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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
28 July 2016 (28.07.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/118883 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01) *G06F 19/22* (2011.01)
C12N 15/10 (2006.01)

(21) International Application Number:
PCT/US2016/014559

(22) International Filing Date:
22 January 2016 (22.01.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/106,967 23 January 2015 (23.01.2015) US

(71) Applicant: **WASHINGTON UNIVERSITY** [US/US]; One Brookings Drive, St. Louis, Missouri 63130 (US).

(72) Inventors: **DRULEY, Todd E.**; c/o WASHINGTON UNIVERSITY, One Brookings Drive, St. Louis, Missouri 63130 (US). **YOUNG, Andrew**; c/o WASHINGTON UNIVERSITY, One Brookings Drive, St. Louis, Missouri 63130 (US).

(74) Agents: **RILEY-VARGAS, Rebecca** et al.; Polsinelli PC, 100 South Fourth Street, Suite 1000, St. Louis, Missouri 63102 (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, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

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

Published:

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

— *with sequence listing part of description (Rule 5.2(a))*

WO 2016/118883 A1

(54) Title: DETECTION OF RARE SEQUENCE VARIANTS, METHODS AND COMPOSITIONS THEREFOR

(57) Abstract: The present disclosure encompasses methods of error corrected sequencing (ECS) that enable detection of very rare mutations well below the error rate of convention next generation sequencing (NGS). Further, the methods disclosed herein enable multiplex targeting of genomic DNA.

DETECTION OF RARE SEQUENCE VARIANTS, METHODS AND COMPOSITIONS THEREFOR

GOVERNMENTAL RIGHTS

[0001] This invention was made with government support under 1K08CA140720-01A1 awarded by the NIH. The government has certain rights in the invention.

CROSS REFERENCE TO RELATED APPLICATIONS

[0002] This application claims the benefit of U.S. Provisional Application number 62/106,967, filed January 23, 2015, the disclosure of which is hereby incorporated by reference in its entirety.

FIELD OF THE INVENTION

[0003] The present disclosure encompasses methods of error corrected sequencing (ECS) that enable detection of very rare mutations well below the error rate of conventional next generation sequencing (NGS). Further, the methods disclosed herein enable multiplex targeting of genomic DNA.

BACKGROUND OF THE INVENTION

[0004] Massively parallel next generation sequencing is a powerful tool for whole genome sequencing. Its low cost relative to prior methods and ease in automation allow for large scale analyses of large genomes or many samples with an error rate of 1%. For many sequencing applications, this is sufficient, however when searching for rare mutations in a heterogeneous population, this 1% error rate can confound the isolation of single base mutations in a small population of cells with technical sequencing errors. Detecting rare mutations at 2-5% variant allele fraction (VAF) using current methods requires costly and time-intensive deep resequencing, and lower-frequency variants are undetectable regardless of sequencing depth.

[0005] Several groups have tried to mitigate this problem through a variety of methods, including counting PCR amplicons (Casbon et al. 2011), large amounts of template and small numbers of cycles combined with statistical analyses (Flaherty et al.

2012), tagging of DNA molecules during initial PCR (Miner et al 2004, Jabara et al. 2011, Smith et al 2014 and Schmitt et al. 2012), and performing hybridization capture reactions in lieu of PCR (Hiatt et al 2013). The methods of Casbon and Flaherty require complex mathematical models on the current data and are unsuitable for high throughput applications.

[0006] Previous implementations of error-corrected next-generation sequencing (NGS) have limitations that have hampered their clinical applicability. First, some methods cannot be targeted and are not compatible with multiplexing, which limits their ability to handle mammalian-sized genomes (Lou et al., 2013; Schmitt et al., 2012). The method of Schmitt also hinges on obtaining sequencing reads of both strands of the same molecule. In theory, this would mean about half of the sequencing power would be lost due to pairing up of data strands, however, due to experimental limitations, nearly three quarters of the sequencing reads are not included in the data analyses. While this is acceptable for some applications, Illumina® sequencing methods are expensive and this method of error correction requires wasting resources on data that are never going to be analyzed.

[0007] Several other targeted methods require large amounts of starting material. Schmitt's method as described uses 3 µg of DNA isolated from a phage library in *Escherichia coli* for library preparation. Jabara used 10,000 RNA molecules from a single HIV strain. Kinde and colleagues (2011) used a DNA library from 100,000 cells to isolate rare mutations using low-efficiency two-dimensional capture arrays. Such amounts of template are not available for sequencing genomic DNA samples of limited quantity.

SUMMARY OF THE INVENTION

[0008] In an aspect, the disclosure provides a method of identifying a genetic mutation in a biological sample comprising nucleic acid obtained from a subject. The method comprises: (a) amplifying one or more regions of interest from the biological sample comprising nucleic acid, wherein a plurality of amplicons for each region of interest are generated; (b) attaching an adapter and a random component to each amplicon generated in (a) and amplifying; (c) sequencing the amplicons

comprising the random component generated in (b), wherein redundant reads are generated and wherein the redundant reads are grouped by the random component and a consensus sequence is identified; and (d) comparing the consensus sequence to a reference sequence, wherein a consensus sequence that differs from the reference sequence comprises a genetic mutation.

[0009] In another aspect, the disclosure provides a method of identifying a genetic mutation in a biological sample comprising nucleic acid obtained from a subject. The method comprises: (a) hybridizing a primer pool comprising one or more primer pairs specific to one or more regions of interest from the biological sample comprising nucleic acid, extending from an upstream primer of the primer pair to a downstream primer of the primer pair, and ligating the extension product to the downstream primer of the primer pair, wherein products comprising the regions of interest flanked by sequences required for amplification are generated; (b) attaching an adapter comprising a random component and attaching an adapter comprising an index sequence to the products from (a) and amplifying; (c) sequencing the products comprising the random component generated in (b), wherein redundant reads are generated and wherein the redundant reads are grouped by the random component and a consensus sequence is identified; and (d) comparing the consensus sequence to a reference sequence, wherein a consensus sequence that differs from the reference sequence comprises a genetic mutation.

[0010] In still another aspect, the disclosure provides a method of detecting minimal residual disease (MRD) in a subject. The method comprises: (a) hybridizing a primer pool comprising one or more primer pairs specific to one or more regions of interest from a biological sample comprising nucleic acid obtained from the subject, extending from an upstream primer of the primer pair to a downstream primer of the primer pair, and ligating the extension product to the downstream primer of the primer pair, wherein products comprising the regions of interest flanked by sequences required for amplification are generated; (b) attaching an adapter comprising a random component and attaching an adapter comprising an index sequence to the products from (a) and amplifying; (c) sequencing the products comprising the random component

generated in (b), wherein redundant reads are generated and wherein the redundant reads are grouped by the random component and a consensus sequence is identified; and (d) comparing the consensus sequence to a reference sequence, wherein a consensus sequence that differs from the reference sequence comprises a genetic mutation and is indicative of MRD.

BRIEF DESCRIPTION OF THE FIGURES

[0011] The application file contains at least one drawing executed in color. Copies of this patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

[0012] **FIG. 1A, FIG. 1B, FIG. 1C, FIG. 1D, FIG. 1E, FIG. 1F, FIG. 1G** and **FIG. 1H** depict graphs showing benchmarking for ECS and the identification of rare pre-leukemic mutations. (**FIG. 1A, FIG. 1B**) DNA extracted from a diagnostic leukemia sample with known mutations in *RUNX1* (**FIG. 1A**) and *IDH2* (**FIG. 1B**) was serially diluted into non-cancer, unrelated human DNA. Two replicates were run per sample/dilution. The coefficient of determination (r^2) between diluted tumor concentration in the sample and VAF in the generated read families was 0.9999 and 0.9991 for *RUNX1* and *IDH2*, respectively. (**FIG. 1C**) The VAF at every nucleotide not expected to contain mutations in the dilution series experiment were analyzed to determine the error profile of the error-corrected consensus sequences compared with conventional deep sequencing. A cumulative distribution function of VAF demonstrated a reduced error profile in read families relative to conventional deep sequenced reads. (**FIG. 1D**) The most frequent class of substitution seen in read families was in G to T (C to A) transversions, which was consistent with oxidative conversion of guanine to 8-oxo-guanine. (**FIG. 1E, FIG. 1F**) The leukemia-specific variants identified in *ASXL1* and *U2AF1* at diagnosis (circled) were not distinguishable from sequencing errors in the same substitution class by conventional deep sequencing. (**FIG. 1G, FIG. 1H**) Targeted error-corrected sequencing identified the *ASXL1* variant in the 2002 banked sample at 0.004 VAF and the *U2AF1* variant in the 2004 banked sample at 0.009 VAF.

[0013] **FIG. 2A, FIG. 2B, FIG. 2B, FIG. 2C, FIG. 2D, FIG. 2E** and **FIG. 2F** depict schematics of the error-corrected sequencing workflow. Schematic depiction of library preparation (**FIG. 2A, FIG. 2B, FIG. 2C**) and bioinformatics analysis (**FIG. 2D, FIG. 2E, FIG. 2F**) for generating read families and error-corrected consensus sequences. First, the region of interest is amplified from genomic DNA (**FIG. 2A**), then the sequencing library is prepared (**FIG. 2B**) generating a sequence library (**FIG. 2C**). From the sequence library, read families are generated (**FIG. 2D**) and an error-corrected consensus sequence (ECCS) is created (**FIG. 2E**). The ECCSs are aligned to identify a variant allele (**FIG. 2F**).

[0014] **FIG. 3** depicts a graph showing the cumulative distribution function of the error profile comparing ECS to conventional deep sequencing. The variant allele fraction for each non-variant position covered in the dilution series experiment was sorted and plotted cumulatively. The variant allele fractions of errors were higher in every nucleotide covered across all substitution types for the raw sequenced reads compared the error-corrected consensus sequences generated from read families.

[0015] **FIG. 4** depicts a graph showing the cumulative distribution function of read family error profile per specific substitution type with and without FPG pretreatment. The error profile of G to T (C to A) substitutions, consistent with guanine oxidation to 8-oxo guanine, was higher than the other classes of mutations. The C to T (G to A) substitutions, consistent with cytosine deamination to uracil, was visible just over the error profile for the remaining 8 types of substitutions (inset). FPG pretreatment did not appreciably change the error profile.

[0016] **FIG. 5A, FIG. 5B, FIG. 5C, FIG. 5D, FIG. 5E** and **FIG. 5F** depict graphs showing *ASXL1* mutations over time in UPN684949. Formalin-fixed paraffin-embedded bone marrow samples were banked over three years (2002, 2003, 2004) from this individual. Conventional deep sequencing (**FIG. 5A, FIG. 5B, FIG. 5C**) only distinguished the *ASXL1* variant from the T to G sequencing errors in the 2003 banked sample at 0.097 VAF (**FIG. 5B**). **FIG. 5A** is samples from 2002 and **FIG. 5C** is samples from 2004. Correcting the sequencing errors with ECS clearly identified the *ASXL1*

variant at 0.0042 VAF in 2002 (**FIG. 5D**), 0.092 VAF in 2003 (**FIG. 5E**) and 0.029 VAF in 2004 (**FIG. 5F**).

[0017] **FIG. 6A, FIG. 6B, FIG. 6C, FIG. 6D, FIG. 6E** and **FIG. 6F** depict graphs showing *U2AF1* mutations over time in UPN684949. Formalin-fixed paraffin-embedded bone marrow samples were banked over three years (2002, 2003, 2004) from this individual. Conventional deep sequencing (**FIG. 6A, FIG. 6B, FIG. 6C**) only distinguished the *U2AF1* variant from the G to T sequencing errors in the 2003 banked sample at 0.036 VAF (**FIG. 6B**). **FIG. 6A** is samples from 2002 and **FIG. 6C** is samples from 2004.Correcting the sequencing errors with ECS did not identify the *U2AF1* variant in 2002 (**FIG. 6D**), but did identify the *U2AF1* variant at 0.031 VAF in 2003 (**FIG. 6E**) and 0.0089 VAF in 2004 (**FIG. 6F**).

[0018] **FIG. 7** depicts a graph showing the error profile observed with increased read family size. Read families generated with 3x or greater coverage (solid line) had a higher cumulative distribution of erroneous substitutions called compared to read families with 5x or greater coverage (dotted line).

[0019] **FIG. 8** depicts a graph showing the representative distribution of read family size. Singletons represent index sequences containing a sequencing error. Excluding singletons, the median read family size was 7x (mean 7.4x). Only read families with 5-20 reads were included in ECS analysis.

[0020] **FIG. 9A** and **FIG. 9B** depict a method of multiplex targeted genomic capture using the error-corrected sequencing methodology. **FIG. 9A** depicts (a) the annealing of primers to genomic DNA, (b) single strand extension, and (c) ligation. **FIG. 9B** depicts (d) the newly minted single-stranded amplicon after capture, (e) attachment of an adapter with a sample-specific index (fixed) via PCR, (f) attachment of an adapter with an ECS index (random) via PCR, and (g) amplifying of this molecule to make read families.

[0021] **FIG. 10A** and **FIG. 10B** depict graphs showing that the amplicon coverage between replicates is correlated. **FIG. 10A** shows to libraries sequenced on the same run with an R^2 of 0.9718 and **FIG. 10B** shows two libraries sequenced on different runs with an R^2 of 0.7536.

[0022] **FIG. 11** depicts a graph showing that the coverage per amplicon is variable.

[0023] **FIG. 12** depicts graphs showing the identification of constitutional and, importantly, rare SNVs in different samples. 49 germline (~0.5/1.0 VAF) SNVs were identified, 5 high VAF (0.14-0.36 VAF) SNVs were identified, and 106 low VAF (<0.1 VAF) SNVs were identified for a total of 160 SNVs identified.

[0024] **FIG. 13** depicts graphs showing that rare subclones are detected longitudinally in the same healthy individual.

[0025] **FIG. 14** depicts a graph showing that total rare subclonal variants detected per individual. The majority of the subclonal variants were detected in exonic regions.

[0026] **FIG. 15A** and **FIG. 15B** depict pie charts showing the classification of detected rare subclonal variants. **FIG. 15A** shows the detected rare variants by function of which the majority are exonic. **FIG. 15B** shows the detected exonic rare variants by function of which the majority are nonsynonymous SNVs.

[0027] **FIG. 16** depicts a graph showing that the detected exonic variants cluster in *DNMT3A* and *TET2*.

[0028] **FIG. 17** depicts a graph showing that intronic variants are more evenly distributed.

[0029] **FIG. 18A** and **FIG. 18B** depict graphs show that the variants are not exclusively called in highly covered amplicons. **FIG. 18A** shows the histogram coverage per amplicon. **FIG. 18B** show the histogram coverage per amplicon with variants called.

[0030] **FIG. 19A** and **FIG. 19B** depict graphs showing the target space per gene does not correlated with SNV calls per gene. **FIG. 19A** shows the exonic mutations per target space in the panel. **FIG. 19B** shows the intronic mutations per target space in the panel.

[0031] **FIG. 20** depicts the distribution of exonic mutations by gene.

[0032] **FIG. 21** depicts the spectrum of *DNMT3A* mutations.

[0033] **FIG. 22** depicts the spectrum of *TET2* mutations.

[0034] **FIG. 23** depicts the distribution of rare subclonal mutations per person.

[0035] **FIG. 24** depicts the COSMIC variants analyzed to validate the ECS methodology disclosed herein with ddPCR.

[0036] **FIG. 25A** and **FIG. 25B** depict the concordance of VAF measured by ECS and ddPCR. **FIG. 25A** shows that VAF measured by ECS is highly correlated with VAF measured by ddPCR (R^2=0.98). **FIG. 25B** shows that even when specifically focusing on a VAF of <0.01, ECS and ddPCR still correlated (R^2=0.72).

[0037] **FIG. 26** depicts the number of singleton ddPCR droplets generated by flow sorting cells from the study participants and extracting genomic DNA from those flow sorted cells.

[0038] **FIG. 27** depicts graphs showing that sublconal mutations are present in multiple lineages in all tested samples.

[0039] **FIG. 28** depicts graphs showing and expanded view of **FIG. 27** focusing on a VAF of <0.01.

DETAILED DESCRIPTION OF THE INVENTION

[0040] The present inventors have developed sequencing methods for identification of rare mutations. Methods of present disclosure can be used to quantify rare somatic mutations, such as, for example, DNA from clinical specimens. Importantly, the limit of detection for the disclosed method is at least two orders of magnitude below the error rate of the Illumina® sequencing platform performed by standard methods.

