci. USA*, 87, 1874 (1990) and W090/06995), selective amplification of target polynucleotide sequences (U.S. Patent No. 6,410,276), consensus sequence primed polymerase chain reaction (CP-PCR) (U.S. Patent No. 4,437,975), arbitrarily primed polymerase chain reaction (AP-PCR) (U.S. Patent Nos. 5,413,909, 5,861,245), rolling circle amplification (RCA) (for example, Fire and Xu, PNAS 92:4641 (1995) and Liu et al., J. Am. Chem. Soc. 118:1587 (1996)) and U.S. Pat. No. 5,648,245, strand displacement amplification (see Lasken and Egholm, *Trends Biotechnol.* 2003 21(12):531-5; Barker et al. *Genome Res.* 2004 May;14(5):901-7; Dean *et al. Proc Natl Acad Sci* U S A. 2002; 99(8):5261-6; Walker et al. 1992, Nucleic Acids Res. 20(7):1691-6, 1992 and Paez, *et al. Nucleic Acids Res.* 2004; 32(9):e71), Qbeta Replicase, described in PCT Patent Application No. PCT/US87/00880 and nucleic acid based sequence amplification (NABSA). (*See*, U.S. Patent Nos. 5,409,818, 5,554,517, and 6,063,603, each of which is incorporated herein by reference), Other amplification methods that

may be used are described in, U.S. Patent Nos. 6,582,938, 5,242,794, 5,494,810, 4,988,617, and US Pub. No. 20030143599 each of which is incorporated herein by reference. DNA may also be amplified by multiplex locus-specific PCR or using Primer ID-ligation and single primer PCR (See Kinzler and Vogelstein, NAR (1989) 17:3645-53. Other available methods of amplification, such as balanced PCR (Makrigiorgos, et al. (2002), *Nat Biotechnol*, Vol. 20, pp.936-9), may also be used.

[0063] Molecular inversion probes ("MIPs") may also be used for amplification of selected targets. MIPs may be generated so that the ends of the pre-circle probe are complementary to regions that flank the region to be amplified. The gap can be closed by extension of the end of the probe so that the complement of the target is incorporated into the MIP prior to ligation of the ends to form a closed circle. The closed circle can be amplified and detected by sequencing or hybridization as previously disclosed in Hardenbol *et al., Genome Res.* 15:269-275 (2005) and in U.S. Patent No. 6,858,412.

[0064] Methods of ligation will be known to those of skill in the art and are described, for example in Sambrook et al. (2001) and the New England BioLabs catalog both of which are incorporated herein by reference for all purposes. Methods include using T4 DNA Ligase which catalyzes the formation of a phosphodiester bond between juxtaposed 5' phosphate and 3' hydroxyl termini in duplex DNA or RNA with blunt and sticky ends; Taq DNA Ligase which catalyzes the formation of a phosphodiester bond between juxtaposed 5' phosphate and 3' hydroxyl termini of two adjacent oligonucleotides which are hybridized to a complementary target DNA; E. coli DNA ligase which catalyzes the formation of a phosphodiester bond between juxtaposed 5'-phosphate and 3'-hydroxyl termini in duplex DNA containing cohesive ends; and T4 RNA ligase which catalyzes ligation of a 5' phosphoryl-terminated nucleic acid donor to a 3' hydroxyl-terminated nucleic acid acceptor through the formation of a 3'→5' phosphodiester bond, substrates include single-stranded RNA and DNA as well as dinucleoside pyrophosphates; or any other methods described in the art. Fragmented DNA may be treated with one or more enzymes, for example, an endonuclease, prior to ligation of Primer IDs to one or both ends to facilitate ligation by generating ends that are compatible with ligation.

[0065] Methods for ligating primers comprising Primer IDs to fragments of nucleic acid are well known. Primers may be double-stranded, single-stranded or partially single-stranded. In some aspects, primers are formed from two oligonucleotides that have a region of

IDT Exhibit 1020.1801

complementarity, for example, about 10 to 30, or about 15 to 40 bases of perfect complementarity, so that when the two oligonucleotides are hybridized together they form a double stranded region. Optionally, either or both of the oligonucleotides may have a region that is not complementary to the other oligonucleotide and forms a single stranded overhang at one or both ends of the primer. Single-stranded overhangs may preferably by about 1 to about 8 bases, and most preferably about 2 to about 4. The overhang may be complementary to the overhang created by cleavage with a restriction enzyme to facilitate "sticky-end" ligation. Primers may include other features, such as primer binding sites and restriction sites. In some aspects the restriction site may be for a Type IIS restriction enzyme or another enzyme that cuts outside of its recognition sequence, such as EcoP151 (see, Mucke *et al. J Mol Biol* 2001, 312(4):687-698 and US 5,710,000 which is incorporated herein by reference in its entirety).

[0066] Methods for using mapping arrays see, for example, Applications of microarrays for SNP genotyping have been described in e.g., U.S. Patent Nos 6,300,063, 6,361,947, 6,368,799 and US Patent Publication Nos. 20040067493, 20030232353, 20030186279, 20050260628, 20070065816 and 20030186280, and Kennedy *et al., Nat. Biotech.* 21:1233-1237 (2003), Matsuzaki *et al., Genome Res.* 14:414-425 (2004), Matsuzaki *et al., Nat. Meth.* 1:109-111 (2004) and U.S. Patent Pub. Nos. 20040146890 and 20050042654, each incorporated herein by reference. Fixed content mapping arrays are available from Affymetrix, for example, the SNP 6.0 array and the AXIOM® array system. Selected panels of SNPs and markers (e.g. copy number markers) can also be interrogated using a panel of locus specific probes in combination with a universal array as described in Hardenbol *et al., Genome Res.* 15:269-275 (2005) and in U.S. Patent No. 6,858,412. Universal tag arrays and reagent kits for performing such locus specific genotyping using panels of custom molecular inversion probes (MIPs) are available from Affymetrix.

[0067] Methods for analyzing chromosomal copy number using mapping arrays are disclosed, for example, in Bignell et al., *Genome Res.* 14:287-95 (2004), Lieberfarb, *et al., Cancer Res.* 63:4781-4785 (2003), Zhao *et al., Cancer Res.* 64:3060-71 (2004), Huang *et al., Hum Genomics* 1:287-299 (2004), Nannya *et al., Cancer Res.* 65:6071-6079 (2005), Slater *et al., Am. J. Hum. Genet.* 77:709-726 (2005) and Ishikawa *et al., Biochem. and Biophys. Res. Comm.*, 333:1309-1314 (2005). Computer implemented methods for estimation of copy number based on hybridization intensity are disclosed in U.S. Patent Pub. Nos. 20040157243, 20050064476,

IDT Exhibit 1020.1802

20050130217, 20060035258, 20060134674 and 20060194243.

[0068] Methods for conducting polynucleotide hybridization assays have been well developed in the art. Hybridization assay procedures and conditions will vary depending on the application and are selected in accordance with known general binding methods, including those referred to in: Maniatis et al. *Molecular Cloning: A Laboratory Manual* (2nd Ed. Cold Spring Harbor, N.Y, 1989); Berger and Kimmel *Methods in Enzymology*, Vol. 152, *Guide to Molecular Cloning Techniques* (Academic Press, Inc., San Diego, CA, 1987); Young and Davis, *P.N.A.S*, 80: 1194 (1983). Methods and apparatus for carrying out repeated and controlled hybridization reactions have been described in U.S. Patent Nos. 5,871,928, 5,874,219, 6,045,996 and 6,386,749, 6,391,623 each of which are incorporated herein by reference.

[0069] The present invention also contemplates signal detection of hybridization between ligands in certain preferred embodiments. See U.S. Patent Nos. 5,143,854, 5,578,832, 5,631,734, 5,834,758, 5,936,324, 5,981,956, 6,025,601, 6,141,096, 6,185,030, 6,201,639, 6,218,803, and 6,225,625 in U.S. Patent Pub. No. 20040012676 and in PCT Application PCT/US99/06097 (published as WO99/47964), each of which also is hereby incorporated by reference in its entirety for all purposes.

[0070] Methods and apparatus for signal detection and processing of intensity data are disclosed in, for example, U.S. Patents Nos. 5,143,854, 5,547,839, 5,578,832, 5,631,734, 5,800,992, 5,834,758, 5,856,092, 5,902,723, 5,936,324, 5,981,956, 6,025,601, 6,090,555, 6,141,096, 6,185,030, 6,201,639; 6,218,803; and 6,225,625, in U.S. Patent Pub. Nos. 20040012676 and 20050059062 and in PCT Application PCT/US99/06097 (published as WO99/47964), each of which also is hereby incorporated by reference in its entirety for all purposes.

[0071] The practice of the present invention may also employ conventional biology methods, software and systems. Computer software products of the invention typically include computer readable medium having computer-executable instructions for performing the logic steps of the method of the invention. Suitable computer readable medium include floppy disk, CD-ROM/DVD/DVD-ROM, hard-disk drive, flash memory, ROM/RAM, magnetic tapes, etc. The computer-executable instructions may be written in a suitable computer language or combination of several languages. Basic computational biology methods are described in, for example, Setubal and Meidanis et al., *Introduction to Computational Biology Methods* (PWS Publishing

IDT Exhibit 1020.1803

Company, Boston, 1997); Salzberg, Searles, Kasif, (Ed.), *Computational Methods in Molecular Biology*, (Elsevier, Amsterdam, 1998); Rashidi and Buehler, *Bioinformatics Basics: Application in Biological Science and Medicine* (CRC Press, London, 2000) and Ouelette and Bzevanis *Bioinformatics: A Practical Guide for Analysis of Gene and Proteins* (Wiley & Sons, Inc., 2nd ed., 2001). See also US 6,420,108.

[0072] The present invention may also make use of various computer program products and software for a variety of purposes, such as probe design, management of data, analysis, and instrument operation. See, U.S. Patent Nos. 5,593,839, 5,795,716, 5,733,729, 5,974,164, 6,066,454, 6,090,555, 6,185,561, 6,188,783, 6,223,127, 6,229,911 and 6,308,170. Computer methods related to genotyping using high density microarray analysis may also be used in the present methods, see, for example, US Patent Pub. Nos. 20050250151, 20050244883, 20050108197, 20050079536 and 20050042654. Additionally, the present disclosure may have preferred embodiments that include methods for providing genetic information over networks such as the Internet as shown in U.S. Patent Pub. Nos. 20030097222, 20020183936, 20030100995, 20030120432, 20040002818, 20040126840, and 20040049354.

[0073] An allele refers to one specific form of a genetic sequence (such as a gene) within a cell, an individual or within a population, the specific form differing from other forms of the same gene in the sequence of at least one, and frequently more than one, variant sites within the sequence of the gene. The sequences at these variant sites that differ between different alleles are termed “variances”, “polymorphisms”, or “mutations”. At each autosomal specific chromosomal location or “locus" an individual possesses two alleles, one inherited from one parent and one from the other parent, for example one from the mother and one from the father. An individual is "heterozygous" at a locus if it has two different alleles at that locus. An individual is "homozygous" at a locus if it has two identical alleles at that locus.

[0074] The term “polymorphism” as used herein refers to the occurrence of two or more genetically determined alternative sequences or alleles in a population. A polymorphic marker or site is the locus at which divergence occurs. In some instances, the polymorphic markers occur at a frequency of less than 0.5%. In some instances, the polymorphic markers occur at a frequency of less than 1%. In some instances, the polymorphic markers occur at a frequency of less than 2%. In some instances, the polymorphic markers occur at a frequency of less than 5%. In some instances, the polymorphic markers occur at a frequency of greater than 1%. In some instances,

the polymorphic markers occur at a frequency of greater than 5%. In some instances, the polymorphic markers occur at a frequency of greater than 10%. In some instances, the polymorphic markers occur at a frequency of greater than 20%. In some instances, the polymorphic markers occur at a frequency of greater than 30%. In some instances, preferred markers have at least two alleles, each occurring at frequency of greater than 1%, and more preferably greater than 10% or 20% in a selected population. In some instances, preferred polymorphic markers comprise viral or bacterial sequences and occur at a frequency of less than 5%, and more preferably, less than 1% in a selected population. A polymorphism may comprise one or more base changes, an insertion, a repeat, or a deletion of one or more bases. Copy number variants (CNVs), transversions and other rearrangements are also forms of genetic variation. Polymorphic markers include restriction fragment length polymorphisms, variable number of tandem repeats (VNTR's), hypervariable regions, minisatellites, dinucleotide repeats, trinucleotide repeats, tetranucleotide repeats, simple sequence repeats, and insertion elements such as Alu. The allelic form occurring most frequently in a selected population is sometimes referred to as the wildtype form. Diploid organisms may be homozygous or heterozygous for allelic forms. A diallelic polymorphism has two forms. A triallelic polymorphism has three forms. Single nucleotide polymorphisms (SNPs) are a form of polymorphisms. SNPs are a common type of human genetic variation and are useful in the performance of genome wide association studies (GWAS). GWAS may be used, for example for the analysis of biological pathways, see Wang and Hakonarson, Nat. Rev. Genet. 2010, 11:843-854.

[0075] The term genotyping refers to the determination of the genetic information an individual carries at one or more positions in the genome. For example, genotyping may comprise the determination of which allele or alleles an individual carries for a single SNP or the determination of which allele or alleles an individual carries for a plurality of SNPs or CNVs. A diploid individual may be homozygous for each of the two possible alleles (for example, AA or BB) or heterozygous (for example, AB). For additional information regarding genotyping and genome structure see, *Color Atlas of Genetics*, Ed. Passarge, Thieme, New York, NY (2001), which is incorporated by reference.

[0076] Normal cells that are heterozygous at one or more loci may give rise to tumor cells that are homozygous at those loci. This loss of heterozygosity (LOH) may result from structural deletion of normal genes or loss of the chromosome carrying the normal gene, mitotic

IDT Exhibit 1020.1805

recombination between normal and mutant genes, followed by formation of daughter cells homozygous for deleted or inactivated (mutant) genes; or loss of the chromosome with the normal gene and duplication of the chromosome with the deleted or inactivated (mutant) gene. LOH may be copy neutral or may result from a deletion or amplification.

[0077] The term "array" as used herein refers to an intentionally created collection of molecules which can be prepared either synthetically or biosynthetically. The molecules in the array can be identical or different from each other. The array can assume a variety of formats, *for example*, libraries of soluble molecules; libraries of compounds tethered to resin beads, silica chips, microparticles, nanoparticles or other solid supports.

[0078] The term "complementary" as used herein refers to the hybridization or base pairing between nucleotides or nucleic acids, such as, for instance, between the two strands of a double stranded DNA molecule or between an oligonucleotide primer and a primer binding site on a single stranded nucleic acid to be sequenced or amplified. See, M. Kanehisa Nucleic Acids Res. 12:203 (1984), incorporated herein by reference.

[0079] The term "copy number variation" or "CNV" refers to differences in the copy number of genetic information. In many aspects it refers to differences in the per genome copy number of a genomic region. For example, in a diploid organism the expected copy number for autosomal genomic regions is 2 copies per genome. Such genomic regions should be present at 2 copies per cell. For a recent review see Zhang *et al. Annu. Rev. Genomics Hum. Genet.* 2009. 10:451-81. CNV is a source of genetic diversity in humans and can be associated with complex disorders and disease, for example, by altering gene dosage, gene disruption, or gene fusion. They can also represent benign polymorphic variants. CNVs can be large, for example, larger than 1 Mb, but many are smaller, for example between 100 bp and 1 Mb. More than 38,000 CNVs greater than 100 bp (and less than 3 Mb) have been reported in humans. Along with SNPs these CNVs account for a significant amount of phenotypic variation between individuals. In addition to having deleterious impacts, e.g. causing disease, they may also result in advantageous variation.

[0080] Digital PCR is a technique where a limiting dilution of the sample is made across a large number of separate PCR reactions so that most of the reactions have no template molecules and give a negative amplification result. Those reactions that are positive at the reaction endpoint are counted as individual template molecules present in the original sample in a 1 to 1 relationship. See Kalina et al. NAR 25:1999-2004 (1997) and Vogelstein and Kinzler, PNAS 96:9236-9241

IDT Exhibit 1020.1806

(1999). This method is an absolute counting method where solutions are partitioned into containers until there is an average probability of one molecule per two containers or when, $P_0 = (I\text{-}e^{-n/c}) = ½$; where *n* is the number of molecules and *c* is the number of containers, or *n/c* is 0.693. Quantitative partitioning is assumed, and the dynamic range is governed by the number of containers available for stochastic separation. The molecules are then detected by PCR and the number of positive containers is counted. Each successful amplification is counted as one molecule, independent of the actual amount of product. PCR-based techniques have the additional advantage of only counting molecules that can be amplified, *e.g.* that are relevant to the massively parallel PCR step in the sequencing workflow. Because digital PCR has single molecule sensitivity, only a few hundred library molecules are required for accurate quantification. Elimination of the quantification bottleneck reduces the sample input requirement from micrograms to nanograms or less, opening the way for minute and/or precious samples onto the next-generation sequencing platforms without the distorting effects of pre-amplification. Digital PCR has been used to quantify sequencing libraries to eliminate uncertainty associated with the construction and application of standard curves to PCR-based quantification and enable direct sequencing without titration runs. See White *et al. BMC Genomics* 10: 116 (2009). To vary dynamic range, micro-fabrication can be used to substantially increase the number of containers. See, Fan et al. *Am J Obstet Gynecol* 200, 543 el (May, 2009).

[0081] Similarly, in stochastic labeling, the same statistical conditions are met when $P_0 = (1\text{-}e^{-n/m}) = ½$; where *m* is the number of Primer IDs, and one half of the Primer IDs will be used at least once when $n/m = 0.693$. The dynamic range is governed by the number of Primer IDs used, and the number of Primer IDs can be easily increased to extend the dynamic range. The number of containers in digital PCR plays the same role as the number of Primer IDs in stochastic labeling and by substituting containers for Primer IDs identical statistical equations may be applied. Using the principles of physical separation, digital PCR stochastically expands identical molecules into *physical space*, whereas the principle governing stochastic labeling is identity based and expands identical molecules into *identity space*. See PCT Application PCT/US11/65291, which also is hereby incorporated by reference in its entirety for all purposes

[0082] The term "hybridization" as used herein refers to the process in which two single-stranded polynucleotides bind non-covalently to form a stable double-stranded polynucleotide; triple-stranded hybridization is also theoretically possible. The resulting (usually) double-

IDT Exhibit 1020.1807

stranded polynucleotide is a "hybrid." The proportion of the population of polynucleotides that forms stable hybrids is referred to herein as the "degree of hybridization." Hybridizations may be performed under stringent conditions, for example, at a salt concentration of no more than 1 M and a temperature of at least 25°C. For example, conditions of 5X SSPE (750 mM NaCl, 50 mM Na Phosphate, 5 mM EDTA, pH 7.4) and a temperature of 25-30°C are suitable for allele-specific probe hybridizations. For stringent conditions, see, for example, Sambrook, Fritsche and Maniatis. "Molecular Cloning A laboratory Manual" 2nd Ed. Cold Spring Harbor Press (1989) which is hereby incorporated by reference in its entirety for all purposes above. In some aspects salt concentrations for hybridization are preferably between about 200 mM and about 1M or between about 200 mM and about 500 mM. Hybridization temperatures can be as low as 5°C , but are typically greater than 22°C, more typically greater than about 30°C, and preferably in excess of about 37°C. Longer fragments may require higher hybridization temperatures for specific hybridization. As other factors may affect the stringency of hybridization, including base composition and length of the complementary strands, presence of organic solvents and extent of base mismatching, the combination of parameters is more important than the absolute measure of any one alone.

[0083] The term "mRNA" or sometimes refer by "mRNA transcripts" as used herein, include, but not limited to pre-mRNA transcript(s), transcript processing intermediates, mature mRNA(s) ready for translation and transcripts of the gene or genes, or nucleic acids derived from the mRNA transcript(s). Transcript processing may include splicing, editing and degradation. As used herein, a nucleic acid derived from an mRNA transcript refers to a nucleic acid for whose synthesis the mRNA transcript or a subsequence thereof has ultimately served as a template. Thus, a cDNA reverse transcribed from an mRNA, an RNA transcribed from that cDNA, a DNA amplified from the cDNA, an RNA transcribed from the amplified DNA, *etc.*, are all derived from the mRNA transcript and detection of such derived products is indicative of the presence and/or abundance of the original transcript in a sample. Thus, mRNA derived samples include, but are not limited to, mRNA transcripts of the gene or genes, cDNA reverse transcribed from the mRNA, cRNA transcribed from the cDNA, DNA amplified from the genes, RNA transcribed from amplified DNA, and the like.

[0084] Other classes of RNAs are also expressed, including, for example, ribosomal RNA, snRNA, miRNA and siRNA. Recent evidence suggests that the human transcriptome contains

IDT Exhibit 1020.1808

many functional RNA transcripts that are not translated into proteins. These non-coding RNAs have been recognized as important to a more complete understanding of biology. Mature miRNAs are relatively small (21-23 nucleotides) RNA duplexes that act as translational repressors of protein expression. The guide strand of a miRNA interacts with proteins to form RNA-Induced Silencing Complexes (RISC) in the cell. These sequence-specific ribonucleoprotein complexes bind target mRNAs typically in the 3'UTR and can subsequently silence gene expression either through directed mRNA degradation or by simply sequestering the target mRNA in an ineffectual form (Lee et al., Cell (1993), 75: 843-854; Bartel, Cell (2009), 136: 215-233). It has been demonstrated that miRNA based regulation plays a significant role in routine cellular processes including metabolism (Esau et al, Cell Met. 2006, v.3, p 87-98), development (Carthew et al., Cell 2009, v.137, p. 273-282), and even apoptosis (Cheng et al, Nucl. Acids Res. 2005, v.33, p1290-1297). Further research has revealed that miRNAs play critical roles in diverse disease processes such as hepatitis C (Jopling et al., Science 2005, v.309, p. 1577-1581), diabetes (Poy et al., Nature 2004, v.432, p. 226-230), and most notably multiple cancer types (Hammond, Can. Chemo. Pharma. 2006 v.58, s63-s68; Calin et al., Cancer Res. 2006, v.66, p. 7390-7394) including leukemia (Calin et al., PNAS 2002, v.101, p. 2999-3004) and glioma (Corsten et al., Cancer Res. 2007, v.67, p. 8994-9000). Over one thousand miRNAs have now been identified in animals, but only a few individual miRNAs have been linked to specific functions. Methods of the invention disclosed herein can be used for tagging of relatively short regulatory non-coding RNAs, such as micro RNAs (miRNAs), Piwi-interacting RNAs (piRNAs), snoRNAs, snRNAs, moRNAs, PARs, sdRNAs, tel-sRNAs, crasiRNAs, and small interfering RNAs (siRNAs). Methods of the invention can also be used for tagging long non-coding RNAs (long ncRNAs), traditional non-coding tRNAs and ribosomal RNA (rRNA).

[0085] The term "nucleic acid" as used herein refers to a polymeric form of nucleotides of any length, either ribonucleotides, deoxyribonucleotides or peptide nucleic acids (PNAs), that comprise purine and pyrimidine bases, or other natural, chemically or biochemically modified, non-natural, or derivatized nucleotide bases. The backbone of the polynucleotide can comprise sugars and phosphate groups, as may typically be found in RNA or DNA, or modified or substituted sugar or phosphate groups. A polynucleotide may comprise modified nucleotides, such as methylated nucleotides and nucleotide analogs. The sequence of nucleotides may be interrupted by non-nucleotide components. Thus the terms nucleoside, nucleotide,

deoxynucleoside and deoxynucleotide generally include analogs such as those described herein. These analogs are those molecules having some structural features in common with a naturally occurring nucleoside or nucleotide such that when incorporated into a nucleic acid or oligonucleoside sequence, they allow hybridization with a naturally occurring nucleic acid sequence in solution. Typically, these analogs are derived from naturally occurring nucleosides and nucleotides by replacing and/or modifying the base, the ribose or the phosphodiester moiety. The changes can be tailor made to stabilize or destabilize hybrid formation or enhance the specificity of hybridization with a complementary nucleic acid sequence as desired.

[0086] The term "oligonucleotide" or sometimes refer by "polynucleotide" as used herein refers to a nucleic acid ranging from at least 2, preferable at least 8, and more preferably at least 20 nucleotides in length or a compound that specifically hybridizes to a polynucleotide. Polynucleotides of the present invention include sequences of deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) that may be isolated from natural sources, recombinantly produced or artificially synthesized and mimetics thereof. A further example of a polynucleotide of the present invention may include non natural analogs that may increase specificity of hybridization, for example, peptide nucleic acid (PNA) linkages and Locked Nucleic Acid (LNA) linkages. The LNA linkages are conformationally restricted nucleotide analogs that bind to complementary target with a higher melting temperature and greater mismatch discrimination. Other modifications that may be included in probes include: 2'OMe, 2'OAllyl, 2'O-propargyl, 2'O-alkyl, 2' fluoro, 2' arabino, 2' xylo, 2' fluoro arabino, phosphorothioate, phosphorodithioate, phosphoroamidates, 2'Amino, 5-alkyl-substituted pyrimidine, 5-halo-substituted pyrimidine, alkyl-substituted purine, halo-substituted purine, bicyclic nucleotides, 2'MOE, LNA-like molecules and derivatives thereof. The invention also encompasses situations in which there is a nontraditional base pairing such as Hoogsteen base pairing which has been identified in certain tRNA molecules and postulated to exist in a triple helix. "Polynucleotide" and "oligonucleotide" are used interchangeably in this application.

[0087] The term "primer" as used herein refers to a double-stranded, single-stranded, or partially single-stranded oligonucleotide. In some embodiments, primers are capable of acting as a point of initiation for template-directed nucleic acid synthesis under suitable conditions for example, buffer and temperature, in the presence of four different nucleoside triphosphates and an agent for polymerization, such as, for example, DNA or RNA polymerase or reverse

IDT Exhibit 1020.1810

transcriptase. The length of the primer, in any given case, depends on, for example, the intended use of the primer, and generally ranges from 15 to 100 nucleotides. Short primer molecules generally require cooler temperatures to form sufficiently stable hybrid complexes with the template. A primer need not reflect the exact sequence of the template but must be sufficiently complementary to hybridize with such template. The primer site is the area of the template to which a primer hybridizes. The primer pair is a set of primers including a 5' upstream primer that hybridizes with the 5' end of the sequence to be amplified and a 3' downstream primer that hybridizes with the complement of the 3' end of the sequence to be amplified. As used herein, the primer may comprise a target specific sequence and a Primer ID sequence. The primer may further comprise a barcode sequence. The barcode sequence may be used to identify the presence of a Primer ID sequence. The primer may also comprise a PCR priming sequence. The PCR priming sequence may be used to initiate amplification of a tagged nucleic acid molecule.

[0088] The term "probe" as used herein refers to a surface-immobilized molecule that can be recognized by a particular target. See U.S. Patent No. 6,582,908 for an example of arrays having all possible combinations of probes with 10, 12, and more bases. Examples of probes that can be investigated by this invention include, but are not restricted to, agonists and antagonists for cell membrane receptors, toxins and venoms, viral epitopes, hormones (for example, opioid peptides, steroids, etc.), hormone receptors, peptides, enzymes, enzyme substrates, cofactors, drugs, lectins, sugars, oligonucleotides, nucleic acids, oligosaccharides, proteins, and monoclonal antibodies.

[0089] The term "solid support", "support", and "substrate" as used herein are used interchangeably and refer to a material or group of materials having a rigid or semi-rigid surface or surfaces. In many embodiments, at least one surface of the solid support will be substantially flat, although in some embodiments it may be desirable to physically separate synthesis regions for different compounds with, for example, wells, raised regions, pins, etched trenches, or the like. According to other embodiments, the solid support(s) will take the form of beads, resins, gels, microspheres, or other geometric configurations. See U.S. Patent No. 5,744,305 and US Patent Pub. Nos. 20090149340 and 20080038559 for exemplary substrates.

[0090] The term "Primer ID" as used herein to refer to the information that is added to. Libraries of primer having a diversity of unique Primer IDs, for example about 1,000, about 5,000, about 10,000, about 100,000 or more than 100,000 may be used to uniquely identify

IDT Exhibit 1020.1811

occurrences of target species thereby marking each species with an identifier that can be used to distinguish between two otherwise identical or nearly identical targets. For example, each Primer ID may be a short string of nucleotides that can be attached to different copies of an mRNA, for example, a first Primer ID may be 5'GCATCTTC3' and a second may be 5'CAAGTAAC3'. Each has a unique identity that can be determined by determining the identity and order of the bases in the Primer ID.

[0091] Although nucleic acids are used throughout as a preferred embodiment of Primer ID, one of skill in the art will appreciate that a number of types of molecules or items that can be generated with the diversity needed may be used as Primer IDs. Primer IDs should be compounds, structures or elements that are amenable to at least one method of detection that allows for discrimination between different Primer IDs and should be associable in some means with the elements to be counted. For example, a pool of Primer IDs may be comprised of a collection of different semiconductor nanocrystals, metal compounds, peptides, antibodies, small molecules, isotopes, particles or structures having different shapes, colors, or diffraction patterns associated therewith or embedded therein, strings of numbers, random fragments of proteins or nucleic acids, or different isotopes (see, Abdelrahman, A.I. et al. *Journal of Analytical Atomic Spectrometry* 25 (3):260-268, 2010 for use of metal containing polystyrene beads as standards for mass cytometry, incorporated herein by reference). Pools of Primer IDs may be partitioned into distinct sets that can be attached to separate sample mixtures and then combined for later analysis. For example, a set of 1,000,000 different Primer IDs could be physically divided into 10 sets of 100,000 different Primer IDs and each could be used to Primer ID a different mixture. The identity of the Primer IDs in each set can be used as an indication of the original source. Counting of multiple samples in parallel can be facilitated.

[0092] In one embodiment the Primer ID may also be used in conjunction with a barcode, which may be 2-10 nucleotides, e.g., 2, 3, 4, 5, 6, 7, 8, 9, or 10 nucleotides. The barcode may be directly linked to the Primer ID or there may be an intervening sequence between the barcode and the Primer ID. The barcode may represent an analysis date, time or location; a clinical trial; a collection date, time or location; a patient number; a sample number; a species; a subspecies; a subtype; a therapeutic regimen; or a tissue type. In one non-limiting embodiment, both the Primer ID and the barcode are single-stranded. A 3 nucleotide barcode representing different study dates is exemplified herein.

IDT Exhibit 1020.1812

[0093] The term "detectable label" as used herein refers to any chemical moiety attached to a nucleotide, nucleotide polymer, or nucleic acid binding factor, wherein the attachment may be covalent or non-covalent. Preferably, the label is detectable and renders the nucleotide or nucleotide polymer detectable to the practitioner of the invention. Detectable labels that may be used in combination with the methods disclosed herein include, for example, a fluorescent label, a chemiluminescent label, a quencher, a radioactive label, biotin and gold, or combinations thereof. Detectable labels include luminescent molecules, fluorochromes, fluorescent quenching agents, colored molecules, radioisotopes or scintillants. Detectable labels also include any useful linker molecule (such as biotin, avidin, streptavidin, HRP, protein A, protein G, antibodies or fragments thereof, Grb2, polyhistidine, Ni^2+, FLAG tags, myc tags), heavy metals, enzymes (examples include alkaline phosphatase, peroxidase and luciferase), electron donors/acceptors, acridinium esters, dyes and calorimetric substrates. It is also envisioned that a change in mass may be considered a detectable label, as is the case of surface plasmon resonance detection. The skilled artisan would readily recognize useful detectable labels that are not mentioned above, which may be employed in the operation of the present invention. In some instances, detectable labels are used with primers. In some instances, detectable labels are used with the Primer IDs. In some instances, detectable labels are used with the nucleic acid template molecule. In some instances, detectable labels are used to detect the tagged amplicons. In some instances, detectable labels are used to detect the nucleic acid template molecule.

[0094] The term "consensus sequence" as used herein refers to a sequence formed from two or more sequences containing an identical Primer ID. In some instances, a consensus sequence is the most common variant of a nucleic acid molecule.

[0095] Disclosed herein is a method for determining genetic diversity of a sample comprising: (a) providing a sample comprising a nucleic acid template molecule; (b) attaching a primer comprising a Primer ID to each nucleic acid template molecule to be analyzed to generate a tagged nucleic acid template, wherein each tagged nucleic acid template is attached to a unique Primer ID; (c) amplifying the tagged nucleic acid template to produce tagged amplicons; and (d) detecting the tagged amplicons, thereby determining genetic diversity of a sample. In some embodiments, the Primer ID comprises a degenerate sequence. In some embodiments, the Primer ID comprises a semi-degenerate sequence. In some embodiments, the Primer ID comprises a mixed sequence. In some embodiments, the Primer ID comprises an ambiguous sequence. In

some embodiments, the Primer ID comprises a wobble sequence. In some embodiments, the Primer ID comprises a random sequence. In some embodiments, the Primer ID comprises a predetermined sequence. In some embodiments, the Primer ID is attached to the template by ligation. In some embodiments, the Primer ID is attached to the template by hybridization. In some embodiments, the Primer ID is attached to the template through PCR. In some embodiments, at least one template molecule is analyzed. In some embodiments, at least two different template molecules are analyzed. In some embodiments, detecting the tagged amplicons comprises sequencing the tagged amplicons. Sequencing of the tagged amplicons may occur by a variety of methods, including, but not limited to the Maxam-Gilbert sequencing method, the Sanger dideoxy sequencing method, dye-terminator sequencing method, pyrosequencing, multiple-primer DNA sequencing, shotgun sequencing, and primer walking. In some embodiments, sequencing comprises pyrosequencing. In some embodiments, detecting the tagged amplicons further comprises counting a number of different Primer IDs associated with the tagged amplicons, wherein the number of different Primer IDs associated with the tagged amplicons reflects the number of templates sampled. In some embodiments, the method further comprises forming a consensus sequence for tagged amplicons comprising the same Primer ID. In some embodiments, the nucleic acid template comprises a DNA template. In some embodiments, the nucleic acid template comprises an RNA template. In some embodiments, amplifying comprises a PCR-based method. In some embodiments, the PCR-based method comprises PCR. In some embodiments, the PCR-based method comprises quantitative PCR. In some embodiments, the PCR-based method comprises emulsion PCR. In some embodiments, the PCR-based method comprises droplet PCR. In some embodiments, the PCR-based method comprises hot start PCR. In some embodiments, the PCR-based method comprises in situ PCR. In some embodiments, the PCR-based method comprises inverse PCR. In some embodiments, the PCR-based method comprises multiplex PCR. In some embodiments, the PCR-based method comprises Variable Number of Tandem Repeats (VNTR) PCR. In some embodiments, the PCR-based method comprises asymmetric PCR. In some embodiments, the PCR-based method comprises long PCR. In some embodiments, the PCR-based method comprises nested PCR. In some embodiments, the PCR-based method comprises a hemi-nested PCR. In some embodiments, the PCR-based method comprises touchdown PCR. In some embodiments, the PCR-based method comprises assembly PCR. In some embodiments, the PCR-based method

IDT Exhibit 1020.1814

comprises colony PCR. In some embodiments, amplifying comprises a non-PCR-based method. In some embodiments, the non-PCR-based method comprises multiple displacement amplification (MDA). In some embodiments, the non-PCR-based method comprises transcription-mediated amplification (TMA). In some embodiments, the non-PCR-based method comprises nucleic acid sequence-based amplification (NASBA). In some embodiments, the non-PCR-based method comprises strand displacement amplification (SDA). In some embodiments, the non-PCR-based method comprises real-time SDA. In some embodiments, the non-PCR-based method comprises rolling circle amplification. In some embodiments, the non-PCR-based method comprises circle-to-circle amplification.

[0001] Suitable next generation sequencing technologies are widely available for use in connection with the methods described herein. Examples include the 454 Life Sciences platform (Roche, Branford, CT) (Margulies *et al.* 2005 *Nature*, 437, 376-380); lllumina's Genome Analyzer, GoldenGate Methylation Assay, or Infinium Methylation Assays, i.e., Infinium HumanMethylation 27K BeadArray or VeraCode GoldenGate methylation array (Illumina, San Diego, CA; Bibkova *et al.*, 2006, *Genome Res.* 16, 383-393; U.S. Pat. Nos. 6,306,597 and 7,598,035 (Macevicz); 7,232,656 (Balasubramanian *et al.*)); or DNA Sequencing by Ligation, SOLiD System (Applied Biosystems/Life Technologies; U.S. Pat. Nos. 6,797,470, 7,083,917, 7,166,434, 7,320,865, 7,332,285, 7,364,858, and 7,429,453 (Barany *et al.*); or the Helicos True Single Molecule DNA sequencing technology (Harris *et al.*, 2008 *Science*, 320, 106-109; U.S. Pat. Nos. 7,037,687 and 7,645,596 (Williams *et al.*); 7,169,560 (Lapidus *et al.*); 7,769,400 (Harris)), the single molecule, real-time (SMRT™) technology of Pacific Biosciences, and sequencing (Soni and Meller, 2007, *Clin. Chem.* 53, 1996-2001) which are incorporated herein by reference in their entirety. These systems allow the sequencing of many nucleic acid molecules isolated from a specimen at high orders of multiplexing in a parallel fashion (Dear, 2003, *Brief Funct. Genomic Proteomic*, 1(4), 397-416 and McCaughan and Dear, 2010, *J. Pathol.*, 220, 297-306). Each of these platforms allow sequencing of clonally expanded or non-amplified single molecules of nucleic acid fragments. Certain platforms involve, for example, (i) sequencing by ligation of dye-modified probes (including cyclic ligation and cleavage), (ii) pyrosequencing, and (iii) single-molecule sequencing.

[0002] Pyrosequencing is a nucleic acid sequencing method based on sequencing by synthesis, which relies on detection of a pyrophosphate released on nucleotide incorporation.

IDT Exhibit 1020.1815

Generally, sequencing by synthesis involves synthesizing, one nucleotide at a time, a DNA strand complimentary to the strand whose sequence is being sought. Study nucleic acids may be immobilized to a solid support, hybridized with a sequencing primer, incubated with DNA polymerase, ATP sulfurylase, luciferase, apyrase, adenosine 5' phosphsulfate and luciferin. Nucleotide solutions are sequentially added and removed. Correct incorporation of a nucleotide releases a pyrophosphate, which interacts with ATP sulfurylase and produces ATP in the presence of adenosine 5' phosphsulfate, fueling the luciferin reaction, which produces a chemiluminescent signal allowing sequence determination. Machines for pyrosequencing and methylation specific reagents are available from Qiagen, Inc. (Valencia, CA). See also Tost and Gut, 2007, *Nat. Prot.* 2 2265-2275. An example of a system that can be used by a person of ordinary skill based on pyrosequencing generally involves the following steps: ligating an adaptor nucleic acid to a study nucleic acid and hybridizing the study nucleic acid to a bead; amplifying a nucleotide sequence in the study nucleic acid in an emulsion; sorting beads using a picoliter multiwell solid support; and sequencing amplified nucleotide sequences by pyrosequencing methodology (e.g., Nakano *et al.*, 2003, *J. Biotech.* 102, 117-124). Such a system can be used to exponentially amplify amplification products generated by a process described herein, e.g., by ligating a heterologous nucleic acid to the first amplification product generated by a process described herein.

[0003] Certain single-molecule sequencing embodiments are based on the principal of sequencing by synthesis, and utilize single-pair Fluorescence Resonance Energy Transfer (single pair FRET) as a mechanism by which photons are emitted as a result of successful nucleotide incorporation. The emitted photons often are detected using intensified or high sensitivity cooled charge-couple-devices in conjunction with total internal reflection microscopy (TIRM). Photons are only emitted when the introduced reaction solution contains the correct nucleotide for incorporation into the growing nucleic acid chain that is synthesized as a result of the sequencing process. In FRET based single-molecule sequencing or detection, energy is transferred between two fluorescent dyes, sometimes polymethine cyanine dyes Cy3 and Cy5, through long-range dipole interactions. The donor is excited at its specific excitation wavelength and the excited state energy is transferred, non-radiatively to the acceptor dye, which in turn becomes excited. The acceptor dye eventually returns to the ground state by radiative emission of a photon. The two dyes used in the energy transfer process represent the "single pair", in single pair FRET. Cy3

IDT Exhibit 1020.1816

often is used as the donor fluorophore and often is incorporated as the first labeled nucleotide. Cy5 often is used as the acceptor fluorophore and is used as the nucleotide label for successive nucleotide additions after incorporation of a first Cy3 labeled nucleotide. The fluorophores generally are within 10 nanometers of each other for energy transfer to occur successfully. Bailey *et al.* recently reported a highly sensitive (15pg methylated DNA) method using quantum dots to detect methylation status using fluorescence resonance energy transfer (MS-qFRET)(Bailey *et al.* 2009, *Genome Res.* 19(8), 1455-1461, which is incorporated herein by reference in its entirety).

[0004] An example of a system that can be used based on single-molecule sequencing generally involves hybridizing a primer to a study nucleic acid to generate a complex; associating the complex with a solid phase; iteratively extending the primer by a nucleotide tagged with a fluorescent molecule; and capturing an image of fluorescence resonance energy transfer signals after each iteration (e.g., Braslavsky *et al.*, PNAS 100(7): 3960-3964 (2003); U.S. Pat. No. 7,297,518 (Quake *et al.*) which are incorporated herein by reference in their entirety). Such a system can be used to directly sequence amplification products generated by processes described herein. In some embodiments the released linear amplification product can be hybridized to a primer that contains sequences complementary to immobilized capture sequences present on a solid support, a bead or glass slide for example. Hybridization of the primer-released linear amplification product complexes with the immobilized capture sequences, immobilizes released linear amplification products to solid supports for single pair FRET based sequencing by synthesis. The primer often is fluorescent, so that an initial reference image of the surface of the slide with immobilized nucleic acids can be generated. The initial reference image is useful for determining locations at which true nucleotide incorporation is occurring. Fluorescence signals detected in array locations not initially identified in the "primer only" reference image are discarded as non-specific fluorescence. Following immobilization of the primer-released linear amplification product complexes, the bound nucleic acids often are sequenced in parallel by the iterative steps of, a) polymerase extension in the presence of one fluorescently labeled nucleotide, b) detection of fluorescence using appropriate microscopy, TIRM for example, c) removal of fluorescent nucleotide, and d) return to step a with a different fluorescently labeled nucleotide.

[0096] In some embodiments, at least 2 different nucleic acid template molecules are

IDT Exhibit 1020.1817

analyzed. In some embodiments, at least 3 different nucleic acid template molecules are analyzed. In some embodiments, at least 4 different nucleic acid template molecules are analyzed. In some embodiments, at least 5 different nucleic acid template molecules are analyzed. In some embodiments, at least 6 different nucleic acid template molecules are analyzed. In some embodiments, at least 7 different nucleic acid template molecules are analyzed. In some embodiments, at least 8 different nucleic acid template molecules are analyzed. In some embodiments, at least 9 different nucleic acid template molecules are analyzed. In some embodiments, at least 10 different nucleic acid template molecules are analyzed. In some embodiments, at least 15 different nucleic acid template molecules are analyzed. In some embodiments, at least 20 different nucleic acid template molecules are analyzed. In some embodiments, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 85, at least 90, at least 95, at least 100, at least 125, at least 150, at least 175, at least 200, at least 250, at least 300, at least 350, or at least 400 different nucleic acid template molecules are analyzed. In some embodiments, at least 500 different nucleic acid template molecules are analyzed. In some embodiments, at least 1,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 5,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 10,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 20,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 30,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 40,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 50,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 60,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 70,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 80,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 90,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 100,000 different nucleic acid template molecules are analyzed. In some embodiments, the Primer IDs are attached to the nucleic acid template molecules simultaneously. In some embodiments, the Primer IDs are attached to the nucleic acid template molecules sequentially. In some embodiments, the nucleic acid template molecules are amplified and/or detected simultaneously. In some embodiments, the nucleic acid template molecules are

amplified and/or detected sequentially. In some embodiments, the Primer ID comprises a nucleic acid sequence. In some embodiments, the Primer ID comprises a deoxyribonucleic acid sequence. In some embodiments, the Primer ID comprises a ribonucleic acid sequence. In some embodiments, the Primer ID comprises 5-100 nucleotides. In some embodiments, the Primer ID comprises 5-50 nucleotides. In some embodiments, the Primer ID comprises at least 6 nucleotides. In some embodiments, the Primer ID comprises at least 7 nucleotides. In some embodiments, the Primer ID comprises at least 8 nucleotides. In some embodiments, the Primer ID comprises at least 9 nucleotides. In some embodiments, the Primer ID comprises at least 10 nucleotides. In some embodiments, the Primer ID comprises at least 12 nucleotides. In some embodiments, the Primer ID comprises at least 15 nucleotides. In some embodiments, the Primer ID comprises at least 20 nucleotides. In some embodiments, the Primer ID comprises at least 25 nucleotides. In some embodiments, the Primer ID comprises at least 35 nucleotides.

[0097] Further provided is a method for detecting genetic variants comprising: (a) providing a sample comprising a nucleic acid template molecule; (b) attaching a primer comprising a Primer ID to each nucleic acid template molecule to be analyzed to generate a tagged nucleic acid template, wherein each tagged nucleic template is attached to a unique Primer ID; (c) amplifying the tagged nucleic acid template to produce tagged amplicons; and (d) detecting the tagged amplicons, thereby detecting genetic variants. In some embodiments, detecting the genetic variants comprises determining the prevalence of mutations. In some embodiments, detecting the genetic variants comprises forming a consensus sequence for tagged nucleic acid templates comprising the same Primer ID. In some embodiments, detecting genetic variants comprises sequencing the tagged amplicons. Sequencing of the tagged amplicons may occur by a variety of methods, including, but not limited to the Maxam-Gilbert sequencing method, the Sanger dideoxy sequencing method, dye-terminator sequencing method, pyrosequencing, multiple-primer DNA sequencing, shotgun sequencing, primer walking. In some embodiments, sequencing comprises pyrosequencing. In some embodiments, detecting genetic variants comprises counting a number of different tagged amplicons. In some embodiments, the genetic variant comprises a polymorphism. In some embodiments, the polymorphism comprises a single nucleotide polymorphism. In some instances, the polymorphism occurs at a frequency of less than 0.5%. In some instances, the polymorphism occurs at a frequency of less than 1%. In some instances, the polymorphism occurs at a frequency of less than 2%. In some instances, the

IDT Exhibit 1020.1819

polymorphism occurs at a frequency of less than 5%. In some instances, the polymorphism occurs at a frequency of greater than 1%. In some instances, the polymorphism occurs at a frequency of greater than 5%. In some instances, the polymorphism occurs at a frequency of greater than 10%. In some instances, the polymorphism occurs at a frequency of greater than 20%. In some instances, the polymorphism occurs at a frequency of greater than 30%. In some embodiments, the genetic variant comprises a mutation. In some embodiments, the genetic variant comprises a deletion. In some embodiments, the genetic variant comprises an insertion. In some embodiments, the Primer ID comprises a degenerate sequence. In some embodiments, the Primer ID comprises a semi-degenerate sequence. In some embodiments, the Primer ID comprises a mixed sequence. In some embodiments, the Primer ID comprises an ambiguous sequence. In some embodiments, the Primer ID comprises a wobble sequence. In some embodiments, the Primer ID comprises a random sequence. In some embodiments, the Primer ID comprises a predetermined sequence. In some embodiments, the Primer ID is attached to the template by ligation. In some embodiments, the Primer ID is attached to the template by hybridization. In some embodiments, the Primer ID is attached to the template through PCR. In some embodiments, at least one template molecule is analyzed. In some embodiments, at least two different template molecules are analyzed. In some embodiments, detecting the tagged amplicons further comprises counting a number of different Primer IDs associated with the tagged amplicons, wherein the number of different Primer IDs associated with the tagged amplicons reflects the number of templates sampled. In some embodiments, the method further comprises forming a consensus sequence for tagged amplicons comprising the same Primer ID. In some embodiments, amplifying comprises a PCR-based method. In some embodiments, the PCR-based method comprises PCR. In some embodiments, the PCR-based method comprises quantitative PCR. In some embodiments, the PCR-based method comprises emulsion PCR. In some embodiments, the PCR-based method comprises droplet PCR. In some embodiments, the PCR-based method comprises hot start PCR. In some embodiments, the PCR-based method comprises in situ PCR. In some embodiments, the PCR-based method comprises inverse PCR. In some embodiments, the PCR-based method comprises multiplex PCR. In some embodiments, the PCR-based method comprises Variable Number of Tandem Repeats (VNTR) PCR. In some embodiments, the PCR-based method comprises asymmetric PCR. In some embodiments, the PCR-based method comprises long PCR. In some embodiments, the PCR-based method

IDT Exhibit 1020.1820

comprises nested PCR. In some embodiments, the PCR-based method comprises hemi-nested PCR. In some embodiments, the PCR-based method comprises touchdown PCR. In some embodiments, the PCR-based method comprises assembly PCR. In some embodiments, the PCR-based method comprises colony PCR. In some embodiments, amplifying comprises a non-PCR-based method. In some embodiments, the non-PCR-based method comprises multiple displacement amplification (MDA). In some embodiments, the non-PCR-based method comprises transcription-mediated amplification (TMA). In some embodiments, the non-PCR-based method comprises nucleic acid sequence-based amplification (NASBA). In some embodiments, the non-PCR-based method comprises strand displacement amplification (SDA). In some embodiments, the non-PCR-based method comprises real-time SDA. In some embodiments, the non-PCR-based method comprises rolling circle amplification. In some embodiments, the non-PCR-based method comprises circle-to-circle amplification. In some embodiments, the nucleic acid template comprises a DNA template. In some embodiments, the nucleic acid template comprises an RNA template. In some embodiments, at least 2 different nucleic acid template molecules are analyzed. In some embodiments, at least 3 different nucleic acid template molecules are analyzed. In some embodiments, at least 4 different nucleic acid template molecules are analyzed. In some embodiments, at least 5 different nucleic acid template molecules are analyzed. In some embodiments, at least 6 different nucleic acid template molecules are analyzed. In some embodiments, at least 7 different nucleic acid template molecules are analyzed. In some embodiments, at least 8 different nucleic acid template molecules are analyzed. In some embodiments, at least 9 different nucleic acid template molecules are analyzed. In some embodiments, at least 10 different nucleic acid template molecules are analyzed. In some embodiments, at least 15 different nucleic acid template molecules are analyzed. In some embodiments, at least 20 different nucleic acid template molecules are analyzed. In some embodiments, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 85, at least 90, at least 95, at least 100, at least 125, at least 150, at least 175, at least 200, at least 250, at least 300, at least 350, or at least 400 different nucleic acid template molecules are analyzed. In some embodiments, at least 500 different nucleic acid template molecules are analyzed. In some embodiments, at least 1,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 5,000 different nucleic acid template molecules are

IDT Exhibit 1020.1821

analyzed. In some embodiments, at least 10,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 20,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 30,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 40,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 50,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 60,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 70,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 80,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 90,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 100,000 different nucleic acid template molecules are analyzed. In some embodiments, the Primer IDs are attached to the nucleic acid template molecules simultaneously. In some embodiments, the Primer IDs are attached to the nucleic acid template molecules sequentially. In some embodiments, the nucleic acid template molecules are amplified and/or detected simultaneously. In some embodiments, the nucleic acid template molecules are amplified and/or detected sequentially. In some embodiments, the Primer ID comprises 5-100 nucleotides. In some embodiments, the Primer ID comprises 5-50 nucleotides. In some embodiments, the Primer ID comprises at least 6 nucleotides. In some embodiments, the Primer ID comprises at least 7 nucleotides. In some embodiments, the Primer ID comprises at least 8 nucleotides. In some embodiments, the Primer ID comprises at least 9 nucleotides. In some embodiments, the Primer ID comprises at least 10 nucleotides. In some embodiments, the Primer ID comprises at least 12 nucleotides. In some embodiments, the Primer ID comprises at least 15 nucleotides. In some embodiments, the Primer ID comprises at least 20 nucleotides. In some embodiments, the Primer ID comprises at least 25 nucleotides. In some embodiments, the Primer ID comprises at least 35 nucleotides.

[0098] Also provided herein is a method for determining or screening for drug-resistant variants comprising: (a) providing a sample comprising a nucleic acid template molecule; (b) attaching a primer comprising a Primer ID to each nucleic acid template molecule to be analyzed to generate a tagged nucleic acid template, wherein each tagged nucleic acid template is attached to a unique Primer ID; (c) amplifying the tagged nucleic acid template to produce tagged amplicons; and (d) detecting the tagged amplicons, thereby determining or screening for drug-resistant variants. In some embodiments, detecting the tagged amplicons comprises sequencing

IDT Exhibit 1020.1822

the tagged amplicons. Sequencing of the tagged amplicons may occur by a variety of methods, including, but not limited to the Maxam-Gilbert sequencing method, the Sanger dideoxy sequencing method, dye-terminator sequencing method, pyrosequencing, multiple-primer DNA sequencing, shotgun sequencing, primer walking. In some embodiments, sequencing comprises pyrosequencing. In some embodiments, detecting the tagged amplicons further comprises forming a consensus sequence for the tagged amplicons comprising the same Primer ID. In some embodiments, the nucleic acid template molecule comprises a viral sequence. In some embodiments, the nucleic acid template molecule comprises a bacterial sequence. In some embodiments, the sample is from an individual suffering from a viral infection. In some embodiments, the sample is from an individual suffering from a bacterial infection. In some embodiments, the sample is from an individual suffering from cancer. In some embodiments, the sample is from an individual suffering from an autoimmune disorder. In some embodiments, the Primer ID comprises a degenerate sequence. In some embodiments, the Primer ID comprises a semi-degenerate sequence. In some embodiments, the Primer ID comprises a mixed sequence. In some embodiments, the Primer ID comprises an ambiguous sequence. In some embodiments, the Primer ID comprises a wobble sequence. In some embodiments, the Primer ID comprises a random sequence. In some embodiments, the Primer ID comprises a predetermined sequence. In some embodiments, the Primer ID is attached to the template by ligation. In some embodiments, the Primer ID is attached to the template by hybridization. In some embodiments, the Primer ID is attached to the template through PCR. In some embodiments, at least one template molecule is analyzed. In some embodiments, at least two different template molecules are analyzed. In some embodiments, detecting the tagged amplicons further comprises counting a number of different Primer IDs associated with the tagged amplicons, wherein the number of different Primer IDs associated with the tagged amplicons reflects the number of templates sampled. In some embodiments, amplifying comprises a PCR-based method. In some embodiments, the PCR-based method comprises PCR. In some embodiments, the PCR-based method comprises quantitative PCR. In some embodiments, the PCR-based method comprises emulsion PCR. In some embodiments, the PCR-based method comprises droplet PCR. In some embodiments, the PCR-based method comprises hot start PCR. In some embodiments, the PCR-based method comprises in situ PCR. In some embodiments, the PCR-based method comprises inverse PCR. In some embodiments, the PCR-based method comprises multiplex PCR. In some embodiments,

IDT Exhibit 1020.1823

the PCR-based method comprises Variable Number of Tandem Repeats (VNTR) PCR. In some embodiments, the PCR-based method comprises asymmetric PCR. In some embodiments, the PCR-based method comprises long PCR. In some embodiments, the PCR-based method comprises nested PCR. In some embodiments, the PCR-based method comprises hemi-nested PCR. In some embodiments, the PCR-based method comprises touchdown PCR. In some embodiments, the PCR-based method comprises assembly PCR. In some embodiments, the PCR-based method comprises colony PCR. In some embodiments, amplifying comprises a non-PCR-based method. In some embodiments, the non-PCR-based method comprises multiple displacement amplification (MDA). In some embodiments, the non-PCR-based method comprises transcription-mediated amplification (TMA). In some embodiments, the non-PCR-based method comprises nucleic acid sequence-based amplification (NASBA). In some embodiments, the non-PCR-based method comprises strand displacement amplification (SDA). In some embodiments, the non-PCR-based method comprises real-time SDA. In some embodiments, the non-PCR-based method comprises rolling circle amplification. In some embodiments, the non-PCR-based method comprises circle-to-circle amplification. In some embodiments, at least 2 different nucleic acid template molecules are analyzed. In some embodiments, at least 3 different nucleic acid template molecules are analyzed. In some embodiments, at least 4 different nucleic acid template molecules are analyzed. In some embodiments, at least 5 different nucleic acid template molecules are analyzed. In some embodiments, at least 6 different nucleic acid template molecules are analyzed. In some embodiments, at least 7 different nucleic acid template molecules are analyzed. In some embodiments, at least 8 different nucleic acid template molecules are analyzed. In some embodiments, at least 9 different nucleic acid template molecules are analyzed. In some embodiments, at least 10 different nucleic acid template molecules are analyzed. In some embodiments, at least 15 different nucleic acid template molecules are analyzed. In some embodiments, at least 20 different nucleic acid template molecules are analyzed. In some embodiments, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 85, at least 90, at least 95, at least 100, at least 125, at least 150, at least 175, at least 200, at least 250, at least 300, at least 350, or at least 400 different nucleic acid template molecules are analyzed. In some embodiments, at least 500 different nucleic acid template molecules are analyzed. In some embodiments, at least

IDT Exhibit 1020.1824

1,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 5,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 10,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 20,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 30,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 40,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 50,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 60,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 70,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 80,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 90,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 100,000 different nucleic acid template molecules are analyzed. In some embodiments, the Primer IDs are attached to the nucleic acid template molecules simultaneously. In some embodiments, the Primer IDs are attached to the nucleic acid template molecules sequentially. In some embodiments, the nucleic acid template molecules are amplified and/or detected simultaneously. In some embodiments, the nucleic acid template molecules are amplified and/or detected sequentially. In some embodiments, the Primer ID comprises 5-100 nucleotides. In some embodiments, the Primer ID comprises 5-50 nucleotides. In some embodiments, the Primer ID comprises at least 6 nucleotides. In some embodiments, the Primer ID comprises at least 7 nucleotides. In some embodiments, the Primer ID comprises at least 8 nucleotides. In some embodiments, the Primer ID comprises at least 9 nucleotides. In some embodiments, the Primer ID comprises at least 10 nucleotides. In some embodiments, the Primer ID comprises at least 12 nucleotides. In some embodiments, the Primer ID comprises at least 15 nucleotides. In some embodiments, the Primer ID comprises at least 20 nucleotides. In some embodiments, the Primer ID comprises at least 25 nucleotides. In some embodiments, the Primer ID comprises at least 35 nucleotides.

[0099] Further disclosed herein is a method for determining PCR resampling in an amplification reaction comprising: (a) providing a sample comprising a nucleic acid template molecule; (b) attaching a primer comprising a Primer ID to each nucleic acid template molecule to be analyzed to generate a tagged nucleic acid template, wherein each tagged nucleic acid template is attached to a unique Primer ID; (c) amplifying the tagged nucleic acid template to

IDT Exhibit 1020.1825

produce tagged amplicons; and (d) detecting the tagged amplicons, thereby determining PCR resampling in an amplification reaction. In some embodiments, the Primer ID comprises a degenerate sequence. In some embodiments, the Primer ID comprises a semi-degenerate sequence. In some embodiments, the Primer ID comprises a mixed sequence. In some embodiments, the Primer ID comprises an ambiguous sequence. In some embodiments, the Primer ID comprises a wobble sequence. In some embodiments, the Primer ID comprises a random sequence. In some embodiments, the Primer ID comprises a predetermined sequence. In some embodiments, the Primer ID is attached to the template by ligation. In some embodiments, the Primer ID is attached to the template by hybridization. In some embodiments, the Primer ID is attached to the template through PCR. In some embodiments, at least one template molecule is analyzed. In some embodiments, at least two different template molecules are analyzed. In some embodiments, detecting the tagged amplicons comprises sequencing the tagged amplicons. Sequencing of the tagged amplicons may occur by a variety of methods, including, but not limited to the Maxam-Gilbert sequencing method, the Sanger dideoxy sequencing method, dye-terminator sequencing method, pyrosequencing, multiple-primer DNA sequencing, shotgun sequencing, primer walking. In some embodiments, sequencing comprises pyrosequencing. In some embodiments, detecting the tagged amplicons further comprises counting a number of different Primer IDs associated with the tagged amplicons, wherein the number of different Primer IDs associated with the tagged amplicons reflects the number of templates sampled. In some embodiments, the method further comprises forming a consensus sequence for tagged amplicons comprising the same Primer ID. In some embodiments, the nucleic acid template comprises a DNA template. In some embodiments, the nucleic acid template comprises an RNA template. In some embodiments, amplifying comprises a PCR-based method. In some embodiments, the PCR-based method comprises PCR. In some embodiments, the PCR-based method comprises quantitative PCR. In some embodiments, the PCR-based method comprises emulsion PCR. In some embodiments, the PCR-based method comprises droplet PCR. In some embodiments, the PCR-based method comprises hot start PCR. In some embodiments, the PCR-based method comprises in situ PCR. In some embodiments, the PCR-based method comprises inverse PCR. In some embodiments, the PCR-based method comprises multiplex PCR. In some embodiments, the PCR-based method comprises Variable Number of Tandem Repeats (VNTR) PCR. In some embodiments, the PCR-based method comprises asymmetric PCR. In some

IDT Exhibit 1020.1826

embodiments, the PCR-based method comprises long PCR. In some embodiments, the PCR-based method comprises nested PCR. In some embodiments, the PCR-based method comprises hemi-nested PCR. In some embodiments, the PCR-based method comprises touchdown PCR. In some embodiments, the PCR-based method comprises assembly PCR. In some embodiments, the PCR-based method comprises colony PCR. In some embodiments, amplifying comprises a non-PCR-based method. In some embodiments, the non-PCR-based method comprises multiple displacement amplification (MDA). In some embodiments, the non-PCR-based method comprises transcription-mediated amplification (TMA). In some embodiments, the non-PCR-based method comprises nucleic acid sequence-based amplification (NASBA). In some embodiments, the non-PCR-based method comprises strand displacement amplification (SDA). In some embodiments, the non-PCR-based method comprises real-time SDA. In some embodiments, the non-PCR-based method comprises rolling circle amplification. In some embodiments, the non-PCR-based method comprises circle-to-circle amplification. In some embodiments, at least 2 different nucleic acid template molecules are analyzed. In some embodiments, at least 3 different nucleic acid template molecules are analyzed. In some embodiments, at least 4 different nucleic acid template molecules are analyzed. In some embodiments, at least 5 different nucleic acid template molecules are analyzed. In some embodiments, at least 6 different nucleic acid template molecules are analyzed. In some embodiments, at least 7 different nucleic acid template molecules are analyzed. In some embodiments, at least 8 different nucleic acid template molecules are analyzed. In some embodiments, at least 9 different nucleic acid template molecules are analyzed. In some embodiments, at least 10 different nucleic acid template molecules are analyzed. In some embodiments, at least 15 different nucleic acid template molecules are analyzed. In some embodiments, at least 20 different nucleic acid template molecules are analyzed. In some embodiments, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 85, at least 90, at least 95, at least 100, at least 125, at least 150, at least 175, at least 200, at least 250, at least 300, at least 350, or at least 400 different nucleic acid template molecules are analyzed. In some embodiments, at least 500 different nucleic acid template molecules are analyzed. In some embodiments, at least 1,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 5,000 different nucleic acid template molecules are analyzed. In some embodiments, at least

IDT Exhibit 1020.1827

10,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 20,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 30,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 40,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 50,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 60,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 70,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 80,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 90,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 100,000 different nucleic acid template molecules are analyzed. In some embodiments, the Primer IDs are attached to the nucleic acid template molecules simultaneously. In some embodiments, the Primer IDs are attached to the nucleic acid template molecules sequentially. In some embodiments, the nucleic acid template molecules are amplified and/or detected simultaneously. In some embodiments, the nucleic acid template molecules are amplified and/or detected sequentially. In some embodiments, the Primer ID comprises 5-100 nucleotides. In some embodiments, the Primer ID comprises 5-50 nucleotides. In some embodiments, the Primer ID comprises at least 6 nucleotides. In some embodiments, the Primer ID comprises at least 7 nucleotides. In some embodiments, the Primer ID comprises at least 8 nucleotides. In some embodiments, the Primer ID comprises at least 9 nucleotides. In some embodiments, the Primer ID comprises at least 10 nucleotides. In some embodiments, the Primer ID comprises at least 12 nucleotides. In some embodiments, the Primer ID comprises at least 15 nucleotides. In some embodiments, the Primer ID comprises at least 20 nucleotides. In some embodiments, the Primer ID comprises at least 25 nucleotides. In some embodiments, the Primer ID comprises at least 35 nucleotides.

[00100] Further disclosed herein is a method for determining PCR errors and/or sequencing errors comprising: (a) providing a sample comprising a nucleic acid template molecule; (b) attaching a primer comprising a Primer ID to each nucleic acid template molecule to be analyzed to generate a tagged nucleic acid template, wherein each tagged nucleic acid template is attached to a unique Primer ID; (c) amplifying the tagged nucleic acid template to produce tagged amplicons; and (d) detecting the tagged amplicons, thereby determining PCR errors and/or sequencing errors. In some embodiments, determining the PCR errors and/or sequencing error

IDT Exhibit 1020.1828

comprises determining fidelity of a polymerase. In some embodiments, determining the PCR error and/or sequencing error comprises determining accuracy of oligonucleotides synthesized in vitro. In some embodiments, determining the PCR error and/or sequencing error comprises determining the PCR error and/or sequencing error comprises determining accuracy of the sequencing reaction. In some embodiments, the Primer ID comprises a degenerate sequence. In some embodiments, the Primer ID comprises a semi-degenerate sequence. In some embodiments, the Primer ID comprises a mixed sequence. In some embodiments, the Primer ID comprises an ambiguous sequence. In some embodiments, the Primer ID comprises a wobble sequence. In some embodiments, the Primer ID comprises a random sequence. In some embodiments, the Primer ID comprises a predetermined sequence. In some embodiments, the Primer ID is attached to the template by ligation. In some embodiments, the Primer ID is attached to the template by hybridization. In some embodiments, the Primer ID is attached to the template through PCR. In some embodiments, at least one template molecule is analyzed. In some embodiments, at least two different template molecules are analyzed. In some embodiments, detecting the tagged amplicons further comprises counting a number of different Primer IDs associated with the tagged amplicons, wherein the number of different Primer IDs associated with the tagged amplicons reflects the number of templates sampled. In some embodiments, the method further comprises forming a consensus sequence for tagged amplicons comprising the same Primer ID. In some embodiments, the nucleic acid template comprises a DNA template. In some embodiments, the nucleic acid template comprises an RNA template. In some embodiments, amplifying comprises a PCR-based method. In some embodiments, the PCR-based method comprises PCR. In some embodiments, the PCR-based method comprises quantitative PCR. In some embodiments, the PCR-based method comprises emulsion PCR. In some embodiments, the PCR-based method comprises droplet PCR. In some embodiments, the PCR-based method comprises hot start PCR. In some embodiments, the PCR-based method comprises in situ PCR. In some embodiments, the PCR-based method comprises inverse PCR. In some embodiments, the PCR-based method comprises multiplex PCR. In some embodiments, the PCR-based method comprises Variable Number of Tandem Repeats (VNTR) PCR. In some embodiments, the PCR-based method comprises asymmetric PCR. In some embodiments, the PCR-based method comprises long PCR. In some embodiments, the PCR-based method comprises nested PCR. In some embodiments, the PCR-based method comprises hemi-nested

IDT Exhibit 1020.1829

PCR. In some embodiments, the PCR-based method comprises touchdown PCR. In some embodiments, the PCR-based method comprises assembly PCR. In some embodiments, the PCR-based method comprises colony PCR. In some embodiments, amplifying comprises a non-PCR-based method. In some embodiments, the non-PCR-based method comprises multiple displacement amplification (MDA). In some embodiments, the non-PCR-based method comprises transcription-mediated amplification (TMA). In some embodiments, the non-PCR-based method comprises nucleic acid sequence-based amplification (NASBA). In some embodiments, the non-PCR-based method comprises strand displacement amplification (SDA). In some embodiments, the non-PCR-based method comprises real-time SDA. In some embodiments, the non-PCR-based method comprises rolling circle amplification. In some embodiments, the non-PCR-based method comprises circle-to-circle amplification. In some embodiments, at least 2 different nucleic acid template molecules are analyzed. In some embodiments, at least 3 different nucleic acid template molecules are analyzed. In some embodiments, at least 4 different nucleic acid template molecules are analyzed. In some embodiments, at least 5 different nucleic acid template molecules are analyzed. In some embodiments, at least 6 different nucleic acid template molecules are analyzed. In some embodiments, at least 7 different nucleic acid template molecules are analyzed. In some embodiments, at least 8 different nucleic acid template molecules are analyzed. In some embodiments, at least 9 different nucleic acid template molecules are analyzed. In some embodiments, at least 10 different nucleic acid template molecules are analyzed. In some embodiments, at least 15 different nucleic acid template molecules are analyzed. In some embodiments, at least 20 different nucleic acid template molecules are analyzed. In some embodiments, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 85, at least 90, at least 95, at least 100, at least 125, at least 150, at least 175, at least 200, at least 250, at least 300, at least 350, or at least 400 different nucleic acid template molecules are analyzed. In some embodiments, at least 500 different nucleic acid template molecules are analyzed. In some embodiments, at least 1,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 5,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 10,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 20,000 different nucleic acid template molecules are analyzed. In some embodiments, at least

IDT Exhibit 1020.1830

30,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 40,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 50,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 60,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 70,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 80,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 90,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 100,000 different nucleic acid template molecules are analyzed. In some embodiments, the Primer IDs are attached to the nucleic acid template molecules simultaneously. In some embodiments, the Primer IDs are attached to the nucleic acid template molecules sequentially. In some embodiments, the nucleic acid template molecules are amplified and/or detected simultaneously. In some embodiments, the nucleic acid template molecules are amplified and/or detected sequentially. In some embodiments, the Primer ID comprises 5-100 nucleotides. In some embodiments, the Primer ID comprises 5-50 nucleotides. In some embodiments, the Primer ID comprises at least 6 nucleotides. In some embodiments, the Primer ID comprises at least 7 nucleotides. In some embodiments, the Primer ID comprises at least 8 nucleotides. In some embodiments, the Primer ID comprises at least 9 nucleotides. In some embodiments, the Primer ID comprises at least 10 nucleotides. In some embodiments, the Primer ID comprises at least 12 nucleotides. In some embodiments, the Primer ID comprises at least 15 nucleotides. In some embodiments, the Primer ID comprises at least 20 nucleotides. In some embodiments, the Primer ID comprises at least 25 nucleotides. In some embodiments, the Primer ID comprises at least 35 nucleotides.