[0041] In various embodiments, methods of the present disclosure involve PCR amplification of multiple regions of interest in the genome, attaching adaptors comprising a random component and/or index sequences to the amplified DNA, performing sequencing, creating read families of the same index sequence and comparing reads in the same family. By these methods, true variations in the sequence can be distinguished from technical artifacts.

[0042] The methodology disclosed herein is described in greater detail below.

(a) sample preparation

[0043] The disclosure encompasses a method of identifying a rare sequence in a sample comprising nucleic acid. In an embodiment, the disclosure encompasses a method of identifying a genetic mutation in a sample comprising nucleic acid. Specifically, the disclosure encompasses a method of identifying a genetic mutation in a biological sample comprising nucleic acid obtained from a subject. A first iteration of the method may be used to query about 1 to about 20 genomic loci (e.g. region of interest). A second iteration of the method may be used to query about 1 to about 600 or more genomic loci.

[0044] A region of interest may be any nucleic acid amenable to standard PCR. Non-limiting examples of a region of interest may be a nucleic acid used to identify a rare mutation or low levels associated with drug-resistance, graft rejection, residual disease, tumors, immune diseases, fetal DNA, and microbial infection or contamination. With respect to microbial infection or contamination, a region of interest may be a nucleic acid used to identify a bacterial strain. It is known in the art that 16S nucleic acid is a good, widely used nucleic acid to identify a bacterial strain. In an embodiment, the primer pair comprises sequences complementary to a 16S nucleic acid sequence. In another embodiment, the region of interest may be one or more nucleic acids used to diagnose cancer, wherein a mutation within that region of interest is indicative of cancer. Specifically, the region of interest may be one or more nucleic acids used to diagnose leukemia. For example, the region of interest may be any nucleic acid known to be mutated in leukemia.

[0045] The sample comprising nucleic acid may be a sample from a subject, the environment, a laboratory, or any sample in which nucleic acid is present. When the sample is from a subject, the sample may be from stool, sputum, urine, plasma, peripheral blood, serum, bone marrow, tissue, and other bodily fluids. The tissue sample may be a tissue biopsy. The biopsied tissue may be fixed, embedded in paraffin or plastic, and sectioned, or the biopsied tissue may be frozen and cryosectioned. Alternatively, the biopsied tissue may be processed into individual cells or an explant, or processed into a homogenate, a cell extract, a membranous fraction,

or a protein extract. The sample may be used “as is” or the nucleic acid may be purified from the sample prior to sample preparation.

[0046] The subject may be a rodent, a human, a livestock animal, a companion animal, or a zoological animal. In one embodiment, the subject may be a rodent, e.g. a mouse, a rat, a guinea pig, etc. In another embodiment, the subject may be a livestock animal. Non-limiting examples of suitable livestock animals may include pigs, cows, horses, goats, sheep, llamas and alpacas. In still another embodiment, the subject may be a companion animal. Non-limiting examples of companion animals may include pets such as dogs, cats, rabbits, and birds. In yet another embodiment, the subject may be a zoological animal. As used herein, a “zoological animal” refers to an animal that may be found in a zoo. Such animals may include non-human primates, large cats, wolves, and bears. In a preferred embodiment, the subject is a human.

i. first iteration

[0047] In an aspect, a method of the disclosure comprises, in part, amplifying one or more regions of interest from a biological sample comprising nucleic acid. The amplification generates a plurality of amplicons for each region of interest.

[0048] Amplification takes place in the presence of one or more primer pairs. A first primer of the primer pair comprises a sequence complementary to an upstream portion of the region of interest and a second primer of the primer pair comprises a sequence complementary to a downstream portion of the region of interest. The primer pairs are designed to anneal to complementary strands of nucleic acid (i.e. on primer of the primer pair anneals to the sense strand and one primer of the primer pair anneals to the antisense strand). The complementary sequence may be altered based on the region of interest to be amplified. The complementary sequences of the primer pair may comprise about 10 to about 100 nucleotides complementary to the region of interest. For example, the complementary sequences of the primer pair may comprise about 15 to about 50 nucleotides complementary to the region of interest. In an embodiment, the complementary sequences of the primer pair may comprise about 20 to about 40 nucleotides complementary to the region of interest. In another

embodiment, the complementary sequences of the primer pair may comprise about 20 to about 35 nucleotides complementary to the region of interest.

[0049] One or more primer pairs is contacted with a sample comprising nucleic acid. Nucleic acid may be, for example, RNA or DNA. Modified forms of RNA or DNA may be used. In an exemplary embodiment, the nucleic acid is genomic DNA. The amount of nucleic acid in the sample may be about 200 to about 1000 ng or more. For example, the amount of nucleic acid in the sample may be about 400 to about 800 ng. In certain embodiment, the amount of nucleic acid in the sample is about 200 ng, about 300 ng, about 400 ng, about 500 ng, about 600 ng, about 700 ng, about 800 ng, about 900 ng or about 1000 ng or more. In some embodiments, the amount of nucleic acid in the sample may be about 1 µg, about 5 µg, about 10 µg, about 20 µg, about 30 µg, about 40 µg, or about 50 µg. It is important to note that as the amount of nucleic acid increases, the amount of random components (described below) must proportionally increase to ensure that the same random component is not utilized twice. A person of skill in the art would understand how to scale the methodology based on the amount of nucleic acid used.

[0050] In general, amplification of the region of interest is carried out using polymerase chain reaction (PCR). A PCR reaction may comprise sample comprising nucleic acid, one or more primer pairs, polymerase, water, buffer, and deoxynucleotide triphosphates (dNTPs) in a single reaction vial. PCR may be performed according to standard methods in the art. By way of non-limiting example, the PCR reaction may comprise denaturation, followed by about 15 to about 30 cycles of denaturation, annealing and extension, followed by a final extension. In an exemplary embodiment, the PCR reaction comprises denaturation at about 98°C for about 30 seconds, followed by about 15 to about 30 cycles of (about 98°C for about 10 seconds, about 62-72°C for about 30 seconds, about 72°C for about 30 seconds), followed by a final extension at about 72°C for about 2 minutes.

[0051] In certain embodiments, a single reaction vial is used per primer pair. In other embodiments, a single reaction vial comprises more than one primer pair such that more than one region of interest is amplified per reaction vial. More

specifically, amplification of a region of interest may be multiplexed. Accordingly, a single reaction comprises primer pairs sufficient to amplify about 1-5, about 5-10, about 10-15, or about 15-20 regions of interest. In other embodiments, a single reaction comprises primer pairs sufficient to amplify 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, or 20 regions of interest.

[0052] In a different embodiment, a single reaction vial may comprise more than one primer pair and the amplification may be carried out for about 10-20 cycles, or about 15-20 cycles. Then, the amplicons may be separated into more than one reaction vial, one primer pair is then added to each reaction vial, and the reaction may be carried out for an additional about 10-20 cycles, or about 15-20 cycles.

[0053] Optionally, the amplicons may be purified prior to attaching an adapter, random component, and/or index sequence (described below in **Section I(b)**). Methods of purifying amplicons are known in the art. For example, AMPure bead cleanup may be used.

ii. second iteration

[0054] In an aspect, a method of the disclosure comprises, in part, hybridizing a primer pool comprising one or more primer pairs specific to one or more regions of interest from the biological sample comprising nucleic acid, extending from an upstream primer of the primer pair to a downstream primer of the primer pair, and ligating the extension product to the downstream primer of the primer pair. The hybridization, extension and ligation generates products comprising the regions of interest flanked by sequences required for amplification.

[0055] The primer pool comprises one or more primer pairs designed to anneal to the same strand of nucleic acid. A first primer of the primer pair comprises a sequence complementary to an upstream portion of the region of interest and a second primer of the primer pair comprises a sequence complementary to a downstream portion of the region of interest. The complementary sequence may be altered based on the region of interest to be amplified. The complementary sequences of the primer pair may comprise about 10 to about 100 nucleotides complementary to the region of interest. For example, the complementary sequences of the primer pair may comprise

about 15 to about 50 nucleotides complementary to the region of interest. In an embodiment, the complementary sequences of the primer pair may comprise about 20 to about 40 nucleotides complementary to the region of interest. In another embodiment, the complementary sequences of the primer pair may comprise about 20 to about 35 nucleotides complementary to the region of interest. In an exemplary embodiment, the primer pool is the TruSight® Myeloid Sequencing Panel (Illumina).

[0056] The primer pool is contacted with a sample comprising nucleic acid. Nucleic acid may be, for example, RNA or DNA. Modified forms of RNA or DNA may be used. In an exemplary embodiment, the nucleic acid is genomic DNA. The amount of nucleic acid in the sample may be about 200 to about 1000 ng or more. For example, the amount of nucleic acid in the sample may be about 400 to about 800 ng. In certain embodiment, the amount of nucleic acid in the sample is about 200 ng, about 300 ng, about 400 ng, about 500 ng, about 600 ng, about 700 ng, about 800 ng, about 900 ng or about 1000 ng or more. In some embodiments, the amount of nucleic acid in the sample may be about 1 µg, about 5 µg, about 10 µg, about 20 µg, about 30 µg, about 40 µg, or about 50 µg. It is important to note that as the amount of nucleic acid increases, the amount of random components (described below) must proportionally increase to ensure that the same random component is not utilized twice. A person of skill in the art would understand how to scale the methodology based on the amount of nucleic acid used.

[0057] Hybridization of the primer pool to the nucleic acid may be done via methods standard in the art. For example, the primer pool and nucleic acid may be incubated at elevated temperature for about 1 to about 2 hours. More specifically, the primer pool and nucleic acid may be incubated at about 95^{o}C for about 1 minute and then the temperature may be allowed to decrease to about 40^{o}C for about 80 minutes.

[0058] Following hybridization, a polymerase extends from the upstream primer through the region of interest, followed by ligation to the 5' end of the downstream primer using ligase. This process results in the formation of products comprising the regions of interest flanked by sequences required for amplification. If the nucleic acid is DNA, the polymerase may be any suitable DNA polymerase known in the

art. Further, if the nucleic acid is DNA, the ligase may be any suitable DNA ligase known in the art. Extension and ligation may be carried out via methods standard in the art and dependent upon the polymerase and ligase utilized. For example, extension and ligation may be conducted at about 37°C for about 45 minutes.

[0059] Optionally, following hybridization, the unbound primers may be washed away prior to proceeding to the extension and ligation step. Methods of washing away unbound primers are known in the art.

[0060] A single reaction vial comprises the entire primer pool such that more than one region of interest per reaction vial may be amplified downstream in the method. More specifically, the method enables multiplex targeting from genomic DNA. Accordingly, a single reaction comprises primer pairs sufficient to hybridize to about 1-5 or about 5-10, about 10-20, about 20-30, about 30-40, about 40-50, about 50-60, about 60-70, about 70-80, about 80-90, about 90-100, about 100-150, about 150-200, about 200-250, about 250-300, about 300-350, about 350-400, about 400-450, about 450-500, about 500-550, about 550-600, about 600-700, about 700-800, about 800-900, or about 900-1000 regions of interest. Alternatively, a single reaction comprises primer pairs sufficient to hybridize to more than 100, more than 150, more than 200, more than 250, more than 300, more than 350, more than 400, more than 450, more than 500, more than 550, more than 600, more than 650, more than 700, more than 750, more than 800, more than 850, more than 900, more than 950, more than 1000, more than 1050, more than 1100, more than 1150, more than 1200, more than 1250, more than 1300, more than 1350, more than 1400, more than 1450, more than 1500, more than 1550, more than 1600, more than 1650, more than 1700, more than 1750, more than 1800, more than 1850, more than 1900, more than 1950, or more than 2000 regions of interest. In certain embodiments, a single reaction comprises primer pairs sufficient to hybridize to about 100, about 150, about 200, about 250, about 300, about 350, about 400, about 450, about 500, about 550, about 600, about 650, about 700, about 750, about 800, about 850, about 900, about 950, about 1000, about 1050, about 1100, about 1150, about 1200, about 1250, about 1300, about 1350, about 1400, about 1450, about 1500, about 1550, about 1600, about 1650, about 1700, about 1750, about 1800,

about 1850, about 1900, about 1950, or about 2000 regions of interest. In other embodiments, a single reaction comprises primer pairs sufficient to hybridize to about 5, about 10, about 15, about 20, about 25, about 30, about 35, about 40, about 45, about 50, about 55, about 60, about 65, about 70, about 75, about 80, about 85, about 90, about 95, or about 100 regions of interest. In still other embodiments, a single reaction comprises primer pairs sufficient to hybridize to about 500, about 510, about 515, about 520, about 525, about 530, about 535, about 540, about 545, about 550, about 555, about 560, about 565, about 570, about 575, about 580, about 585, about 590, about 595, or about 600 regions of interest.

(b) error-corrected sequencing library preparation

[0061] A method of the disclosure comprises, in part, attaching an adapter, random component, and/or index sequence to each amplicon or product generated in **Section I(a)**.

[0062] As used herein, an "adapter" is a sequence that permits universal amplification. A key feature of the adapter is to enable the unique amplification of the amplicon or product only without the need to remove existing template nucleic acid or purify the amplicons or products. This feature enables an "add only" reaction with fewer steps and ease of automation. The adapter is attached to the 5' and 3' end of the amplicon or product. The adapter may be Y-shaped, U-shaped, hairpin-shaped, or a combination thereof. In a specific embodiment, the adaptor is Y-shaped. In an exemplary embodiment, the adapter may be an Illumina adapter for Illumina sequencing.

[0063] As used herein, a "random component" is composed of random nucleotides to generate a complexity of random components far greater than the number of unique amplicons or products to be sequenced. This ensures that having the same random component attached to multiple amplicons or products is an extremely statistically improbable event. A random component may also be referred to as a barcode. The random component design can theoretically generate $9.1x10^8$ to $1.4x10^{10}$ unique random components. This complexity can easily be expanded by increasing the length of the random regions in the random component. In an embodiment, the random

component may be about 5 to about 100 nucleotides. In another embodiment, the random component may be about 10 to about 25 nucleotides. For example, the random component may be about 15 to about 20 nucleotides. In still another embodiment, the random component is about 16 to about 18 nucleotides. Accordingly, the random component may be 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24 or 25 or more nucleotides. The random component is attached to the 5' or 3' end of the amplicon or product. In a specific embodiment, the random component is attached to the 5' end of the amplicon or product.

[0064] In addition to a random component and an adapter, an index sequence may also be attached to each amplicon or product generated. The addition of an index sequence allows pooling of multiple samples into a single sequencing run. This greatly increases experimental scalability, while maintaining extremely low error rates and conserving read length. The index sequence may be about 5 to about 10 nucleotides. Accordingly, the index sequence may be 5, 6, 7, 8, 9 or 10 or more nucleotides. In an embodiment, the index sequence is about 6 nucleotides.

[0065] In a specific embodiment, an adapter, a random component and an index sequence are attached to each amplicon or product. In an embodiment, a nucleotide sequence comprising an adaptor and a random component is attached to the 5' end of each amplicon or product and a nucleotide sequence comprising an adaptor and an index sequence is attached to the 3' end. In another embodiment, a nucleotide sequence comprising an adaptor and a random component is attached to the 3' end of each amplicon or product and a nucleotide sequence comprising an adaptor and an index sequence is attached to the 5' end. In still another embodiment, a nucleotide sequence comprising an adaptor, a random component and an index sequence is attached to the 5' end and a nucleotide sequence comprising an adaptor is attached to the 3' end. In still yet another embodiment, a nucleotide sequence comprising an adaptor, a random component and an index sequence is attached to the 3' end and a nucleotide sequence comprising an adaptor is attached to the 5' end. In an exemplary embodiment, a nucleotide sequence comprising SEQ ID NO:1 is attached to each amplicon or product at the 5' end and a nucleotide sequence comprising SEQ ID NO:2

is attached to each amplicon or product at the 3' end. In another exemplary embodiment, a nucleotide sequence comprising SEQ ID NO:2 is attached to each amplicon or product at the 5' end and a nucleotide sequence comprising SEQ ID NO:1 is attached to each amplicon or product at the 3' end.

[0066] The nucleotide sequence comprising an adapter, a random component and/or an index sequence may be attached to the amplicon or product via methods known in the art. In certain embodiments, the nucleotide sequence comprising an adapter, a random component and/or an index sequence is ligated to an amplicon or product via methods standard in the art. For example, the nucleotide sequence is annealed at about 95°C for about 5 minutes, then the temperature is decreased by about 1°C, about every 30 seconds until about 4°C. Enrichment of the properly ligated products is then carried out. Methods of enriching properly ligated products are known in the art. For example, PCR amplification is carried out using the ligation product and appropriate primers. In an exemplary embodiment, the PCR is carried out as follows: about 98°C for about 30 seconds, followed by about 6 cycles of about 98°C for about 10 seconds, about 57°C for about 30 seconds, and about 72°C for about 30 seconds, finishing with an extension at about 72°C for about 2 minutes.