[00101] Further disclosed herein is a method for correcting PCR errors and/or sequencing errors comprising: (a) providing a sample comprising a nucleic acid template molecule; (b) attaching a primer comprising a Primer ID to each nucleic acid template molecule to be analyzed to generate a tagged nucleic acid template, wherein each tagged nucleic acid template is attached to a unique Primer ID; (c) amplifying the tagged nucleic acid template to produce tagged amplicons; and (d) detecting the tagged amplicons, thereby correcting PCR error and/or sequencing error. In some embodiments, the Primer ID comprises a degenerate sequence. In some embodiments, the Primer ID comprises a semi-degenerate sequence. In some embodiments, the Primer ID comprises a mixed sequence. In some embodiments, the Primer ID

IDT Exhibit 1020.1831

comprises an ambiguous sequence. In some embodiments, the Primer ID comprises a wobble sequence. In some embodiments, the Primer ID comprises a random sequence. In some embodiments, the Primer ID comprises a predetermined sequence. In some embodiments, the Primer ID is attached to the template by ligation. In some embodiments, the Primer ID is attached to the template by hybridization. In some embodiments, the Primer ID is attached to the template through PCR. In some embodiments, at least one template molecule is analyzed. In some embodiments, at least two different template molecules are analyzed. In some embodiments, detecting the tagged amplicons further comprises counting a number of different Primer IDs associated with the tagged amplicons, wherein the number of different Primer IDs associated with the tagged amplicons reflects the number of templates sampled. In some embodiments, the method further comprises forming a consensus sequence for tagged amplicons comprising the same Primer ID. In some embodiments, the nucleic acid template comprises a DNA template. In some embodiments, the nucleic acid template comprises an RNA template. In some embodiments, amplifying comprises a PCR-based method. In some embodiments, the PCR-based method comprises PCR. In some embodiments, the PCR-based method comprises quantitative PCR. In some embodiments, the PCR-based method comprises emulsion PCR. In some embodiments, the PCR-based method comprises droplet PCR. In some embodiments, the PCR-based method comprises hot start PCR. In some embodiments, the PCR-based method comprises in situ PCR. In some embodiments, the PCR-based method comprises inverse PCR. In some embodiments, the PCR-based method comprises multiplex PCR. In some embodiments, the PCR-based method comprises Variable Number of Tandem Repeats (VNTR) PCR. In some embodiments, the PCR-based method comprises asymmetric PCR. In some embodiments, the PCR-based method comprises long PCR. In some embodiments, the PCR-based method comprises nested PCR. In some embodiments, the PCR-based method comprises hemi-nested PCR. In some embodiments, the PCR-based method comprises touchdown PCR. In some embodiments, the PCR-based method comprises assembly PCR. In some embodiments, the PCR-based method comprises colony PCR. In some embodiments, amplifying comprises a non-PCR-based method. In some embodiments, the non-PCR-based method comprises multiple displacement amplification (MDA). In some embodiments, the non-PCR-based method comprises transcription-mediated amplification (TMA). In some embodiments, the non-PCR-based method comprises nucleic acid sequence-based amplification (NASBA). In some

IDT Exhibit 1020.1832

embodiments, the non-PCR-based method comprises strand displacement amplification (SDA). In some embodiments, the non-PCR-based method comprises real-time SDA. In some embodiments, the non-PCR-based method comprises rolling circle amplification. In some embodiments, the non-PCR-based method comprises circle-to-circle amplification. In some embodiments, at least 2 different nucleic acid template molecules are analyzed. In some embodiments, at least 3 different nucleic acid template molecules are analyzed. In some embodiments, at least 4 different nucleic acid template molecules are analyzed. In some embodiments, at least 5 different nucleic acid template molecules are analyzed. In some embodiments, at least 6 different nucleic acid template molecules are analyzed. In some embodiments, at least 7 different nucleic acid template molecules are analyzed. In some embodiments, at least 8 different nucleic acid template molecules are analyzed. In some embodiments, at least 9 different nucleic acid template molecules are analyzed. In some embodiments, at least 10 different nucleic acid template molecules are analyzed. In some embodiments, at least 15 different nucleic acid template molecules are analyzed. In some embodiments, at least 20 different nucleic acid template molecules are analyzed. In some embodiments, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 85, at least 90, at least 95, at least 100, at least 125, at least 150, at least 175, at least 200, at least 250, at least 300, at least 350, or at least 400 different nucleic acid template molecules are analyzed. In some embodiments, at least 500 different nucleic acid template molecules are analyzed. In some embodiments, at least 1,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 5,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 10,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 20,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 30,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 40,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 50,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 60,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 70,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 80,000 different nucleic acid template molecules are analyzed. In some embodiments, at least 90,000 different nucleic acid template molecules are analyzed. In some embodiments, at least

IDT Exhibit 1020.1833

100,000 different nucleic acid template molecules are analyzed. In some embodiments, the Primer IDs are attached to the nucleic acid template molecules simultaneously. In some embodiments, the Primer IDs are attached to the nucleic acid template molecules sequentially. In some embodiments, the nucleic acid template molecules are amplified and/or detected simultaneously. In some embodiments, the nucleic acid template molecules are amplified and/or detected sequentially. In some embodiments, the Primer ID comprises 5-100 nucleotides. In some embodiments, the Primer ID comprises 5-50 nucleotides. In some embodiments, the Primer ID comprises at least 6 nucleotides. In some embodiments, the Primer ID comprises at least 7 nucleotides. In some embodiments, the Primer ID comprises at least 8 nucleotides. In some embodiments, the Primer ID comprises at least 9 nucleotides. In some embodiments, the Primer ID comprises at least 10 nucleotides. In some embodiments, the Primer ID comprises at least 12 nucleotides. In some embodiments, the Primer ID comprises at least 15 nucleotides. In some embodiments, the Primer ID comprises at least 20 nucleotides. In some embodiments, the Primer ID comprises at least 25 nucleotides. In some embodiments, the Primer ID comprises at least 35 nucleotides.

[00102] Disclosed herein is a method to analyze nucleic acid sequences, comprising (a) attaching a Primer ID to a first end of each of a plurality of nucleic acid fragments to form tagged-nucleic acid templates; (b) redundantly determining nucleotide sequence of a tagged-nucleic acid template, wherein determined nucleotide sequences which share a Primer ID form a family of members; and (c) identifying a nucleotide sequence as accurately representing an analyte nucleic acid fragment when at least 1 % of members of the family contain the sequence.

[00103] The nucleotide sequence may be identified when at least 5% of members of the family contain the sequence. The nucleotide sequence may be identified when at least 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 92%, 95%, 97%, 98%, 99% or more of members of the family contain the sequence. The nucleotide sequence may be identified when at least about 75% to about 99% of members of the family contain the sequence. The nucleotide sequence may be identified when at least about 85% to about 99% of members of the family contain the sequence. The nucleotide sequence may be identified when at least about 92% to about 98% of members of the family contain the sequence.

[00104] A first universal priming site may be attached to a second end of each of a plurality of analyte nucleic acid fragments.

IDT Exhibit 1020.1834

[00105] At least two cycles of polymerase chain reaction may be performed such that a family may be formed of tagged-nucleic acid templates that have a Primer ID on the first end and a first universal priming site on a second end.

[00106] The Primer ID may be covalently linked to a second universal priming site. The Primer ID may be attached to the 5' end of a nucleic acid fragment and the second universal priming site may be 5' to the Primer ID. The Primer ID may be attached to the 3' end of a nucleic acid fragment and the second universal priming site may be 3' to the Primer ID.

[00107] The nucleic acid fragments may be formed by applying a shear force to a nucleic acid. Alternatively, the nucleic acid fragments may be formed by one or more restriction endonucleases.

[00108] The method may further comprise, prior to the step of redundantly determining, amplifying the tagged-nucleic acid templates. The method may further comprise, prior to the step of redundantly determining, amplifying the tagged-nucleic acid templates by using a pair of primers which may be complementary to the first and the second universal priming sites, respectively.

[00109] The method may further comprise, prior to the step of redundantly determining, amplifying the tagged-nucleic acid templates, and wherein prior to said amplification, a single strand-specific exonuclease may be used to digest excess primers used to attach the Primer ID the nucleic acid fragments.

[00110] The method may further comprise, prior to the step of redundantly determining, amplifying the tagged-nucleic acid templates, and wherein prior to said amplification, the single strand-specific exonuclease may be inactivated, inhibited, or removed. The single strand-specific exonuclease may be inactivated by heat treatment.

[00111] The primers used in said amplification may comprise one or more chemical modifications rendering them resistant to exonucleases. The primers used in said amplification may comprise one or more phosphorothioate linkages.

[00112] The method may further comprise, prior to the amplification step, treating the DNA with bisulfite to convert unmethylated cytosine bases to uracil.

[00113] The method may further comprise the step of comparing number of families representing a first DNA fragment to number of families representing second DNA fragment to determine a relative concentration of a first DNA fragment to a second DNA fragment in the

IDT Exhibit 1020.1835

plurality of DNA fragments.

[00114] Disclosed herein is a method to analyze nucleic acid sequences, comprising (a) attaching a Primer ID to a first end of each of a plurality of DNA fragments using at least two cycles of amplification with first and second primers to form tagged-DNA fragments, wherein the Primer IDs are in excess of the DNA fragments during amplification, wherein the first primers comprise (i) a first segment complementary to a desired amplicon; (ii) a second segment containing the Primer ID; and (iii) a third segment containing a universal priming site for subsequent amplification; and wherein the second primers comprise a universal priming site for subsequent amplification; wherein each cycle of amplification attaches one universal priming site to a strand; (b) amplifying the tagged-DNA fragments to form a family of tagged-DNA fragments from each tagged-DNA fragment; and (c) determining nucleotide sequences of a plurality of members of the family.

[00115] The method may further comprise the steps of (d) comparing sequences of a family of tagged-DN A fragments; and (e) identifying a nucleotide sequence as accurately representing an DNA fragment when at least 1 % of members of the family contain the sequence.

[00116] The nucleotide sequence may be identified when at least 5% of members of the family contain the sequence. The nucleotide sequence may be identified when at least 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 92%, 95%, 97%, 98%, 99% or more of members of the family contain the sequence. The nucleotide sequence may be identified when at least about 75% to about 99% of members of the family contain the sequence. The nucleotide sequence may be identified when at least about 85% to about 99% of members of the family contain the sequence. The nucleotide sequence may be identified when at least about 92% to about 98% of members of the family contain the sequence.

[00117] The second primers may each comprise a Primer ID.

[00118] The Primer IDs may be from 2 to 4000 bases or base pairs inclusive. The Primer IDs may be from 20 to 100 bases or base pairs inclusive. The Primer IDs may be from 20 to 80 bases or base pairs inclusive. The Primer IDs may be from 20 to 60 bases or base pairs inclusive. The Primer IDs may be at least about 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100 or more bases or base pairs. The Primer IDs may be at least about 125, 150, 175, 200, 250, 300, 350, 400, 450, 500 or more bases or base pairs. The Primer IDs may be less than about 400, 300, 200, 100, 90, 80, 70, 60 or fewer bases or base pairs.

IDT Exhibit 1020.1836

[00119] The method may further comprise, prior to the step of amplifying the tagged-DNA fragments, digesting excess primers used to attach the Primer ID the DNA fragments with a single strand-specific exonuclease. The method may further comprise, prior to the step of amplifying, inactivating, inhibiting or removing the single strand-specific exonuclease. The single strand-specific exonuclease may be inactivated by heat treatment. The primers used in the step of amplifying may comprise one or more phosphorothioate linkages.

[00120] The method may further comprise, prior to the amplification step, treating the DNA with bisulfite to convert unmethylated cytosine bases to uracil. The method may further comprise the step of comparing number of families representing a first DNA fragment to number of families representing second DNA fragment to determine a relative concentration of a first DNA fragment to a second DNA fragment in the plurality of DNA fragments.

[00121] Disclosed herein is a method to analyze DNA using endogenous unique identifiers, comprising (a) attaching adapter oligonucleotides to ends of fragments of DNA of between 30 to 2000 bases, inclusive, to form adapted fragments, wherein each end of a fragment before said attaching is an endogenous unique identifier for the fragment; (b) amplifying the adapted fragments using primers complementary to the adapter oligonucleotides to form families of adapted fragments; (c) determining nucleotide sequence of a plurality of members of a family; comparing nucleotide sequences of the plurality of members of the family; and (d) identifying a nucleotide sequence as accurately representing an DNA fragment when at least 1% of members of the family contain the sequence.

[00122] The method may further comprise enriching for fragments representing one or more selected genes by means of capturing a subset of the fragments using capture oligonucleotides complementary to selected genes in the DNA.

[00123] The method may further comprise enriching for fragments representing one or more selected genes by means of amplifying fragments complementary to selected genes.

[00124] The step of attaching may be prior to the step of enriching.

[00125] The fragments may be formed by shearing. The fragments may be formed by digestion with one or more restriction enzymes.

[00126] The nucleotide sequence may be identified when at least 5% of members of the family contain the sequence. The nucleotide sequence may be identified when at least 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 92%,

IDT Exhibit 1020.1837

95%, 97%, 98%, 99% or more of members of the family contain the sequence. The nucleotide sequence may be identified when at least about 75% to about 99% of members of the family contain the sequence. The nucleotide sequence may be identified when at least about 85% to about 99% of members of the family contain the sequence. The nucleotide sequence may be identified when at least about 92% to about 98% of members of the family contain the sequence.

[00127] The method may further comprise, prior to the amplification step, treating the DNA with bisulfite to convert unmethylated cytosine bases to uracil.

[00128] The method may further comprise the step of comparing number of families representing a first DNA fragment to number of families representing second DNA fragment to determine a relative concentration of a first DNA fragment to a second DNA fragment in the plurality of DNA fragments.

[00129] Disclosed herein is a population of primer pairs, wherein each pair comprises a first and second primer for amplifying and identifying a gene or gene portion, wherein (a) the first primer comprises a first portion of 10-100 nucleotides complementary to the gene or gene portion and a second portion of 10 to 100 nucleotides comprising a site for hybridization to a third primer; (b) the second primer comprises a first portion of 10-100 nucleotides complementary to the gene or gene portion and a second portion of 10 to 100 nucleotides comprising a site for hybridization to a fourth primer, wherein interposed between the first portion and the second portion of the second primer is a third portion consisting of 2 to 4000 nucleotides forming a Primer ID, wherein the Primer IDs in the population have at least 4 different sequences, wherein the first and second primers are complementary to opposite strands of the gene or gene portion.

[00130] The first portion of the first primer and/or the first portion of the second primer may comprise at least about 10, 15, 20, 25, 30 or more nucleotides complementary to the gene or gene portion. The first portion of the first primer and/or the first portion of the second primer may comprise less than about 80, 70, 60, 50 or fewer nucleotides complementary to the gene or gene portion. The first portion of the first primer and/or the first portion of the second primer may comprise between about 10 to about 90, between about 10 to about 80, between about 10 to about 70, between about 10 to about 60 nucleotides complementary to the gene or gene portion.

[00131] The first primer may further comprise a Primer ID.

[00132] The Primer IDs in the population can have at least at least 10, 15, 20, 25, 30, 35, 40,

IDT Exhibit 1020.1838

45, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000 or more different sequences. The Primer IDs in the population can have at least at least 2,000; 3,000; 4,000; 5,000; 6,000; 7,000; 8,000; 9,000; 10,000; 15,000; 20,000; 25,000; 30,000; 35,000; 40,000; 45,000; 50,000; 60,000; 70,000; 80,000; 90,000; 100,000; 200,000; 300,000; 400,000; 500,000; 600,000; 700,000; 800,000; 900,000; 1,000,000 or more different sequences.

[00133] Further disclosed herein is a kit comprising a population of primer primers, wherein each pair comprises a first and second primer for amplifying and identifying a gene or gene portion, wherein (a) the first primer comprises a first portion of 10-100 nucleotides complementary to the gene or gene portion and a second portion of 10 to 100 nucleotides comprising a site for hybridization to a third primer; (b) the second primer comprises a first portion of 10-100 nucleotides complementary to the gene or gene portion and a second portion of 10 to 100 nucleotides comprising a site for hybridization to a fourth primer, wherein interposed between the first portion and the second portion of the second primer is a third portion consisting of 2 to 4000 nucleotides forming a Primer ID, wherein the Primer IDs in the population have at least 4 different sequences, wherein the first and second primers are complementary to opposite strands of the gene or gene portion.

[00134] The kit may further comprise third and fourth primers complementary to the second portions of each of the first and second primers. The first portion of the first primer and/or the first portion of the second primer may comprise at least about 10, 15, 20, 25, 30 or more nucleotides complementary to the gene or gene portion. The first portion of the first primer and/or the first portion of the second primer may comprise less than about 80, 70, 60, 50 or fewer nucleotides complementary to the gene or gene portion. The first portion of the first primer and/or the first portion of the second primer may comprise between about 10 to about 90, between about 10 to about 80, between about 10 to about 70, between about 10 to about 60 nucleotides complementary to the gene or gene portion.

[00135] The first primer may further comprise a Primer ID. The Primer IDs in the population can have at least at least 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000 or more different sequences. The Primer IDs in the population can have at least at least 2,000; 3,000; 4,000; 5,000; 6,000; 7,000; 8,000; 9,000; 10,000; 15,000; 20,000; 25,000; 30,000; 35,000; 40,000; 45,000; 50,000; 60,000; 70,000; 80,000; 90,000; 100,000; 200,000; 300,000; 400,000; 500,000; 600,000; 700,000; 800,000; 900,000; 1,000,000 or more

IDT Exhibit 1020.1839

different sequences.

[0010] The invention provides kits for detecting and/or measuring types and/or levels. In one non-limiting embodiment, kits for carrying out diagnostic assays of the invention typically include, a suitable container means, (i) a probe that comprises a nucleic acid sequence (including Primer ID and optionally barcode) that specifically binds to a polynucleotide of interest; (ii) a label for detecting the presence of the probe; and (iii) instructions for how to use and/or interpret the results. The container means of the kits will generally include at least one vial, test tube, flask, bottle, syringe and/or other container into which a nucleic acid specific for one of a polynucleotide of interest of the present invention may be placed and/or suitably aliquoted. Where a second and/or third and/or additional component is provided, the kit will also generally contain a second, third and/or other additional container into which this component may be placed. Alternatively, a container may contain a mixture of more than one nucleic acid reagent, each reagent specifically binding a different marker in accordance with the present invention. The kits of the present invention will also typically include means for containing the nucleic acid probes in close confinement for commercial sale. Such containers may include injection and/or blow-molded plastic containers into which the desired vials are retained.

[00136] The kits may further comprise positive and negative controls, as well as instructions for the use of kit components contained therein, in accordance with the methods of the present invention. The kit may also include a package insert with instructions in connection with an approved indication.

[00137] The Primer ID may comprise randomly selected sequences. The Primer ID may comprise pre-defined nucleotide sequences. The Primer ID may comprise both randomly selected sequences and pre-defined nucleotides.

[00138] The method may further comprise, prior to the amplification step, treating the DNA with bisulfite to convert unmethylated cytosine bases to uracil.

[00139] The method may further comprise the step of comparing number of families representing a first DNA fragment to number of families representing second DNA fragment to determine a relative concentration of a first DNA fragment to a second DNA fragment in the plurality of DNA fragments. See also, Jabara *et al.* 2011 PNAS 20166-20171, the contents of which are hereby incorporated in its entirety.

IDT Exhibit 1020.1840

[0005] Throughout the specification the word "comprising," or variations such as "comprises" or "comprising," will be understood to imply the inclusion of a stated element, integer or step, or group of elements, integers or steps, but not the exclusion of any other element, integer or step, or group of elements, integers or steps. The present invention may suitably "comprise", "consist of", or "consist essentially of", the steps, elements, and/or reagents described in the claims.

[0006] It is further noted that the claims may be drafted to exclude any optional element. As such, this statement is intended to serve as antecedent basis for use of such exclusive terminology as "solely", "only" and the like in connection with the recitation of claim elements, or the use of a "negative" limitation.

[0007] The following Examples further illustrate the invention and are not intended to limit the scope of the invention. In particular, it is to be understood that this invention is not limited to particular embodiments described, as such may, of course, vary. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to be limiting, since the scope of the present invention will be limited only by the appended claims.

EXAMPLES

Example 1. Use of Primer IDs for the analysis of a viral template molecule

[00140] High throughput sequencing allows the acquisition of large amounts of sequence data that can encompass entire genomes. With sufficient amounts of starting DNA, PCR is not needed before the library preparation step of the sequencing protocol. Sequencing miscalls inherent in high throughput sequencing approaches are resolved using multiple reads over a given base.

[00141] Deep sequencing can also capture the genetic diversity of viral populations, including intrahost populations derived from clinical samples. This approach offers the opportunity to view population diversity and dynamics and viral evolution in unprecedented detail. One place where the presence of minor variants is of immediate practical importance is in the detection of drug-resistant variants. Standard bulk sequencing methods typically miss allelic variants below 20% in frequency within a population. Alternative assays can detect less abundant variants that confer drug resistance, but require a priori selection of sites and variants. Thus, deep sequencing approaches offer the opportunity to identify minor variants associated with resistance de novo

IDT Exhibit 1020.1841

with the goal of understanding their role in therapy failure.

[00142] Although screening for drug-resistant variants is a practical application of the deep sequencing technology, this technology also addresses broader questions of sequence diversity and structure for a complex population like HIV-1. However, the relatively high sequencing error rates of these technologies artificially increase genetic diversity, which confounds the detection of natural genetic variation especially when sequencing a highly heterogeneous viral population. Moreover, the use of PCR to amplify the amount of material before starting the sequencing protocol adds the potential for several serious artifacts: First, nucleotide misincorporation by the polymerase during many rounds of amplification artificially increases sequence diversity; second, artifactual recombination during amplification occurs when premature termination products prime a subsequent round of synthesis, which can obscure the linkage of two sequence polymorphisms; third, differential amplification can skew allelic frequencies; and fourth, PCR amplification can create a significant mass of DNA from a small number of starting templates, which obscures the true sampling of the original population as these few starting templates/genomes get resampled in the PCR product, creating sequence resampling rather than the observation of independent genomes. Overall, these biases artificially decrease true diversity while introducing artifactual diversity and also skew allelic frequencies, which can lead to incongruence between the real and observed viral populations. Most investigators use statistical tools to attempt to control for the types of sequencing errors that are associated with each sequencing platform.

[00143] To make deep sequencing useful for complex populations, it is necessary to overcome PCR resampling, which is mistaken for sampling of the original population, and PCR and sequencing errors, which can be mistaken for diversity. As nucleotide misincorporation is largely random across sites and template switching/recombination is more likely to occur in the later cycles of a PCR, strategies that create a bulk or consensus sequence for each sampled template will call the correct base at each position. One approach to sampling highly heterogeneous populations, such as the HIV-1 env gene, is through endpoint dilution titration of the template before nested PCR, such that a single template is present in each PCR amplification. In addition to masking the misincorporations, PCR-mediated recombination produces recombinant templates identical to the parental sequence. Although highly accurate, this technique is labor-intensive and, as population sampling is dependent on the number of templates sequenced, this methodology

IDT Exhibit 1020.1842

does not lend itself to the identification of minor variants or to understanding the structure of a complex population, nor is it easily adaptable to a high throughput approach.

[00144] We have developed a high throughput technique for directly resolving the genetic diversity of a viral population. This technique avoids the recording of PCR and sequencing errors that create artificial diversity, and corrects for artificial allelic skewing and PCR resampling, revealing the original genomes in the population. This is accomplished by embedding a degenerate block of nucleotides within the primer used in the first round of cDNA synthesis. This creates a random library of sequences within the primer population. As primers are individually used out of this library, each viral template is copied such that the complement (cDNA) now includes a unique sequence tag, or Primer ID. This Primer ID is carried through all of the subsequent manipulations to mark all sequences that derive from each independent templating event, and PCR resampling then becomes over-coverage for each template to create a consensus sequence of that template. Using this approach, we were able to directly remove error, correct for PCR resampling, and capture the fluctuation of minor variants in the viral population within a host. We also resolved minor drug-resistant variants below 1% in frequency before the initiation of antiretroviral therapy, and were able to correlate these variants with the emergence of drug resistance.

[00145] **Results**

[00146] A cDNA Synthesis Primer Containing a Primer ID Can Be Used to Track Individual Viral Templates. A population of cDNA synthesis primers was designed to prime DNA synthesis downstream of the HIV-1 protease (pro) gene, with the primer containing two additional blocks of identifying information (Fig. 1A). The first block was a string of eight degenerate nucleotides that created 65,536 distinct sequence combinations, or Primer IDs. This region was flanked by an a priori selected three nucleotide barcode, creating a sample identification block so that multiple samples could be pooled together in a sequencing run. A designed sequence at the 5′ end of the cDNA primer was used for subsequent amplification of the cDNA sequences by nested PCR.

[00147] Viral RNA was extracted from three longitudinal blood plasma samples from an individual infected with subtype B HIV-1 who was participating in a protease inhibitor efficacy trial (M94-247) (Fig. 4). Approximately 10,000 copies of viral RNA from each sample were used in a reverse transcription reaction for cDNA synthesis and tagging using the Primer ID. The cDNA product was separated from the unused cDNA primers, and then the viral sequences were

IDT Exhibit 1020.1843

amplified by nested PCR and sequenced on the 454 GS FLX Titanium. Our data were distilled from total reads of 20,429, 24,658, and 27,075 for the three time points (T1, T2, and T3, respectively). Raw sequence reads were assessed for the cDNA tagging primer and a full length pro gene sequence (297 nucleotides long representing 99 codons), and when three or more sequences within a sample contained an identical Primer ID, a consensus sequence was formed to represent one sequence/ genome in the population (Fig. 1B and 1C and Fig. 5).

[00148] With these manipulations we generated 857, 1,609, and 2,213 consensus sequences, respectively, for the three time points (Fig. 1C). The median number of reads per Primer ID was 6, ranging from 1 to 96 (Fig. 6A). The distribution of identical Primer IDs did not form a normal distribution as would be expected if all templates were amplified equally. We saw a higher than expected number of single reads of Primer IDs; although we do not know the reason for this, such a result is consistent with different cDNA templates entering the PCR at different cycles. Because each template is individually tagged the different number of reads is an indication of allelic skewing, as noted this can be nearly 100-fold. In an analysis of a number of low abundant variants we saw a 20-fold range of representation through allelic skewing, with half of the variants up to 2- to 3-fold more abundant than the mean, and the other half up to 5- to 10-fold less abundant (Fig. 7).

[00149] We conservatively estimate the combined in vitro error rate of the cDNA synthesis step by reverse transcriptase (RT) and the first strand synthesis by the Taq polymerase to be on the order of 1 mutation in 10,000 bases, or approximately one mutation per 33 pro gene sequences, based on an RT error rate of 1 in 22,000 nucleotides (38) and a Taq polymerase error rate of 1.1 in 10,000 nucleotides (39) but reduced by half because only the first round of synthesis is relevant and a misincorporation at this step gives a mixture. Later rounds of Taq polymerase errors should be largely lost through the creation of the consensus sequence. Thus, we would expect 139 sequence misincorporations to be present in the data set of 4,679 total sequences representing T1+T2+T3, and with an excess of transitions. These would be expected to occur as 113 single copy single-nucleotide polymorphisms (SNPs) and 13 SNPs that appeared twice. We observed 98 single copy SNPs in the data set with a threefold excess of transitions, and with three-fourths of them being coding changes, which is consistent with random mutations. We expect there to be low frequency SNPs in the viral population from rare but persistent variants that are fortuitously sampled, and from the intrinsic error rate of viral replication (the

IDT Exhibit 1020.1844

error rate during one round of viral replication would represent approximately one mutation per 150 pro gene sequences). However, we cannot distinguish real polymorphisms from the inferred background error rate associated with the first and second rounds of in vitro DNA synthesis. Thus, we have limited the analysis of population diversity to SNPs that appeared at least twice in the data set (e.g.,linked to at least two separate Primer IDs), either at the same time point or at multiple time points in the overall data set (Table 1). We have not corrected the data set for the presumed 13 SNPs that appeared twice that are expected to be present due to error even though this represents 33% of all of the SNPs that appeared twice (13 of 39). Overall, 80% of the SNPs (e.g.,any sequence change from the consensus that appeared at least once) in the total data set of 72,162 sequence reads were removed as error. Also, 60–65% of the sequence reads were revealed as resampling. Finally, allelic skewing of up to nearly 100 fold was corrected (Fig. 7).

[00150] Longitudinal Sequencing of the HIV-1 Protease (pro) Gene in an Untreated Individual Reveals Dynamic Changes in Genetic Variation. We analyzed the sequences of the pro gene populations to assess allelic frequency at the two sampled time points, separated by 6 mo and before ritonavir (37) drug selection (Fig. 4). The combined sequence population from the two time points (T1 and T2) before therapy consisted of 492 unique pro gene sequences with 155 SNPs. About 4% (e.g.,21) of these unique gene sequences were above 0.5% abundance, and these 21 unique gene sequences represented 67% of all sampled genomes, with the genome representing the overall consensus sequence comprising 21% of the total population (Fig. 8A and 8B). The relatively small number of unique gene sequences above 0.5% frequency in the population contained only 7% of the 155 detected SNPs. Thus, a large proportion of the viral population's diversity was associated with a large number of pro gene sequences that were present at low abundance (Fig. 8A and 8C); conversely, the majority of the population consisted of a small number of SNPs. Similarly, Tajima's D statistic for T1 and T2 in this individual were –2.35 and –2.31, respectively (Table 2), indicative of a population structure that has an excess of low frequency polymorphisms. This pattern is consistent with but more extreme than that observed in a prior shallow intrahost survey in which a metapopulation model was proposed to explain the pattern of Tajima's D statistic (40). Fig. 2A-2B shows the encoded amino acid variability and synonymous nucleotide variability present in two or more individual genomes across the 99 codons in the pro gene for these samples.

[00151] Synonymous variability. There were 57 codons (with 63 variants/ SNPs) that

IDT Exhibit 1020.1845

contained synonymous diversity that appeared in both pretherapy time points, and 30 codons (with 31 variants) that appeared in only one time point. Taken together, 75 of the 99 codons contained some level of synonymous diversity (Fig. 2A-2B and Table 1). Of the 63 variants that were present in both untreated time points, 92% were transitions. Of the 31 variants that appeared in only one of the time points, 71% were transitions, representing a significantly smaller fraction of transitions than among the synonymous variants that appeared at both time points (P = 0.012; Fisher's exact test). This suggests that synonymous transversions are selected against over time.

[00152] Nonsynonymous variability. There were 26 codons (28 variants) that contained coding variability that appeared in both pretherapy time points, and an additional 28 codons (33 variants) with nonsynonymous changes found in only one of the time points. Taken together, 49 of the 99 codons contained some level of nonsynonymous diversity (Fig. 2A-2B and Table 1). For the 28 nonsynonymous variants detected at both time points, 22 were transitions, and these mostly represented conservative amino acid changes. In the case of synonymous mutations two-thirds of the variants were present at both time points, whereas in the case of nonsynonymous mutations, less than half were present at both time points (P = 0.012; Fisher's exact test). This observation suggests that, at this level of sequence sampling, we are able to see a difference in stability within the population in comparing synonymous and nonsynonymous substitutions.

[00153] Genetic fluctuation. We compared the stability of minor SNPs present at both T1 and T2. A total of 14 of the 91 SNPs (synonymous and nonsynonymous that appeared at both time points) had significant changes in abundance between the two time points ($\chi 2$ test with a false discovery rate of 0.05). Of the 14 SNPs with significant changes in abundance, 11 had a decrease in the abundance, with an average decrease around 7.5-fold. There were three SNPs that had a significant increase in abundance, all of which were synonymous, ranging from a 4- to 47-fold increase. Although a majority of SNPs that changed in abundance had a decrease in the frequency between T1 and T2, on a population level, there was not a large change in diversity between the two time points (T1 $\pi = 0.0080$, T2 $\pi = 0.0079$; Table 2). However, the trend of increased abundance at the three sites may be driven by selection of cryptic epitopes in an alternative reading frame (see Discussion).

[00154] Significance of rare variants. We observed two extremes in terms of biological relevance in the untreated population among variants detected as at least two independent

IDT Exhibit 1020.1846

sequences across the three time points. At one extreme was the detection of nonviable genomes in the form of a coding variant at position 25, which mutates the active site of the protease, and the detection of termination codons at positions 42 and 61 (Table 1). At the other extreme was the detection of the L90M and V82A variants (at time points 1 and 2, respectively) that became the major resistance populations after ritonavir therapy was initiated (see below, Fig. 3); in addition, V82I and V82L were detected at T2. We found two more examples of primary resistance mutations at low abundance, K20R at all three time points and M46I at two time points, but these did not grow out in the presence of ritonavir (Fig. 3 and Table 1). Similarly, fitness compensatory mutations were also detected at low abundance (L10F, M36I, L63P, A71T, and V77I), all below 1%, and only L63P increased (modestly) in abundance after exposure to ritonavir. More generally, of the 28 substitutions most closely associated with protease inhibitor drug resistance, we found 10 such variants, half of which were detected at both pretherapy time points (Table 1).

[00155] Assessment of Linkage Disequilibrium (LD) Within the HIV-1 pro Gene Population. We measured LD for the sequences in the T1 and T2 populations. We identified very few examples of LD at these two time points using the Fisher's exact test with a Bonferroni correction. Of the 103 polymorphic sites in T1, only three pairs were in significant LD. Similarly, in T2 with 118 polymorphic sites, only and before ritonavir (37) drug selection (Fig. 4). The combined sequence population from the two time points (T1 and T2) before therapy consisted of 492 unique pro gene sequences with 155 SNPs. About 4% (e.g.,21) of these unique gene sequences were above 0.5% abundance, and these 21 unique gene sequences represented 67% of all sampled genomes, with the genome representing the overall consensus sequence comprising 21% of the total population (Fig. 8A and 8B). The relatively small number of unique gene sequences above 0.5% frequency in the population contained only 7% of the 155 detected SNPs. Thus, a large proportion of the viral population's diversity was associated with a large number of pro gene sequences that were present at low abundance (Fig. 8A and 8C); conversely, the majority of the population consisted of a small number of SNPs. Similarly, Tajima's D statistic for T1 and T2 in this individual were –2.35 and –2.31, respectively (Table 2), indicative of a population structure that has an excess of low frequency polymorphisms. This pattern is consistent with but more extreme than that observed in a prior shallow intrahost survey in which a metapopulation model was proposed to explain the pattern of Tajima's D statistic. Fig. 2A-2B

IDT Exhibit 1020.1847

shows the encoded amino acid variability and synonymous nucleotide variability present in two or more individual genomes across the 99 codons in the pro gene for these samples.

[00156] Synonymous variability. There were 57 codons (with 63 variants/ SNPs) that contained synonymous diversity that appeared in both pretherapy time points, and 30 codons (with 31 variants) that appeared in only one time point. Taken together, 75 of the 99 codons contained some level of synonymous diversity (Fig. 2A-2B and Table 1). Of the 63 variants that were present in both untreated time points, 92% were transitions. Of the 31 variants that appeared in only one of the time points, 71% were transitions, representing a significantly smaller fraction of transitions than among the synonymous variants that appeared at both time points ($P = 0.012$; Fisher's exact test). This suggests that synonymous transversions are selected against over time.

[00157] Nonsynonymous variability. There were 26 codons (28 variants) that contained coding variability that appeared in both pretherapy time points, and an additional 28 codons (33 variants) with nonsynonymous changes found in only one of the time points. Taken together, 49 of the 99 codons contained some level of nonsynonymous diversity (Fig. 2A-2B and Table 1). For the 28 nonsynonymous variants detected at both time points, 22 were transitions, and these mostly represented conservative amino acid changes. In the case of synonymous mutations two-thirds of the variants were present at both time points, whereas in the case of nonsynonymous mutations, less than half were present at both time points ($P = 0.012$; Fisher's exact test). This observation suggests that, at this level of sequence sampling, we are able to see a difference in stability within the population in comparing synonymous and nonsynonymous substitutions.

[00158] Genetic fluctuation. We compared the stability of minor SNPs present at both T1 and T2. A total of 14 of the 91 SNPs (synonymous and nonsynonymous that appeared at both time points) had significant changes in abundance between the two time points ($\chi 2$ test with a false discovery rate of 0.05). Of the 14 SNPs with significant changes in abundance, 11 had a decrease in the abundance, with an average decrease around 7.5-fold. There were three SNPs that had a significant increase in abundance, all of which were synonymous, ranging from a 4- to 47-fold increase. Although a majority of SNPs that changed in abundance had a decrease in the frequency between T1 and T2, on a population level, there was not a large change in diversity between the two time points (T1 $\pi = 0.0080$, T2 $\pi = 0.0079$; Table 2). However, the trend of increased abundance at the three sites may be driven by selection of cryptic epitopes in an

IDT Exhibit 1020.1848

alternative reading frame.

[00159] Significance of rare variants. We observed two extremes in terms of biological relevance in the untreated population among variants detected as at least two independent sequences across the three time points. At one extreme was the detection of nonviable genomes in the form of a coding variant at position 25, which mutates the active site of the protease, and the detection of termination codons at positions 42 and 61 (Table 1). At the other extreme was the detection of the L90M and V82A variants (at time points 1 and 2, respectively) that became the major resistance populations after ritonavir therapy was initiated (see below, Fig. 3); in addition, V82I and V82L were detected at T2. We found two more examples of primary resistance mutations at low abundance, K20R at all three time points and M46I at two time points, but these did not grow out in the presence of ritonavir (Fig. 3 and Table 1). Similarly, fitness compensatory mutations were also detected at low abundance (L10F, M36I, L63P, A71T, and V77I), all below 1%, and only L63P increased (modestly) in abundance after exposure to ritonavir. More generally, of the 28 substitutions most closely associated with protease inhibitor drug resistance, we found 10 such variants, half of which were detected at both pretherapy time points (Table 1).

[00160] Assessment of Linkage Disequilibrium (LD) Within the HIV-1 pro Gene Population. We measured LD for the sequences in the T1 and T2 populations. We identified very few examples of LD at these two time points using the Fisher's exact test with a Bonferroni correction. Of the 103 polymorphic sites in T1, only three pairs were in significant LD. Similarly, in T2 with 118 polymorphic sites, only the number of pathogen genomes in the sample is limited, and the use of PCR can obscure the quality of the sampling by creating a large amount of DNA from a relatively small number of starting templates. This can create artificial homogeneity, inflate estimates of segregating genetic variation, skew the distribution of alleles in the population, and introduce artificial diversity.

[00161] We have developed a strategy that allows each sampled template to be tagged with a unique ID by a primer that has a degenerate sequence tag incorporated during the primer oligonucleotide synthesis (Fig. 10). This tag can then be followed through the PCR and the deep sequencing protocol to identify sequencing over-coverage (resampling) of the individual viral templates. Because the Primer ID allows for the identification of over-coverage, this can then be used to create a consensus sequence for each template, avoiding both PCR-related errors and

IDT Exhibit 1020.1849

sequencing errors (Fig. 11). In addition, the number of different Primer IDs reflects the number of templates that were actually sampled. This allows a realistic assessment of the depth of population sampling and makes it possible to apply a more rigorous analysis of minor variants by correcting the allelic skewing during the PCR.