[0067] In other embodiments, the nucleotide sequence comprising an adapter, a random component and/or an index sequence is attached to an amplicon or product via PCR. For example, the amplicon or product may be contacted with a nucleotide sequence comprising an adaptor and an index sequence and a PCR reaction is conducted. Then, this product is contacted with a nucleotide sequence comprising an adaptor and a random component and a PCR reaction is conducted. The resulting product is a nucleotide sequence comprising an adaptor, a random component, a region of interest, an index sequence and a downstream adaptor. Alternatively, the amplicon or product may be contacted with a nucleotide sequence comprising an adaptor and a random component and a PCR reaction is conducted. Then, this product is contacted with a nucleotide sequence comprising an adaptor and an index sequence and a PCR reaction is conducted. The resulting product is a nucleotide sequence comprising an

adaptor, an index sequence, a region of interest, a random component and a downstream adaptor.

[0068] The products or amplicons comprising an adapter, a random component and/or an index sequence are then subjected to exponential PCR. In an embodiment, an exponential PCR reaction may comprise the products or amplicons comprising an adapter, a random component and/or an index sequence, primers, polymerase, water, buffer, and deoxynucleotide triphosphates (dNTPs) in a single reaction vial. Exponential PCR may be performed according to standard methods in the art. By way of non-limiting example, the exponential PCR reaction may comprise denaturation, followed by about 15-30 cycles of denaturation, annealing and extension, followed by a final extension. In a specific embodiment, the exponential PCR reaction comprises denaturation at about 95°C for about 3 minutes, followed by about 16-33 cycles of (about 95°C for about 30 seconds, about 62-72°C for about 30 seconds, about 72°C for about 60 seconds), followed by a final extension at about 72°C for about 5 minutes.

[0069] Upon performing exponential PCR, the products or amplicons comprising an adapter, a random component and/or an index sequence are amplified. The exponential PCR products comprise: an adapter, a random component, a region of interest, a downstream adapter and an index sequence.

[0070] Optionally, the products or amplicons comprising an adapter, a random component and/or an index sequence may be purified prior to exponential PCR. Methods of purifying products or amplicons are known in the art. For example, AMPure bead cleanup may be used.

(c) error-corrected sequencing

[0071] A method of the disclosure comprises, in part, sequencing the exponential PCR product. According to the method of the disclosure, sequencing of the exponential PCR product generates redundant reads. The redundant reads are grouped by random component and a consensus sequence is identified such that the redundant reads mitigate sequence errors.

[0072] Sequencing may be performed according to standard methods in the art. Sequencing is preferably performed on a massively parallel sequencing platform, many of which are commercially available including, but not limited to Illumina, Roche/454, Ion Torrent, and PacBIO. In an exemplary embodiment, Illumina sequencing is used.

[0073] Reads may be separated by the index sequence and trimmed to remove primer sequences. Reads may be grouped by the random component. In certain embodiments, groups of reads with less than three, less than four, or less than 5 reads may be removed. To eliminate ambiguous sequences, the random components may be sorted by abundance and clustered at an identity of about 85%. Alternatively, the random components may be sorted by abundance and clustered at an identity of about 65% to about 95%. The random components may be clustered from most abundant to least abundant. Given that most sequencing errors are random and that the correct sequence should occur more often than a variant with sequencing errors, the abundance-weighted clustering provides a means to eliminate spurious random components that are most likely due to sequencing errors while retaining the more abundant (and most likely true positive) random components.

[0074] This redundant sequencing of each amplicon or product allows the error-correction of each amplicon or product. For example, a consensus sequence is generated for each random component group by scoring and weighing the nucleotide at each base position. Sequences with a consensus sequence that is identical to the most abundant sequence associated with the same random component are kept, this process is called quality filtering. Specifically, at every position, the nucleotides called by each sequence read are compared and a consensus nucleotide is called if there is at least about 90% agreement between the reads. If there is less than about 90% agreement, an “N” is called in the consensus sequence at that position.

[0075] The inventors demonstrated that the methodology disclosed herein was 99% specific to detect variants above 0.0034 VAF for G to T (C to A) substitutions, 0.00020 VAF for C to T (G to A) substitutions, and 0.000079 VAF for the other eight possible substitutions.

(d) comparison to reference sequence

[0076] After an error-corrected consensus sequence (ECCS) has been identified, the ECCS may be compared to a reference sequence to determine the presence of one or more mutations. A reference sequence may be a sequence without any known mutations. A reference sequence without any known mutations may be referred to as a wild-type sequence. In certain embodiments, a reference sequence is a human sequence.

[0077] Comparison of the ECCSs to a reference sequence may identify clinically silent single nucleotide variations (SNVs). Specifically, the method disclosed herein may identify a genetic mutation that is present at a frequency of less than 1 in 1,000 in the sample (0.1%). For example, the method disclosed herein may identify a genetic mutation that is present at a frequency of less than 1 in 1,000, less than 1 in 2,000, less than 1 in 3,000, less than 1 in 4,000, less than 1 in 5,000, less than 1 in 6,000, less than 1 in 7,000, less than 1 in 8,000, less than 1 in 9,000, or less than 1 in 10,000 in the sample. In a specific embodiment, the method disclosed herein may identify a genetic mutation that is present at a frequency of less than 1 in 10,000 in the sample (0.01%).

II. METHODS OF USE

[0078] A method of the invention may be used to quantitate as well as to determine a sequence. For example, the relative abundance of two or more analyte nucleic acid fragments may be compared. A method of the invention may be used to identify rare mutants in a population of DNA templates, to measure polymerase error rates, or to judge the reliability of oligonucleotide synthesis. Additionally, a method of the invention may be used to diagnose, treat or prevent a disease in a subject. Identification of a rare mutation could facilitate the diagnosis of a disease, enable the proper methodology, such as a therapeutic, to treat the disease, or prevent the onset of disease by administration of prophylactic therapies. Still further, a method of the invention may be used to detect genetic mutations involved in cancer or other diseases, such as immune-mediated diseases. In another embodiment, a method of the invention may be used to identify and quantify a microbial infection of a subject. The knowledge

gained may be used to assess the health of the subject. Further, a method of the invention may be used as a quality control measurement in clinical labs or in synthetic biology to determine microbial contamination.

[0079] The results described in the examples below describe a method of identifying ultra-rare pre-leukemic clones using the methodology described above. The methodology disclosed herein substantially improves the accuracy and depth of massively parallel sequencing. Thus, the methodology results in an assay to determine a VAF of 1:10,000 molecules in individuals at high depth with high precision. The methodology disclosed herein may be applied to virtually any sample preparation workflow or sequencing platform. As demonstrated here, the approach can easily be used to identify rare or low abundant mutations indicative of disease, such as leukemia.

EXAMPLES

[0080] The following examples are included to demonstrate preferred embodiments of the invention. It should be appreciated by those of skill in the art that the techniques disclosed in the examples that follow represent techniques discovered by the inventors to function well in the practice of the invention, and thus can be considered to constitute preferred modes for its practice. However, those of skill in the art should, in light of the present disclosure, appreciate that many changes can be made in the specific embodiments which are disclosed and still obtain a like or similar result without departing from the spirit and scope of the invention.

Example 1. Quantifying ultra-rare pre-leukemic clones via targeted error-corrected sequencing.

[0081] The quantification of rare clonal and subclonal populations from a heterogeneous DNA sample has multiple clinical and research applications for the study and treatment of leukemia. Specifically, in the hematopoietic compartment, recent reports demonstrate the presence of subclonal variation in normal and malignant hematopoiesis,[1,2] and leukemia is now recognized as an oligoclonal disease.[3] Currently, clonal heterogeneity in leukemia is studied using next-generation sequencing (NGS) targeting subclone-specific mutations. With this method, detecting mutations at 2–5%

variant allele fraction (VAF) requires costly and time-intensive deep resequencing and identifying lower frequency variants is impractical regardless of sequencing depth. Recently, various methods have been developed to circumvent the error rate of NGS.[4,5] These methods tag individual DNA molecules with unique oligonucleotide indexes, which enable error correction after sequencing.

[0082] Here we present a direct application of error-corrected sequencing (ECS) to study clonal heterogeneity during leukemogenesis and validate the accuracy of this method with a series of benchmarking experiments. Specifically, we demonstrate the ability of ECS to identify leukemia-associated mutations in banked pre-leukemic blood and bone marrow from patients with either therapy-related acute myeloid leukemia (t-AML) or therapy-related myelodysplastic syndrome (t-MDS). T-AML/t-MDS occurs in 1–10% of individuals who receive alkylator- or epipodophyllotoxin-based chemotherapy or radiation to treat a primary malignancy.[6] For the seven individuals surveyed in this study, matched leukemia/normal whole-genome sequencing identified the t-AML/t-MDS-specific somatic mutations present at diagnosis. We applied our method for ECS to identify leukemia-specific mutations in four individuals from DNA extracted from blood and bone marrow samples collected years before diagnosis. In a separate study into the role of *TP53* mutations in t-AML/t-MDS leukemogenesis, this method was used to identify leukemia-associated mutations at low frequency in samples banked years before diagnosis.[7] In two cases, subclones were identified below the 1% threshold of detection governed by conventional NGS. These results highlight the ability of targeted ECS to identify clinically silent single-nucleotide variations (SNVs).

[0083] We employed ECS by tagging individual DNA molecules with adapters containing 16 bp random oligonucleotide molecular indexes in a manner similar to other reports.[4,5,8] Our implementation of ECS easily targets loci of interest by single or multiplex PCR and inserts seamlessly into the standard NGS library preparation (**FIG. 2, Methods for Example 1**). Our deviations from the standard protocol are ligation of customized adapters containing random indexes instead of the manufacturer's supplied adapters and a quantitative PCR (qPCR) quantification step before sequencing (**Table**

2). Following sequencing, sequence reads containing the same index and originating from the same molecule are grouped into read families. Sequencing errors are identified by comparing reads within a read family and removed to create an error-corrected consensus sequence (ECCS). We performed a dilution series experiment to assess bias during library preparation and determine the limit of detection for ECS. For this experiment, we spiked DNA from a t-AML sample into control human DNA, which was serially diluted over five orders of magnitude. The experiment was comprised of two technical replicates targeting two separate mutations (20 total independent libraries). The results demonstrate that ECS is quantitative to a VAF of 1:10 000 molecules and provides a highly reproducible digital readout of tumor DNA prevalence in a heterogeneous DNA sample (r^2 of 0.9999 and 0.9991, **FIG. 1A, FIG. 1B**). We next characterized the error profile based on the wild-type nucleotides included in the dilution series experiment. Variant identification using the ECCSs was 99% specific at a VAF of 0.0016 versus 0.0140 for deep sequencing alone (**FIG. 1C**). We noticed that ECCS errors were heavily biased towards G to T transversions and to a lesser degree C to T transitions (**FIG. 1D, FIG. 3**), as previously observed.[4,9] When separated by substitution type, variants identified from the ECCSs were 99% specific at a VAF of 0.0034 for G to T (C to A) mutations, 0.00020 for C to T (G to A) mutations and 0.000079 for the other eight possible substitutions. Although excess G to T mutations are a known consequence of DNA oxidation leading to 8-oxo-guanine conversion,[4] the pretreatment of samples with formamidopyrimidine-DNA glycosylase before PCR amplification did not appreciably improve the error profile of G to T mutations (**FIG. 4**).

[0084] As proof of principle, we applied ECS to study rare pre-leukemic clonal hematopoiesis in seven individuals who later developed t-AML/t-MDS. Leukemia/normal whole-genome sequencing at diagnosis was used to identify the leukemia-specific somatic mutations in each patient's malignancy (**Table 3**). We applied targeted ECS to query these 18 different loci in 10 cryopreserved or formalin-fixed paraffin-embedded blood and bone marrow samples that were 9–22-year old and banked up to 12 years before diagnosis (**Table 4**).

[0085] We generated ~ 25 Gb of 150 bp paired-end reads from six Illumina (San Diego, CA, USA) MiSeq runs. We targeted 1–7 somatic mutations per individual (25 mutations spanning 5.5 kb from 15 genes in total) and identified leukemia-specific subclonal populations in four individuals up to 12 years before diagnosis (**Table 1**). For each sequencing library, we tagged ~2.5 million locus-specific amplicons generated from genomic DNA using high-fidelity PCR with randomly indexed custom adapters. Sequencing errors were removed to create ECCSs as described above. Each ECCS was then aligned to the reference genome for variant calling (**FIG. 2**).

[0086] Using conventional deep sequencing, we detected t-AML/t-MDS-specific mutations in prior banked samples at variant allele fractions between 0.03 and 0.87 (data not shown). In one individual (UPN 684949), deep sequencing alone was insufficient to distinguish known *ASXL1* and *U2AF1* mutations from the sequencing errors in samples banked 5 and 3 years before t-MDS diagnosis, respectively (**FIG. 1E, FIG. 1F**). However, ECS identified the L866* nonsense mutation in *ASXL1* at a VAF of 0.004 (**FIG. 1G**) and the S34Y missense mutation in *U2AF1* at a VAF of 0.009 (**FIG. 1H**. In addition, ECS was able to temporally quantify these mutations from three pre-t-MDS samples banked yearly from 3 to 5 years before diagnosis (**FIG. 5, FIG. 6**). In two cases (UPN643006 and UPN942008), only a subset of the variants identified at diagnosis were present in the prior banked sample (**Table 1**). Specifically, in the UPN643006 sample, banked 12 years before diagnosis, a single-nucleotide deletion in *ASXL1* was present at VAF 0.03. But, the G to T substitution in *ASXL1*, CTT deletion in *GATA2* and G to T substitution in *U2AF1* were not detectable in this prior banked sample.

[0087] Here we present a practical and clinically oriented application for targeted error-corrected NGS utilizing single molecule indexing. This method easily integrates into existing NGS library preparation protocols and enables the quantification of previously undetectable mutations in heterogeneous DNA samples. A modification to the standard NGS library preparation is the replacement of the stock adapters with our randomly indexed adapters and the addition of a qPCR step before sequencing. The qPCR step limits the number of molecules sequenced, ensuring adequate coverage for each read family. With these two modifications, we achieve

highly specific detection for rare mutations. The bioinformatics analysis is straightforward and does not require proprietary algorithms or tools (**Methods for Example 1**). Our results highlight the ability of this method to identify rare subclonal populations in a heterogeneous biological sample. As applied to t-AML/t-MDS, we show these previously undetectable mutations are present years before diagnosis and fluctuate in prevalence over time.

[0088] A clinical application of ECS is to quantify minimal residual disease (MRD). As the genomic characterization of leukemia becomes more readily available, identifying causative genetic lesions and rare therapy-resistant subclones will become increasingly useful for risk stratification, therapeutic selection and disease monitoring. Already, whole-genome sequencing of AML has demonstrated that nearly every case of AML harbors one or more somatic SNVs.[10] These SNVs are more reliable clonal markers of malignancy than cell surface markers, which can change over time. Leveraging this information, conventional NGS was implemented retrospectively to detect MRD harboring leukemia-specific insertions/deletions (indels) as rare as 0.00001 VAF in *NPM1*[11] and 0.0001 VAF in *RUNX1*.[12] This was possible because indels are only rarely generated erroneously by NGS. Unfortunately, measuring rare leukemia-associated substitutions is limited owing to the relatively high error profile of conventional NGS.[13] However, ECS can achieve the 1:10,000 limit of detection featured by conventional MRD platforms.[14] For patients whose leukemia lacks suitable markers for conventional MRD, ECS could offer an alternative with comparable sensitivity and specificity that is easy to implement in a clinical sequencing lab. Furthermore, the ability to multiplex targets for ECS enables the surveillance of known mutations and the simultaneous discovery of new somatic mutations. Ongoing work will directly compare gold-standard MRD methods with targeted ECS in patients with and without relapsed leukemia.

Methods for Example 1.

[0089] *Study Design:* Blood and bone marrow samples from patients treated for t-AML/t-MDS at Washington University were banked or accessed following informed consent under Human Research Protection Protocol #201011766. Patients included in this study underwent matched leukemia and non-cancer (skin) whole

genome sequencing on the Illumina HiSeq 2500 platform, which identified tumor-specific somatic coding mutations in leukemia samples. Our study focused on identifying these known mutations from matched blood or bone marrow samples banked 1-12 years prior to the initial diagnosis of t-AML/t-MDS.