[00162] We tested the Primer ID approach by sequencing the HIV-1 protease coding domain at three time points in a subject who was intermittently exposed to a protease inhibitor between the second and third time points. A key feature of our approach is the removal of fortuitous errors and accounting for resampling, which results in a dramatic reshaping of the original data set of 72,162 reads. Other approaches that rely on statistical modeling have been developed to deal with the problem of high sequencing error rates associated with deep sequencing technologies. The use of the Primer ID to create consensus sequences resulted in the removal of 80% of the unique sequence polymorphisms (defined as a change in the consensus without regard to frequency of appearance) in the data set. Similarly, allelic skewing was dramatic among the sampled sequences, in most cases ranging from 2- to 15-fold but going up to nearly 100-fold. Although the Primer ID reveals such skewing and helps correct it, this is clearly a poorly controlled feature of PCR amplifications that can dramatically affect the observed abundance of complex populations, especially the minor variants. Allelic skewing may still persist if the cDNA primer or the upstream PCR primer binds differentially among the templates, or if cDNAs enter the PCR amplification in later rounds and are discarded because they do not result in at least three reads to allow a consensus sequence to be formed. Also, residual misincorporation errors by RT and in the first round of PCR synthesis still limit the interpretation of mutations that occur in the range of 0.01–0.1%. This problem is not overcome with larger numbers of sequences. Given the low diversity in these samples, we removed all substitutions that appeared once because their number approximated the expected number of residual sequence errors, and this resulted in a sensitivity of detection in the range of 0.1% for SNPs that appeared above the frequency of the residual sequence error rate.

[00163] Using the Primer ID approach, we were able to describe a number of features of the protease sequence population. First, a pooled analysis of two time points six months apart showed that the variants present at greater than 0.5% in abundance made up two-thirds of the total population but represented only 4% of unique genome sequences and contained only 7% of the total unique sequence polymorphisms. About 60% of the diversity was stable over both time

IDT Exhibit 1020.1850

points, with synonymous SNPs maintained at a significantly higher proportion in the two time points than nonsynonymous SNPs. Only 18% of the total diversity represented nonsynonymous SNPs that were present at both time points. However, our ability to assess persistence of these sequences is limited by the depth of sampling, although we feel we are approaching the practical limit of sampling with this technology. We observed nonviable substitutions and estimate that most of the SNPs that appeared once were the result of remaining method error. We found no pattern of conserved linkage among these SNPs, consistent with high levels of recombination across the population.

[00164] Although the overall measurement of diversity (π) was similar between the first two time points, we noted that the biggest changes in SNP abundance between the two time points were in three synonymous codon positions (L24L, K70K, and G73G). These dynamic increases made these SNPs part of a larger group of SNPs that accounted for 51% of the total sequences that were otherwise identical to the consensus sequence (Q18Q, L19I, L24L, K70K, G73G, and Q18Q/L19I/L24L'). These SNPs also overlapped the major SNPs that defined subgroups of the resistant variants (L19I; L19V; G16G/L19V). We considered the possibility that there was a unifying feature of these SNPs. We found such a feature in that all of these SNPs, both coding and noncoding, result in changes in two relatively large alternative ORFs that lie at the 5′ and 3′ ends of the pro gene. Alternative reading frames have been suggested to generate cryptic CTL epitopes. In this scenario, these abundant SNPs would represent various escape mutants. Such selective pressures could explain the dynamic behavior of several of these SNPs between the first two time points.

[00165] After intermittent exposure to the protease inhibitor ritonavir, we were able to identify six independent lineages of drug resistance mutations. With the intermittent exposure in this particular subject, it was possible to see the major V82A lineage most often seen with ritonavir resistance, but also significant populations of I84V and L90M. We also saw minor populations of V82I, V82L, and V82F. This mixed population of resistant lineages likely represents the early stages of the evolution of resistance, a conclusion supported by the minor appearance of the L63P compensatory mutation and the complete absence of I54V, which is an often seen compensatory mutation for V82A. We saw few examples of genomes with multiple resistance mutations, although these would be expected after more extensive selection. We and others have previously examined viral sequences that have been collected in large databases.

Typically, these sequences represent the single predominant sequence within an individual, and the use of these sequences allows for assessment of interperson diversity. In the future, it will be an interesting exercise to compare the conclusions reached by examining viral diversity within a person to viral diversity between people; however more intraperson diversity needs to be measured at this level of detail to allow comparison of inter- versus intraperson diversity.

[00166] The presence of preexisting drug-resistant variants and their role in therapy failure is of great interest, and accurate, deep sampling of a viral population can add significantly to our understanding of this question. We were able to detect several examples of drug-resistance mutations but only at a very low level. Our ability to reliably detect these mutations is limited to those that appear at a frequency of 0.1–0.2%, limited in part by the low overall diversity in the population. We were able to see examples of mutations that are typically seen only in the presence of drug selection. However, the detection was usually as one genome at two time points or two genomes at one time point. This was also the level of detection of active site mutations in the protease and of termination codons, which must represent either transient viral genomes or residual misincorporation errors. In two cases, we were able to observe the resistance mutation (V82A and L90M) at pretherapy time points linked to the same polymorphisms that were present on the variant that grew out during drug exposure. Thus, although it is likely that we are detecting relevant preexisting drug-resistant variants, these are at the limit of detection and, if they are maintained at a steady-state level, it is well under 0.5% abundance.

[00167] Most protocols of high throughput sequencing technologies still require an initial quantity of DNA that necessitates an upfront PCR step for many applications. The use of a Primer ID will help clarify the sequencing products in any strategy that uses an initial PCR step with its attendant error rate, recombination, and resampling. In an independent effort Kinde et al. have described an analogous approach in another deep sequencing of individual templates before PCR and subsequent sequence analysis will be essential for understanding the true complexity and diversity of genetically dynamic populations.

[00168] Materials and Methods

[00169] Viral RNA was isolated from blood plasma using the QIAmp Viral RNA kit (Qiagen). cDNA was generated using SuperScript III Reverse Transcriptase (Invitrogen) using the primer (with Primer ID) as described. Following the reaction, RNA in hybrid was removed by RNaseH treatment (Invitrogen). Unincorporated cDNAprimer was removed, and the cDNAproduct

IDT Exhibit 1020.1852

amplified by PCR. Sequencing was done using the 454 platform (Roche).

[00170] vRNA Extraction and cDNA Synthesis. Viral RNA was extracted from three plasma samples taken longitudinally from an individual infected with subtype B HIV-1 who was participating in a protease inhibitor efficacy trial (M94-247). Two samples were collected at ~6 mo before and immediately before the addition of the protease inhibitor ritonavir to a failed therapy regimen (plasma viral loads of 285,360 copies of viral RNA/mL and 321,100 copies of viral RNA/mL, respectively), and one sample was collected during ritonavir therapy (at approximately 2 months on therapy, 349,920 copies of viral RNA/mL) but during a time of apparent intermittent compliance. For each plasma sample, vRNA was extracted from pelleted (25,000 × g for 2 h) viral particles using the QiaAMP Viral RNA kit (Qiagen). Approximately 10,000 copies of viral RNA from each sample were present in the cDNA synthesis reaction as described. The tagging primer used was, 5′-GCCTTGCCAGCACGCTCAGGCCTTGCA(BARCODE)CGNNNNNNNNTCCTGGCTTTAATTTTACTGGTACAGT-3′. (SEQ ID NO. 2) The barcode represented TCA, GTA, and TAT for study days 58, 248, and 303, respectively. The 3′ end of the tagging primer targeted downstream of the protease coding domain (HXB2 2568–2594). The oligonucleotides were purchased from IDT and were purified by standard desalting.

[00171] Amplification of Tagged Sequences. The single-stranded cDNA was column purified using the PureLink PCR Purification Kit (Invitrogen), using Binding Buffer HC (high cutoff) and three washes to remove the cDNA primer. Primer removal was verified by electropherogram analysis using an Experion HighSense RNA microfluidic chip (Bio-Rad Laboratories). Samples were amplified by nested PCR, using upstream primers 5′-GAGAGACAGGCTAATTTTTTAGG-3′ (HXB2 2071–2093) (SEQ ID NO. 3) and 5′-ATAGACAAGGAACTGTATCC-3′ (HXB2 2224–2243) (SEQ ID NO. 4); the downstream primers targeted the 5′ portion of the cDNA tagging primer 5-GCCTTGCCAGCACGCTCAGGC-3′ (SEQ ID NO. 5) then 5′-CCAGCACGCTCAGGCCTTGCA-3′(SEQ ID NO. 6). The PCR was done using Platinum Taq DNA Polymerase High Fidelity (Invitrogen). Each reaction contained 1× High Fidelity PCR Buffer, 0.2 mM of each dNTP, 2 mM MgCl2, 0.2 μM of each primer, 1.5 units of Platinum Taq DNA Polymerase. The purified cDNA template was split to 2 × 50 μl for the first round PCR, and 1 μl of the purified first round product was used for nested PCR. Samples were denatured at

IDT Exhibit 1020.1853

94 °C for 2 min, followed by 30 cycles of 94 °C for 15 s, 55 °C for 30 s, 68 °C for 1 min, and a final extension at 68 °C for 5 min.

[00172] Samples were column purified after the first round of PCR using the MinElute PCR Purification Kit (Qiagen), and eluted into 30 μl of buffer EB. Second round PCR product was gel purified using a 2% agarose gel and QIAquick gel extraction kit (Qiagen), with incubation of the solubilization buffer at room temperature. DNA was quantified by Qubit fluorometer using dsDNA High Sense assay (Invitrogen). Product generation, quality, and primer removal for both PCR rounds was verified using an Experion DNA microfluidic chip (Bio-Rad).

[00173] 454 Pyrosequencing. Tagged samples from the three time points were combined and sequenced on the 454 GS FLX platform with XLR70 Titanium sequencing chemistry as per the manufacturer's instructions (Roche) but with under-loaded beads to minimize signal crosstalk. Sequences were processed from two independent 454 GS FLX Titanium runs (1/8th of a plate each).

[00174] Bioinformatic Pipeline for Raw Sequence Processing. A suite of programs was written to filter and parse raw 454 sequencing reads. In short, first, each sequence was placed in the correct orientation compared with a reference pro gene sequence. This alignment was then used to identify insertions or deletions caused by the 454 sequencing of homopolymers. When there was an insertion, the extraneous base was excised from the sequence. Deletions retained were largely resolved in the construction of the consensus sequence. Second, they were evaluated for the presence of the cDNA primer 5′ tail, with the encoded information (barcode and Primer ID) exactly spaced. Third, individual sequences were binned by their barcodes, and then by their Primer ID. Fourth, sequences were trimmed to the protease coding domain (pro gene). Within a barcode bin, when three sequences contained an identical Primer ID, a consensus sequence was called by majority rule. Ambiguous nucleotide designations were used when there was a tie (Fig. 6B). Sequences are available under GenBank accession numbers JN820319-JN824997.

[00175] Population Analyses. A $\chi 2$ test was used to test for significance changes in allele frequency between the two untreated time points. To control for multiple testing, collective assessment of significance was based on False Discovery Rate analysis (FDR = 0.05). Tests for linkage disequilibrium were computed by DnaSP v.5.10.01 (4). These tests were done on filtered populations devoid of sequences containing ambiguities or gaps. Tests for neutrality were

IDT Exhibit 1020.1854

computed by DnaSP and R (5) on filtered populations devoid of sequences containing ambiguities. Gaps and alleles represented by a single sequence were reverted to the consensus. Beta P values were calculated against the null hypothesis that D = 0, assuming that D follows a beta distribution after rescaling on [0, 1].

[00176] Diversity across and within populations was computed through customized bioinformatics suites. Unfiltered sequences were used in the analysis, and ambiguities, gaps, and alleles represented by a single sequence were removed from the final tabulation (Fig. 2A-2B and Table 1).

[00177] SNPs were graphically displayed through the Highlighter tool (www.hiv.lanl.gov).

[00178] Phylogenetic Resolution of Sequences. The phylogeny for the population of consensus sequences from all three time points was resolved using two alternative methods and on populations devoid of sequences containing gaps or ambiguities. When only one example of a SNP was present across all sequences, it was converted to the consensus on the assumption that it was likely generated by residual method error. First, the Neighbor-Joining tree using the Kimura translation for pairwise distance and a bootstrap of 100 iterations was constructed with QuickTree v.1.1.

[00179] Second, Maximum likelihood phylogeny was inferred using the PHYLIP package, version 3.69, and the calculated phylogeny is available upon request. The PHYLIP program seqboot was used to create 100 bootstraps. Resulting bootstraps were submitted to the PHYLIP program dnamlk for maximum likelihood inference subject to a strict molecular clock. The consensus tree of all boostrap results was constructed using the PHYLIP program consense.

[00180] Both phylogenetic trees were visualized by a customized modification of Figtree v.1.3.1.

[00181] Additional Considerations. Degenerate base synthesis in the cDNA primer. The degenerate bases (Primer ID) in the cDNA synthesis primer were randomized using machine mixing during oligonucleotide synthesis. All four DNA phosphoramidite monomer bases are introduced to the column at the same time, but due to slight differences in binding or delivery, a strict equimolar ratio of dA, dT, dC, and dG may not be realized, resulting in a Primer ID bias (Fig. 10). When there is a Primer ID bias, there is an increased probability that a particular Primer ID will tag multiple templates because sequence tags with over-represented nucleotides will be more abundant than sequence tags with under-represented nucleotides. Because the bias

IDT Exhibit 1020.1855

is amplified over the length of the Primer ID the skewing can be significant. We observed a bias of ~40% dC in one of our Primer ID syntheses, and at the extreme dC_8 would be present at a 40-fold excess over the sequence frequency expected if all nucleotides were present at equal concentration. Similarly, we observed 15% dA in one synthesis which would result in a 60-fold decrease in the expected frequency of dA_8. This appears to be the result in variation in primer synthesis because the bias varied in the different barcode bins and therefore was not a constant feature of the cDNA synthesis step. However, this phenomenon is somewhat mitigated when a consensus sequence is formed, as whatever template was resampled to the greater extent within a mixed Primer ID population would be recorded.

[00182] Frameshift mutations. Pyrosequencing commonly miscalls homopolymers, resulting in a frameshift mutation by either calling too few or too many nucleotides in the homopolymer run. The HIV-1 pro gene contains several homopolymeric stretches. We took advantage of a known length (conserved in a coding region) to align individual reads against a reference sequence. Given this bias we removed the insertions to retain the correct length of the homopolymer run. Deletions were retained. Through consensus sequence generation, the deleted base was often recovered when the other resampled reads contained the missing base. Although consensus sequence generation reduced the spread and frequency of deletions in the final resolved, consensus reads, it did not eliminate deletions altogether (Fig. 11).

Example 2. Accurate sampling and deep sequencing HIV-1 protease using a Primer ID

[00183] Viruses can create complex genetic populations within a host, and Deep sequencing technologies offer the opportunity to extensively sample these populations. However, features of these techniques limit their application, particularly when a polymerase chain reaction (PCR) step precedes the sequencing protocol.

[00184] Typically, an unknown number of templates are utilized in initiating the PCR amplification and this can lead to unrecognized sequence resampling. PCR-mediated recombination can create artifactual linkage and disrupt real linkage. Finally, misincorporation during PCR and errors during the sequencing protocol can create artificial diversity.

[00185] We have solved this by including a random sequence tag in the initial primer such that each template receives a Primer ID. After sequencing, repetitive identification of the Primer ID reveals sequence resampling, which can then be used to create an accurate consensus

IDT Exhibit 1020.1856

sequence for each template. The resulting population of consensus sequences directly identify the initial sampled templates. The use of Primer IDs can directly correct for unrecognized sequence resampling, PCR-mediated recombination, differential template amplification, polymerase nucleotide misincorporation, and sequencing error.

[00186] We applied this approach to the HIV-1 protease (pro) gene to view the distribution of sequence variation within a complex population. We identified major and minor polymorphisms within coding and noncoding positions. In addition, we observed dynamic changes across the population during intermittent drug exposure, including the emergence of resistant alleles.

[00187] **Methods:** A population of cDNA synthesis primers were designed to contain a string of eight degenerate nucleotides (65,536 distinct sequence combinations, or Primer IDs), and an *a priori* selected three nucleotide barcode (Fig. 12A). After cDNA synthesis, the nonspecific 5' end of the primer was used for enrichment of tagged sequences by nested PCR.

[00188] HIV-1 RNA was extracted from blood plasma. Protease from two pre-Ritonaivr therapy and 1 post-intermittent Ritonavir therapy was tagged and sequenced (Fig. 12B). Approximately 10,000 copies of HIV-1 viral RNA were tagged downstream of proteases, amplified, and sequenced on the 454 GS FLX Titanium. FIG 13B displays a summary of the resolved sequences.

[00189] A bioinformatics pipeline was developed to assess raw sequences for tagged, full-length protease (Fig. 14), and when three or more sequences contained an identical Primer ID, a consensus sequence was generated (Fig. 15). After direct error correction through filtering and processing of Primer IDs, the allelic frequency of the *pro* gene populations were assessed (Fig. 13A).

[00190] **Discussion:** We have developed a strategy that allows each sampled template to be tagged with a sequence ID by a primer that has a degenerate sequence tag incorporated during synthesis. This tag can then be followed through the deep sequencing protocol to identify sequencing over-coverage of the individual templates. The over-coverage can be used to create a consensus sequence for each template, avoiding both PCR-related errors and sequencing errors. In addition, the number of different Primer IDs reflects the number templates that were actually sampled, allowing a realistic assessment of the quality of sampling which will make it possible to apply a more rigorous analysis of minor variants. In many settings, especially when working with pathogenic agents in clinical samples, the number of templates can be limiting and the use

IDT Exhibit 1020.1857

of PCR can obscure the limited quality of the sampling. This problem is solved by tagging each template as the first step then simply counting the number of templates that were actually used as part of the last step.

[00191] Accurately resolving viral templates comes at a cost of sequencing depth, as depth is a direct function of template number, but we have documented the nature and degree of allelic fluctuation throughout protease down to the 0.1% level of resolution. We showed synonymous and nonsynonymous fluctuation over time in an untreated environment, and when in an environment of intermittent selection from the protease inhibitor, Ritonavir. We detected the drug resistant allele, V82A, as a minor allele in the untreated population, and detected this exact variant comprising over a quarter of the rebounding population (Fig. 16). These results demonstrate the utility of applying this approach to the detection of minor variants in the context of HIV-1 treatment, and more generally to the question of minor variants in the context of a large genetically complex population.

[00192] **Table 1. Frequency of nonconsensus codons per position**

AApos[a]	AAc[b]	Cc[c]	Cm[d]	Am[e]	T1[f]	T2[g]	T3[h]	T3s[i]	T3r[j]	Cm[k]	T1[l]	T2[m]	T3[n]	T3s[o]	T3r[p]
4	T	ACT	GCT	A		0.06	0.05		0.09						
5	L	CTT	CCT	P	0.12		0.05	0.14							
7	Q	CAA								CAG	0.35	0.12	0.09	0.14	0.09
8	R	CGA								CGG	0.12		0.05	0.14	
9	P	CCC													
10	L	CTC	TTC	F		0.19				CTT		0.19			
11	V	GTC	ATC	I	0.23	0.25				GTT		0.12			
14	K	AAG	AGG	R		0.12				AAA	1.17	0.19	0.59	0.29	0.72
15	I	ATA	GTA	V	1.17	0.12	0.14	0.14	0.18	ATC			0.09		0.18
16	G	GGG	AGG	R		0.06	0.05		0.09	GGA	2.22	3.54	38.86	17.70	45.97
17	G	GGG	AGG	R			0.09	0.29		GGA	0.35	0.19	0.18	0.43	0.09
18	Q	CAA	GAA	E	0.23	0.12				CAG	18.55	21.75	6.46	12.81	3.53
19	L	CTA	ACA	T	0.47										
			ATA	I	19.25	19.83	20.42	19.28	24.98	TTA	0.12	0.19	0.09	0.29	
			GTA	V	3.38	5.66	46.00	25.61	52.76						
20	K	AAG	AGG	R	0.12	0.12	0.05		0.09	AAA		0.31	0.86	0.29	1.27
21	E	GAA								GAG	0.12	0.06	0.05	0.14	
22	A	GCT								GCC	0.47	0.44	0.27	0.58	0.18
										GCG	0.23				
23	L	CTA								CTG		0.19			
24	L	TTA								CTA	0.35	5.72	1.31	2.16	0.63

IDT Exhibit 1020.1858

AApos[a]	AAc[b]	Cc[c]	Cm[d]	Am[e]	T1[f]	T2[g]	T3[h]	T3s[i]	T3r[j]	Cm[k]	T1[l]	T2[m]	T3[n]	T3s[o]	T3r[p]
										TTG	12.49	0.81	0.59	1.01	0.27
25	D	GAT	GGT	G	0.12	0.12				GAC	0.23	0.93	0.05	0.14	
26	T	ACA	GCA	A		0.12									
27	G	GGA								GGG	0.12	0.06			
28	A	GCA								GCG	0.12		0.09	0.14	
29	D	GAT	AAT	N	0.12		0.05		0.09	GAC	0.23	0.19			
30	D	GAT								GAC		0.06	0.09	0.14	0.09
31	T	ACA								ACG		0.12			
32	V	GTA								GTG		0.25			
33	L	TTA	GTA	V	0.47	0.06				CTA		0.25	0.14	0.29	0.09
										TTG	0.35	0.12	0.14	0.43	
34	E	GAA	GGA	G		0.12	0.05		0.09	GAG	0.12		0.05	0.14	
			CAA				0.09								
35	E	GAA	AAA	K	0.12	0.06	0.09	0.14							
36	M	ATG	ATA	I	0.82	0.81	0.27	0.43	0.27						
37	N	AAT	AGT	S		0.19	0.05			AAC		0.06	0.05	0.14	
			GAT	D	2.33	2.30	0.95	0.86	1.27						
38	L	TTG								TTA	0.23	0.62	0.05		0.09
39	P	CCA								CCT	0.23				
40	G	GGA								GGG	0.12	0.12			
41	K	AAA	AGA	R		0.06	0.18	0.14	0.27	AAG	4.08	1.43	0.50	1.15	0.27
42	W	TGG	CGG	R	0.12	0.06									
			TAG	_	0.12		0.05		0.09						
			TGA	_			0.14		0.27						
43	K	AAA	AGA	R		0.06	0.05		0.09	AAG	0.35		0.14	0.14	0.18
44	P	CCA								CCG		0.06	0.23	0.43	0.18
45	K	AAA	AGA	R	0.12	0.12	0.05		0.09	AAG	0.58	0.99	0.41	1.29	
46	M	ATG	ATA	I		0.12	0.09	0.14	0.09						
48	G	GGA	GAA	E			0.14	0.14	0.18	GGG	0.35	0.19			
49	G	GGA	GAA	E	0.12	0.06	0.05		0.09	GGG	0.23	0.12			
50	I	ATT								ATC	0.12	0.12			
51	G	GGA								GGG	0.12	0.06			
52	G	GGT	AGT	S	0.12	0.06	0.05	0.14		GGA		0.06	0.05	0.14	
										GGC	0.12	0.31	0.09	0.14	0.09
										GGG			0.14	0.43	
53	F	TTT								TTC	0.70		0.05	0.14	
54	I	ATC	ACC	T	0.12	0.06	0.05		0.09	ATT	0.35	0.06	0.14	0.14	
55	K	AAA	AGA	R	0.12		0.05		0.09	AAG	0.12	0.06			
56	V	GTA	ATA	I	0.12		0.05	0.14		GTG		0.75	0.14	0.14	0.18
57	R	AGA	AAA	K	0.23					AGG	0.23	0.87	0.14	0.14	0.18

IDT Exhibit 1020.1859

AApos[a]	AAc[b]	Cc[c]	Cm[d]	Am[e]	T1[f]	T2[g]	T3[h]	T3s[i]	T3r[j]	Cm[k]	T1[l]	T2[m]	T3[n]	T3s[o]	T3r[p]
58	Q	CAG	TAG	_			0.05		0.09	CAA	0.93	0.50	0.23	0.29	0.27
60	D	GAT	AAT	N		0.12									
			GGT	G		0.12									
61	Q	CAA	CGA	R	0.12	0.06	0.05	0.14		CAG		0.19	0.23	0.58	
			TAA	_	0.12	0.06	0.05		0.09						
62	I	ATA	GTA	V	0.35	0.06									
63	L	CTC	CCC	P	0.12		0.41	0.58	0.36	CTT	11.32	5.41	1.27	2.88	0.45
64	I	ATA	GTA	V	1.05	0.06	0.09		0.18						
			ATG	M	0.23		0.05	0.14							
65	E	GAA	AAA	K			0.09	0.14	0.09	GAG	0.35	0.06	0.05		
66	I	ATC								ATA		0.25	0.18	0.58	
										ATT	1.98	0.19			
67	C	TGT								TGC	0.35	0.12	0.05	0.14	
68	G	GGA								GGG	0.23	0.12	0.05	0.14	
69	H	CAT	TAT	Y	0.23	0.06	0.09	0.14		CAC	0.82	0.31	0.14	0.29	0.09
70	K	AAA	CAA	Q	0.47	0.12	0.41	1.29		AAG	3.27	10.88	15.27	6.62	25.34
71	A	GCT	ACT	T		0.12	0.09								
72	I	ATA	GTA	V	0.12	0.12									
73	G	GGT								GGC	0.47	18.09	7.05	15.68	3.62
74	T	ACA								ACG	0.23	0.12			
75	V	GTA	ATA	I	0.23	0.06	0.05			GTG	1.87	0.99	0.27	0.43	0.27
			GCA	A			0.09		0.18						
76	L	TTA								CTA		0.12	0.09		0.18
										TTG	0.93	0.62	0.27	0.43	0.18
77	V	GTA	ATA	I	0.23	0.56	0.72	2.01	0.18	GTG	0.82	0.62	0.23	0.58	
			CTA	L			0.14								
78	G	GGA								GGG	1.17	1.24	0.09	0.14	
79	P	CCT								CCC	1.17	0.31	0.54	1.29	0.18
81	P	CCT								CCC	0.12	0.19			
										CCG	1.52	0.44			
82	V	GTC	ATC	I		0.06	1.27	3.60		GTA	0.35	0.31	0.05		
			CTC	L		0.06	1.08	3.45		GTT	1.05	0.75	0.41	1.01	
			GCC	A		0.12	49.89		99.91						
			TTC	F			0.14	0.43							
83	N	AAC	AGC	S	0.12		0.05		0.09	AAT	8.17	6.40	3.62	4.75	4.16
84	I	ATA	GTA	V			5.15								
85	I	ATT								ATA		0.12	0.05	0.14	
										ATC	0.12	0.12	0.05		
86	G	GGA								GGG		0.12			

IDT Exhibit 1020.1860

AApos[a]	AAc[b]	Cc[c]	Cm[d]	Am[e]	T1[f]	T2[g]	T3[h]	T3s[i]	T3r[j]	Cm[k]	T1[l]	T2[m]	T3[n]	T3s[o]	T3r[p]
										GGT	0.12	0.06			
87	R	AGA	AAA	K	0.12	0.06	0.05		0.09	AGG	0.58	0.37	0.05	0.14	
			GGA	G		0.06	0.09	0.14	0.09						
88	N	AAT								AAC	0.35	0.93			
89	L	CTA	ATA	I		0.12				CTG	1.17	0.68	1.36	1.87	1.54
										TTA	1.98	0.56	1.27	0.14	2.44
90	L	TTG	ATG	M	0.12		13.56		0.09	CTG	0.47		0.09	0.14	0.09
			TCG	S	0.12		0.05		0.09	TTA	0.47	0.19	0.14	0.43	
91	T	ACT	GCT	A		0.06	0.05		0.09	ACC	0.12	0.06	0.09	0.14	0.09
										ACG	0.12	0.12	0.77		1.54
92	Q	CAG								CAA	0.23	0.19	0.14		
93	I	ATT	CTT	L	0.12	0.06				ATC	0.23		0.09	0.14	0.09
94	G	GGT	GAT	D	0.12	0.06				GGA	0.23				
										GGC	1.28	0.25	0.50	1.29	0.18
										GGG	0.23	0.06	0.09	0.14	
95	C	TGC								TGT	0.70	0.12	0.14		0.27
96	T	ACT								ACA	0.12		0.09	0.14	0.09
										ACC	0.70	0.12	0.23	0.43	0.09
										ACG		0.06	0.05	0.14	
97	L	TTA								CTA	0.58		0.05	0.14	
										TTG	0.12	0.25	0.27	0.43	0.27

Only positions of diversity and SNPs that were represented by more than 1 sequence are shown.

Consensus

AApos[a] Amino acid position, protease.

AAc[b] Consensus amino acid in untreated population.

Cc[c] Consensus codon in untreated population.

Nonsynomous

Cm[d] Coding nonconsensus amino acid.

AAm[e] Coding nonconsensus codon.

T1[f] Frequency of SNP in first untreated time point.

T2[g] Frequency of SNP in second untreated time point.

T3[h] Frequency of SNP in third time point, treated.

T3[i] Frequency of SNP in third time point, treated, susceptible population (not V82A, I84V, L90M).

T3r[j] Frequency of SNP in third time point, treated, population containing major ritonavir resistant variant V82A.

Synomous

Cm[k] Silent nonconsensus codon.

T1[l] Frequency of SNP in first untreated time point.

T2[m] Frequency of SNP in second untreated time point.

T3[n] Frequency of SNP in third time point, treated.

T3s[o] Frequency of SNP in third time point, treated, susceptible population (not V82A, I84V, L90M).

$T3r^p$Frequency of SNP in third time point, treated, population containing major ritonavir resistant variant V82A.

[00193] Table 2. Summary of nucleotide variation in sampled time points

Variable	T1	T2	T3	T3s	T3r
No.of sequences	810	1449	1925	547	970
No.of polymorphic	104	115	110	71	69
(segregating)sites					
Total number of mutations	115	129	121	75	73
Average number nt differences, k	2.38809	2.33683	3.08838	2.43819	2.05962
Nucleotide diversity,π	0.00804	0.00787	0.01040	0.00821	0.00693
Theta (per sequence)	14.29822	14.63943	13.51412	10.31864	9.25678
Theta (per site)	0.04814	0.04929	0.04550	0.03474	0.03117
Tajima'sD	−2.3541	−2.3164	−2.0937	−2.1606	−2.1209
Beta P value	0.0013	0.0014	0.0070	0.0065	0.0071

T1 and T2 are untreated populations, and T3 is a population intermittently exposed to ritonavir monotherapy. Within T3, T3s represents the sensitive (not V82A, I84V, or L90M) portion of the population. T3r represents the major drug resistance clade V82A.

IDT Exhibit 1020.1862

What is claimed is:

1. A method for analyzing a plurality of nucleic acid molecules comprising:
   (a) attaching a plurality of primers comprising a Primer ID to a plurality of nucleic acid molecules in a sample to generate tagged-nucleic acid templates, wherein
      (i) the plurality of nucleic acid molecules comprises 10 or more nucleic acid templates, and
      (j) each tagged nucleic acid template is attached to a unique Primer ID;
   (b) amplifying the tagged nucleic acid templates to produce tagged amplicons; and
   (c) detecting the tagged amplicons, thereby analyzing the plurality of nucleic acid molecules.
2. The method of claim 1, wherein analyzing the plurality of nucleic acid molecules comprises determining, with the aid of a computer processor, PCR resampling in an amplification reaction.
3. The method of claim 1, wherein analyzing the plurality of nucleic acid molecules comprises determining, with the aid of a computer processor, the PCR error rate and/or sequencing error rate.
4. The method of claim 1, wherein analyzing the plurality of nucleic acid molecules comprises determining, with the aid of a computer processor, genetic diversity of the plurality of nucleic acid molecules in the sample.
5. The method of claim 1, wherein analyzing the plurality of nucleic acid molecules comprises identifying, with the aid of a computer processor, genetic variants of the nucleic acid templates.
6. The method of claim 1, wherein analyzing the plurality of nucleic acid molecules comprises determining or screening, with the aid of a computer processor, drug-resistant variants.
7. The method of claim 3, further comprising correcting PCR errors and/or sequencing errors.
8. The method of claim 3, wherein determining the PCR error rate and/or sequencing error rate comprises determining fidelity of a polymerase.
9. The method of claim 3, wherein determining the PCR error rate and/or sequencing error rate comprises determining accuracy of oligonucleotides synthesized in vitro.

IDT Exhibit 1020.1863

10. The method of claim 3, wherein determining the PCR error rate and/or sequencing error rate comprises determining accuracy of the sequencing reaction.
11. The method of claim 1, wherein detecting the tagged amplicons comprises sequencing the tagged amplicons.
12. The method of claim 1, wherein detecting the tagged amplicons further comprises counting a number of different Primer IDs associated with the tagged amplicons, wherein the number of different Primer IDs associated with the tagged amplicons reflects the number of templates sampled.
13. The method of claim 1, further comprising forming a consensus sequence for tagged amplicons comprising the same Primer ID.
14. The method of claim 1, wherein the nucleic acid template comprises a DNA template.
15. The method of claim 1, wherein the nucleic acid template comprises an RNA template.
16. The method of claim 1, wherein the nucleic acid template comprises a non-viral template.
17. The method of claim 4, wherein detecting the genetic variants comprises determining the prevalence of mutations.
18. The method of claim 4, wherein the genetic variant comprises a polymorphism
19. The method of claim 18, wherein the polymorphism comprises a single nucleotide polymorphism.
20. The method of claim 4, wherein the genetic variant comprises a mutation.
21. The method of claim 4, wherein the genetic variant comprises a deletion.
22. The method of claim 4, wherein the genetic variant comprises an insertion.
23. The method of claim 5, wherein the nucleic acid template molecule comprises a viral sequence.
24. The method of claim 5, wherein the nucleic acid template molecule comprises a bacterial sequence.
25. The method of claim 5, wherein the sample is from an individual suffering from a viral infection.
26. The method of claim 5, wherein the sample is from an individual suffering from a bacterial infection.
27. The method of claim 5, wherein the sample is from an individual suffering from cancer.
28. The method of claim 5, wherein the sample is from an individual suffering from an

IDT Exhibit 1020.1864

autoimmune disorder.

29. The method of claim 1, wherein the Primer ID comprises a degenerate sequence.
30. The method of claim 1, wherein the Primer ID is attached to the template by ligation.
31. The method of claim 1, wherein the Primer ID is attached to the template by hybridization.
32. The method of claim 1, wherein the Primer ID is attached to the template through PCR.
33. The method of claim 1, wherein amplifying comprises a PCR-based method.
34. The method of claim 33, wherein the PCR-based method comprises PCR.
35. The method of claim 33, wherein the PCR-based method comprises quantitative PCR.
36. The method of claim 33, wherein the PCR-based method comprises emulsion PCR.
37. The method of claim 33, wherein the PCR-based method comprises droplet PCR.
38. The method of claim 33, wherein the PCR-based method comprises hot start PCR.
39. The method of claim 33, wherein the PCR-based method comprises multiplex PCR.
40. The method of claim 33, wherein the PCR-based method comprises nested PCR.
41. The method of claim 33, wherein the PCR-based method comprises semi-nested PCR.
42. The method of claim 1, wherein amplifying comprises a non-PCR-based method.
43. The method of claim 42, wherein the non-PCR-based method comprises multiple displacement amplification (MDA).
44. The method of claim 42, wherein the non-PCR-based method comprises transcription-mediated amplification (TMA).
45. The method of claim 42, wherein the non-PCR-based method comprises nucleic acid sequence-based amplification (NASBA).
46. The method of claim 42, wherein the non-PCR-based method comprises strand displacement amplification (SDA).
47. The method of claim 1, wherein the Primer ID comprises a nucleic acid sequence.
48. The method of claim 1, wherein the Primer ID comprises a deoxyribonucleic acid sequence.
49. The method of claim 1, wherein the Primer ID comprises a ribonucleic acid sequence.
50. The method of claim 1, wherein the Primer ID comprises 5-100 nucleotides.
51. The method of claim 1, wherein the Primer ID comprises 5-50 nucleotides.
52. The method of claim 1, wherein the Primer ID comprises at least 8 nucleotides.