[0090] *Sample Preparation:* Genomic DNA was generated from either FFPE or cryopreserved peripheral blood or bone marrow samples using the QIAamp DNA FFPE Tissue or DNA Mini Kit (Qiagen). PCR primers were designed using primer3[1] to amplify regions harboring individual leukemia-specific mutations from the banked biological samples (**Table 5**). The concentration of each purified DNA sample was determined using the Qubit dsDNA HS Assay Kit (Life Technologies). Genomic DNA (400-800 ng) was amplified using the Q5 High-Fidelity 2X Master Mix (New England Biolabs) in a 25 uL reaction with 0.5 uM primers (**FIG. 2A**). The following conditions were used: 98C for 30s; 16-30 cycles of 98C for 10s, 62-72C (based on a separate optimization) for 30s and 72C for 30s; 72C for 2m; hold 10C. The PCR reactions were purified using the Agencourt AMPure XP (Beckman Coulter) bead-based protocol without modification.

[0091] For a few of the patient samples, the amount of input genomic DNA was limited. In these cases, modifications were made to the protocol to amplify multiple leukemia-specific mutations from the same biological sample (multiplex PCR). Patient-specific primers were pooled during a first round of PCR and amplified for roughly 16 cycles, similar to pre-amplification described in TAm- Seq[2]. After purification the DNA was split into a single PCR reaction per patient-specific SNVs and amplified using only that specific primer pair, again for roughly 16 cycles. This allowed us to generate diverse amplicon pools for multiple loci using only 400-800 ng of starting DNA.

[0092] *ECS Library Preparation:* The concentration of the purified PCR products was measured using the Qubit dsDNA HS Assay Kit (Life Technologies). NGS libraries were prepared from 800 ng of amplicons for each sample/mutation using the Illumina TruSeq DNA Sample Preparation Kit (Illumina). We replaced the Illumina-provided Y-shaped adapters with custom adapters containing a random 16 base pair oligonucleotide index sequence (**Table 2**). Adapters were diluted to 40 uM in Tris-EDTA

with 5 nM NaCl and annealed using the following conditions: 95C for 5m then decreased by 1C every 30s to 4C. Aside from the custom adapters used for ligation, the library preparation protocol from Illumina was mostly unchanged (**FIG. 2B**). Enrichment for correctly ligated products was completed using a 50 uL Q5 PCR amplification with 2 uL of ligation product and 0.5 uM Illumina specific primers under the following conditions: 98C for 30s; 6 cycles of 98C for 10s, 57C for 30s and 72C for 30s; 72C for 2m; hold 10C The PCR reaction was purified using a modified Ampure bead cleanup, which increased the size range of purification to remove adapter dimers. 100 uL of beads were washed twice with ddH2O to remove the stock polyethylene glycol (PEG) solution. The solution was replaced with 25.5 uL 50% wt/vol PEG (Sigma), 37.5 uL 5M NaCl and 37 uL ddH2O. The PCR reaction was added to this solution and purified per the standard Ampure protocol.

[0093] *Quantification by qPCR:* We sought to generate read families from a single randomly-indexed molecule with roughly seven-fold coverage. Given the bandwidth of a single Illumina MiSeq run was roughly 15-18 million read pairs, we sought to generate sequencing libraries from roughly 2.5 million molecules. To achieve this, we quantified the concentration of each library using the qPCR NGS Library Quantification Kit, Illumina GA (Agilent Technologies). Based on the measured concentration, each library was diluted to 0.4 pM such that a 10 uL volume of the diluted library would contain ~2.5 million molecules. The 10 uL aliquot of diluted sequencing library was then amplified for 16-20 cycles and purified with the same Q5 and modified Ampure bead protocol used for the previous enrichment PCR step. The final library was visualized on a 2% SYBR Safe gel (Life Technologies) and quantified using Qubit dsDNA HS Assay Kit. When multiplexing samples on a single lane of sequencing, individual sequencing libraries were combined in equimolar amounts after enrichment PCR and the pooled sample was diluted and quantified using qPCR as stated previously. However, we also found it possible to pool amplicons in equimolar amounts after the initial genomic DNA amplification and make a single sequencing library. Up to 7 different amplicons were multiplexed on a single MiSeq run. Multiplexing was only

possible with mutations in different genes or within different exons of the same gene because the samples were demultiplexed by alignment.

[0094] *Sequencing:* Each library was sequenced on the Illumina MiSeq instrument as specified by the manufacturer (**FIG. 2C**). Approximately, 5-10% of PhiX control DNA was spiked into each sequencing experiment. Each completed sequencing run contained roughly 15-18M paired-end 150 bp reads. Raw sequence reads were aligned to the PhiX genome using Bowtie 2[3]. Sequence reads aligning to PhiX were removed from further analysis. The remaining sequence reads were aligned to UCSC hg19/GRCh37 using Bowtie 2 for comparison against error-corrected consensus sequences (ECCS) derived from read families (below).

[0095] *Error Corrected Consensus Sequences:* Sequence reads containing the same index sequence (originated from the same randomly-indexed molecule) were aligned to each other to generate read families in a fashion similar to previously published methods[4,5] (**FIG. 2D**). Previous studies used a minimum read family size of three[5]. We found using a more stringent cutoff of five reduced the error rate in the read families (**FIG. 7**). The median read family size was seven reads per index (**FIG. 8**). Paired-end reads within a read family were error corrected in a stepwise fashion (**FIG. 2E**). First, at every position, the nucleotides called by each sequence read were compared and a consensus nucleotide was called if there was at least 90% agreement between the reads. If there was less than 90% agreement, an N was called in the consensus sequence at that position. Errors that occurred during library preparation and sequencing were removed because they were not shared between different reads within a read family. Second, an ECCS was thrown out if less than 90% of the 300 nucleotides comprising the paired-end read were assigned a non-N nucleotide. These ECCSs were locally aligned to UCSC hg19/GRCh30 using Bowtie2[3] (**FIG. 2F**). The aligned ECCSs were processed with Mpileup[6] using the parameters *–BQ0 –d 10000000000000*. This removed the coverage thresholds to ensure that all of the pileup output was returned regardless of variant allele fraction (VAF) or coverage. Variant allele factions comprised of both the expected mutations and the background errors for each sample were visualized using IGV[7] and graphically represented using ggplot2[8]. Each known variant

was plotted relative to the error-profile of that specific substitution class (e.g. an expected C to T transition was compared against the C to T error profile). Variants distinguishable from the noise for that specific error class and located at the expected position within the amplicon were called true positives. The threshold for calling true variants varied based on the error profile of that substitution class. Based on our benchmarking studies we were 99% specific to detect variants above 0.0034 VAF for G to T (C to A) substitutions, 0.00020 VAF for C to T (G to A) substitutions and 0.000079 VAF for the other eight possible substitutions.

Table 1. Patient-specific leukemia-associated somatic mutations identified by ECS.

UPN	*Sample ID*	*Years prior*	*Gene*	*Chr*	*Position*	*Mut*	*Amino-acid change*	*Variant RFs*	*Reference RFs*	*VAF*
446294	75.02	1	OBSCN	1	228461129	A to G	H1857R	61 238	156 986	0.2806
			TP53	17	7578271	T to A	H193L	220 551	110 047	0.6671
499258	24.06	2	RUNX1	21	36252865	C to G	R139P	2	486 196	0
574214	26.04	7	DMD	X	32827676	G to A	R187*	7	199 945	0
643006	80.01	12	ASXL1	20	31022448	G to T	G645C	7	85 781	0.0001
			ASXL1	20	31022442	del G	G645fs	2 898	82 245	0.034
			GATA2	3	128200135	del CTT	K390in_fr_del	0	4 187	0
			U2AF1	21	44524456	G to T	S34Y	85	414 613	0.0002
684949	91.01	5	ASXL1	20	31023112	T to G	L866*	3 583	853 598	0.0042
			U2AF1	21	44524456	G to T	S34Y	545	514 410	0.0011
	92.02	4	ASXL1	20	31023112	T to G	L866*	54 074	535 976	0.0916
			U2AF1	21	44524456	G to T	S34Y	11 195	355 276	0.0305
	93.01	3	ASXL1	20	31023112	T to G	L866*	17 319	573 629	0.0293
			U2AF1	21	44524456	G to T	S34Y	827	92 104	0.0089
856024	30.02	1	S100A4	1	153517192	A to G	F27L	0	211 512	0
			IGSF8	1	160062252	G to A	P516S	0	22 614	0
			PLA2R1	2	160798389	A to G	L1431P	2	338 616	0
			POU3F2	6	99282794	C to A	S15R	8	201 240	0
			ANKRD18B	9	33524645	G to A	C53Y	7	214 836	0
			ESR2	14	64701847	G to A	A416V	10	135 861	0.0001
			FBN3	19	8155081	G to A	P2029L	0	152 304	0
942008	33.04	9	IDH2	15	90631934	C to T	R88Q	23 170	236 587	0.0892
			RUNX1	21	36231791	T to C	D171G	40	253 168	0.0002
	107.01	<1	IDH2	15	90631934	C to T	R88Q	138 180	161 371	0.4613
			RUNX1	21	36231791	T to C	D171G	368 438	50 796	0.8788

Abbreviations: ECS, error-corrected sequencing; RFs, read families; VAF, variant allele fraction. Two to seven mutations were queried per individual and the number of read families containing the variant allele or reference allele were reported and used to calculate the variant allele fraction.

Table 2. Random 16-mer molecular indexed adapters. The terminal 5-prime phosphorylation on complementary adapter sequence was used to improve ligation efficiency (*).

Label	Sequence	SEQ ID NO:
16N Index Adapter	AGACGGCATACGAGATNNNNNNNNNNNNNNNNGTGA CTGGAGTTCAGACGTGTGCTCTTCCGATCT	1
Complementary Adapter	*GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGAT CTCGGTGGTCGCCGTATCATT	2

Table 3. Whole-genome sequencing of diagnosis t-AML/t-MDS samples.

UPN	Gene	Chr	Position	Mutation	AA Change	Reference Reads	Variant Reads	VAF
446294	OBSCN	1	228461129	A to G	H1857R	3	5	0.63
	TP53	17	7578271	T to A	H193L	79	106	0.57
499258	RUNX1	21	36252865	C to G	R139P	122	17	0.12
574214	DMD	X	32827676	G to A	R187*	103	73	0.41
643006	ASXL1	20	31022448	G to T	G645C	36	32	0.47
	ASXL1	20	31022442	del G	G645fs	33	32	0.49
	GATA2	3	128200135	del CTT	K390in_frame_de	8	10	0.56
	U2AF1	21	44524456	G to T	S34Y	24	27	0.53
684949	ASXL1	20	31023112	T to G	L866*	75	14	0.16
	U2AF1	21	44524456	G to T	S34Y	57	9	0.14
856024	S100A4	1	153517192	A to G	F27L	103	48	0.32
	IGSF8	1	160062252	G to A	P516S	28	42	0.60
	PLA2R1	2	160798389	A to G	L1431P	45	33	0.42
	POU3F2	6	99282794	C to A	S15R	15	15	0.50
	ANKRD18	9	33524645	G to A	C53Y	26	20	0.43
	ESR2	14	64701847	G to A	A416V	40	22	0.35
	FBN3	19	8155081	G to A	P2029L	54	38	0.41
942008	IDH2	15	90631934	C to T	R88Q	10	10	0.50
	RUNX1	21	36231791	T to C	D171G	15	35	0.70

Table 4. Summary of patient information. The type of primary malignancy, the date of primary malignancy diagnosis, the date and type of blood/bone marrow banked prior to t-AML/t-MDS diagnosis and the date of t-AML/t-MDS diagnosis are included in the table below. At t-AML/t-MDS diagnosis, tumor/normal whole genome sequencing identified leukemia-specific mutations. Some of the prior banked blood/bone marrow samples showed evidence of subclonal populations harboring those leukemia-specific mutations before the clinical detection of disease.

UPN	Primary Malignancy Diagnosis	Date Primary Malignancy	Banked Samples	Banking Type	Date Banked	t-AML/t-MDS Diagnosis	Evidence of Pre-Leukemic Subclones

446294	Breast cancer	2002	75.02	FFPE	07/2005	2006 (t-MDS)	Yes
499258	Hodgkin's	1998	24.06	Cryo	02/2002	2004 (t-MDS)	No
574214	Breast cancer	1998	26.04	Cryo	01/2000	2007 (t-MDS)	No
643006	AML	1989	80.01	FFPE	04/1992	2004 (t-MDS)	Yes
684949	CLL	09/1991	91.01	FFPE	11/2002	2007 (t-MDS)	Yes
			92.02	FFPE	09/2003		Yes
			93.01	FFPE	10/2004		Yes
856024	NHL	11/2004	30.02	Cryo	03/2005	2006 (t-AML)	No
942008	NHL	08/1992	33.04	Cryo	09/1996	2005 (t-AML)	Yes
			107.01	FFPE	11/2005		Yes

Table 5. Primers targeting leukemia-specific variants. Primer sequences used to generate variant-specific amplicons from banked genomic DNA samples.

UPN	Gene	FWD Primer	SEQ ID NO:	Reverse Primer	SEQ ID NO:
446294	OBSCN	GGAGCCTCTGACCCTGCA TCCCTCC	3	CCCGCCTCACAGCTGTAC TCCCCAG	4
	TP53	AGACCTCAGGCGGCTCAT AGGGCAC	5	GGGGCTGGAGAGACGACA GGGCTG	6
499258	RUNX1	TCACTAGAATTTTGAAATG TGGGTTTGTTGCC	7	GCACTCTGGTCACTGTGAT GGCTGGC	8
574214	DMD	GGCGATGTTGAATGCATGT TCCAGT	9	AGGACTATGGGCATTGGT TGTCAAT	10
643006	ASXL1	GGACCCTCGCAGACATTAA AGCCCGT	11	GCCTCACCACCATCACCA CTGCTGC	12
	GATA2	CCACAGGTGCCATGTGTC CAGCCAG	13	CTGTGGCGGGGTGGGAG GAATGTTG	14
	U2AF1	TGAACACAAATGGAAAATA CAACTACGAGAGAAAA	15	CCCAGCAAAATAATCAGCT CTCATTTTCCC	16
684949	ASXL1	CACTATGAAGGATCCTGTA AATGTGACCCC	17	TGGTTTGGGCTGTTTCACT ACCTCA	18
	U2AF1	TGAACACAAATGGAAAATA CAACTACGAGAGAAAA	15	CCCAGCAAAATAATCAGCT CTCATTTTCCC	16
856024	S100A4	CCACGTGGGGACTCACTC AGGCA	19	AATAAGACGGTCTCTGTGC CTCCTG	20
	IGSF8	TGGTACACGCCTTCATCCT CGGG	21	GCTCAGCTCTGTCCCTGC CCAGCT	22
	PLA2R1	ACCCTGGTGTCTGTGGCAT TCTCTG	23	AGTCACAGCATCATTCCTC TTGCGGT	24
	POU3F2	CAAATGCGCGGCTCCTTTA ACCGGA	25	GCGTGGCTGAGCGGGTGT CC	26
	ANKRD18B	TACCACATTCGGGACTGG GAACTGC	27	CTCCCAGGGTCCCGGCGA ACTCC	28
	ESR2	TGGCAATCACCCAAACCAA AGCATCGGT	29	AACCCAGATCACCTCGGA GCAGGCG	30
	FBN3	GGGGACACAGTTCGCAGG GGTC	31	GACTGGGGTGCGGGAGGT CACAGG	32
942008	IDH2	GGCGTGCCTGCCAATGGT GATGGG	33	CCGTCTGGCTGTGTTGTT GCTTGGGG	34

	RUNX1	ACATGGTCCCTGAGTATAC CAGCCT	35	GGCCACCAACCTCATTCT GTTTTGT	36

References for Example 1.

1 Holstege H, Pfeiffer W, Sie D, Hulsman M, Nicholas TJ, Lee CC *et al.* Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. *Genome Res* 2014; **24**: 733–742.

2 Walter MJ, Shen D, Ding L, Shao J, Koboldt DC, Chen K *et al.* Clonal architecture of secondary acute myeloid leukemia. *N Engl J Med* 2012; **366**: 1090–1098.

3 Welch JS, Ley TJ, Link DC, Miller CA, Larson DE, Koboldt DC *et al.* The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell* 2012; **150**: 264–278.

4 Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci USA* 2012; **109**: 14508–14513.

5 Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci USA* 2011; **108**: 9530–9535.

6 Godley LA, Larson RA. Therapy-related myeloid leukemia. *Semin Oncol* 2008; **35**: 418–429.

7 Wong T, Ramsingh G, Young AL, Miller CA, Touma W, Welch JS *et al.* The role of TP53 mutations in the origin and evolution of therapy-related AML. *Nature* 2015; **518**: 552–555.

8 Fu GK, Xu W, Wilhelmy J, Mindrinos MN, Davis RW, Xiao W*et al.* Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. *Proc Natl Acad Sci USA* 2014; **111**: 1891–1896.

9 Lou DI, Hussmann Ja, McBee RM, Acevedo A, Andino R, Press WH *et al.* High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc Natl Acad Sci USA* 2013; **110**: 19872–19877.

10 Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* 2013; **368**: 2059–2074.

11 Salipante SJ, Fromm JR, Shendure J, Wood BL, Wu D. Detection of minimal residual disease in NPM1-mutated acute myeloid leukemia by next-generation sequencing. *Mod Pathol* 2014; **27**: 1438–1446.

12 Kohlmann a, Nadarajah N, Alpermann T, Grossmann V, Schindela S, Dicker F *et al.* Monitoring of residual disease by next-generation deep-sequencing of RUNX1 mutations can identify acute myeloid leukemia patients with resistant disease. *Leukemia* 2014; **28**: 129–137.

13 Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 2012; **30**: 434–439.

14 Hourigan CS, Karp JE. Minimal residual disease in acute myeloid leukaemia. *Nat Rev Clin Oncol* 2013; **10**: 460–471.

References for the Methods for Example 1.

1 Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res* 2012; **40**: e115.

2 Forshew T, Murtaza M, Parkinson C, Gale D, Tsui DWY, Kaper F *et al.* Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci Transl Med* 2012; **4**: 136ra68.

3 Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012; **9**: 357–9.

4 Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci U S A* 2011; **108**: 9530–5.

5 Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb L a. Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A* 2012; **109**: 14508– 13.

6 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078–9.

7 Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013; **14**: 178– 92.

8 Wickham H. *ggplot2*. Springer New York: New York, NY, 2009 doi:10.1007/978-0-387-98141-3.

Example 2. MRD testing in AML using error-corrected sequencing.

[0096] In acute myeloid leukemia (AML), minimal residual disease (MRD) testing following treatment is accomplished using multiparameter flow cytometry, which targets clonal cell surface markers; or qPCR, which targets leukemia-associated chromosomal translocations. While both methods provide prognostic information to a detection limit of 1:10,000 cells, these methods are useful in only a subset of leukemia patients[1–3]. Conversely, leukemia-specific somatic mutations occur in virtually every case of AML and present a potential target for residual disease assessment[4,5]. Our goal is to develop a sequencing-based platform to detect rare leukemic cells by their unique somatic mutation profile. Currently, next-generation sequencing is not sensitive enough to detect rare somatic mutations due to a 1% error rate. Fortunately, we have adapted methods for error-corrected sequencing (ECS) to circumvent this limitation[6–9]. Here, we have extended these methods for ECS with leukemia-specific genomic DNA capture to attempt to detect rare (<1%) persistent leukemic cells regardless of their specific somatic mutation profile.

[0097] Remission blood samples from 15 individuals treated for *de novo* AML were acquired. Subclonal somatic mutations in the remission samples were then identified. These results were used to quantify the burden of persistent leukemia. The results obtained were compared to conventional NGS and clinical findings.

[0098] To facilitate leukemia-specific capture Illumina TruSight Myeloid Panel was used. The Panel captures 54 genes via 568 amplicons frequently mutated in AML and targets 141 kb of genomic DNA (**Table 6**). The Panel method is depicted in **FIG. 9A, FIG. 9B**.

[0099] Future directions involve further development of the TruSight capture/ECS protocol, assessment of persistent AML following treatment using

leukemia-specific somatic mutations, and assessment of the role of rare subclones arising in the hematopoietic compartment of healthy individuals.

Table 6. Coverage Details	
Cumulative target region size	~141 kb
Number of target genes	54
Amplicon size	~250 bp
Number of amplicons	568
Recommended mean coverage	5,000x
Target minimum coverage	500x
Percent exons covered at 500x	95

Example 3. Use of TruSight Myeloid panel and ECS protocol in a clinical study of healthy individuals.

[0100] In collaboration with the Nurses Healthy Study, 20 healthy elderly individuals were enrolled to examine the clinical possibilities of the TruSight Myeloid panel and ECS methodology. Paired buffy coat samples were banked 10 years apart. The average age at collection of the first sample was 57.1 years and the average age at collection of the second sample was 68.5. Samples were prepared in duplicate (80 libraries total) using the Illumina TruSight Myeloid panel and the ECS protocol. The samples were sequence on 10 NextSeq High Output (PE150) runs. **Table 7** presents a summary of the sequencing results. The output per run was ~400M PE reads. **Table 8** shows that the libraries appear to be mixed in equimolar ratios. There are approximately 3M read families per library. **FIG. 10** shows that the amplicon coverage between replicates is correlated. **FIG. 10A** shows that two libraries sequenced on the same run (NHS1) had an R-squared value of 0.9718 and **FIG. 10B** shows that two libraries sequenced on different runs (NHS2, NHS6) had an R-squared value of 0.7536. **FIG. 11**, which presents data from DNMT3A, shows that the coverage per amplicon is variable.

Table 7. Summary of Sequencing Results.	
Library	**Sequenced Reads**
NHS1	364,776,941
NHS2	331,997,319

NHS3	361,510,360
NHS4	387,756,648
NHS5	468,765,873
NHS6	433,606,686
NHS7	435,037,421
NHS8	516,437,915
NHS9	519,524,765
NHS10	495,292,729

Table 8. Summary of Sequencing Results.

Library	Index	Demux Reads	Frac of Total	ECS RFs	Frac of Demux Reads
NHS1	CACCACAC	33,657,209	0.092	2,860,180	0.085
NHS1	ACAGTGGT	33,448,218	0.092	2,804,567	0.084
NHS1	ACAAACGG	33,392,978	0.092	2,784,229	0.083
NHS1	ACCCAGCA	39,268,072	0.108	3,318,173	0.085
NHS1	ATCACGAC	38,483,255	0.105	3,333,626	0.087
NHS1	CCCAACCT	35,157,811	0.096	3,027,607	0.086
NHS1	AACCCCTC	41,984,812	0.115	3,558,516	0.085
NHS1	CAGATCCA	35,427,539	0.097	2,987,274	0.084

[0101] Subclonal single nucleotide variations (SNVs) were then called based on the TruSight-ECS libraries generated. A position-specific binomial error model was used to identify rare subclonal SNVs. For each sample, we generated a position specific error profile from all of the sequenced libraries in the study except for samples sequenced from the same individual (the other replicate from the same time point and both replicates from the other time point). Variants were reported if their binomial p-value was less that 0.05 after Bonferroni correction, the variant was observed in at least 5 ECCSs, the VAF was greater than 0.0001, and the variant was identified in at least two replicates from one of the collection time points. The identification of constitutional and rare SNVs in the samples is presented in **FIG. 12**. Fourty-nine germline SNVs were identified (~0.5/1.0 VAF), 5 high VAF were identified (0.14-0.36 VAF) and 106 low VAF were identified (<0.1 VAF) for a total of 160 SNVs detected. Additionally, rare subclones were detected longitudinally in NHS participants (**FIG. 13**). **FIG. 14** presents the total rare subclonal variants detected per individual. The majority of SNVs were present in the exonic regions. The subclonal variants were then classified. As shown in **FIG. 15A**, the majority of rare variants were present in the exonic regions followed by the intronic regions. Rare variants were occasionally found in ncRNA, splicing region and UTR3. **FIG. 15B** shows that the vast majority of rare variants were nonsynonymous SNVs.

Additionally, detected exonic variants clustered in the *DNMT3A* and *TET2* genes (**FIG. 16**). The intronic variants, in contrast, were more evenly distributed (**FIG. 17**). Notably, variants were not exclusively called in highly covered amplicons. **FIG. 18A** shows the histogram coverage per amplicon and **FIG. 18B** shows the histogram coverage per amplicon with variants called. It is also important to note that the target space per gene does not correlate with the SNV calls per gene. **FIG. 19A** shows that the exonic mutations were distributed throughout the target space in the gene and **FIG. 19B** shows that the intronic mutations were also distributed throughout the target space in the gene. **FIG. 20** presents the distribution of exonic mutations by gene which different depending on the gene. For DNMT3A, mutations were only nonsynonymous or stopgain. For TET2 and BCORL1, mutations were nonsynonymous, stopgain and synonymous.

[0102] Given that mutations in *DNMT3A* and *TET2* were the most prevalent mutations, we analyzed where each of the mutations was found on these two genes. Mapping out the spectrum of *DNMT3A* mutations shows the prevalence of early truncating mutations and the numerous missense mutations in the ZFN and methyltransferase domains (**FIG. 21**). Mapping out the spectrum of *TET2* mutations shows several mutation in the oxygenase domain (**FIG. 22**). We then evaluated the distribution of rare subclonal mutations per person (**FIG. 23**). While the majority of individuals have mutations in *DNMT3A*, mutations in other genes were also detected in combination with the *DNMT3A* mutation.

[0103] In summary, we found that rare subclones harboring mutations in leukemia-associated genes are common in healthy individual (19/20 individuals). We also found that subclones frequently harbor mutations in *DNMT3A* (but not R882) and *TET2*. Additionally, since the samples were taken about 10 years apart, we found that subclones are stable over time. Notably, the detection of subclones is not likely due to coverage or target-space bias.

Example 4. VAF measured with ECS correlates with VAF measured with ddPCR.

[0104] We next sought to validate the COSMIC (Catalogue of Somatic Mutation in Cancer) variants detected using ddPCR (**FIG. 24**). In the digital droplet validation method, 21 probes and 150,000 to 450,000 droplets per sample or control

were used. We found that VAF measured by ECS is highly correlated with VAF measured by ddPCR (R^2=0.98) (**FIG. 25A**). When focusing on the VAF of <0.01, the VAF measured by ECF still correlated with the VAF measured by ddPCR (R^2=0.72) (**FIG. 25B**).

[0105] We then sought to identify the subclones found in various cells types. Accordingly, subclone identification using ddPCR was performed on flow sorted buffy coat samples. Samples were selected from 13 individuals and then pan-leukocyte, myeloid, B-cells and T-cells were sorted. The sorting conditions included the following: pan-leukocyte: BV421 anti-CD45; myeloid: APC anti-CD33; B-cells: FITC anti-CD19; and T-cells: PE-CY7 anti-CD3. Enough DNA was extracted from the sorted samples to perform ddPCR without amplification, however variability in flow yield was detected (**FIG. 26**). We found that subclonal mutations are present in multiple cellular lineages in all tested samples (**FIG. 27**). These results are more apparent when focusing specifically on a VAF of <0.01 (**FIG. 28**).

[0106] In summary, we showed that there is a high concordance between VAF measured with ECS and VAF measured with ddPCR. Additionally, we demonstrated that subclonal mutations are present in distinct hematopoietic lineages. It was also demonstrated that subclone identification is improved with indel calling.

CLAIMS

What is claimed is:

1. A method of identifying a genetic mutation in a biological sample comprising nucleic acid obtained from a subject, the method comprising:
    a) amplifying one or more regions of interest from the biological sample comprising nucleic acid, wherein a plurality of amplicons for each region of interest are generated;
    b) attaching an adapter and a random component to each amplicon generated in (a) and amplifying;
    c) sequencing the amplicons comprising the random component generated in (b), wherein redundant reads are generated and wherein the redundant reads are grouped by the random component and a consensus sequence is identified; and
    d) comparing the consensus sequence to a reference sequence, wherein a consensus sequence that differs from the reference sequence comprises a genetic mutation.

2. The method of claim 1, wherein the biological sample comprises about 400 to about 800 ng nucleic acid.

3. The method of claim 1, wherein one region of interest is amplified in step (a).

4. The method of claim 1, wherein more than one region of interest is amplified in step (a).

5. The method of claim 4, wherein the regions of interest are amplified by PCR for about 15 to about 20 cycles generating a plurality of amplicons, the amplicons are separated into more than one reaction vial, one primer pair is added to each reaction vial, and the region of interest in each reaction vial and amplified for about 15 to about 20 cycles.

6. The method of claim 1, wherein step (b) further comprises attaching an index sequence.

7. The method of claim 6, wherein the adapter, the random component and the index sequence are attached via ligation.

8. The method of claim 1, wherein the adapter is a Y-shaped adapter.

9. The method of claim 1, wherein the adapter is an Illumina adapter.

10. The method of claim 1, wherein the adaptor and the random component are attached via ligation.

11. The method of claim 1, wherein the method identifies clinical silent single-nucleotide variations (SNVs).

12. The method of claim 1, wherein the method identifies a genetic mutation that is at a frequency of less than 1 in 1,000 in the sample.

13. The method of claim 1, wherein the method identifies a genetic mutation that is at a frequency of less than 1 in 5,000 in the sample.

14. The method of claim 1, wherein the method identifies a genetic mutation that is at a frequency of less than 1 in 10,000 in the sample.

15. A method of identifying a genetic mutation in a biological sample comprising nucleic acid obtained from a subject, the method comprising:
    a) hybridizing a primer pool comprising one or more primer pairs specific to one or more regions of interest from the biological sample comprising nucleic acid, extending from an upstream primer of the primer pair to a downstream primer of the primer pair, and ligating the extension product to the downstream primer of the primer pair, wherein products comprising the regions of interest flanked by sequences required for amplification are generated;

b) attaching an adapter comprising a random component and attaching an adapter comprising an index sequence to the products from (a) and amplifying;
c) sequencing the products comprising the random component generated in (b), wherein redundant reads are generated and wherein the redundant reads are grouped by the random component and a consensus sequence is identified; and
d) comparing the consensus sequence to a reference sequence, wherein a consensus sequence that differs from the reference sequence comprises a genetic mutation.

16. The method of claim 15, wherein the biological sample comprises about 400 to about 800 ng nucleic acid.

17. The method of claim 15, wherein more than 500 regions of interest are sequenced.

18. The method of claim 15, wherein unbound primers are washed away prior to proceeding to (b).

19. The method of claim 15, wherein the adapter is a Y-shaped adapter.

20. The method of claim 15, wherein the adapter is an Illumina adapter.

21. The method of claim 15, wherein the adapter comprising a random component and the adaptor comprising an index sequence are attached via PCR.

22. The method of claim 15, wherein the method identifies clinical silent single-nucleotide variations (SNVs).

23. The method of claim 15, wherein the method identifies a genetic mutation that is at a frequency of less than 1 in 1,000 in the sample.

24. The method of claim 15, wherein the method identifies a genetic mutation that is at a frequency of less than 1 in 5,000 in the sample.

25. The method of claim 15, wherein the method identifies a genetic mutation that is at a frequency of less than 1 in 10,000 in the sample.

26. A method of detecting minimal residual disease (MRD) in a subject, the method comprising:
    a) hybridizing a primer pool comprising one or more primer pairs specific to one or more regions of interest from a biological sample comprising nucleic acid obtained from the subject, extending from an upstream primer of the primer pair to a downstream primer of the primer pair, and ligating the extension product to the downstream primer of the primer pair, wherein products comprising the regions of interest flanked by sequences required for amplification are generated;
    b) attaching an adapter comprising a random component and attaching an adapter comprising an index sequence to the products from (a) and amplifying;
    c) sequencing the products comprising the random component generated in (b), wherein redundant reads are generated and wherein the redundant reads are grouped by the random component and a consensus sequence is identified; and
    d) comparing the consensus sequence to a reference sequence, wherein a consensus sequence that differs from the reference sequence comprises a genetic mutation and is indicative of MRD.