IDT Exhibit 1020.1865

53. The method of claim 1, wherein the Primer ID comprises a predetermined sequence.

IDT Exhibit 1020.1866

Fig. 1C

Sample	T1	T2	T3
Ritonavir	-	-	+
Total reads	20,429	24,658	27,075
Consensus sequences	857	1,609	2,213

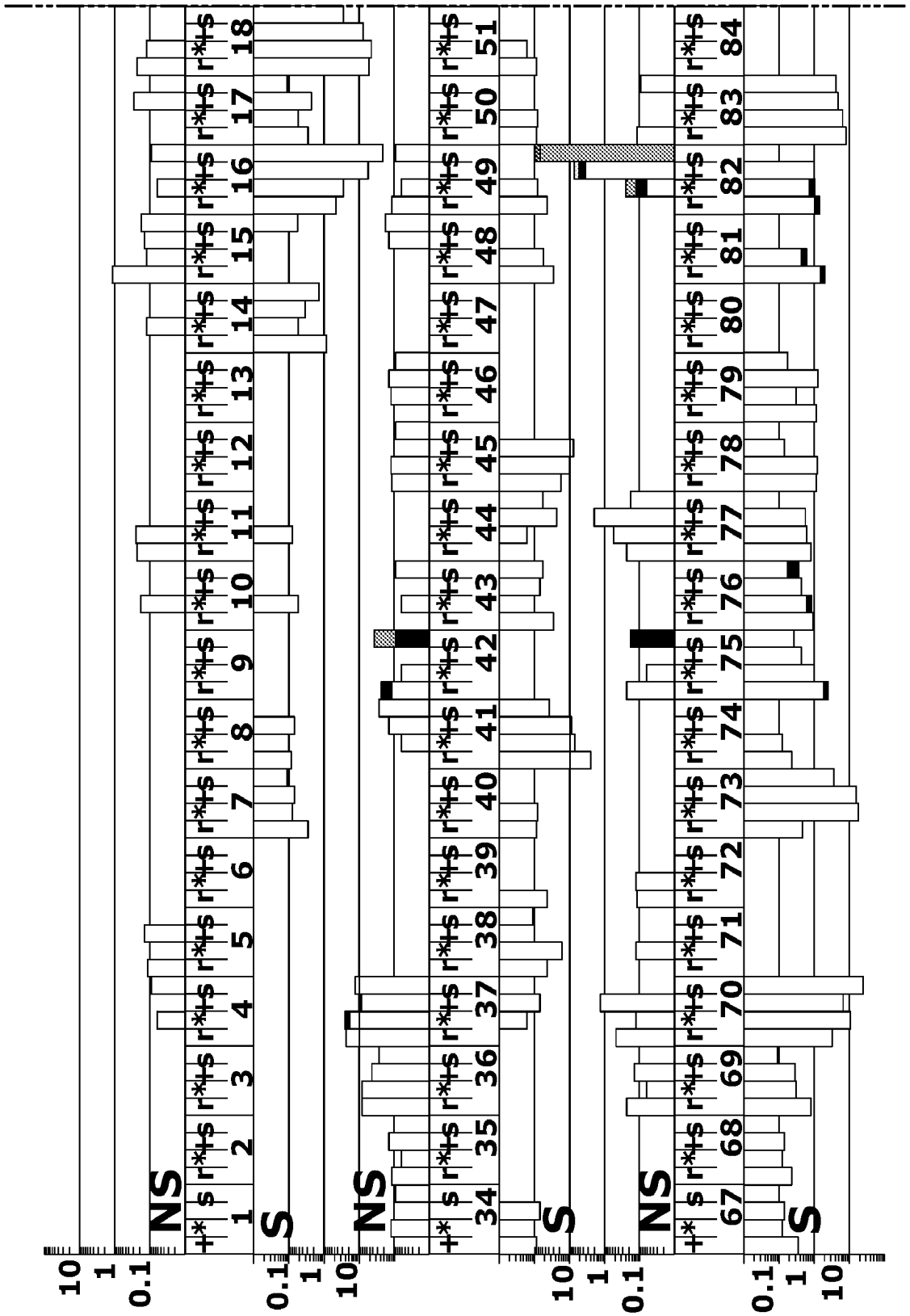


Fig. 2A

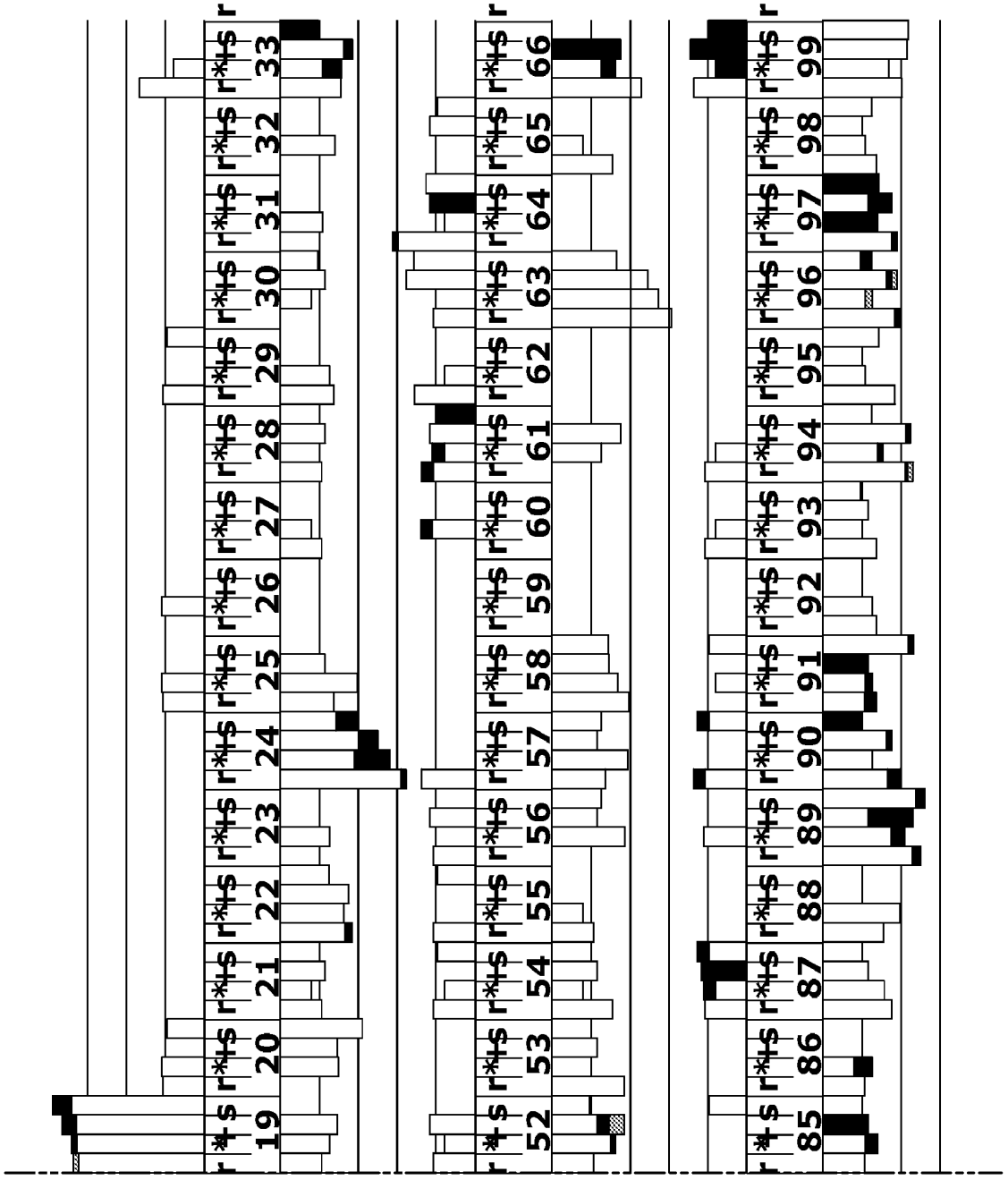


Fig. 2B

IDT Exhibit 1020.1869

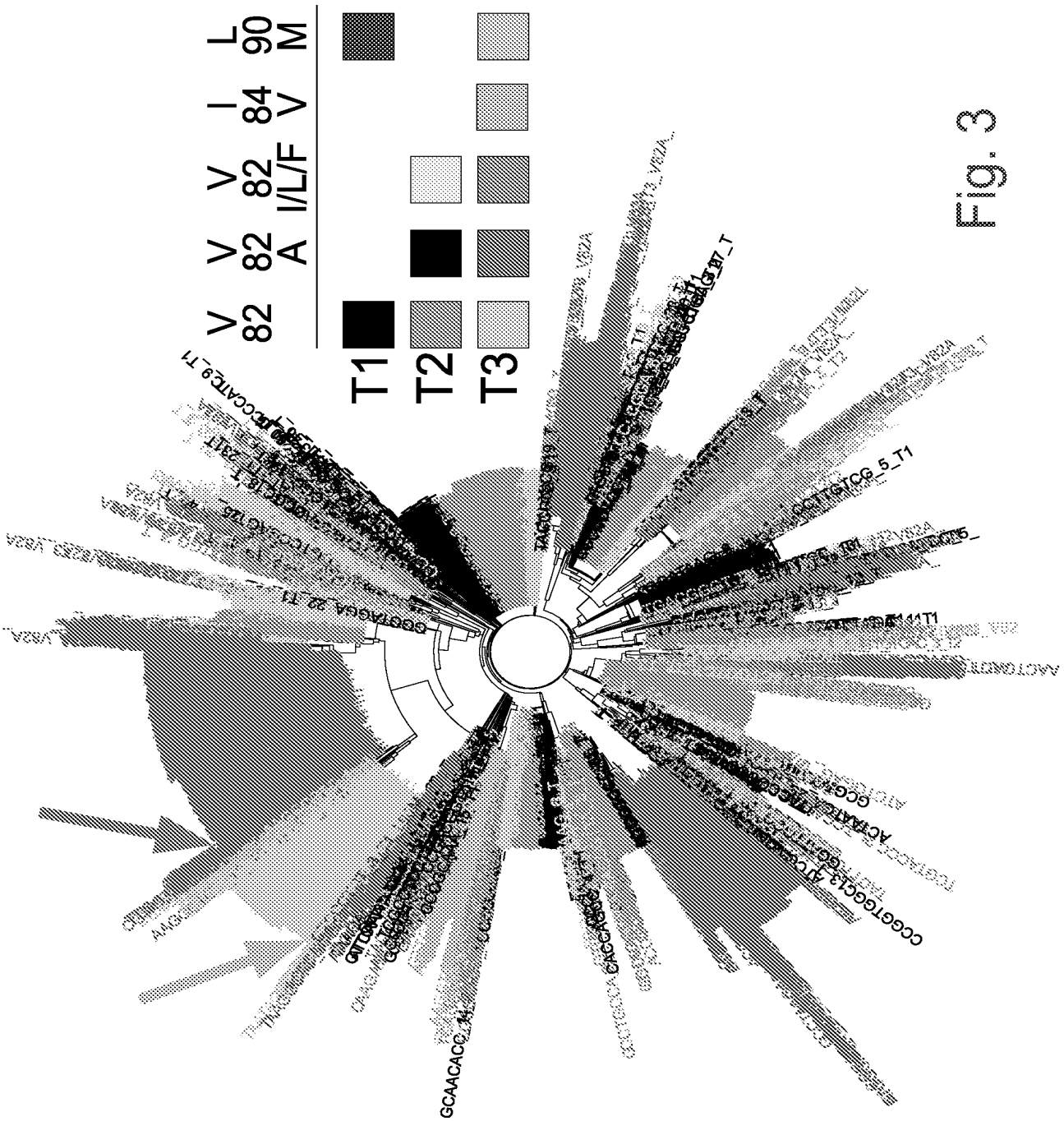


Fig. 3

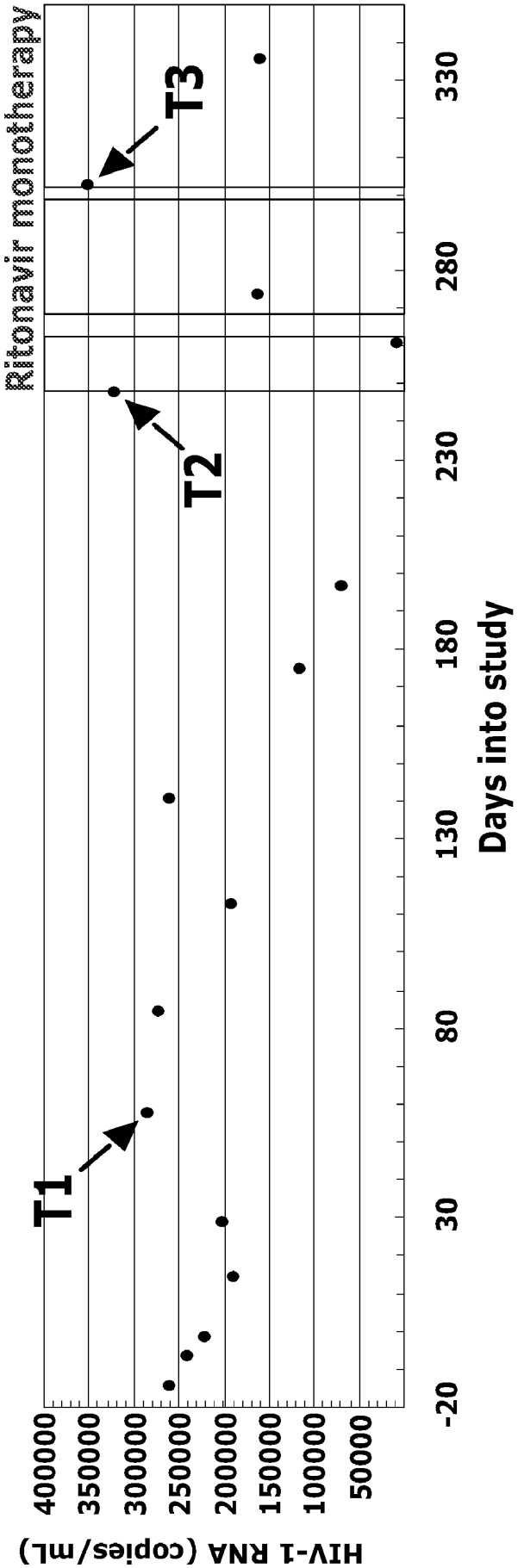

Ritonavir monotherapy
T1
T2
T3
HIV-1 RNA (copies/mL)
400000
350000
300000
250000
200000
150000
100000
50000
-20
30
80
130
180
230
280
330
Days into study

Fig. 4

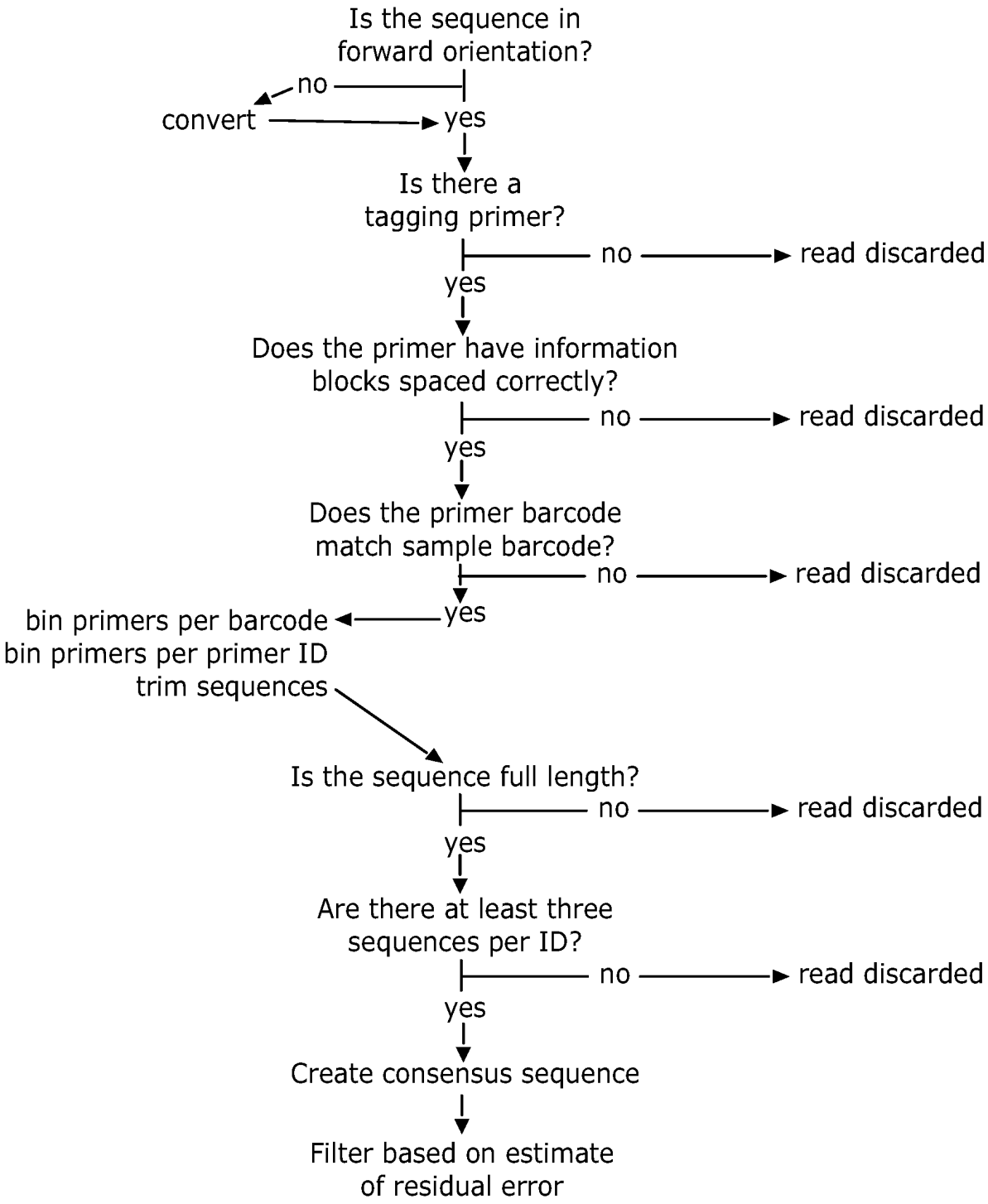

Is the sequence in forward orientation?
no
convert
yes
Is there a tagging primer?
no
read discarded
yes
Does the primer have information blocks spaced correctly?
no
read discarded
yes
Does the primer barcode match sample barcode?
no
read discarded
yes
bin primers per barcode
bin primers per primer ID
trim sequences
Is the sequence full length?
no
read discarded
yes
Are there at least three sequences per ID?
no
read discarded
yes
Create consensus sequence
Filter based on estimate of residual error

Fig. 5

Fig. 6A

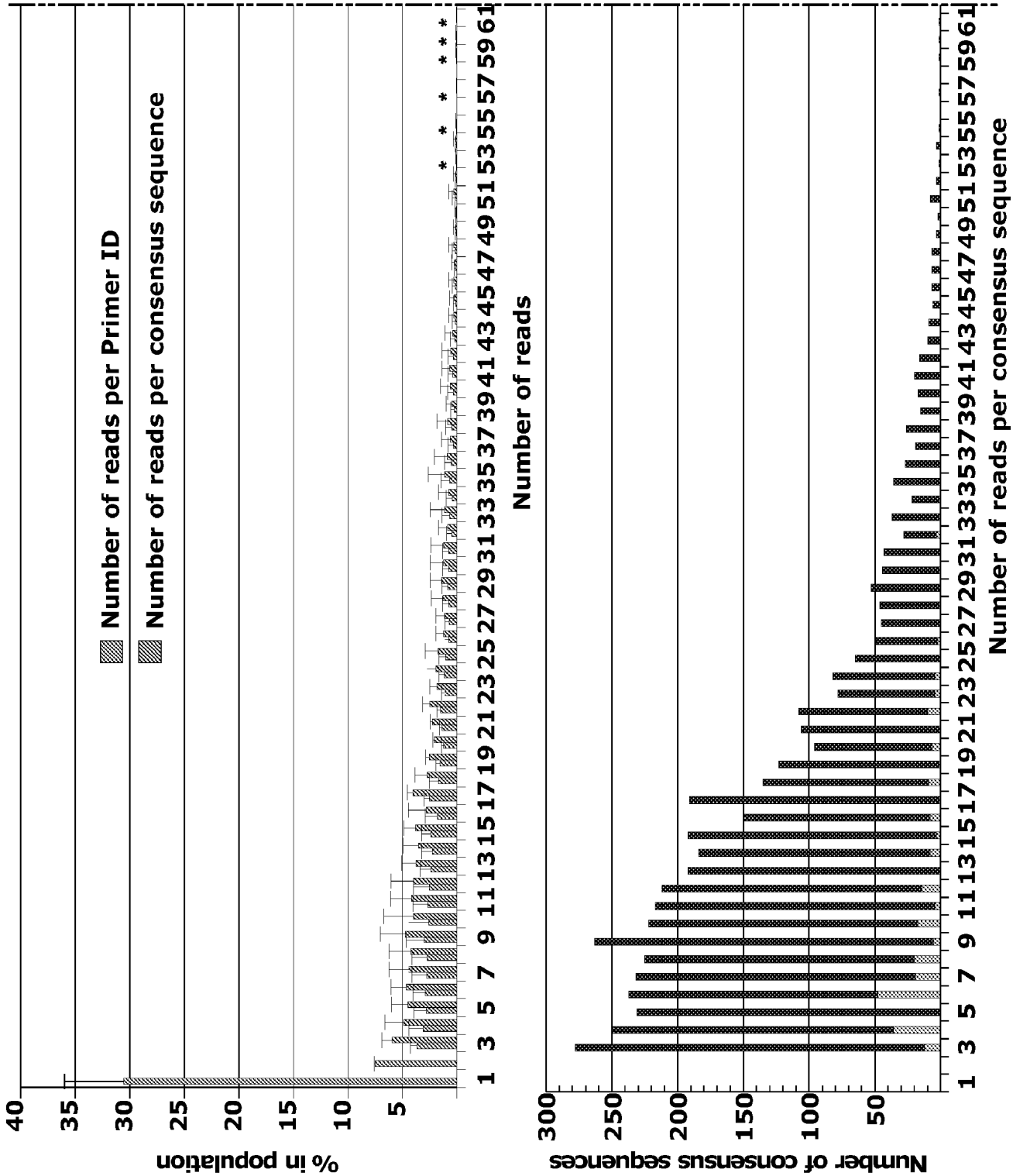
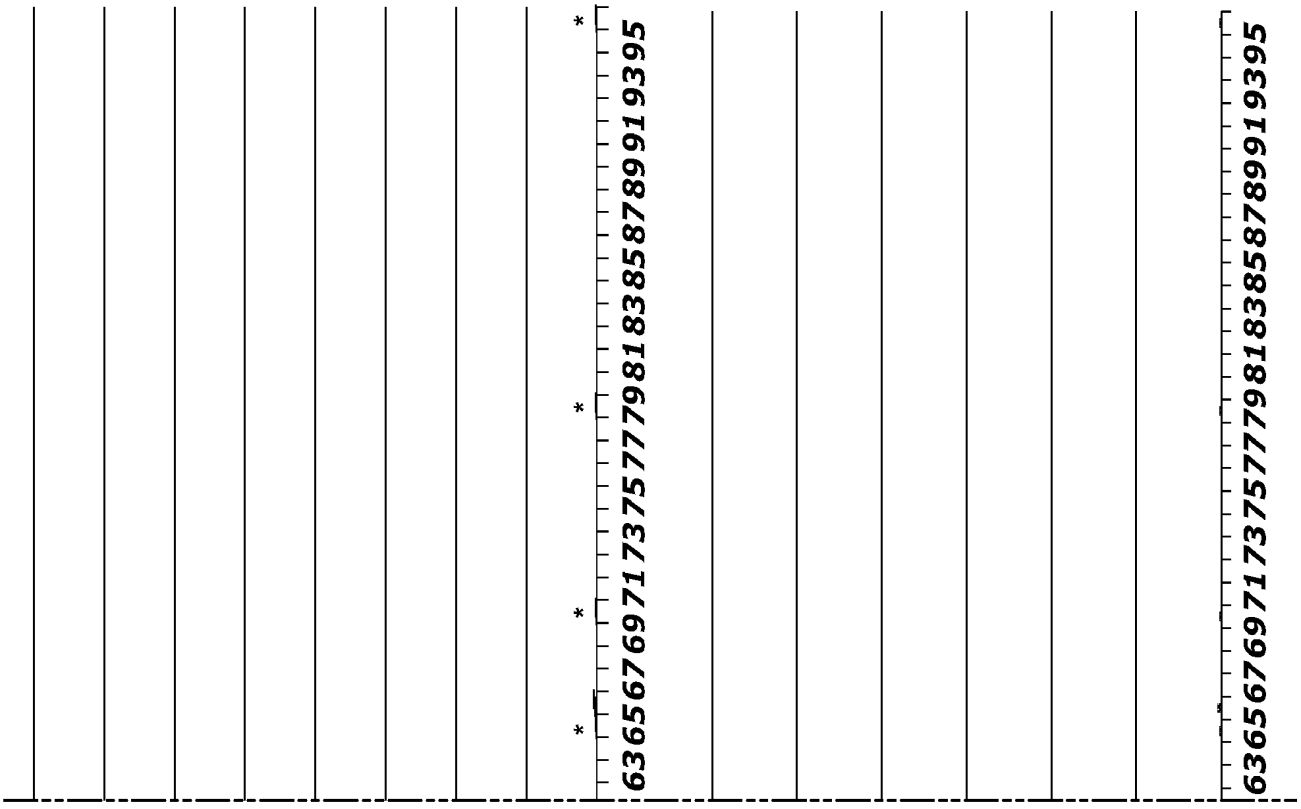

Number of reads per Primer ID
Number of reads per consensus sequence
% in population
Number of reads
Number of consensus sequences
Number of reads per consensus sequence

Fig. 6B



IDT Exhibit 1020.1874

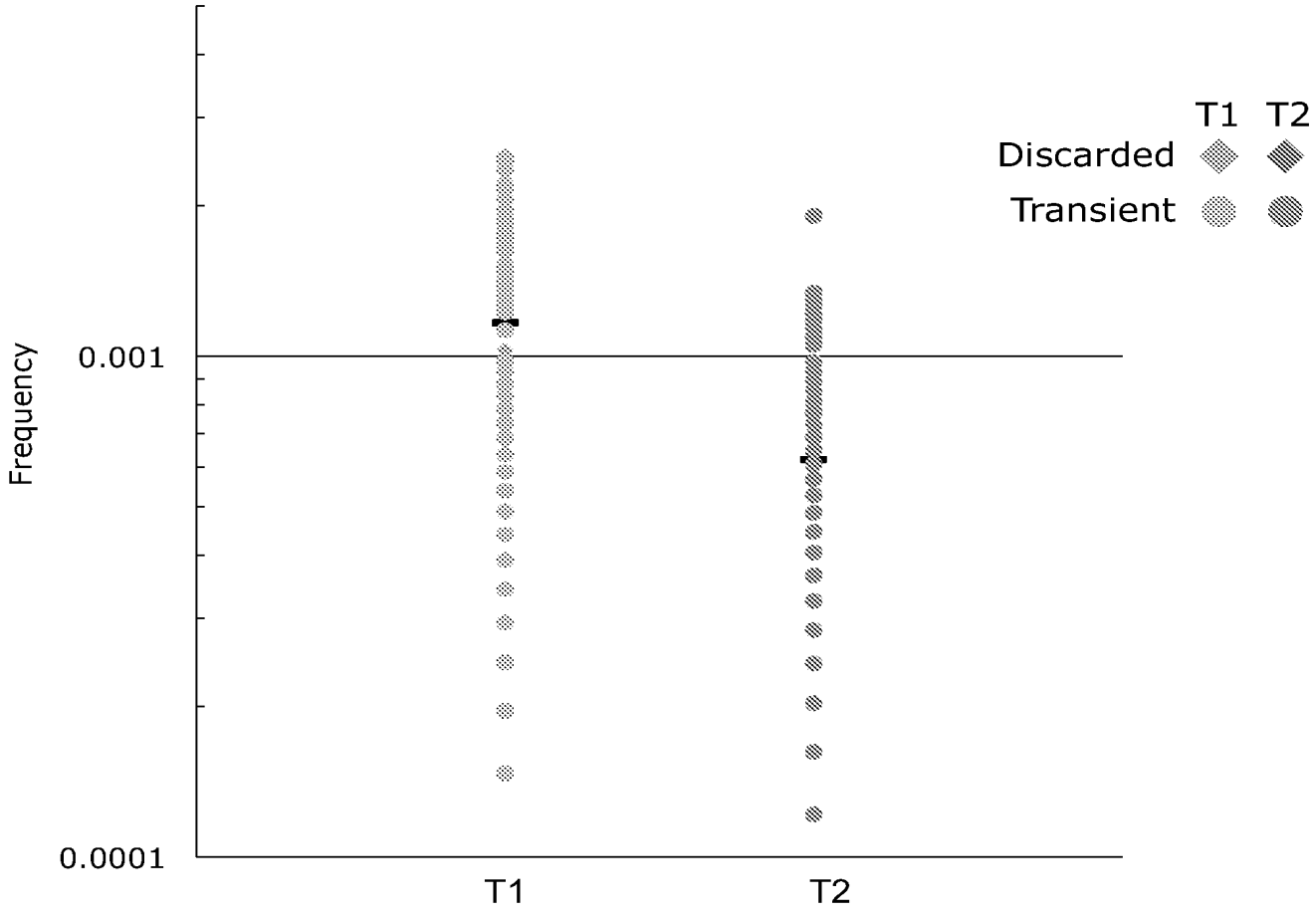


Fig. 7

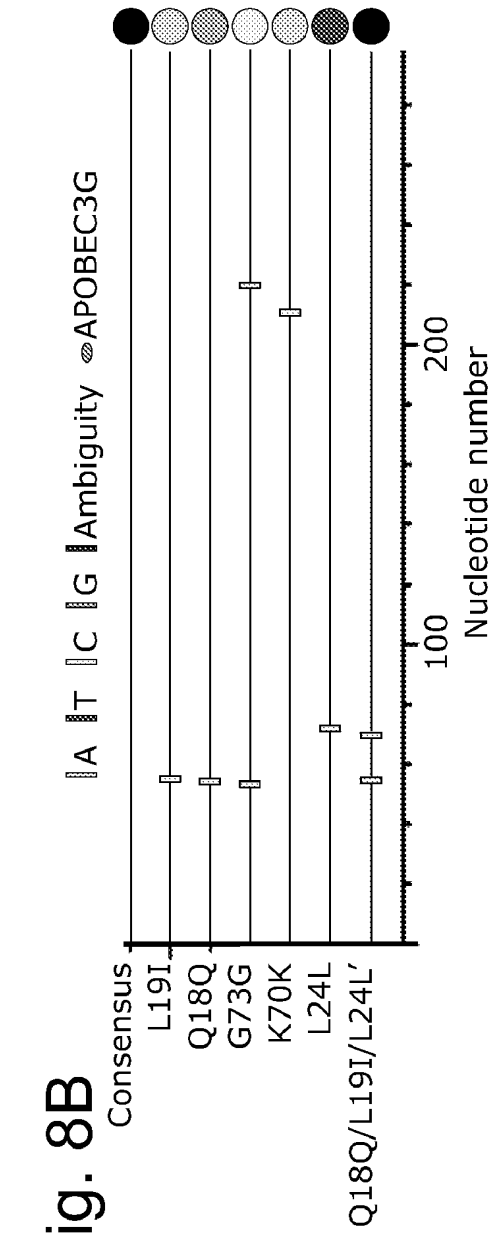


Fig. 8B

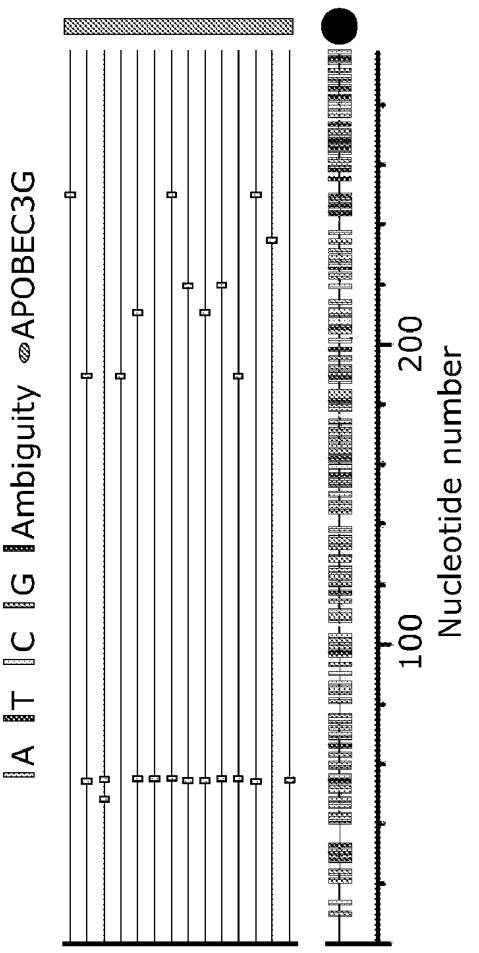


Fig. 8C

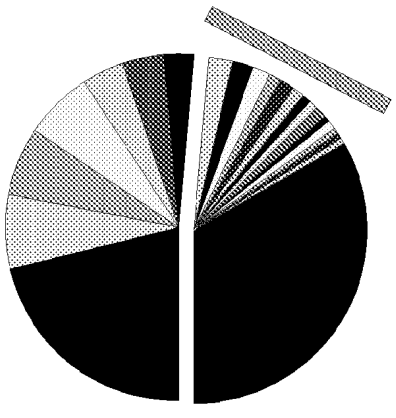
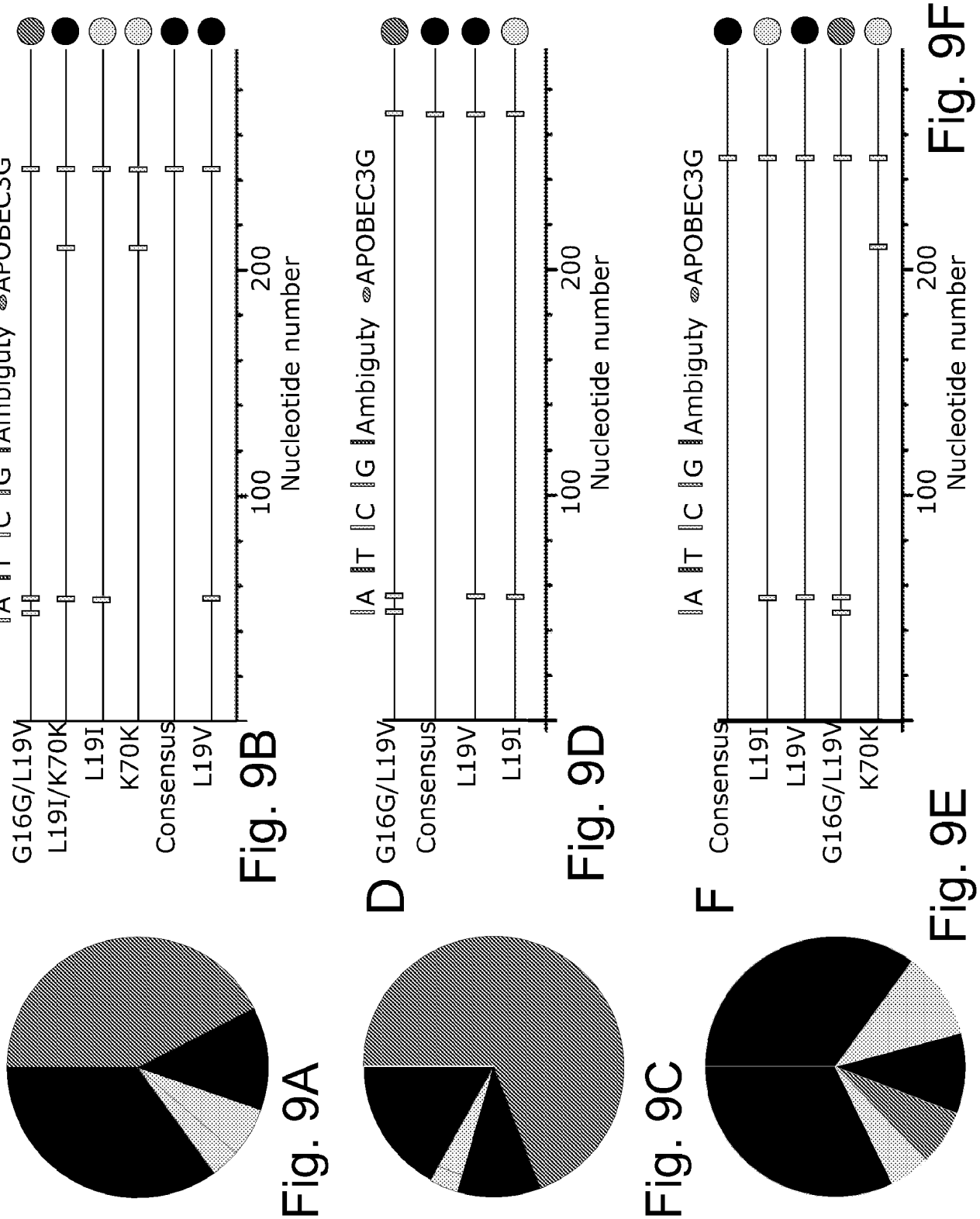


Fig. 8A



IDT Exhibit 1020.1877

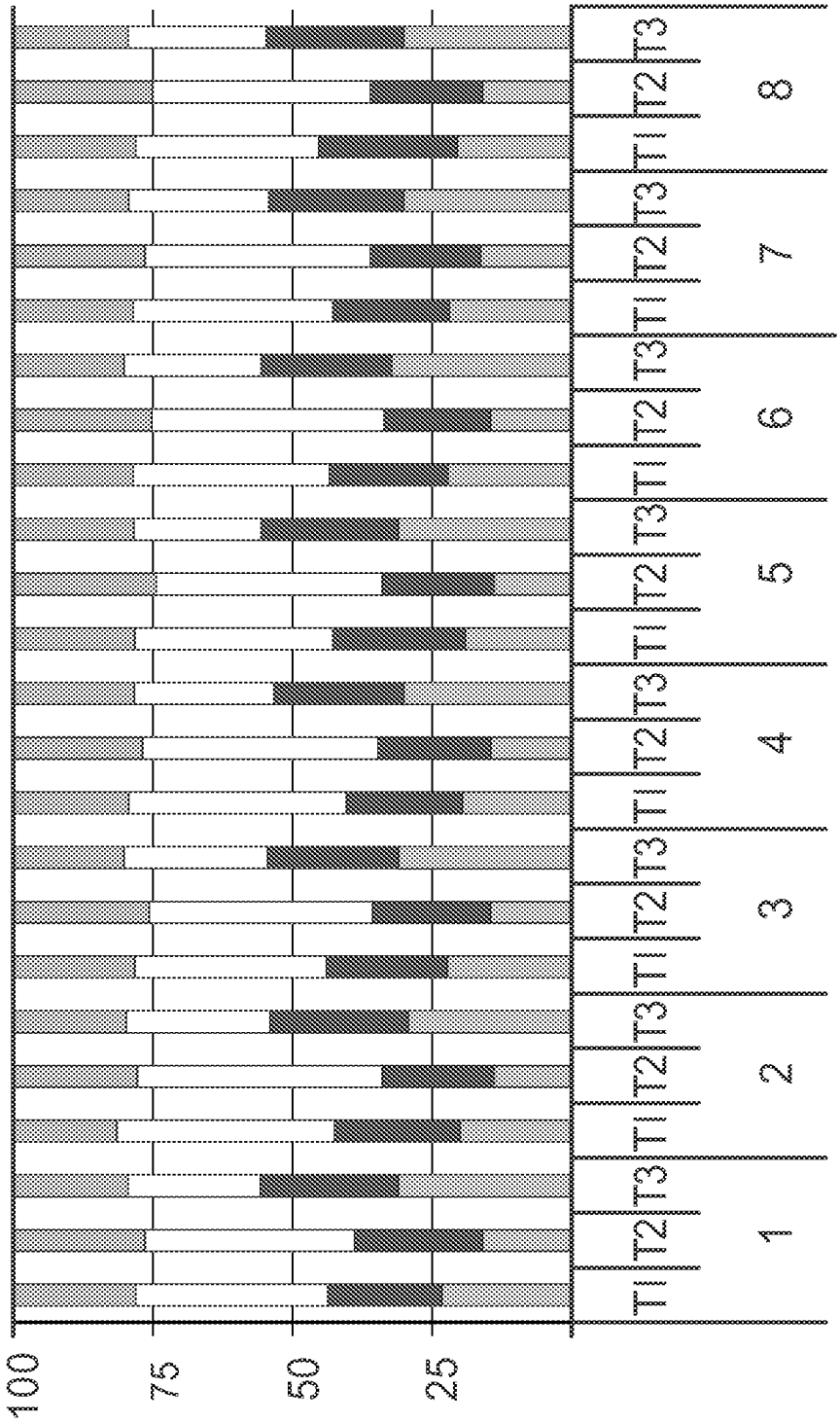

100
75
50
25
T1
T2
T3
1
T1
T2
T3
2
T1
T2
T3
3
T1
T2
T3
4
T1
T2
T3
5
T1
T2
T3
6
T1
T2
T3
7
T1
T2
T3
8

Fig. 10

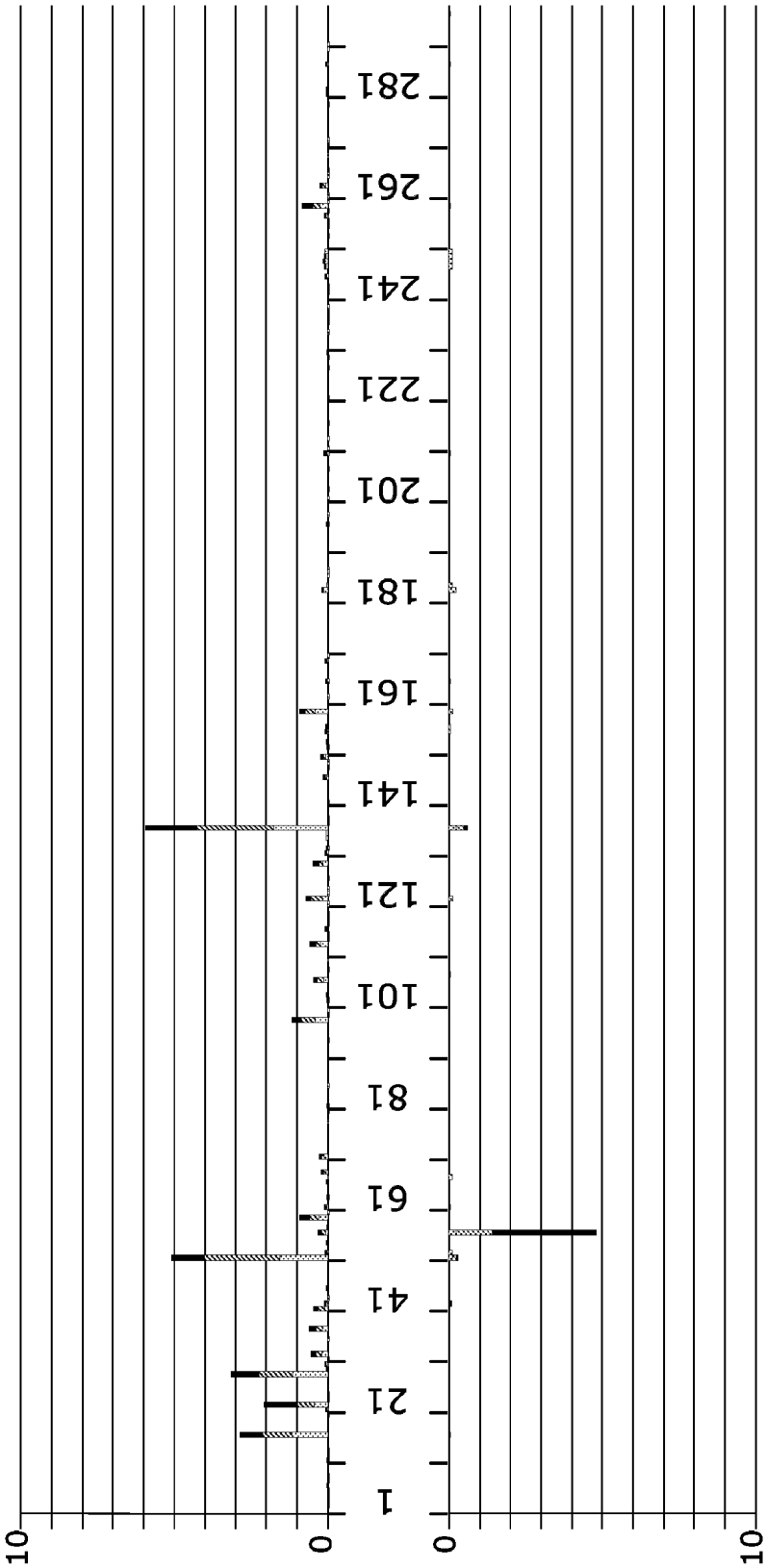


Fig. 11

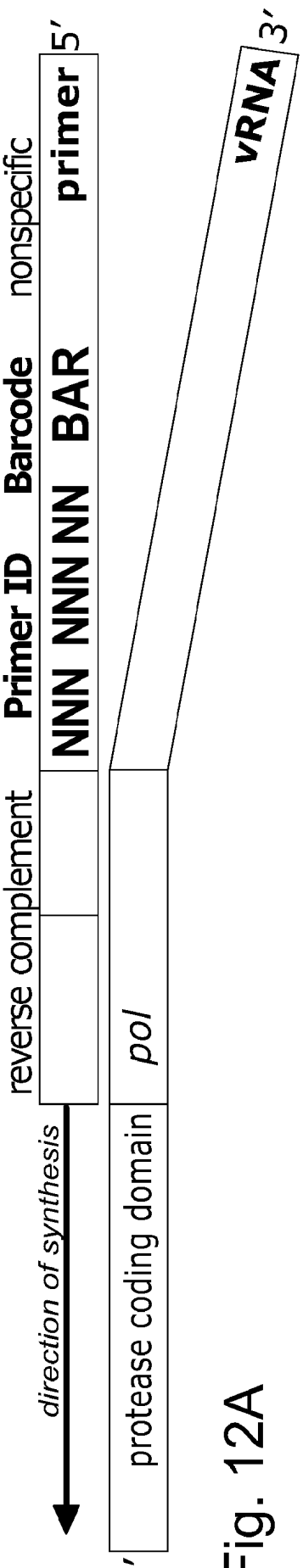


Fig. 12A

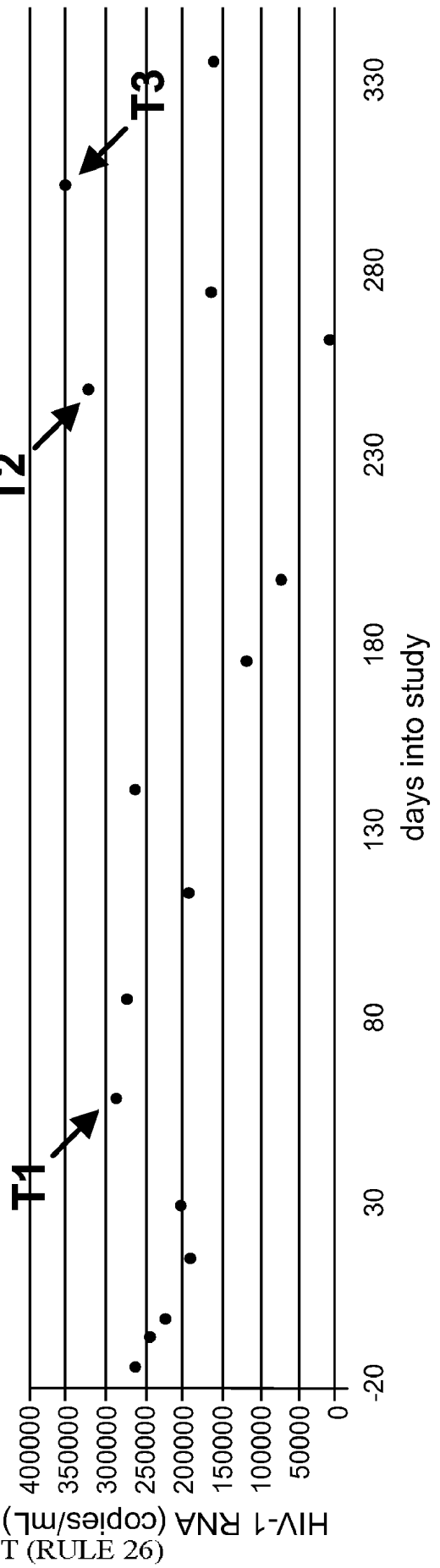


Fig. 12B

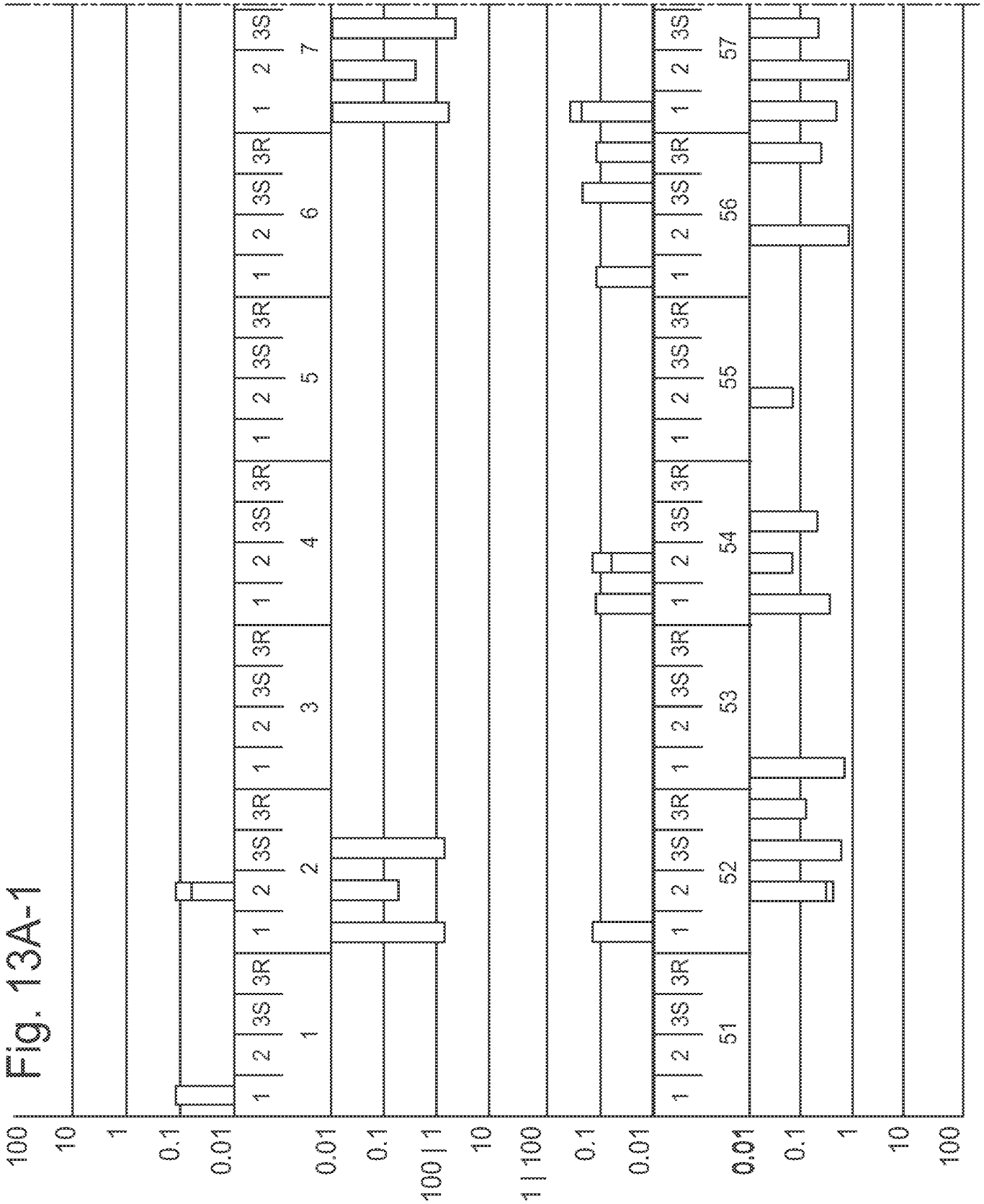

1 2 3S 3R 1 2 3S 3R 1 2 3S 3R 1 2 3S 3R 1 2 3S 3R 1 2 3S 3R 1 2 3S 3R
1 2 3S 3R 1 2 3S 3R 1 2 3S 3R 1 2 3S 3R 1 2 3S 3R 1 2 3S 3R 1 2 3S 3R
100 10 1 0.1 0.01 0.01 0.1 100 1 10 1 100 0.1 0.01 0.01 0.1 1 10 100
1 2 3 4 5 6 7
51 52 53 54 55 56 57

Fig. 13A-1

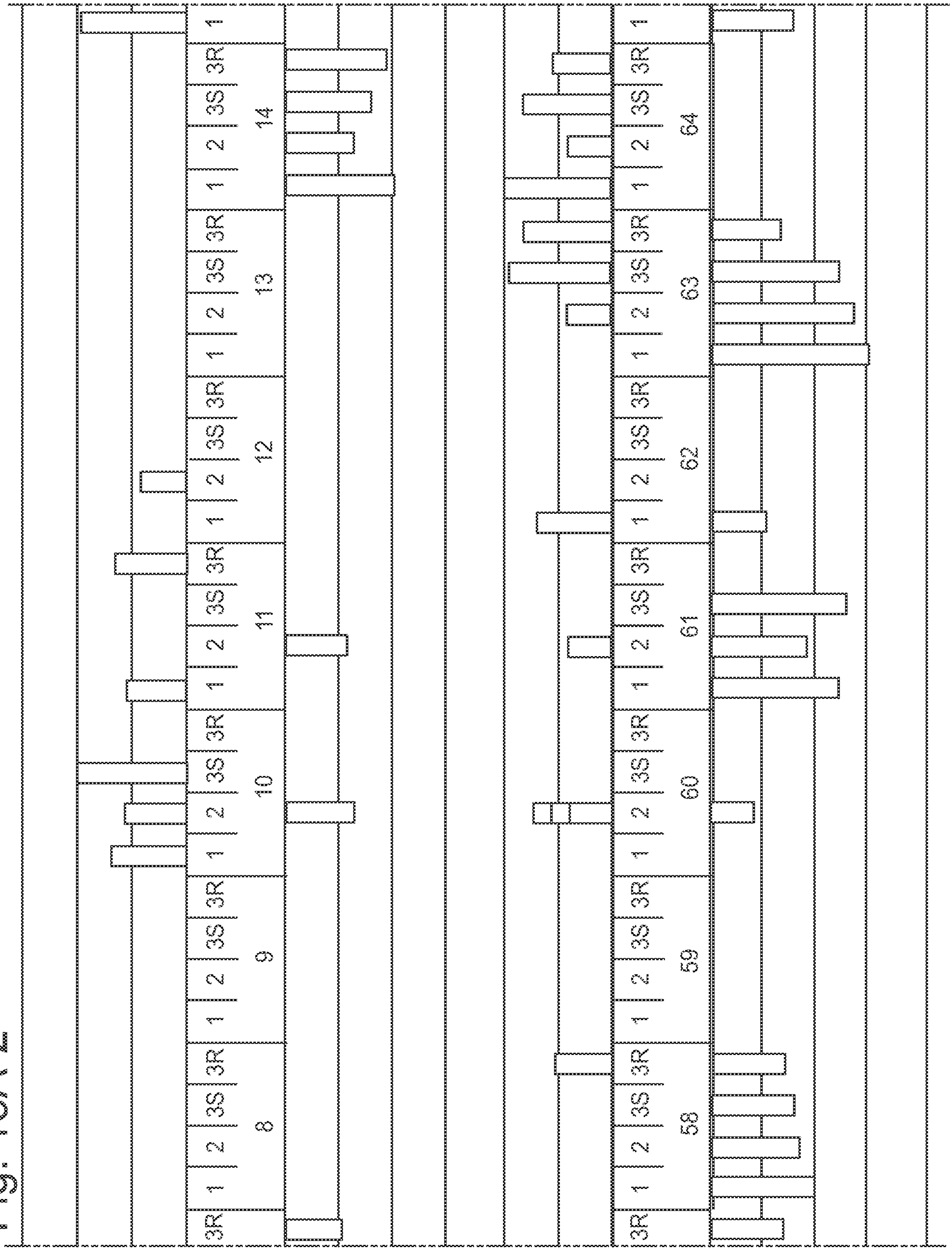


Fig. 13A-2

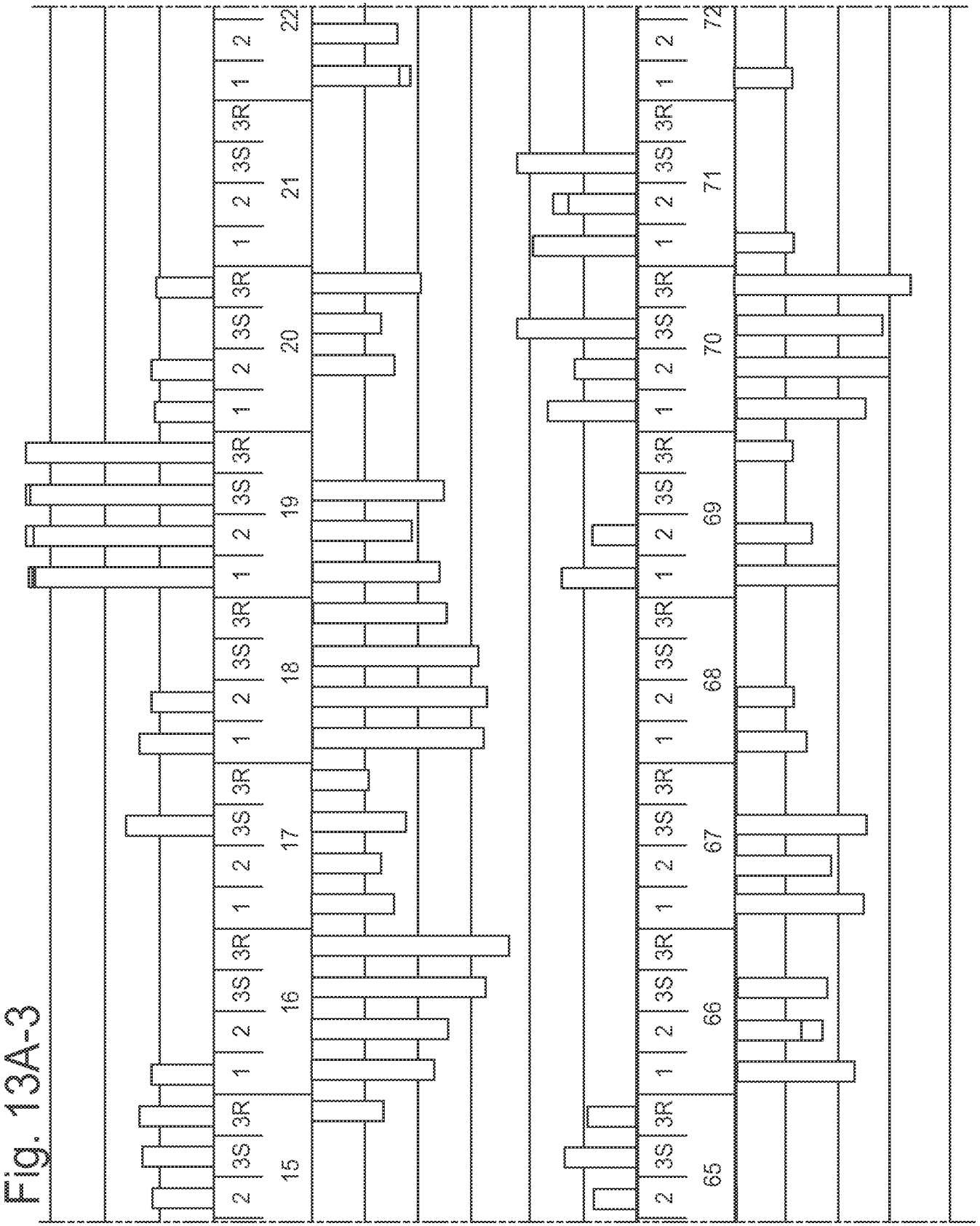


Fig. 13A-3

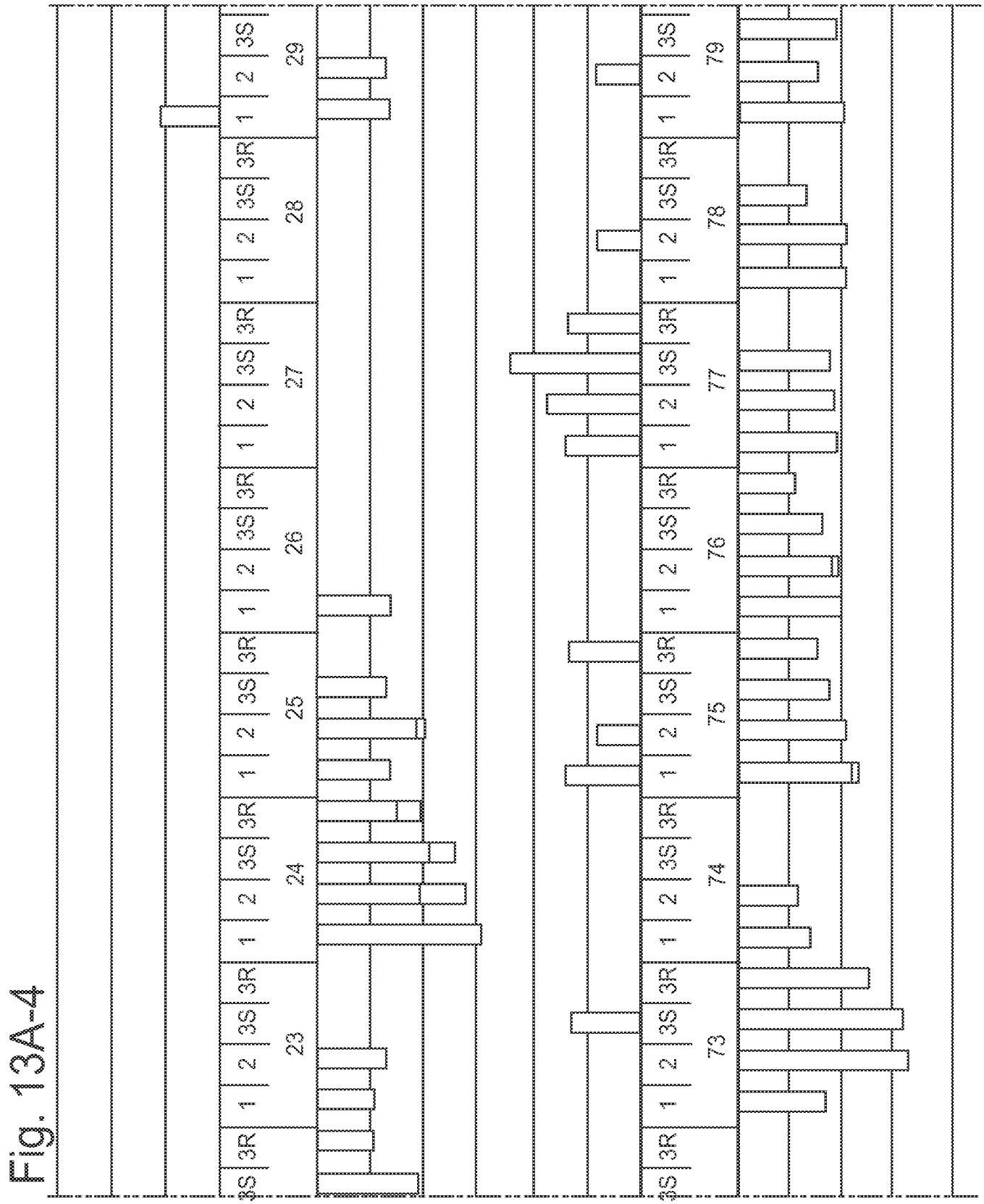

3S
3R
1
2
3S
23
3R
1
2
3S
24
3R
1
2
3S
25
3R
1
2
3S
26
3R
1
2
3S
27
3R
1
2
3S
28
3S
2
1
29
3S
2
1
79
3R
3S
2
1
78
3R
3S
2
1
77
3R
3S
2
1
76
3R
3S
2
1
75
3R
3S
2
1
74
3R
3S
2
1
73
3R

Fig. 13A-4

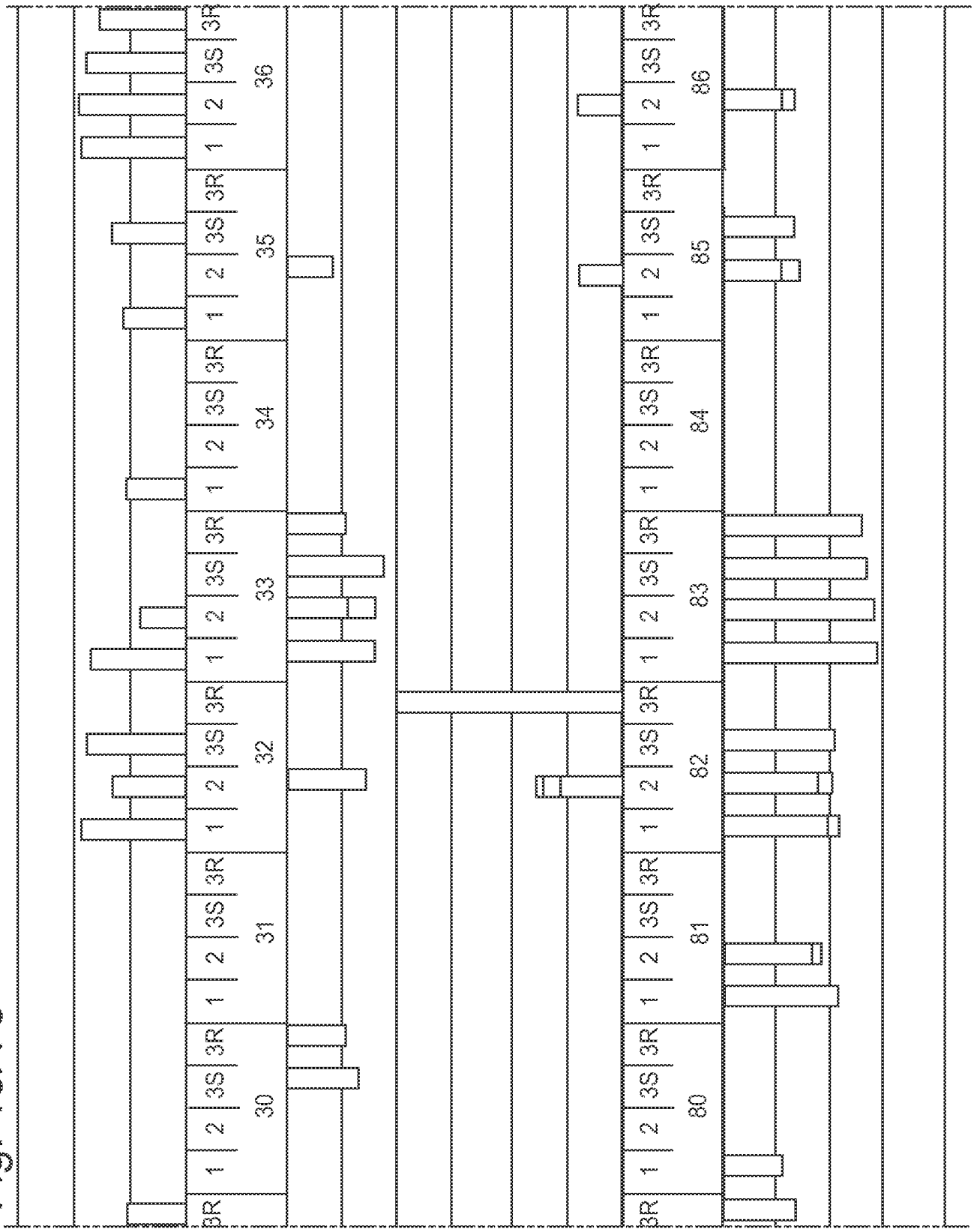

3R
1
2
3S
30
3R
1
2
3S
31
3R
1
2
3S
32
3R
1
2
3S
33
3R
1
2
3S
34
3R
1
2
3S
35
3R
1
2
3S
36
3R
1
2
3S
80
3R
1
2
3S
81
3R
1
2
3S
82
3R
1
2
3S
83
3R
1
2
3S
84
3R
1
2
3S
85
3R
1
2
3S
86

Fig. 13A-5

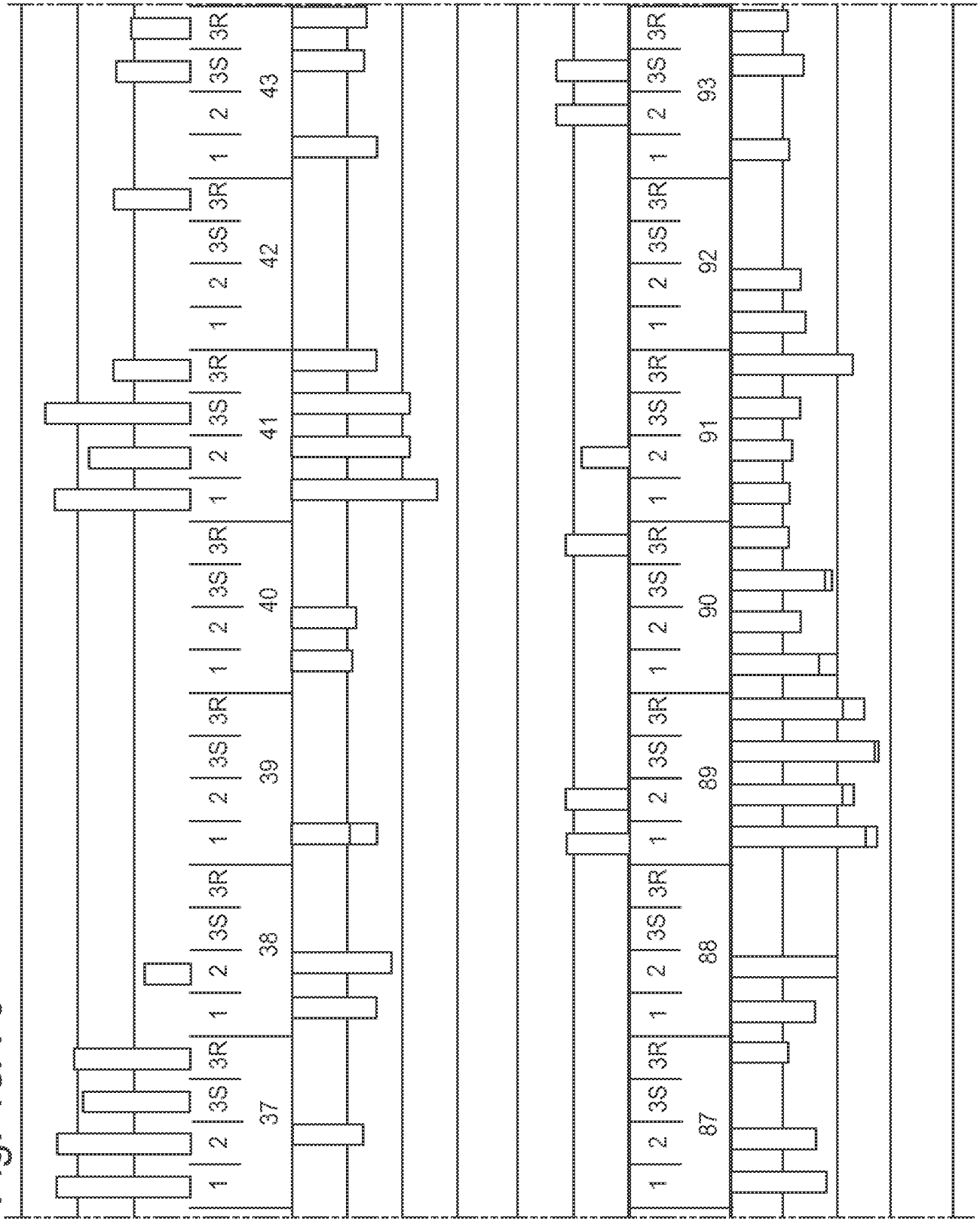
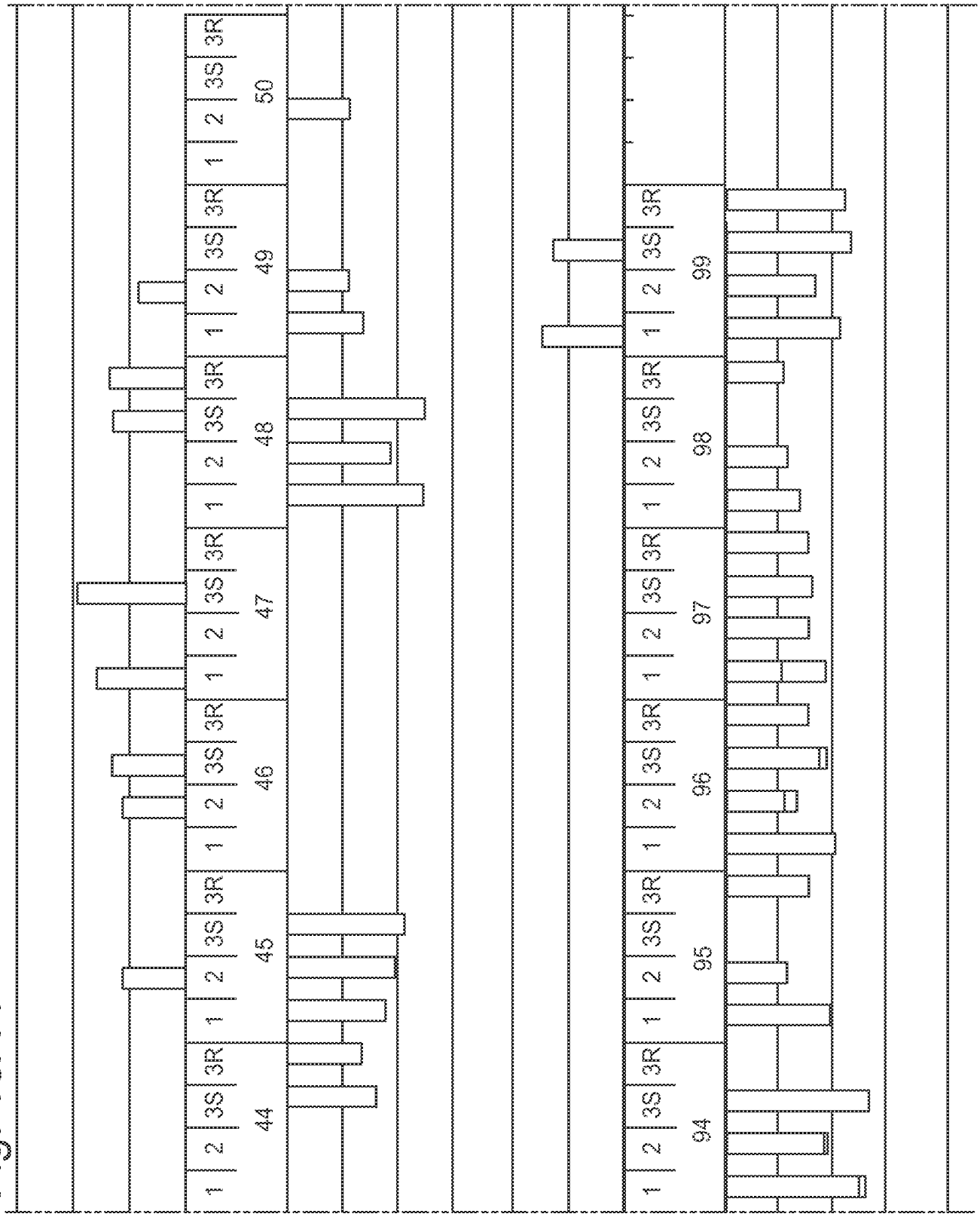