27. The method of claim 26, wherein the subject is treated if a genetic mutation is detected.

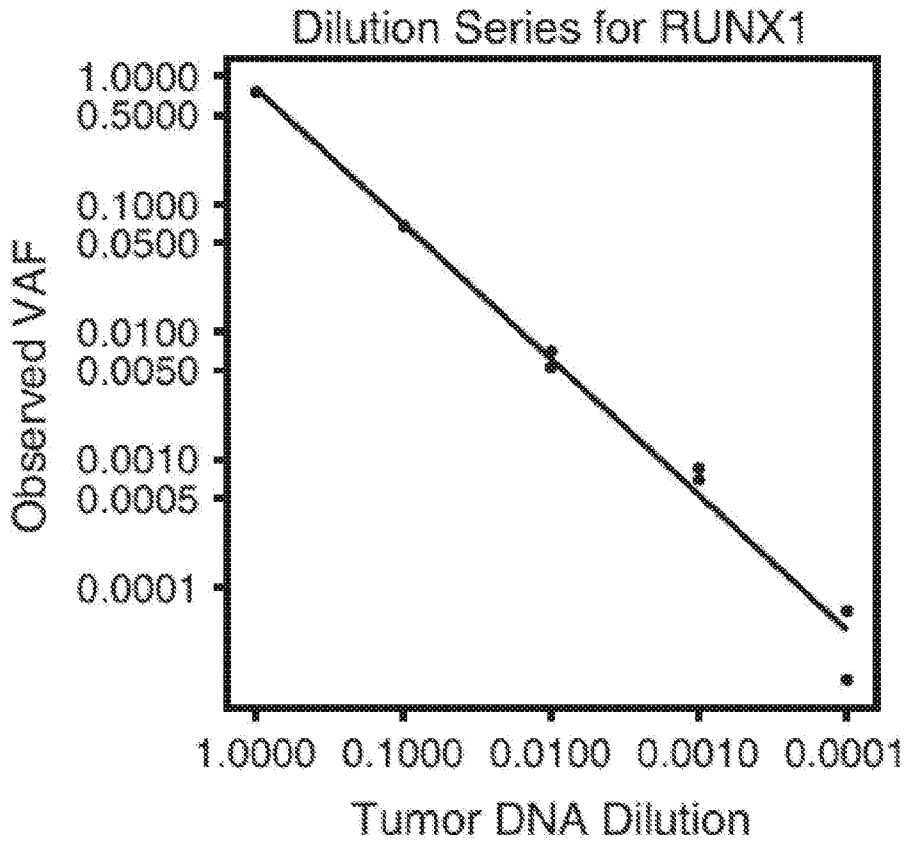
Dilution Series for RUNX1
Observed VAF
1.0000
0.5000
0.1000
0.0500
0.0100
0.0050
0.0010
0.0005
0.0001
1.0000 0.1000 0.0100 0.0010 0.0001
Tumor DNA Dilution

FIG. 1A

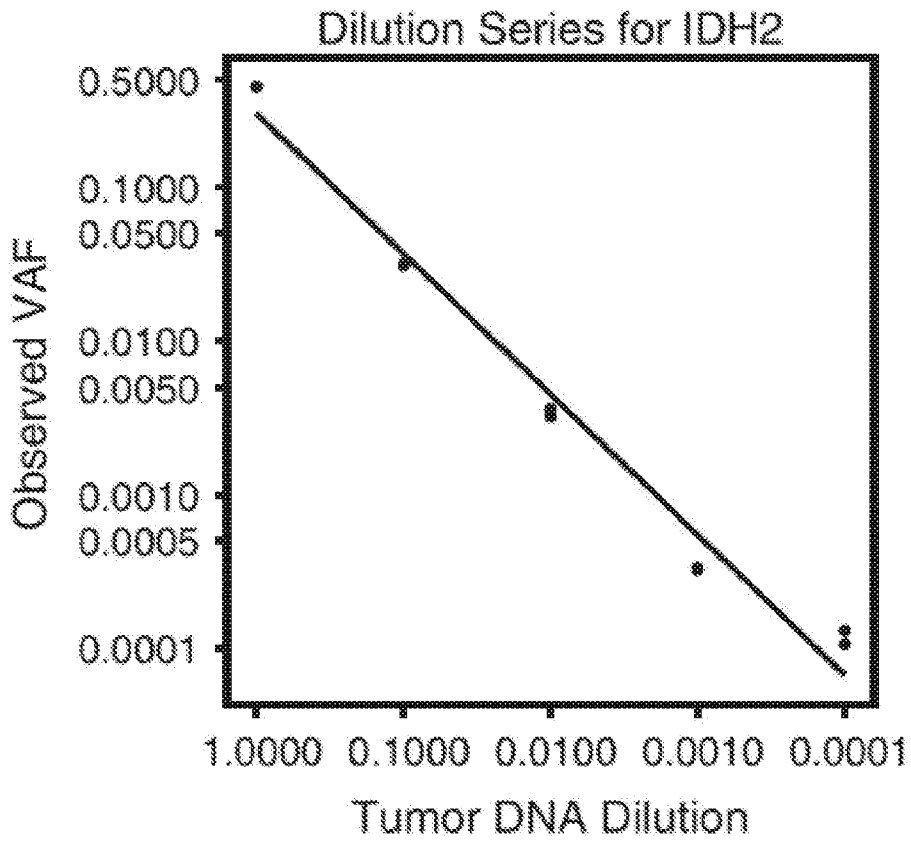
Dilution Series for IDH2
Observed VAF
0.5000
0.1000
0.0500
0.0100
0.0050
0.0010
0.0005
0.0001
1.0000 0.1000 0.0100 0.0010 0.0001
Tumor DNA Dilution

FIG. 1B

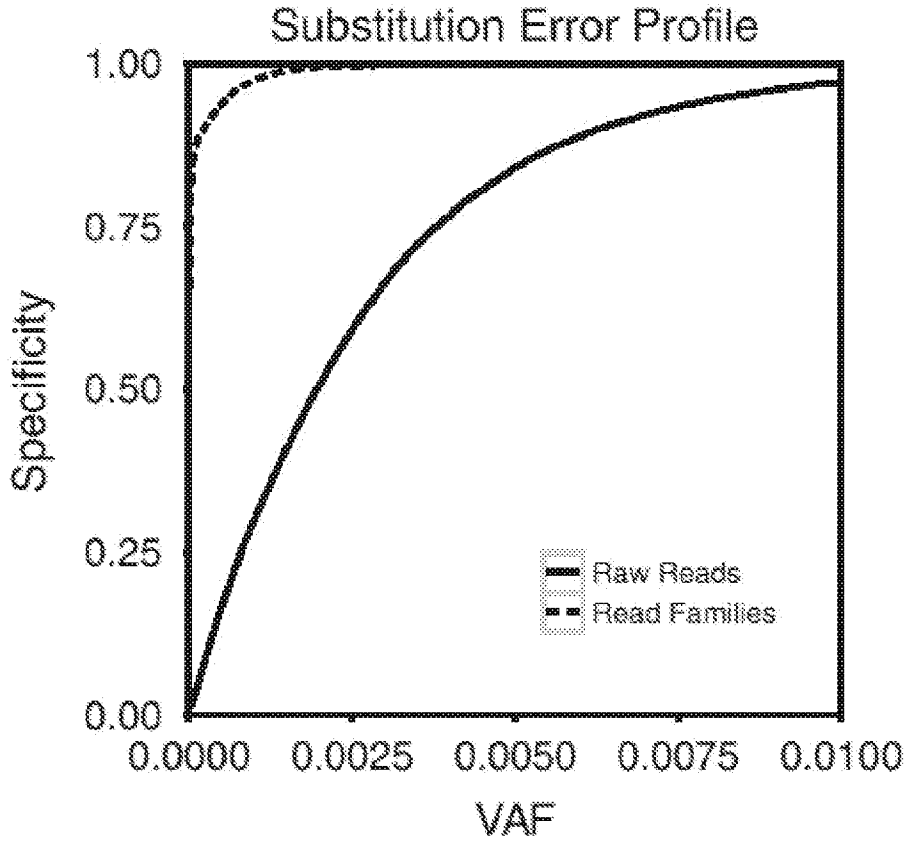
Substitution Error Profile
1.00
0.75
0.50
0.25
0.00
Specificity
0.0000
0.0025
0.0050
0.0075
0.0100
VAF
Raw Reads
Read Families

FIG. 1C

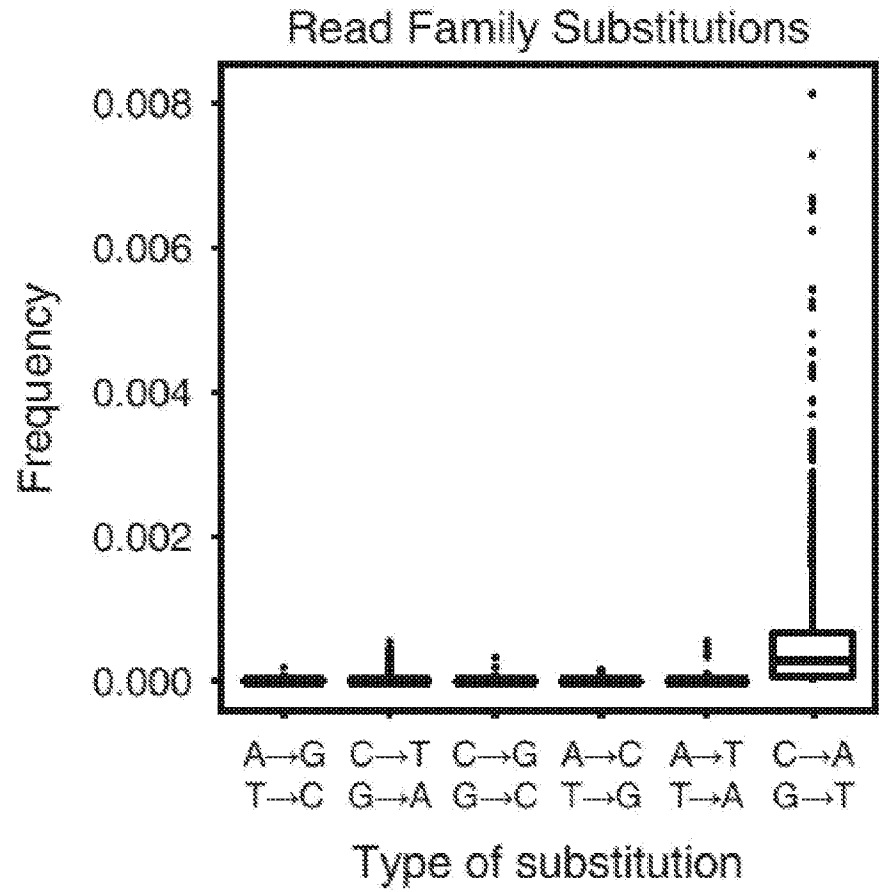
Read Family Substitutions
0.008
0.006
0.004
0.002
0.000
Frequency
A→G
T→C
C→T
G→A
C→G
G→C
A→C
T→G
A→T
T→A
C→A
G→T
Type of substitution

FIG. 1D

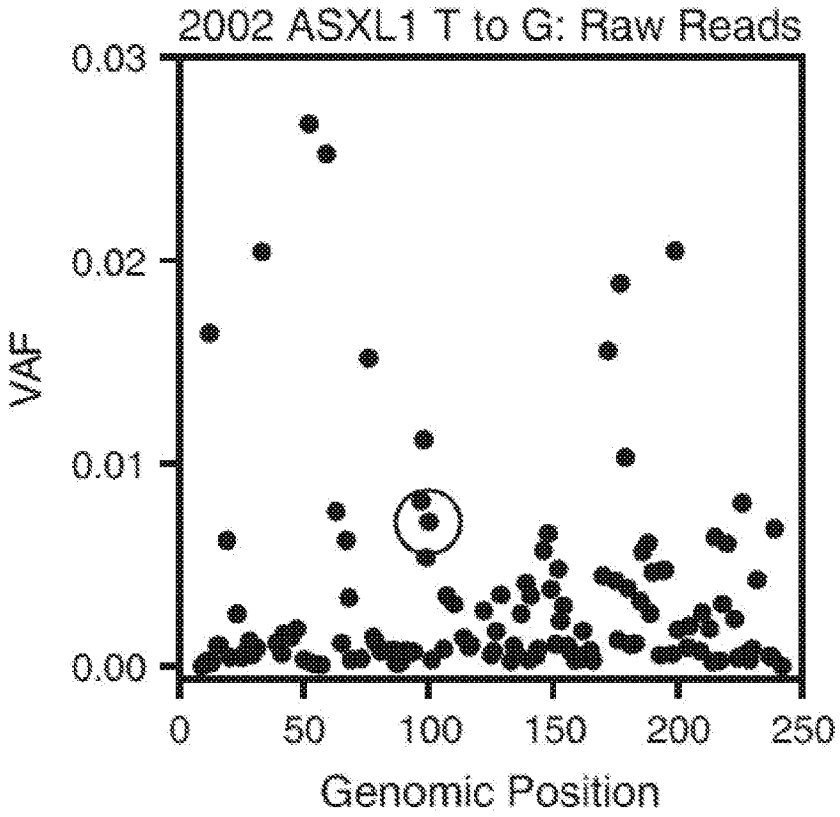
2002 ASXL1 T to G: Raw Reads
0.03
0.02
0.01
0.00
VAF
0
50
100
150
200
250
Genomic Position

FIG. 1E

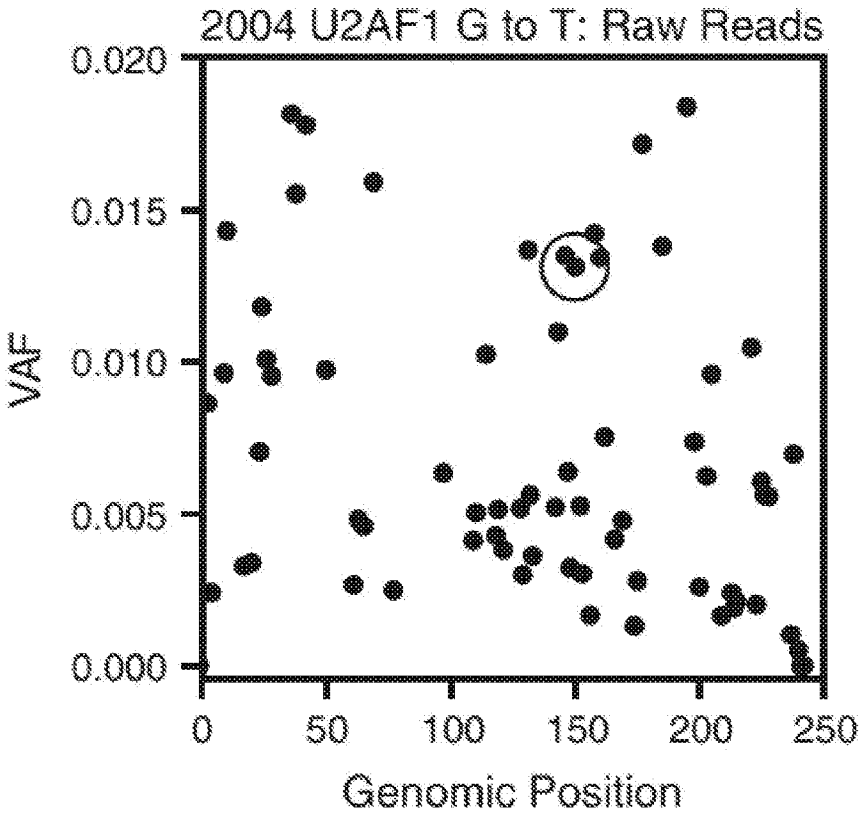
2004 U2AF1 G to T: Raw Reads
0.020
0.015
0.010
0.005
0.000
VAF
0
50
100
150
200
250
Genomic Position

FIG. 1F

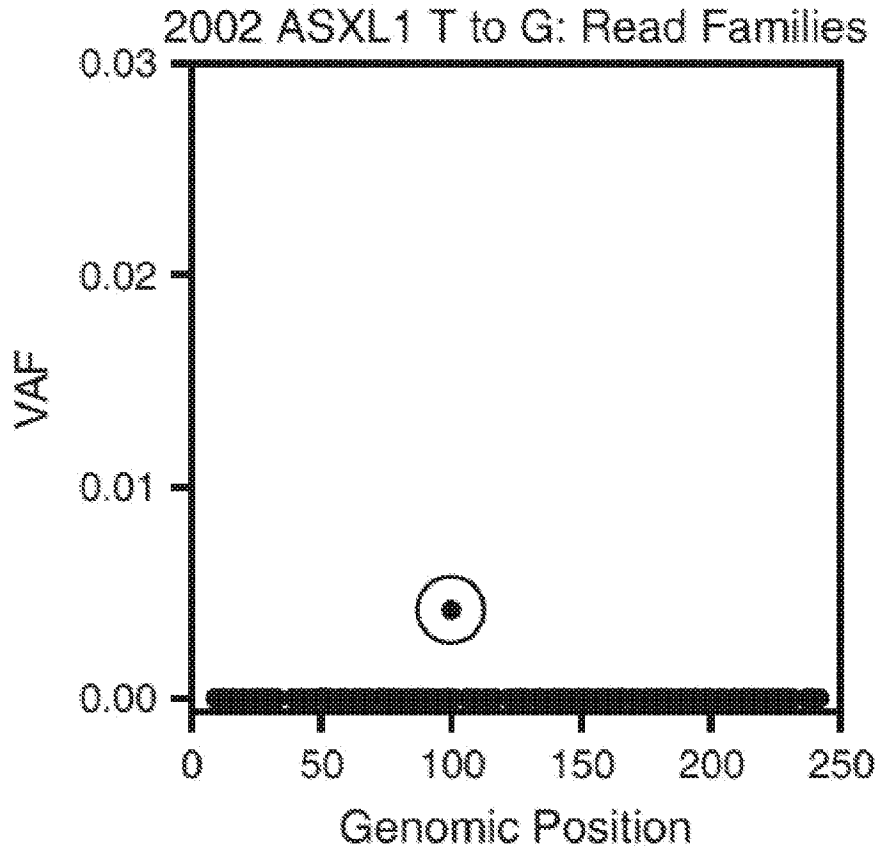
2002 ASXL1 T to G: Read Families
0.03
0.02
0.01
0.00
VAF
0
50
100
150
200
250
Genomic Position