Fig. 13A-6

Fig. 13A-7



Sample	T1	T1	T1
Ritonavir	-	-	+
Total sequences	21,706	25,414	28,264
Consensus sequences	892	1,576	2,283

Fig. 13B

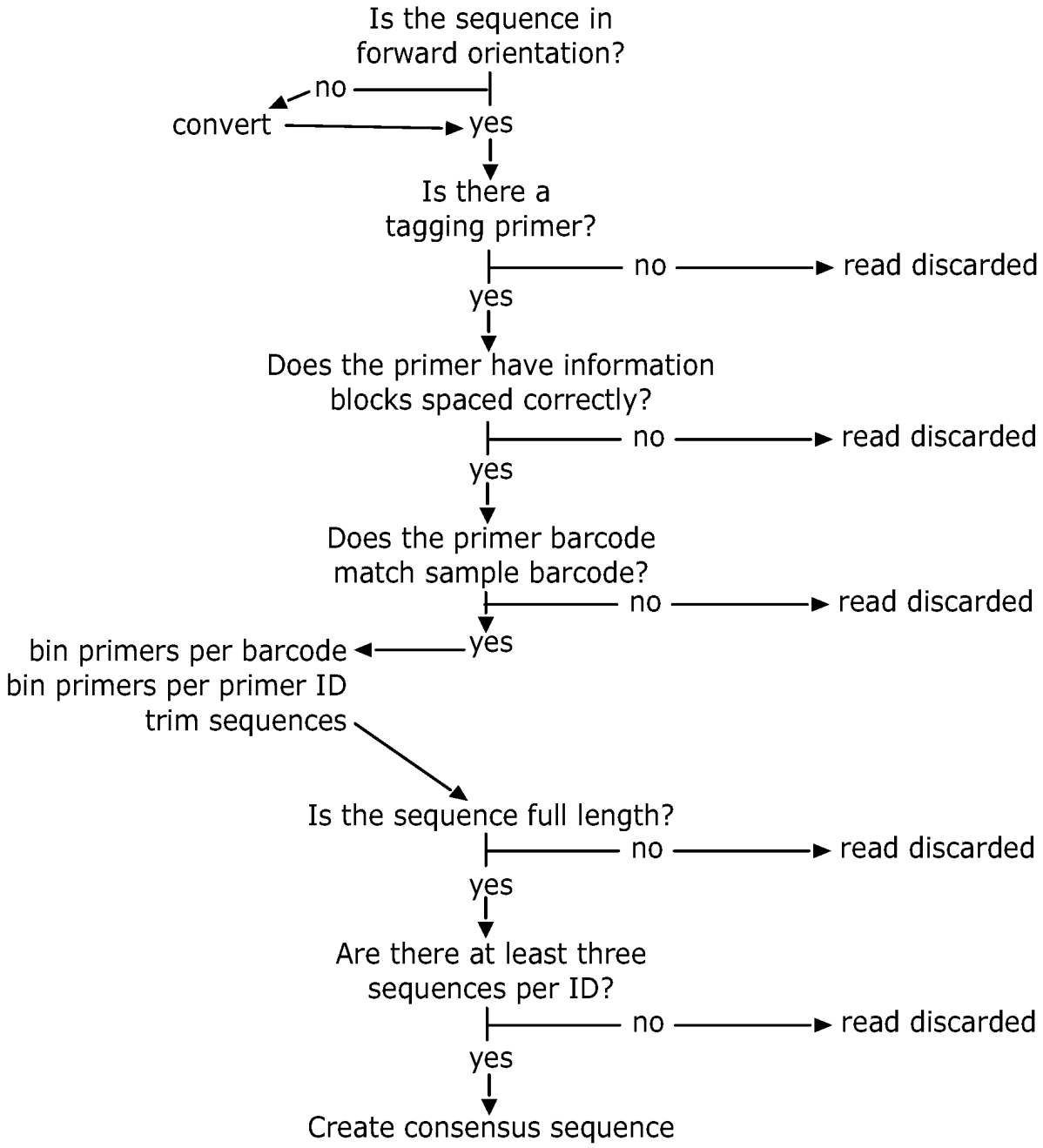

Is the sequence in forward orientation?
no
convert
yes
Is there a tagging primer?
no
read discarded
yes
Does the primer have information blocks spaced correctly?
no
read discarded
yes
Does the primer barcode match sample barcode?
no
read discarded
yes
bin primers per barcode
bin primers per primer ID
trim sequences
Is the sequence full length?
no
read discarded
yes
Are there at least three sequences per ID?
no
read discarded
yes
Create consensus sequence

Fig. 14

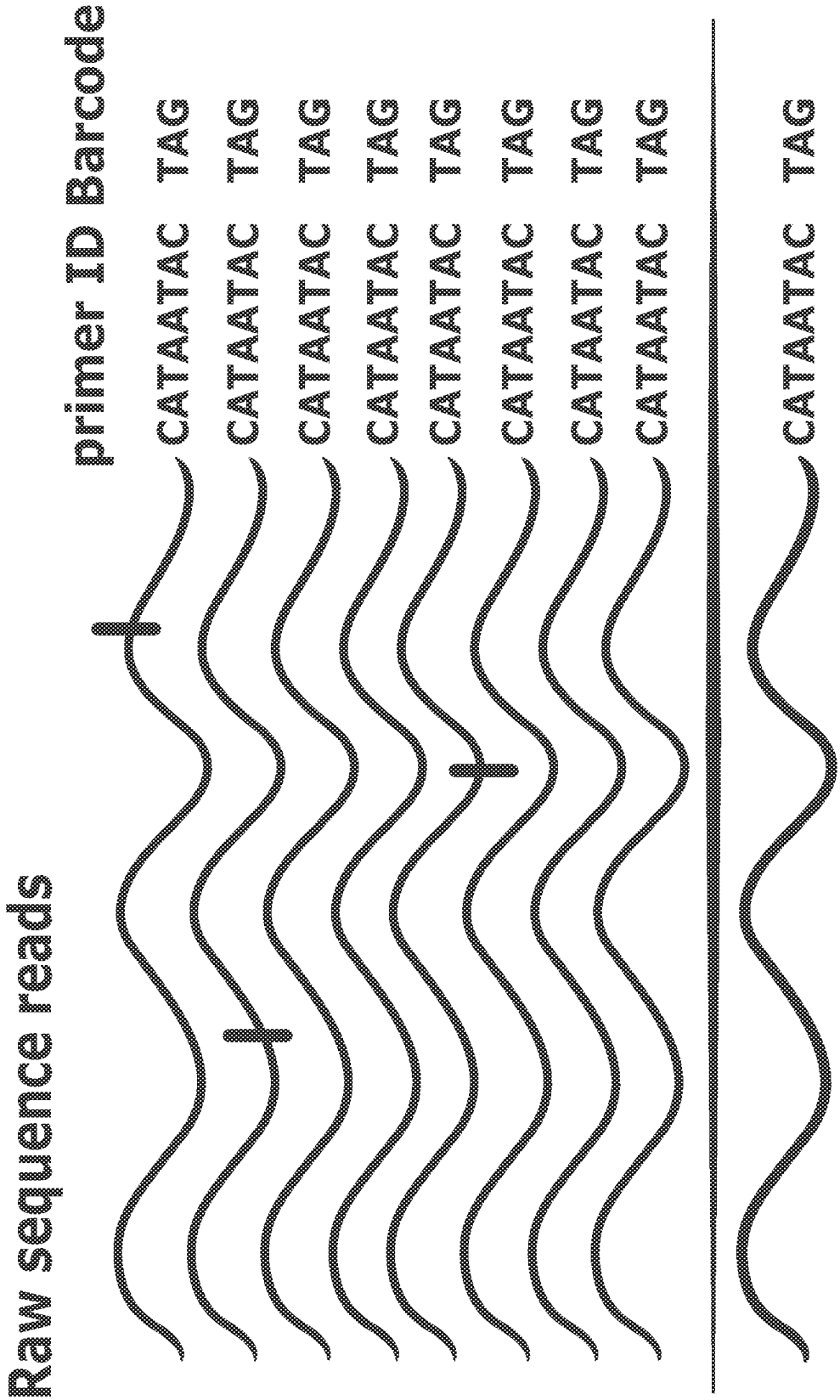

Raw sequence reads
primer ID Barcode
CATAATAC TAG
CATAATAC TAG
CATAATAC TAG
CATAATAC TAG
CATAATAC TAG
CATAATAC TAG
CATAATAC TAG
CATAATAC TAG
Consensus sequence
CATAATAC TAG

Fig. 15

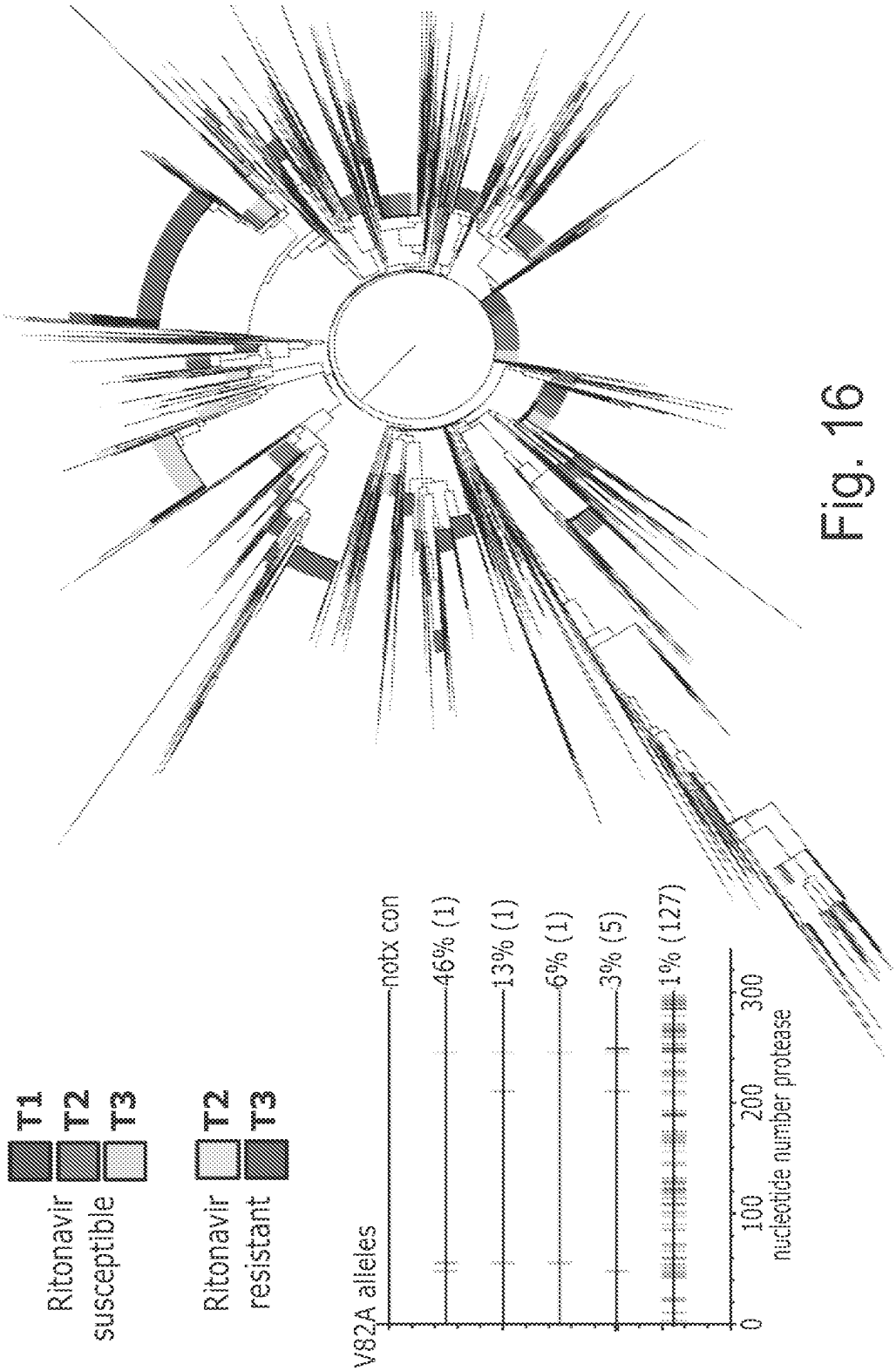

T1
T2
T3
Ritonavir susceptible
T2
T3
Ritonavir resistant
V82A alleles
notx con
46% (1)
13% (1)
6% (1)
3% (5)
1% (127)
0
100
200
300
nucleotide number protease

Fig. 16

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
7 August 2008 (07.08.2008)

PCT

(10) International Publication Number
WO 2008/093098 A2

(51) International Patent Classification:
C12Q 1/68 (2006.01)

(21) International Application Number:
PCT/GB2008/000337

(22) International Filing Date: 1 February 2008 (01.02.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/899,221 2 February 2007 (02.02.2007) US

(71) Applicant *(for all designated States except US)*: **SOLEXA LIMITED** [GB/GB]; Chesterford Research Park, Little Chesterford, Saffron Walden Essex CB10 1XL (GB).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **BIGNALL, Helen** [GB/GB]; SOLEXA LIMITED, Chesterford Research Park, Little Chesterford, Saffron Walden Essex CB10 1XL (GB). **FRASER, Louise** [GB/GB]; SOLEXA LIMITED, Chesterford Research Park, Little Chesterford, Saffron Walden Essex CB10 1XL (GB). **GORMLEY, Niall** [IE/GB]; SOLEXA LIMITED, Chesterford Research Park, Little Chesterford, Saffron Walden Essex CB10 1XL (GB).

(74) Agents: **WHITE, Nina, Louise** et al.; BOULT WADE TENNANT, Verulam Gardens, 70 Gray's Inn Road, London WC1X 8BT (GB).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

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

Published:
— *without international search report and to be republished upon receipt of that report*
— *with sequence listing part of description published separately in electronic form and available upon request from the International Bureau*

WO 2008/093098 A2

(54) **Title:** METHODS FOR INDEXING SAMPLES AND SEQUENCING MULTIPLE NUCLEOTIDE TEMPLATES

(57) **Abstract:** The invention relates to methods for indexing samples during the sequencing of polynucleotide templates, resulting in the attachment of tags specific to the source of each nucleic acid sample such that after a sequencing run, both the source and sequence of each polynucleotide can be determined. Thus, the present invention pertains to analysis of complex genomes (e.g., human genomes), as well as multiplexing less complex genomes, such as those of bacteria, viruses, mitochondria, and the like.

IDT Exhibit 1020.1892

Methods for indexing samples and sequencing multiple polynucleotide templates

Field of the invention

The invention relates to methods for indexing samples during the sequencing of polynucleotide templates, resulting in the attachment of tags specific to the source of each nucleic acid sample such that after a sequencing run, both the source and sequence of each polynucleotide can be determined. Thus, the present invention pertains to analysis of complex genomes (e.g., human genomes), as well as multiplexing less complex genomes, such as those of bacteria, viruses, mitochondria, and the like.

Background to the invention

Several publications and patent documents are referenced in this application in order to more fully describe the state of the art to which this invention pertains. The disclosure of each of these publications and documents is incorporated by reference herein.

Advances in the study of biological molecules have been led, in part, by improvement in technologies used to characterise the molecules or their biological reactions. In particular, the study of the nucleic acids DNA and RNA has benefited from developing technologies used for sequence analysis.

The study of complex genomes, in particular, the search for the genetic basis of disease in humans requires genotyping on a massive scale. Screens for numerous genetic markers performed on populations large enough to yield statistically significant data are needed before associations can be made between a given genotype and a

particular disease. However large-scale genotyping is demanding in terms of the cost of both materials and labour involved, and the time taken to perform the study, especially if the methodology employed involves separate serial analysis of individual DNA samples. One shortcut is to pool DNA from many individuals and to determine parameters such as the ratio of changes at certain positions in the genome. Such measurements of 'allele frequency' in the pool of samples can be used to correlate the relationship between the changes in the genome sequence and the occurrence of a disease. Hence, an association study involving 1000 patients would in theory only necessitate a 'one-pot' reaction for each genetic change. Pooling therefore represents an effective technique for analysing large quantities of samples in a facile manner.

One disadvantage of pooling samples prior to analysis is that information pertaining to individual DNA samples is lost; only global information such as allele frequencies is gathered, as there is no easy method for discerning which individuals gave rise to a particular genotype. An ability to genotype large populations in a small number of reactions, while retaining the information relating to the source of the individual samples, would yield the information content of a full non-pooled population screen in the time and at the cost of a pooled reaction.

Several of the new methods employed for high throughput DNA sequencing (*Nature*. 437, 376-380 (2005); *Science*. 309, 5741, 1728-1732 (2005)) rely on a universal amplification reaction, whereby a DNA sample is randomly fragmented, then treated such that the ends of the different fragments all contain the same DNA sequence. Fragments with universal ends can be amplified in a single reaction with a single pair of

amplification primers. Separation of the library of fragments to the single molecule level prior to amplification ensures that the amplified molecules form discrete populations that can then be further analysed. Such separations can be performed either in emulsions (*Nature*. 437, 376-380 (2005); *Science*. 309, 5741, 1728-1732 (2005)), or on a surface (Nucleic Acids Research 27, e34 (1999); Nucleic Acids Research 15, e87 (2000)).

WO 98/44151 and WO 00/18957 both describe methods of forming polynucleotide arrays based on 'solid-phase' nucleic acid amplification, which is a bridging amplification reaction wherein the amplification products are immobilised on a solid support in order to form arrays comprised of nucleic acid clusters or 'colonies'. Each cluster or colony on such an array is formed from a plurality of identical immobilised polynucleotide strands and a plurality of identical immobilised complementary polynucleotide strands. The arrays so-formed are generally referred to herein as 'clustered arrays' and their general features will be further understood by reference to WO 98/44151 or WO 00/18957, the contents of both documents being incorporated herein in their entirety by reference.

In common with all amplification techniques, solid-phase bridging amplification requires the use of forward and reverse amplification primers which include 'template-specific' nucleotide sequences which are capable of annealing to sequences in the template to be amplified, or the complement thereof, under the conditions of the annealing steps of the amplification reaction. The sequences in the template to which the primers anneal under conditions of the amplification reaction may be referred to herein as 'primer-binding' sequences.

Certain embodiments of the methods described in WO 98/44151 and WO 00/18957 make use of 'universal' primers to amplify templates comprising a variable template portion that it is desired to amplify flanked 5' and 3' by common or 'universal' primer binding sequences. The 'universal' forward and reverse primers include sequences capable of annealing to the 'universal' primer binding sequences in the template construct. The variable template portion, or 'target' may itself be of known, unknown or partially known sequence. This approach has the advantage that it is not necessary to design a specific pair of primers for each target sequence to be amplified; the same primers can be used for amplification of different templates provided that each template is modified by addition of the same universal primer-binding sequences to its 5' and 3' ends. The variable target sequence can therefore be any DNA fragment of interest. An analogous approach can be used to amplify a mixture of templates (targets with known ends), such as a plurality or library of target nucleic acid molecules (e.g. genomic DNA fragments), using a single pair of universal forward and reverse primers, provided that each template molecule in the mixture is modified by the addition of the same universal primer-binding sequences.

DNA from more than one source can be sequenced on an array if each DNA sample is first tagged to enable its identification after it has been sequenced. Many low scale DNA-tag methodologies already exist, for example fluorescent labelling (Haughland, Handbook of Fluorescent Probes and Research Products, Invitrogen/Molecular Probes), but these are limited in scope to less than 10 or so reactions in parallel. DNA tags can be added to the ends of DNA fragments by cloning, of example as described in US 5604097. The tags

consist of eight four base 'words', where each word uses only three bases (A,T and C) in various combinations resulting in a total of 16,777,216 different tags that all have the same base pair composition and melting points. Such tags are used to label target molecules in a sample so that after an amplification reaction, each original molecule in the sample has a unique tag. The tags can then be used to 'sort' the sample onto beads containing sequences complementary to the tags such that each bead contains multiple copies of a single amplified target sequence (Brenner et al., (2000) Nature Biotechnology, 18, 630). In this application the tags are not sequenced, so the method does not provide a method of analysing targets from multiple samples, but rather a method of sorting a mixture of amplified templates from a single sample. The problem with enabling the method for individual samples rather than individual molecules is that the tags are synthesised in a combinatorial manner, meaning that all 16,777,216 different sequences are obtained in a single mixture in the same tube. Whilst this is ideal for treating one sample such that each individual molecule in the sample carries a different tag, it does not permit attachment of the same tag to every molecule in the sample.

DNA samples from multiple sources can, however, be tagged with different nucleic acid tags such that the source of the sample can be identified. Previous application WO05068656 describes the generic concept of indexing samples. In order to utilise this invention on arrays of amplified single molecule templates, for example as described in WO9844151, WO06099579 or WO04069849, it is advantageous to prepare the nucleic acids using the novel method described herein. The optimised DNA sample

preparation techniques described herein are applicable to any method where the samples are amplified prior to sequencing. The DNA sample preparation techniques presented herein describe in detail the optimal placements of the sequencing primers and indexed tags within the DNA constructs to be sequenced.

Summary of the invention

The present inventors have developed methods for indexing samples, wherein the samples are amplified from isolated single template molecules. Using the techniques of the invention, it is possible to prepare a nucleic acid sample for a sequencing reaction wherein both the target sequence and the tag sequence can be determined.

In a first embodiment of the invention, a method is presented for sequencing nucleic acid sequences on an array and identifying subsets of nucleic acid sequences on an array, wherein each subset of nucleic acid sequences is isolated from a different source, wherein the method comprises the steps of:

(a) providing at least two samples of randomly fragmented double stranded nucleic acid targets, wherein each of said randomly fragmented double stranded nucleic acid targets is isolated from a different source;

(b) ligating a universal adaptor to the ends of each target fragment of each sample to generate adaptor-target-adaptors of each sample, wherein each of said adaptor-target-adaptors comprises a target fragment flanked by universal adaptor sequences and said universal adaptor comprises a region of double stranded nucleic acid and at least one region of single stranded nucleic acid;

(c) amplifying adaptor-target-adaptors of each sample with two or more sample specific amplification primers to generate amplified nucleic acids, wherein one of said amplification primers comprises a sample specific tag sequence, and wherein amplified nucleic acids of each sample comprise said sample specific tag sequence and said sample specific tag sequence differentiates amplified nucleic acids originating from different samples;

(d) pooling the amplified nucleic acids of different samples;

(e) immobilising the amplified nucleic acids of different samples on an array to generate an array of immobilised fragments;

(f) sequencing the immobilised fragments on the array to determine a sequence read of each immobilised target fragment and identify the sample specific tag sequence of each immobilised fragment, thereby determining both a nucleic acid sequence of said immobilized fragment and identifying the immobilized fragment as a member of a subset of nucleic acids on the array.

In a second embodiment of the invention, a method is presented for sequencing nucleic acid sequences on an array and identifying subsets of nucleic acid sequences on an array, wherein each subset of nucleic acid sequences is isolated from a different source, wherein the method comprises the steps of:

(a) providing at least two samples of randomly fragmented double stranded nucleic acid targets, wherein each of said randomly fragmented double stranded nucleic acid targets is isolated from a different source;

(b) ligating a sample specific tagged adaptor to the ends of each target fragment of each sample to generate adaptor-

target-adaptors of each sample, wherein each of said adaptor-target-adaptors comprises a target fragment flanked by tagged adaptor sequences and said tagged adaptor comprises a region of double stranded nucleic acid and at least one region of single stranded nucleic acid and a sample specific tag that differentiates adaptor-target-adaptors originating from different samples;

(c) pooling the adaptor-target-adaptors of different samples to generate a pooled sample of adaptor-target-adaptors;

(d) amplifying the pooled sample of adaptor-target-adaptors with a pair of universal amplification primers to generate an amplified pooled sample;

(e) immobilising the amplified pooled sample on an array to generate an array of immobilised fragments;

(f) sequencing the array to determine a sequence read of each immobilised target fragment and identify the sample specific tag sequence of each immobilised fragment, thereby determining both a nucleic acid sequence of said immobilized fragment and identifying the immobilized fragment as a member of a subset of nucleic acids on the array.

In a third embodiment of the invention, a method is presented for sequencing nucleic acid sequences on an array and identifying subsets of nucleic acid sequences on an array, wherein each subset of nucleic acid sequences is isolated from a different source, wherein the method comprises the steps of:

(a) providing at least two samples of randomly fragmented double stranded nucleic acid targets, wherein each of said randomly fragmented double stranded nucleic acid targets is isolated from a different source;

(b) ligating a sample specific tagged adaptor to the ends of each target fragment of each sample to generate adaptor-target-adaptors of each sample, wherein each of said adaptor-target-adaptors comprises a target fragment flanked by tagged adaptor sequences and said tagged adaptor comprises a region of double stranded nucleic acid and at least one region of single stranded nucleic acid and a sample specific tag that differentiates adaptor-target-adaptors originating from different samples, and wherein the sample specific tag sequence is attached directly to the target fragment in each adaptor-target-adaptor with no intervening nucleotides between the sample specific tag sequence and target fragment sequences;

(c) amplifying adaptor-target-adaptors of each sample with a pair of amplification primers complementary to the sample specific tag sequence to generate amplified nucleic acids from each sample;

(d) pooling the amplified nucleic acids from each sample to generate a pool of amplified nucleic acid samples;

(e) immobilising the pool of amplified nucleic acid samples on an array to generate an array of immobilised fragments;

(f) sequencing the array to determine a sequence read of each immobilised target fragment and identify the sample specific tag sequence of each immobilised fragment, thereby determining both a nucleic acid sequence of said immobilized fragment and identifying the immobilized fragment as a member of a subset of nucleic acids on the array.

In accordance with the first or second embodiment of the invention, sequencing can be performed in two separate reads. In accordance with the third embodiment of the invention, sequencing of the target fragment and the sample

specific tag sequence can be performed in a single sequencing read.

With respect to the first, second, or third embodiments, the method encompasses a step wherein the pooled samples are amplified on a solid surface prior to sequencing. In a particular aspect, the solid surface on which amplification is performed is a collection of beads. In a more particular aspect, the collection of beads may be isolated as single beads in an emulsion.

With respect to the first, second, or third embodiments, the method encompasses a step wherein the array of immobilized fragments is amplified prior to sequencing to generate an amplified array. In a particular aspect, an amplified array so produced is a clustered array of amplified single molecules. Such clustered arrays may be formed, for example, by solid-phase nucleic acid amplification or isothermal solid-phase nucleic acid amplification.

The present method encompasses sequencing that involves cycles of ligation with labelled oligonucleotides or cycles of addition of nucleotides. In a particular aspect of the invention, the nucleotides are labelled. In a more particular aspect the nucleotides are labelled with fluorophores.

In an aspect of any of the first, second, or third embodiments of the invention, the sequencing step may be carried out on an array of immobilised beads.

A further aspect of the invention comprises a kit comprising a double stranded nucleic acid adapter and at least two different primer polynucleotides that carry a different tag sequence that does not hybridise to the adapter.

Brief description of the figures

Figure 1 shows a schematic of an indexed sample construct using the first method of the invention, using universal adaptors. Common adaptors are used for all tags, tags are added using one unique amplification primer and sequencing is carried out in two reads.

Figure 2 shows a schematic of an indexed sample construct using the second method of the invention, using tagged adaptors. Tagged adaptors are used for each sample and amplified with one pair of universal primers. Sequencing is carried out in two reads.

Figure 3 shows a schematic of an indexed sample construct using the third method of the invention, using adjacent tag-targets. Two adaptors and two amplification primers are used per tag and sequencing carried out in a single read.

Figure 4 shows numerical data on two sequencing reads obtained from a sample prepared according to the first method of the invention, amplified into an array of clusters and sequenced with two sequencing reads, one for the target and one for the tag.

Figure 5 shows the concept of tagging a sample for use in a paired end reading process. Libraries are amplified at the PCR step using a nested PCR reaction with 3 primers (2 universal and 1 tag oligo.

Figure 6a shows a nested PCR reaction using three amplification primers, one of which carries the tag.

Figure 6b shows the same nested PCR concept as figure 6, but with the benefit of the sequences of the various adapters and primers.

Figure 7 shows the embodiment of the method where the first sequencing read is performed, followed by removing the first sequencing primer, hybridising a second sequencing primer to determine the tag sequence, inverting the strand on the surface using a strand resynthesis protocol, and performing a second read at the opposite end of the fragment to the first read to generate a pair of reads from each fragment along with the tag read.

Figure 8 shows some exemplary data obtained using the method shown in figure 7.

Figure 9 shows the sequence of the various oligonucleotides used in the process

Figure 10 shows a schematic of MJ Research thermocycler, coupled with an 8-way peristaltic pump Ismatec IPC ISM931 equipped with Ismatec tubing (orange/yellow, 0.51 mm ID).

Detailed description of the invention

The present invention is directed to methods that advance techniques for nucleic acid analysis, such that target nucleic acids from a plurality of sources can be tagged with a unique, identifiable sequence of nucleic acid bases and then sequenced on a single array. As described

herein, the presence of a characteristic nucleic acid sequence tag on an immobilised target molecule permits identification of the source of the target, concurrent with a sequence read from the target. This is a dramatic improvement over pre-existing array technologies which generally require a two-step process involving an initial sequencing step for sequencing the pooled nucleic acid targets, followed by a second analysis step wherein the source of the nucleic acid is determined.

The limitations of such pre-existing array technologies are evident in WO 98/44151 and WO 00/18957, for example, which both describe methods of nucleic acid amplification which allow amplification products to be immobilised on a solid support in order to form arrays comprised of clusters or "colonies" formed from a plurality of identical immobilised polynucleotide strands and a plurality of identical immobilised complementary strands. The nucleic acid molecules present in DNA colonies on the clustered arrays prepared according to these methods can provide templates for sequencing reactions, but to date only a single sequencing read can be obtained from each immobilised strand in each colony. The methods described herein allow for two or more reads on a single amplified strand, thereby allowing analysis of the tag sequence independently of the target sequence, and also allowing the possibility of a paired read from the opposite end of each fragment as well as a third read for the tag sequence. The methods also allow the preparation of constructs where the tag sequence is adjacent to the target sequence, and the tag and target are read in a single read, which again can be turned into a paired read methodology if desired.

As will be apparent to the skilled reader, references herein to a particular nucleic acid sequence may, depending on the context, also refer to nucleic acid molecules which comprise the nucleic acid sequence. Sequencing of a target fragment means that a read of the chronological order of bases is established. The bases do not need to be contiguous, although this is preferred, nor does every base on the entire fragment have to be sequenced.

The following passages describe different aspects of the invention in greater detail. Each aspect of the invention may be combined with any other aspect or aspects of the invention unless clearly indicated to the contrary. In particular, any feature indicated as being preferred or advantageous may be combined with any other feature of features indicated as being preferred or advantageous.

The terms 'target nucleic acid sequence', 'target nucleic acid molecule', 'target nucleic acid' and 'target nucleic acid fragment' may be used interchangeably to refer to nucleic acid molecules that it is desired to sequence on an array according to the invention. The target nucleic acid may be essentially any nucleic acid of known or unknown sequence. It may be, for example, a fragment of genomic DNA or cDNA. Sequencing may result in determination of the sequence of the whole, or a part of the target molecule. The targets can be derived from a primary nucleic acid sample that has been randomly fragmented. The targets can be processed into templates suitable for amplification by the placement of universal amplification sequences at the ends of each target fragment. The targets can also be obtained from a primary RNA sample by reverse transcription into cDNA.

As used herein, the term 'polynucleotide' refers to deoxyribonucleic acid (DNA), but where appropriate the skilled artisan will recognise that the method may also be applied to ribonucleic acid (RNA). The terms should be understood to include, as equivalents, analogs of either DNA or RNA made from nucleotide analogs and to be applicable to single stranded (such as sense or antisense) and double stranded polynucleotides. The term as used herein also encompasses cDNA, that is complementary or copy DNA produced from an RNA template, for example by the action of reverse transcriptase.

The primary polynucleotide molecules may originate in double-stranded DNA (dsDNA) form (e.g. genomic DNA fragments, PCR and amplification products and the like) or may have originated in single-stranded form, as DNA or RNA, and been converted to dsDNA form. By way of example, mRNA molecules may be copied into double-stranded cDNAs suitable for use in the method of the invention using standard techniques well known in the art. The precise sequence of the primary polynucleotide molecules is generally not material to the invention, and may be known or unknown.

In a particular embodiment, the primary polynucleotide molecules are RNA molecules. In an aspect of this embodiment, RNA isolated from specific samples is first converted to double-stranded DNA using techniques known in the art. In accordance with the present method, the double-stranded DNA is then tagged or indexed with a sample specific tag. Different preparations of such double-stranded DNA comprising sample specific tags can be generated, in parallel, from RNA isolated from different specific samples. Subsequently, different preparations of double-stranded DNA comprising different sample specific

tags can be mixed, sequenced *en masse*, and the identity of each sequenced fragment determined with respect to the sample from which it was isolated/derived by virtue of the presence of a sample specific tag.

In a particular embodiment, the primary polynucleotide molecules are DNA molecules. More particularly, the primary polynucleotide molecules represent the entire genetic complement of an organism, and are genomic DNA molecules which include both intron and exon sequences (coding sequence), as well as non-coding regulatory sequences such as promoter and enhancer sequences. Although it could be envisaged that particular sub-sets of polynucleotide sequences or genomic DNA could also be used, such as, for example, particular chromosomes. Yet more particularly, the sequence of the primary polynucleotide molecules is not known. Still yet more particularly, the primary polynucleotide molecules are human genomic DNA molecules. The DNA target molecules may be treated chemically or enzymatically either prior to, or subsequent to any random fragmentation processes, and prior to or subsequent to the ligation of the adaptor sequences.

Methods of the invention are especially useful for the parallel sequencing of a large number of relative small samples in the same experiment. In a sequencing platform with a high degree of parallelism, it may be possible to analyse 50-100 million reads in the same experiment, each to a length of 50 bases or more. Thus it may easily be possible to determine more than 5 billion base pairs of sequence from a single experiment. For a genome of 100,000-1 million base pairs, this represents a much larger coverage of each base pair than is actually needed for accurate sequencing. This feature is even more pronounced for viral or mitochondrial

genomes or collections of PCR fragments that may be 10,000 base pairs or less. For a 10 kB sample, sequenced in a run of 5 billion base pairs, each base pair will appear on average 500,000 times. For accurate sequencing, it may be desirable for each base pair to appear only 20 times, and thus it will be possible to sequence 25,000 10kB fragments in the sequencing run. The use of 8 base pair tags gives a possibility of 4^8 (65536) different samples in the same experiment. For the sake of clarity, sequencing of the tags can be performed in a separate read from the sequencing of the target fragments, so the read length or accuracy of the read from the sample is not lowered by the need to sequence the tags.

For genome sizes of 1 mB, at a depth of 20X, it is possible to sequence 250 samples in a 5 billion base pair experiment. The 250 samples can be coded by as few as a four base tag, but extra bases may be added to ensure accurate analysis. Thus, the primary polynucleotide molecules may be a genome of any organism, but the methods of the present invention may be used to particular advantage when analysing smaller genomes, such as those of bacteria, viruses, mitochondria and the like, which have genome sizes in the regions of a few million base pairs or lower.

Random fragmentation refers to the fragmentation of a polynucleotide molecule in a non-ordered fashion by enzymatic, chemical or mechanical means. Such fragmentation methods are known in the art and utilise standard methods (Sambrook and Russell, Molecular Cloning, A Laboratory Manual, third edition). For the sake of clarity, generating smaller fragments of a larger piece of nucleic acid via specific PCR amplification of such smaller fragments is not equivalent to fragmenting the larger piece of nucleic acid

because the larger piece of nucleic acid sequence remains in intact (i.e., is not fragmented by the PCR amplification). Moreover, random fragmentation is designed to produce fragments irrespective of the sequence identity or position of nucleotides comprising and/or surrounding the break. More particularly, the random fragmentation is by mechanical means such as nebulisation or sonication to produce fragments of about 50 base pairs in length to about 1500 base pairs in length, still more particularly 50-700 base pairs in length, yet more particularly 50-400 base pairs in length. Most particularly, the method is used to generate smaller fragments of from 50-150 base pairs in length.

Fragmentation of polynucleotide molecules by mechanical means (nebulization, sonication and Hydroshear for example) results in fragments with a heterogeneous mix of blunt and 3'- and 5'-overhanging ends. It is therefore desirable to repair the fragment ends using methods or kits (such as the Lucigen DNA terminator End Repair Kit) known in the art to generate ends that are optimal for insertion, for example, into blunt sites of cloning vectors. In a particular embodiment, the fragment ends of the population of nucleic acids are blunt ended. More particularly, the fragment ends are blunt ended and phosphorylated. The phosphate moiety can be introduced via enzymatic treatment, for example, using polynucleotide kinase.

In a particular embodiment, the target polynucleotide sequences are prepared with single overhanging nucleotides by, for example, activity of certain types of DNA polymerase such as Taq polymerase or Klenow exo minus polymerase which has a nontemplate-dependent terminal transferase activity that adds a single deoxynucleotide, for example, deoxyadenosine (A) to the 3´ ends of, for example, PCR

products. Such enzymes can be utilised to add a single nucleotide 'A' to the blunt ended 3' terminus of each strand of the target polynucleotide duplexes. Thus, an 'A' could be added to the 3' terminus of each end repaired duplex strand of the target polynucleotide duplex by reaction with Taq or Klenow exo minus polymerase, whilst the adaptor polynucleotide construct could be a T-construct with a compatible 'T' overhang present on the 3' terminus of each duplex region of the adaptor construct. This end modification also prevents self-ligation of both vector and target such that there is a bias towards formation of the combined ligated adaptor-target sequences.

The method of the invention utilises nucleic acid sequence tags as markers characteristic of the source of particular target molecules on the array. A nucleic acid sequence tag characteristic of the source is attached to each of the target molecules in each isolated sample before the sample is immobilised for sequencing. The tag is not itself formed by part of the target nucleic acid molecule, but becomes part of the template for amplification. Generally the tag will be a synthetic sequence of nucleotides which is added to the target as part of the template preparation step. Accordingly, a sample specific tag is a nucleic acid sequence tag which is attached to each of the target molecules of a particular sample, the presence of which is indicative of or is used to identify the sample or source from which the target molecules were isolated.

Preferably the nucleic acid sequence tag may be up to 20 nucleotides in length, more preferably 1-10 nucleotides, and most preferably 4-6 nucleotides in length. A four nucleotide tag gives a possibility of multiplexing 256

samples on the same array, a six base tag enables 4096 samples to be processed on the same array.

In the first embodiment of the invention, the tag sequences are introduced in a two step process involving ligation of an adaptor that is common to all samples and amplification with a tagged amplification primer. See Figures 1 and 6. In the second embodiment the tags are on the adaptor, and the amplification can be performed with universal amplification primers. See Figure 2. In the third embodiment, the tags are on the adaptors, but at the terminus of the duplex region such that the tag and target sequences are contiguous. See Figure 3. The third embodiment allows for the sequencing to be performed in a single read, but requires both strands of the adaptor to be modified and therefore requires more oligonucleotides as two unique adaptor strands are needed for each sample.

In each embodiment, the target nucleic acids of each specific fragmented sample are treated by first ligating identical adaptor polynucleotide molecules ('mismatched adaptors', the general features of which are defined below, and further described in copending application US 20070128624, the contents of which are incorporated herein by reference in their entirety) to the 5' and 3' ends of the target polynucleotide duplexes (which may be of known, partially known or unknown sequence) to form adaptor-target constructs and then carrying out an initial primer extension reaction in which extension products complementary to both strands of each individual adaptor-target construct are formed. The resulting primer extension products, and optionally amplified copies thereof, collectively provide a library of template polynucleotides. The term library refers to the collection of target fragments containing known

common sequences at their 3′ and 5′ ends, and may also be referred to as a 3′ and 5′ modified library.

Depending on the embodiment of the invention, the adaptors may be universal for all samples, or one or both strands of the duplexes may carry the tag sequence to code or track the identity of the samples.

The adaptor polynucleotides used in the method of the invention are referred to herein as 'mismatched' adaptors because, as will be explained in detail herein, it is essential that the adaptors include a region of sequence mismatch, i.e., they must not be formed by annealing of fully complementary polynucleotide strands.

Mismatched adaptors for use in the invention are formed by annealing of two partially complementary polynucleotide strands so as to provide, when the two strands are annealed, at least one double-stranded region and at least one unmatched single-stranded region.

The 'double-stranded region' of the adaptor is a short double-stranded region, typically comprising 5 or more consecutive base pairs, formed by annealing of the two partially complementary polynucleotide strands. This term refers to a double-stranded region of nucleic acid in which the two strands are annealed and does not imply any particular structural conformation.

Generally it is advantageous for the double-stranded region to be as short as possible without loss of function. In this context, 'function' refers to the ability of the double-stranded region to form a stable duplex under standard reaction conditions for an enzyme-catalysed nucleic acid ligation reaction, which will be well known to the skilled reader (e.g. incubation at a temperature in the range of 4°C to 25°C in a ligation buffer appropriate for

the enzyme), such that the two strands forming the adaptor remain partially annealed during ligation of the adaptor to a target molecule. It is not absolutely necessary for the double-stranded region to be stable under the conditions typically used in the annealing steps of primer extension or PCR reactions.

Since identical adaptors are ligated to both ends of each target molecule, the target sequence in each adaptor-target construct will be flanked by complementary sequences derived from the double-stranded region of the adaptors. The longer the double-stranded region, and hence the complementary sequences derived therefrom in the adaptor-target constructs, the greater the possibility that the adaptor-target construct is able to fold back and base-pair to itself in these regions of internal self-complementarity under the annealing conditions used in primer extension and/or PCR. It is, therefore, generally preferred for the double-stranded region to be 20 or less, 15 or less, or 10 or less base pairs in length in order to reduce this effect. The stability of the double-stranded region may be increased, and hence its length potentially reduced, by the inclusion of non-natural nucleotides which exhibit stronger base-pairing than standard Watson-Crick base pairs.

It is preferred, but not absolutely essential, for the two strands of the adaptor to be 100% complementary in the double-stranded region. It will be appreciated that one or more nucleotide mismatches may be tolerated within the double-stranded region, provided that the two strands are capable of forming a stable duplex under standard ligation conditions.

Adaptors for use in the invention will generally include a double-stranded region forming the 'ligatable' end

of the adaptor, i.e. the end that is joined to a target polynucleotide in the ligation reaction. The ligatable end of the adaptor may be blunt or, in other embodiments, short 5′ or 3′ overhangs of one or more nucleotides may be present to facilitate/promote ligation. The 5′ terminal nucleotide at the ligatable end of the adaptor is phosphorylated to enable phosphodiester linkage to a 3′ hydroxyl group on the target polynucleotide.

The term 'unmatched region' refers to a region of the adaptor wherein the sequences of the two polynucleotide strands forming the adaptor exhibit a degree of non-complementarity such that the two strands are not capable of fully annealing to each other under standard annealing conditions for a primer extension or PCR reaction. The unmatched region(s) may exhibit some degree of annealing under standard reaction conditions for an enzyme-catalysed ligation reaction, provided that the two strands revert to single stranded form under annealing conditions in an amplification reaction.

It is to be understood that the 'unmatched region' is provided by different portions of the same two polynucleotide strands which form the double-stranded region(s). Mismatches in the adaptor construct can take the form of one strand being longer than the other, such that there is a single stranded region on one of the strands, or a sequence selected such that the two strands do not hybridise, and thus form a single stranded region on both strands. The mismatches may also take the form of 'bubbles', wherein both ends of the adaptor construct(s) are capable of hybridising to each other and forming a duplex, but the central region is not. The portion of the strand(s) forming the unmatched region are not annealed under conditions in

which other portions of the same two strands are annealed to form one or more double-stranded regions. For avoidance of doubt it is to be understood that a single-stranded or single base overhang at the 3' end of a polynucleotide duplex that subsequently undergoes ligation to the target sequences does not constitute an 'unmatched region' in the context of this invention.

The lower limit on the length of the unmatched region will typically be determined by function, for example, the need to provide a suitable sequence for binding of a primer for primer extension, PCR and/or sequencing. Theoretically there is no upper limit on the length of the unmatched region, except that in general it is advantageous to minimise the overall length of the adaptor, for example, in order to facilitate separation of unbound adaptors from adaptor-target constructs following the ligation step. Therefore, it is generally preferred that the unmatched region should be less than 50, or less than 40, or less than 30, or less than 25 consecutive nucleotides in length.

The precise nucleotide sequence of the adaptors is generally not material to the invention and may be selected by the user such that the desired sequence elements are ultimately included in the common sequences of the library of templates derived from the adaptors to, for example, provide binding sites for particular sets of universal amplification primers and/or sequencing primers. Additional sequence elements may be included, for example, to provide binding sites for sequencing primers which will ultimately be used in sequencing of template molecules in the library, or products derived from amplification of the template library, for example on a solid support.

Although the precise nucleotide sequence of the adaptor is generally non-limiting to the invention, the sequences of the individual strands in the unmatched region should be such that neither individual strand exhibits any internal self-complementarity which could lead to self-annealing, formation of hairpin structures, etc. under standard annealing conditions. Self-annealing of a strand in the unmatched region is to be avoided as it may prevent or reduce specific binding of an amplification primer to this strand.

The mismatched adaptors are preferably formed from two strands of DNA, but may include mixtures of natural and non-natural nucleotides (e.g. one or more ribonucleotides) linked by a mixture of phosphodiester and non-phosphodiester backbone linkages. Other non-nucleotide modifications may be included such as, for example, biotin moieties, blocking groups and capture moieties for attachment to a solid surface, as discussed in further detail below.

The adaptor constructs may contain exonuclease resistant modifications such as phosphorothioate linkages. Such modifications reduce the number of adaptor-dimers present in the library, since the two adaptors can not undergo ligation without removal of their non complementary overhangs. The adaptors can be treated with an exonuclease enzyme prior to the ligation reaction with the target, to ensure that the overhanging ends of the strands can not be removed during the ligation process. Treatment of the adaptors in this manner reduces the formation of the adaptor-dimers at the ligation step.

Ligation methods are known in the art and utilise standard methods (Sambrook and Russell, Molecular Cloning, A Laboratory Manual, third edition). Such methods utilise

ligase enzymes such as DNA ligase to effect or catalyse joining of the ends of the two polynucleotide strands of, in this case, the adaptor duplex construct and the target polynucleotide duplexes, such that covalent linkages are formed. The adaptor duplex construct may contain a 5'-phosphate moiety in order to facilitate ligation to the target 3'-OH. The target contains a 5'-phosphate moiety, either residual from the shearing process, or added using an enzymatic treatment step, and has been end repaired, and optionally extended by an overhanging base or bases, to give a 3'-OH suitable for ligation. In this context, joining means covalent linkage of polynucleotide strands which were not previously covalently linked. In a particular aspect of the invention, such joining takes place by formation of a phosphodiester linkage between the two polynucleotide strands, but other means of covalent linkage (e.g. non-phosphodiester backbone linkages) may be used.

Optionally the combined ligated polynucleotide sequences and unligated adaptor polynucleotide constructs may be purified from any components of the ligation reaction, such as enzymes, buffers, salts and the like. Suitable purification methods are known in the art and utilise standard methods (Sambrook and Russell, Molecular Cloning, A Laboratory Manual, third edition).

In a next step according to the invention an amplification reaction is prepared. The contents of an amplification reaction are known by one skilled in the art and include appropriate substrates (such as dNTPs), enzymes (e.g. a DNA polymerase) and buffer components required for an amplification reaction. Generally amplification reactions require at least two amplification primers, often denoted 'forward' and 'reverse' primers (primer oligonucleotides)

that are capable of annealing specifically to a part of the polynucleotide sequence to be amplified under conditions encountered in the primer annealing step of each cycle of an amplification reaction. In certain embodiments the forward and reverse primers may be identical. Thus the primer oligonucleotides must include an 'adaptor-target specific portion', being a sequence of nucleotides capable of annealing to a part of, that is, a primer-binding sequence, in the polynucleotide molecule to be amplified (or the complement thereof if the template is viewed as a single strand) during the annealing step.

Depending on the embodiment of the invention, the amplification primers may be universal for all samples, or one of the forward or reverse primers may carry the tag sequence that codes for the sample source. The amplification primers may hybridise across the tag region of the ligated adaptor, in which case unique primers will be needed for each sample nucleic acid. The amplification reaction may be performed with more than two amplification primers. In order to prevent the amplification of ligated adapter-adapter dimers, the amplification primers can be modified to contain nucleotides that hybridise across the whole of the ligated adapter and into the ligated template (or the dNTP's attached to the 3' end thereof). This first amplification primer can be modified and treated to help prevent exonucleases digestion of the strands, and thus it may be advantageous to have a first amplification primer that is universal and can amplify all samples rather than modifying and treating each of the tagged primers separately. The tagged primer can be introduced as a sample specific third primer in the amplification reaction, but does not need to be specially modified and treated to lower exonuclease

digestion. The nested PCR approach is shown in figure 6. In the case of this embodiment the third amplification primer that carries the tag needs to contain at sequence the same as at least a portion of the first amplification primer such that it can be used to amplify the duplex resulting from extension of the first amplification primer.

In the context of the present invention, the term 'polynucleotide molecule to be amplified' refers to the original or starting adaptor-target-adaptor sequence added to the amplification reaction. The 'adaptor-target specific portion' in the forward and reverse amplification primers refers to a sequence capable of annealing to the original or initial adaptor-target-adaptor present at the start of the amplification reaction and reference to the length of the 'adaptor-target specific portion' relates to the length of the sequence in the primer which anneals to the starting adaptor-target. It will be appreciated that if the primers contain any nucleotide sequence which does not anneal to the starting adaptor-target in the first amplification cycle then this sequence may be copied into the amplification products (assuming the primer does not contain a moiety which prevents read-through of the polymerase). Hence the amplified template strands produced in the first and subsequent cycles of amplification may be longer than the starting adaptor-target strands.

As the mis-matched adapters can be different lengths, the length of adapter sequence added to the 3' and 5' ends of each strand may be different. The amplification primers may also be of different length to each other, and may hybridise to different lengths of the adapter, and therefore the length added to the ends of each strand can be controlled. In the case of nested PCR, the three or more

amplification primers can be designed to be longer than the primer used to amplify the previous amplicon, so the length of the added nucleotides is fully controllable and may be hundreds of base pairs if desired. In the example shown in figure 6b, the first amplification primer adds 13 bases to the ligated adapter, and the third amplification primer adds a further 27 bases such that one end of the amplicon is 40 bases longer than the short arm of adapter-target construct. The short arm of the adapter is 20 bases in length, meaning that the prepared template comprises the genomic region plus 60 added bases at the end. In figure 6b, the second amplification primer is 25 bases longer than the long arm of adapter, which is 32 bases in length plus the additional T that hybridise across the dATP nucleoside added to the genomic sample. Thus the prepared template comprises the genomic fragment, plus the added dATP, plus 57 known bases. Thus in full, one strand of each template duplex comprises from the 5′ end: 60 known bases, T, the genomic fragment, A, 57 known bases-3′ end. This strand is fully complementary to a sequence: 5′-57 known bases, T, genomic fragment, A, 60 known bases-3′ end. The length 57 and 6 are arbitary, and shown for the purpose of clarification, and should not be taken as limiting. The length of the added sequences may be 20-100 bases or more depending on the desired experimental design.

The forward and reverse primers may be of sufficient length to hybridise to the whole of the adaptor sequence and at least one base of the target sequence (or the nucleotide dNTP added as a 3′-overhang on the target strands). The forward and reverse primers may also contain a region that extends beyond the adaptor construct, and therefore the amplification primers may be at least 20-100 bases in

length. The forward and reverse primers may be of significantly different lengths; for example one may be 20-40 bases, whereas the other one may be 40-100 bases in length. The nucleotide sequences of the adaptor-target specific portions of the forward and reverse primers are selected to achieve specific hybridisation to the adaptor-target sequences to be amplified under the conditions of the annealing steps of the amplification reaction, whilst minimising non-specific hybridisation to any other target sequences present.

Skilled readers will appreciate that it is not strictly required for the adaptor-target specific portion to be 100% complementary, a satisfactory level of specific annealing can be achieved with less than perfectly complementary sequences. In particular, one or two mis-matches in the adaptor-target specific portion can usually be tolerated without adversely affecting specificity for the template. Therefore the term 'adaptor-target specific portion' should not be interpreted as requiring 100% complementarity with the adaptor-target. However, the requirement that the primers do not anneal non-specifically to regions of the adaptor-target other than their respective primer-binding sequences must be fulfilled.

Amplification primers are generally single stranded polynucleotide structures. They may also contain a mixture of natural and non-natural bases and also natural and non-natural backbone linkages, provided that any non-natural modifications do not preclude function as a primer - that being defined as the ability to anneal to a template polynucleotide strand during conditions of the amplification reaction and to act as an initiation point for synthesis of

a new polynucleotide strand complementary to the template strand.

Primers may additionally comprise non-nucleotide chemical modifications, for example phosphorothioates to increase exonuclease resistance, again provided such that modifications do not prevent primer function. Modifications may, for example, facilitate attachment of the primer to a solid support, for example a biotin moiety. Certain modifications may themselves improve the function of the molecule as a primer, or may provide some other useful functionality, such as providing a site for cleavage to enable the primer (or an extended polynucleotide strand derived therefrom) to be cleaved from a solid support.

In an embodiment wherein tags are attached to the adaptors, the amplification can be carried out on either the pooled or unpooled samples. In an embodiment wherein universal adaptors are used, tags must be part of the amplification primers, and therefore, each sample must be amplified independently prior to pooling. The pooled nucleic acid samples can then be processed for sequencing

Preparation of immobilised samples for sequencing

The pooled sample can then be immobilised in preparation for sequencing. Sequencing can be performed as an array of single molecules, or can be amplified prior to sequencing. The amplification can be carried out using one or more immobilised primers. The immobilised primer(s) can be a lawn on a planar surface, or on a pool of beads. The pool of beads can be isolated into an emulsion with a single bead in each 'compartment' of the emulsion. At a concentration of only one template per 'compartment', only a single template is amplified on each bead.

The term 'solid-phase amplification' as used herein refers to any nucleic acid amplification reaction carried out on or in association with a solid support such that all or a portion of the amplified products are immobilised on the solid support as they are formed. In particular, the term encompasses solid-phase polymerase chain reaction (solid-phase PCR) and solid phase isothermal amplification which are reactions analogous to standard solution phase amplification, except that one or both of the forward and reverse amplification primers is/are immobilised on the solid support. Solid phase PCR covers systems such as emulsions, wherein one primer is anchored to a bead and the other is in free solution, and colony formation in solid phase gel matrices wherein one primer is anchored to the surface, and one is in free solution.

Although the invention encompasses 'solid-phase' amplification methods in which only one amplification primer is immobilised (the other primer usually being present in free solution), it is preferred for the solid support to be provided with both the forward and the reverse primers immobilised. In practice, there will be a 'plurality' of identical forward primers and/or a 'plurality' of identical reverse primers immobilised on the solid support, since the amplification process requires an excess of primers to sustain amplification. References herein to forward and reverse primers are to be interpreted accordingly as encompassing a 'plurality' of such primers unless the context indicates otherwise.

As will be appreciated by the skilled reader, any given amplification reaction requires at least one type of forward primer and at least one type of reverse primer specific for the template to be amplified. However, in certain

embodiments the forward and reverse primers may comprise template-specific portions of identical sequence, and may have entirely identical nucleotide sequence and structure (including any non-nucleotide modifications). In other words, it is possible to carry out solid-phase amplification using only one type of primer, and such single-primer methods are encompassed within the scope of the invention. Other embodiments may use forward and reverse primers which contain identical template-specific sequences but which differ in some other structural features. For example one type of primer may contain a non-nucleotide modification which is not present in the other.

In all embodiments of the invention, primers for solid-phase amplification are preferably immobilised by single point covalent attachment to the solid support at or near the 5′ end of the primer, leaving the template-specific portion of the primer free to anneal to its cognate template and the 3′ hydroxyl group free for primer extension. Any suitable covalent attachment means known in the art may be used for this purpose. The chosen attachment chemistry will depend on the nature of the solid support, and any derivatisation or functionalisation applied to it. The primer itself may include a moiety, which may be a non-nucleotide chemical modification, to facilitate attachment. In a particular embodiment, the primer may include a sulphur-containing nucleophile, such as phosphorothioate or thiophosphate, at the 5' end. In the case of solid-supported polyacrylamide hydrogels (as described below), this nucleophile will bind to a bromoacetamide group present in the hydrogel. A more particular means of attaching primers and templates to a solid support is via 5' phosphorothioate attachment to a hydrogel comprised of polymerised acrylamide

and *N*-(5-bromoacetamidylpentyl) acrylamide (BRAPA), as described fully in WO05065814.

Certain embodiments of the invention may make use of solid supports comprised of an inert substrate or matrix (e.g. glass slides, polymer beads, etc) which has been "functionalised", for example by application of a layer or coating of an intermediate material comprising reactive groups which permit covalent attachment to biomolecules, such as polynucleotides. Examples of such supports include, but are not limited to, polyacrylamide hydrogels supported on an inert substrate such as glass. In such embodiments, the biomolecules (e.g. polynucleotides) may be directly covalently attached to the intermediate material (e.g. the hydrogel), but the intermediate material may itself be non-covalently attached to the substrate or matrix (e.g. the glass substrate). The term "covalent attachment to a solid support" is to be interpreted accordingly as encompassing this type of arrangement.

The pooled samples may be amplified on beads wherein each bead contains a forward and reverse amplification primer. In a particular embodiment, the library of templates prepared according to the first, second or third aspects of the invention is used to prepare clustered arrays of nucleic acid colonies, analogous to those described in US20050100900, US7115400, WO 00/18957 and WO 98/44151, the contents of which are incorporated herein by reference in their entirety, by solid-phase amplification and more particularly solid phase isothermal amplification. The terms 'cluster' and 'colony' are used interchangeably herein to refer to a discrete site on a solid support comprised of a plurality of identical immobilised nucleic acid strands and a plurality of identical immobilised complementary nucleic

acid strands. The term 'clustered array' refers to an array formed from such clusters or colonies. In this context the term 'array' is not to be understood as requiring an ordered arrangement of clusters.

The term solid phase, or surface, is used to mean either a planar array wherein primers are attached to a flat surface, for example, glass, silica or plastic microscope slides or similar flow cell devices; beads, wherein either one or two primers are attached to the beads and the beads are amplified; or an array of beads on a surface after the beads have been amplified.

Clustered arrays can be prepared using either a process of thermocycling, as described in WO9844151, or a process whereby the temperature is maintained as a constant, and the cycles of extension and denaturing are performed using changes of reagents. Such isothermal amplification methods are described in patent application numbers WO0246456and US20080009420 (Isothermal methods for creating clonal single molecule arrays), which are incorporated herein by reference in their entirety. Due to the lower temperatures required in the isothermal process, this is particularly preferred.

Use in sequencing/methods of sequencing

The invention also encompasses methods of sequencing amplified nucleic acids generated by solid-phase amplification. Thus, the invention provides a method of nucleic acid sequencing comprising amplifying a pool of nucleic acid templates using solid-phase amplification as described above and carrying out a nucleic acid sequencing reaction to determine the sequence of the whole or a part of at least one amplified nucleic acid strand produced in the

solid-phase amplification reaction, as well as the sequence of the tag that codes for the identity of the sample.

Sequencing can be carried out using any suitable sequencing technique, wherein nucleotides are added successively to a free 3' hydroxyl group, resulting in synthesis of a polynucleotide chain in the 5' to 3' direction. The nature of the nucleotide added is preferably determined after each nucleotide addition. Sequencing techniques using sequencing by ligation, wherein not every contiguous base is sequenced, and techniques such as massively parallel signature sequencing (MPSS) where bases are removed from, rather than added to the strands on the surface are also within the scope of the invention, as are techniques using detection of pyrophosphate release (pyrosequencing). Such pyrosequencing based techniques are particularly applicable to sequencing arrays of beads where the beads have been amplified in an emulsion such that a single template from the library molecule is amplified on each bead.

The initiation point for the sequencing reaction may be provided by annealing of a sequencing primer to a product of the solid-phase amplification reaction. In this connection, one or both of the adaptors added during formation of the template library may include a nucleotide sequence which permits annealing of a sequencing primer to amplified products derived by whole genome or solid-phase amplification of the template library.

Depending on the embodiment of the invention, the tag sequence and fragment sequence may be determined in a single read from a single sequencing primer, or in multiple reads from two sequencing primers. In the case of two reads from two sequencing primers, the 'tag read' and the 'fragment

read' can be performed in either order, with a suitable denaturing step to remove the annealed primer after the first sequencing read is completed. Suitable denaturing steps may include formamide, hydroxide or heat as described elsewhere.

The products of solid-phase amplification reactions wherein both forward and reverse amplification primers are covalently immobilised on the solid surface are so-called 'bridged' structures formed by annealing of pairs of immobilised polynucleotide strands and immobilised complementary strands, both strands being attached to the solid support at the 5' end. Arrays comprised of such bridged structures provide inefficient templates for nucleic acid sequencing, since hybridisation of a conventional sequencing primer to one of the immobilised strands is not favoured compared to annealing of this strand to its immobilised complementary strand under standard conditions for hybridisation.

In order to provide more suitable templates for nucleic acid sequencing, it is preferred to remove substantially all or remove or displace at least a portion of one of the immobilised strands in the 'bridged' structure in order to generate a template which is at least partially single-stranded. The portion of the template which is single-stranded will thus be available for hybridisation to a sequencing primer. The process of removing all or a portion of one immobilised strand in a 'bridged' double-stranded nucleic acid structure may be referred to herein as 'linearisation', and is described in further detail in WO07010251, the contents of which are incorporated herein by reference in their entirety.

Bridged template structures may be linearised by cleavage of one or both strands with a restriction endonuclease or by cleavage of one strand with a nicking endonuclease. Other methods of cleavage can be used as an alternative to restriction enzymes or nicking enzymes, including *inter alia* chemical cleavage (e.g. cleavage of a diol linkage with periodate), cleavage of abasic sites by cleavage with endonuclease (for example 'USER', as supplied by NEB, part number M5505S), or by exposure to heat or alkali, cleavage of ribonucleotides incorporated into amplification products otherwise comprised of deoxyribonucleotides, photochemical cleavage or cleavage of a peptide linker.

It will be appreciated that a linearization step may not be essential if the solid-phase amplification reaction is performed with only one primer covalently immobilised and the other in free solution.

Following the cleavage step, regardless of the method used for cleavage, the product of the cleavage reaction may be subjected to denaturing conditions in order to remove the portion(s) of the cleaved strand(s) that are not attached to the solid support. Suitable denaturing conditions, for example sodium hydroxide solution, formamide solution or heat, will be apparent to the skilled reader with reference to standard molecular biology protocols (Sambrook *et al.*, 2001, *Molecular Cloning, A Laboratory Manual*, 3rd Ed, Cold Spring Harbor Laboratory Press, Cold Spring Harbor Laboratory Press, NY; Current Protocols, eds Ausubel *et al.*). Denaturation results in the production of a sequencing template which is partially or substantially single-stranded. A sequencing reaction may then be initiated by hybridisation of a sequencing primer to the single-stranded portion of the template.

Thus, the invention encompasses methods wherein the nucleic acid sequencing reaction comprises hybridising a sequencing primer to a single-stranded region of a linearised amplification product, sequentially incorporating one or more nucleotides into a polynucleotide strand complementary to the region of amplified template strand to be sequenced, identifying the base present in one or more of the incorporated nucleotide(s) and thereby determining the sequence of a region of the template strand.

One preferred sequencing method which can be used in accordance with the invention relies on the use of modified nucleotides having removable 3′ blocks, for example as described in WO04018497 and US7057026, the contents of which are incorporated herein by reference in their entirety. Once the modified nucleotide has been incorporated into the growing polynucleotide chain complementary to the region of the template being sequenced there is no free 3'-OH group available to direct further sequence extension and therefore the polymerase can not add further nucleotides. Once the nature of the base incorporated into the growing chain has been determined, the 3' block may be removed to allow addition of the next successive nucleotide. By ordering the products derived using these modified nucleotides, it is possible to deduce the DNA sequence of the DNA template. Such reactions can be done in a single experiment if each of the modified nucleotides has a different label attached thereto, known to correspond to the particular base, to facilitate discrimination between the bases added during each incorporation step. Alternatively, a separate reaction may be carried out containing each of the modified nucleotides separately.

The modified nucleotides may carry a label to facilitate their detection. A fluorescent label, for example, may be used for detection of modified nucleotides. Each nucleotide type may thus carry a different fluorescent label, for example, as described in US Provisional Application No. 60/801,270 (Novel dyes and the use of their labelled conjugates), published as WO07135368, the contents of which are incorporated herein by reference in their entirety. The detectable label need not, however, be a fluorescent label. Any label can be used which allows the detection of an incorporated nucleotide.

One method for detecting fluorescently labelled nucleotides comprises using laser light of a wavelength specific for the labelled nucleotides, or the use of other suitable sources of illumination. The fluorescence from the label on the nucleotide may be detected by a CCD camera or other suitable detection means. Suitable instrumentation for recording images of clustered arrays is described in US Provisional Application No. 60/788,248 (Systems and devices for sequence by synthesis analysis), published as WO07123744, the contents of which are incorporated herein by reference in their entirety.

The invention is not intended to be limited to use of the sequencing method outlined above, as essentially any sequencing methodology which relies on successive incorporation of nucleotides into a polynucleotide chain can be used. Suitable alternative techniques include, for example, Pyrosequencing™, FISSEQ (fluorescent in situ sequencing), MPSS and sequencing by ligation-based methods, for example as described is US6306597.

The nucleic acid sample may be further analysed to obtain a second read from the opposite end of the fragment.

Methodology for sequencing both ends of a cluster are described in co-pending applications WO07010252 and PCTGB2007/003798, the contents of which are incorporated by reference herein in their entirety. In one example, the series of steps may be performed as follows; generate clusters, linearise, hybridise first sequencing primer and obtain first sequencing read. The first sequencing primer can be removed, a second primer hybridised and the tag sequenced. The nucleic acid strand may then be 'inverted' on the surface by synthesising a complementary copy from the remaining immobilised primers used in cluster amplification. This process of strand resynthesis regenerates the double stranded cluster. The original template strand can be removed, to linearise the resynthesised strand that can then be annealed to a sequencing primer and sequenced in a third sequencing run.

In the cases where strand resynthesis is employed, both strands must be immobilised to the surface in a way that allows subsequent release of a portion of the immobilised strand. This can be achieved through a number of mechanisms as described in WO07010251, the contents of which are incorporated herein by reference in their entirety. For example, one primer can contain a uracil nucleotide, which means that the strand can be cleaved at the uracil base using the enzymes uracil glycosylase (UDG) which removes the nucleoside base, and endonuclease VIII that excises the abasic nucleotide. This enzyme combination is available as USER™ from New England Biolabs (NEB part number M5505). The second primer may comprise an 8-oxoguanine nucleotide, which is then cleavable by the enzyme FPG (NEB part number M0240). This design of primers give complete control of which primer is cleaved at which point in the process, and also where in the cluster the cleavage occurs. The primers may also be chemically modified, for example with a disulfide or diol

modification that allows chemical cleave at specific locations.

Kits

The invention also relates to kits for use in preparing libraries of tagged polynucleotides using the method of the first, second or third aspects of the invention.

Embodiments of the kit comprise at least a supply of a mismatched adaptor as defined herein, plus a supply of at least one amplification primer which is capable of annealing to the mismatched adaptor and priming synthesis of an extension product, which extension product would include any target sequence ligated to the adaptor when the adaptor is in use.

Particular features of the 'mismatch' adaptors for inclusion in the kit are as described elsewhere herein in relation to other aspects of the invention. The structure and properties of amplification primers are well known to those skilled in the art. Suitable primers of appropriate nucleotide sequence for use with the adaptors included in the kit can be readily prepared using standard automated nucleic acid synthesis equipment and reagents in routine use in the art. The kit may include a supply of one single type of primer or separate supplies (or even a mixture) of two different primers, for example a pair of amplification primers suitable for PCR or isothermal amplification of templates modified with the mismatched adaptor in solution phase and/or on a suitable solid support (i.e. solid-phase amplification). The kit may comprise a mismatch double stranded adapter for ligation to a sample of interest, plus

at least two different amplification primers that carry a different tag sequence, where the tag sequence does not hybridise to the adapter. This kit can be used to amplify at least two different samples where each sample is amplified using a single tagged primer, and then pooled after the individual amplification reactions.

In one embodiment the kit may include supplies of different primer-pairs for use in solution phase and solid phase PCR and more particularly isothermal amplification. In this context the 'different' primer-pairs may be of substantially identical nucleotide sequence but differ with respect to some other feature or modification, such as for example surface-capture moieties, etc. In other embodiments the kit may include a supply of primers for use in an initial primer extension reaction and a different primer-pair (or pairs) for solution and/or solid phase amplification.

Adaptors and/or primers may be supplied in the kits ready for use, or more preferably as concentrates requiring dilution before use, or even in a lyophilised or dried form requiring reconstitution prior to use. If required, the kits may further include a supply of a suitable diluent for dilution or reconstitution of the primers. Optionally, the kits may further comprise supplies of reagents, buffers, enzymes, dNTPs, etc., for use in carrying out PCR or isothermal amplification. Suitable (but non-limiting) examples of such reagents are as described in the Materials and Methods sections of the accompanying Examples. Further components which may optionally be supplied in the kit include 'universal' sequencing primers suitable for sequencing templates prepared using the mismatched adaptors and primers.

The invention will be further understood with reference to the following non-limiting experimental examples:

Examples

The following are examples of general techniques which may be applied in carrying out the method of the invention.

Example 1: Preparation and conventional sequencing of tagged libraries:

The following experimental details describe the complete exposition of one embodiment of the invention as described above.

Two libraries were made, one using purified human BAC DNA (140k human chromosome 6 insert cloned into a pTARBAC vector) and the other purified PhiX174 RF1 DNA (NEB, N3021L). Each library was prepared separately. The DNA was first prepared for ligation to forked adaptors by: fragmentation of the DNA by nebulisation, end repair of the DNA ends to render them blunt-ended and phosphorylated, then the addition of a single 'A' nucleotide onto the 3' ends of the DNA fragments. The ligation reaction was performed with the prepared fragmented DNA and adaptors pre-formed by annealing 'Oligo A' and 'Oligo B' (sequences given below). The same adaptors were used for each library. The product of the reaction was isolated/purified from unligated adaptor by gel electrophoresis. Finally, the product of the ligation reaction was subjected to cycles of PCR to selectively amplify ligated product that contained genomic DNA with adaptor at both ends of the fragments. During the PCR a unique tag was added to each library using a unique PCR