FIG. 1G

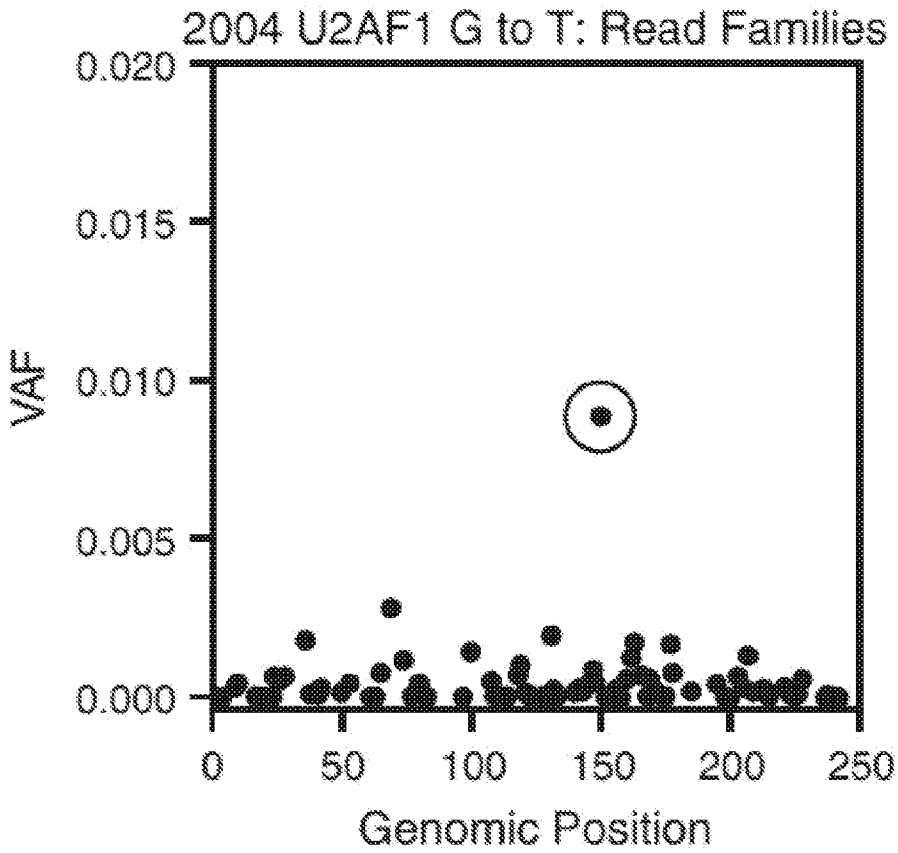
2004 U2AF1 G to T: Read Families
0.020
0.015
0.010
0.005
0.000
VAF
0
50
100
150
200
250
Genomic Position

FIG. 1H

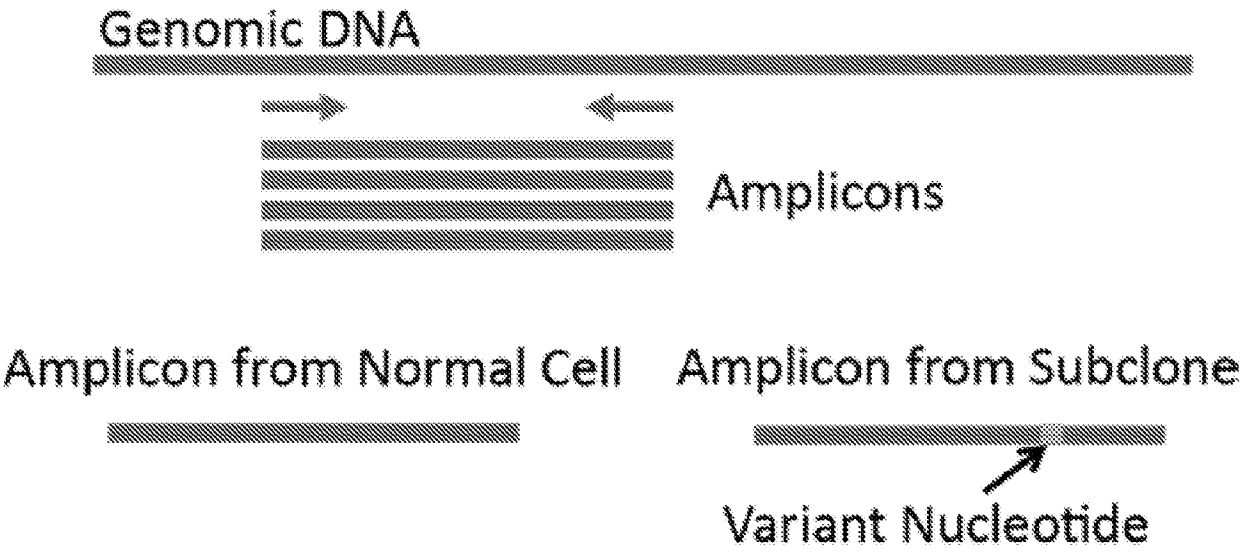

FIG. 2A

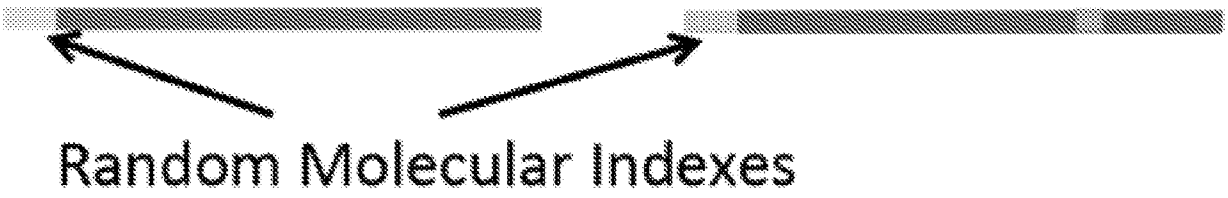

FIG. 2B

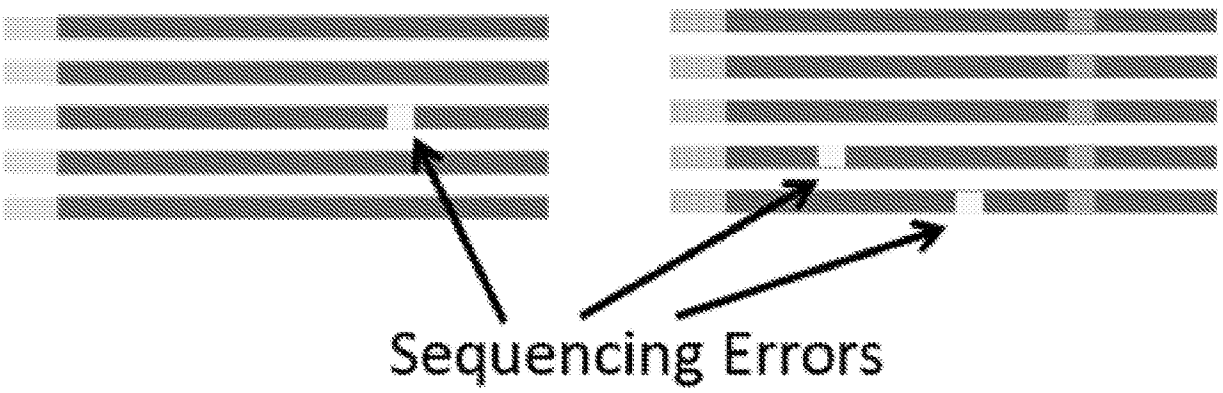

FIG. 2C

FIG. 2D

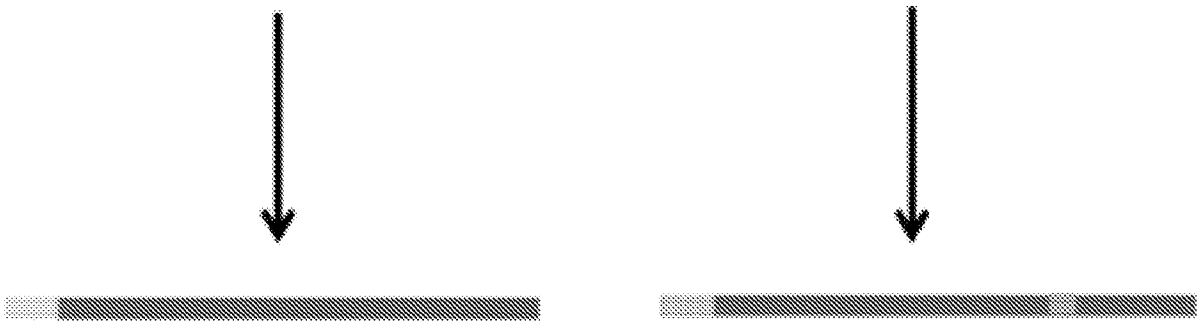

FIG. 2E

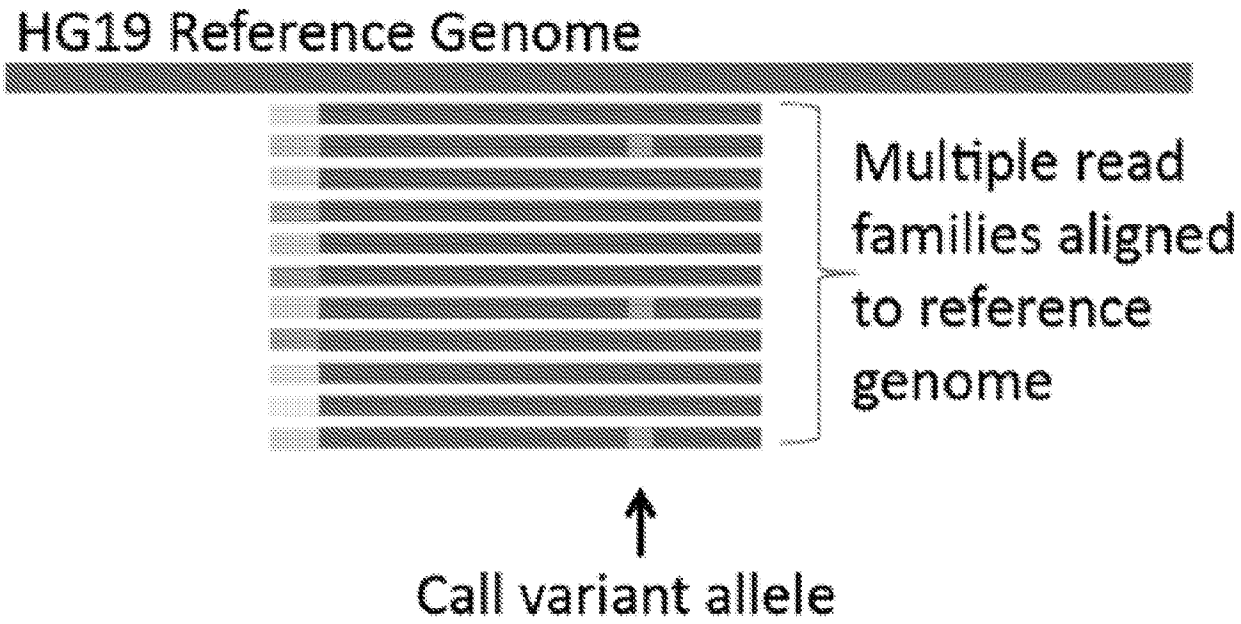
HG19 Reference Genome
Multiple read families aligned to reference genome
Call variant allele

FIG. 2F

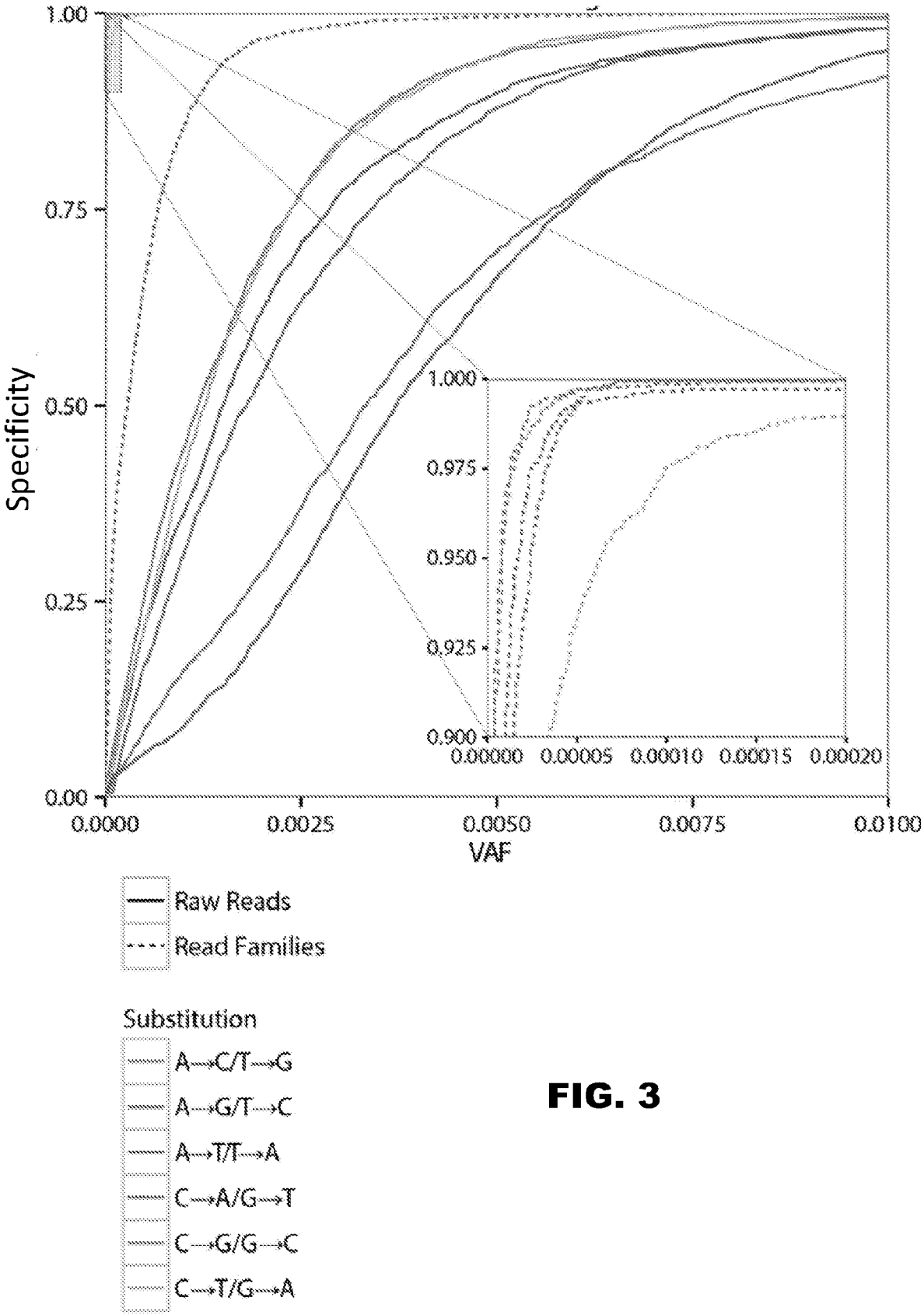
1.00
0.75
0.50
0.25
0.00
Specificity
0.0000
0.0025
0.0050
0.0075
0.0100
VAF
1.000
0.975
0.950
0.925
0.900
0.00000
0.00005
0.00010
0.00015
0.00020
Raw Reads
Read Families
Substitution
A→C/T→G
A→G/T→C
A→T/T→A
C→A/G→T
C→G/G→C
C→T/G→A

FIG. 3

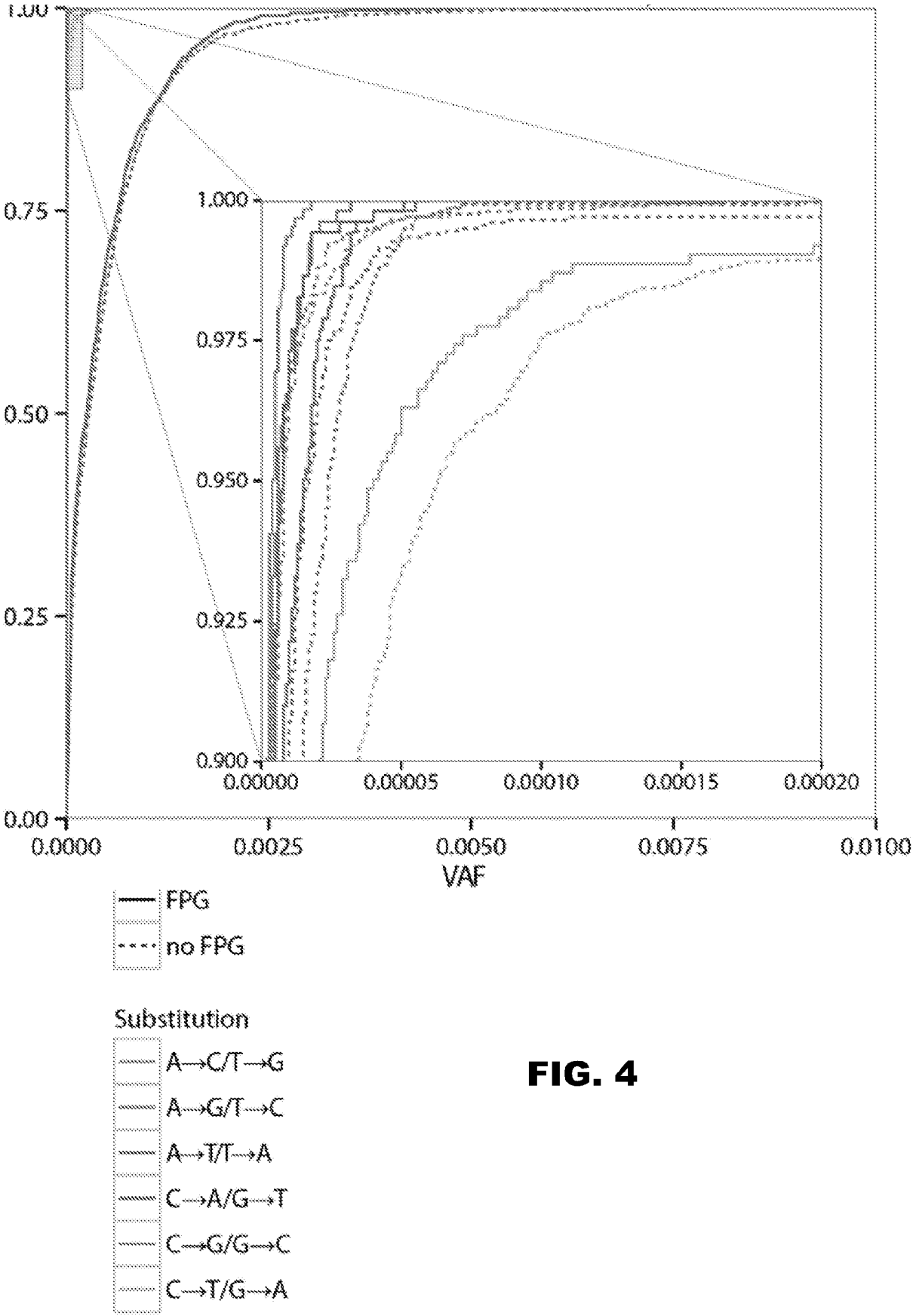
1.00
0.75
0.50
0.25
0.00
0.0000
0.0025
0.0050
0.0075
0.0100
VAF
1.000
0.975
0.950
0.925
0.900
0.00000
0.00005
0.00010
0.00015
0.00020
FPG
no FPG
Substitution
A→C/T→G
A→G/T→C
A→T/T→A
C→A/G→T
C→G/G→C
C→T/G→A

FIG. 4

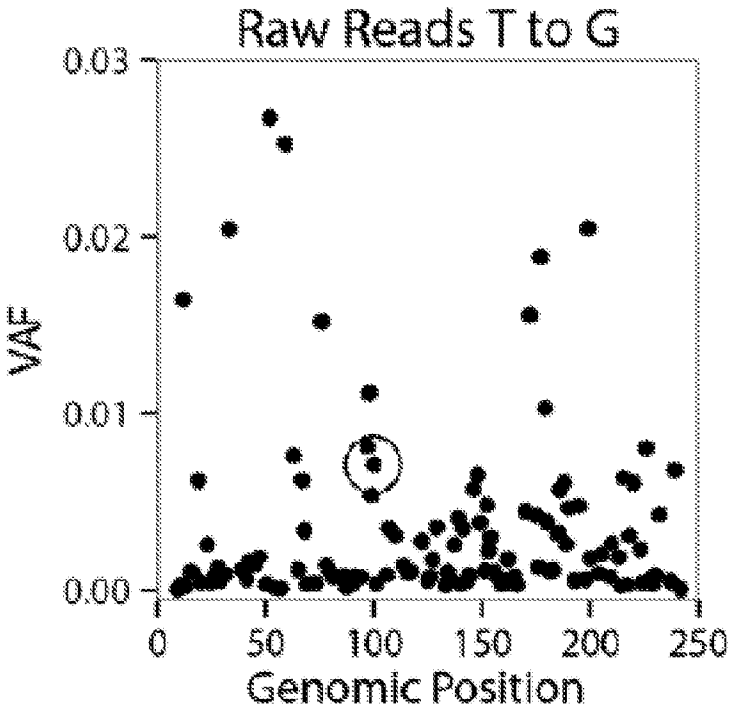
Raw Reads T to G
0.03
0.02
0.01
0.00
VAF
0
50
100
150
200
250
Genomic Position

FIG. 5A

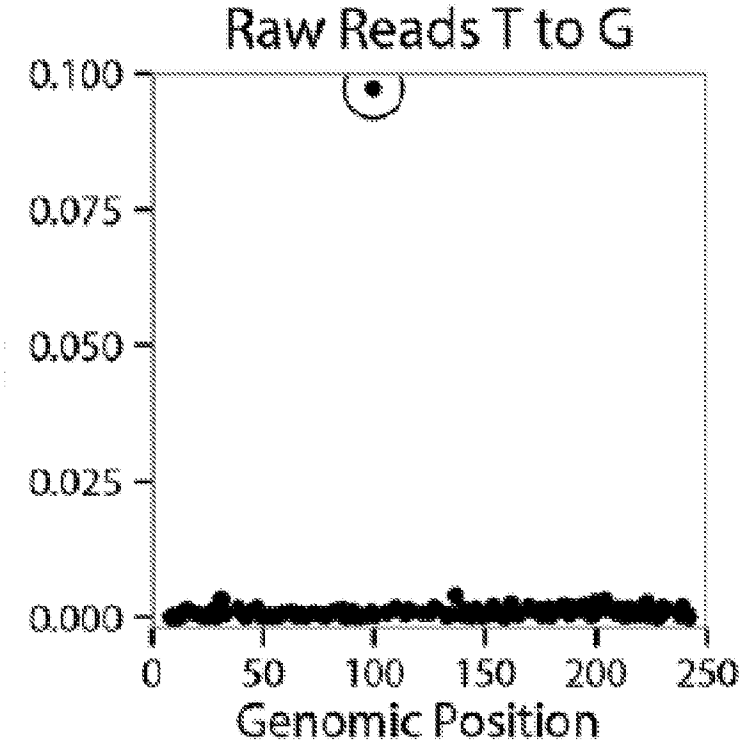
Raw Reads T to G
0.100
0.075
0.050
0.025
0.000
0
50
100
150
200
250
Genomic Position

FIG. 5B

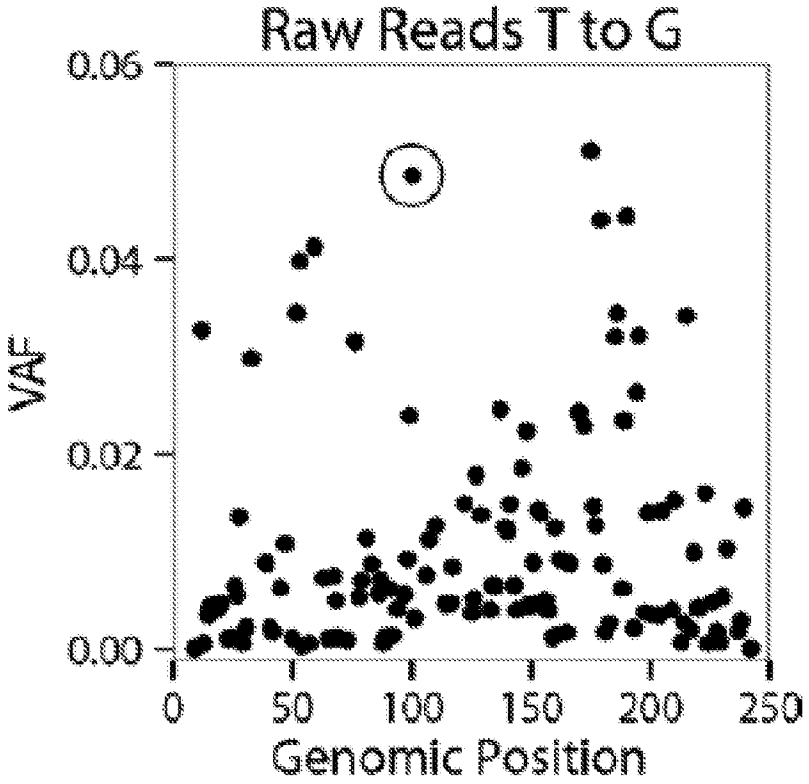
Raw Reads T to G
0.06
0.04
0.02
0.00
VAF
0
50
100
150
200
250
Genomic Position

FIG. 5C

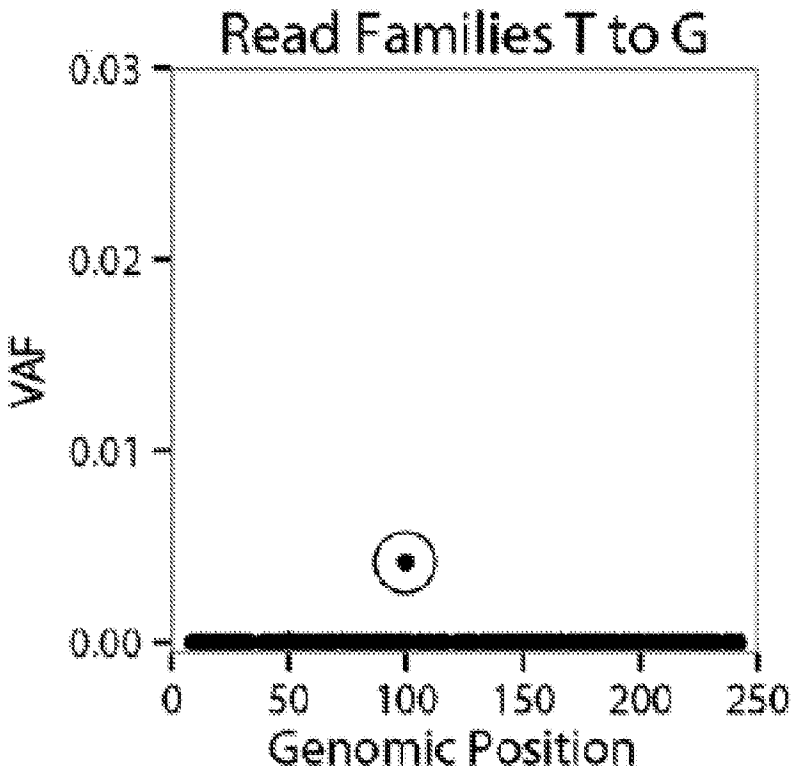
Read Families T to G
0.03
0.02
0.01
0.00
VAF
0
50
100
150
200
250
Genomic Position

FIG. 5D

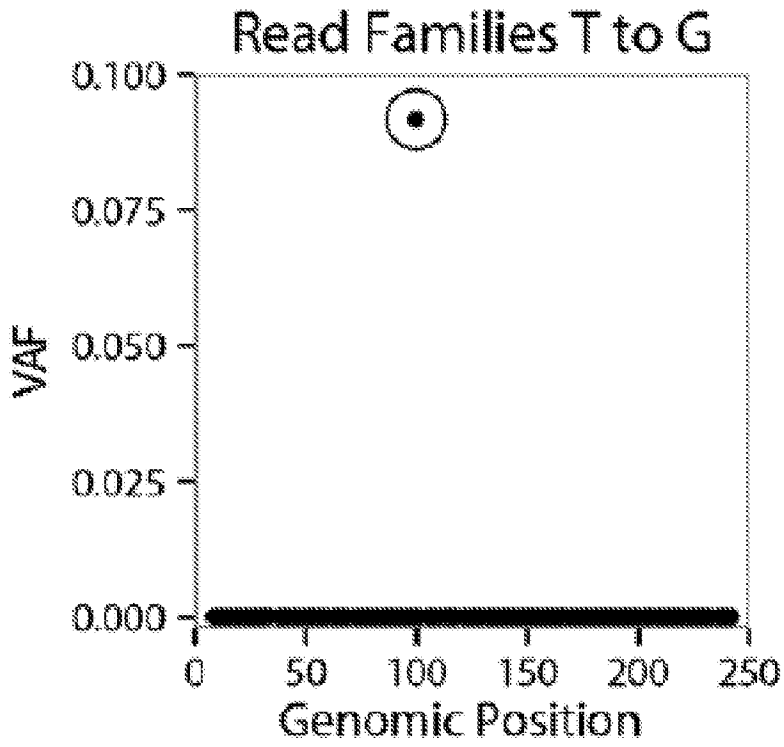
Read Families T to G
0.100
0.075
0.050
0.025
0.000
VAF
0
50
100
150
200
250
Genomic Position

FIG. 5E

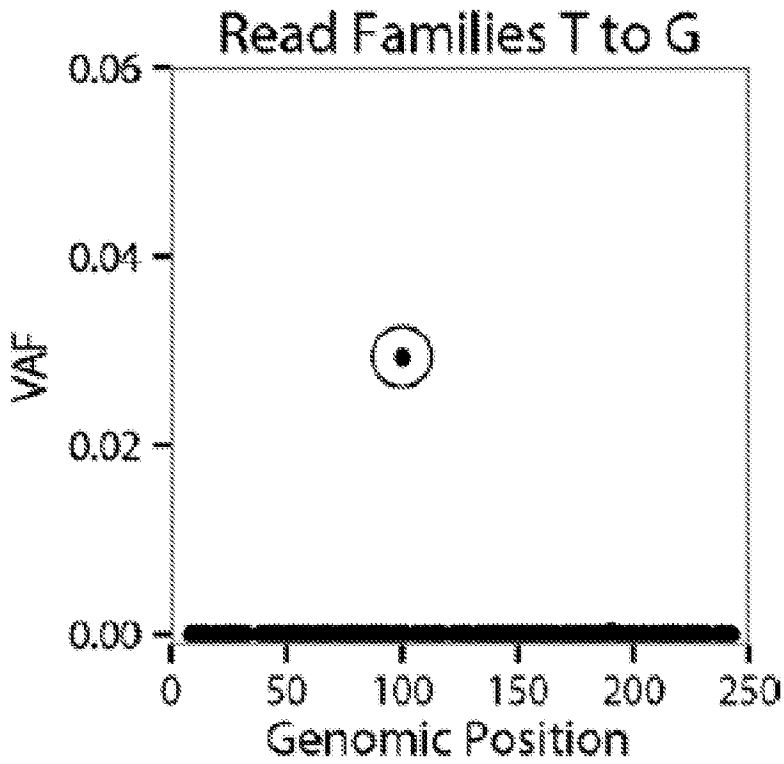
Read Families T to G
0.06
0.04
0.02
0.00
VAF
0
50
100
150
200
250
Genomic Position

FIG. 5F

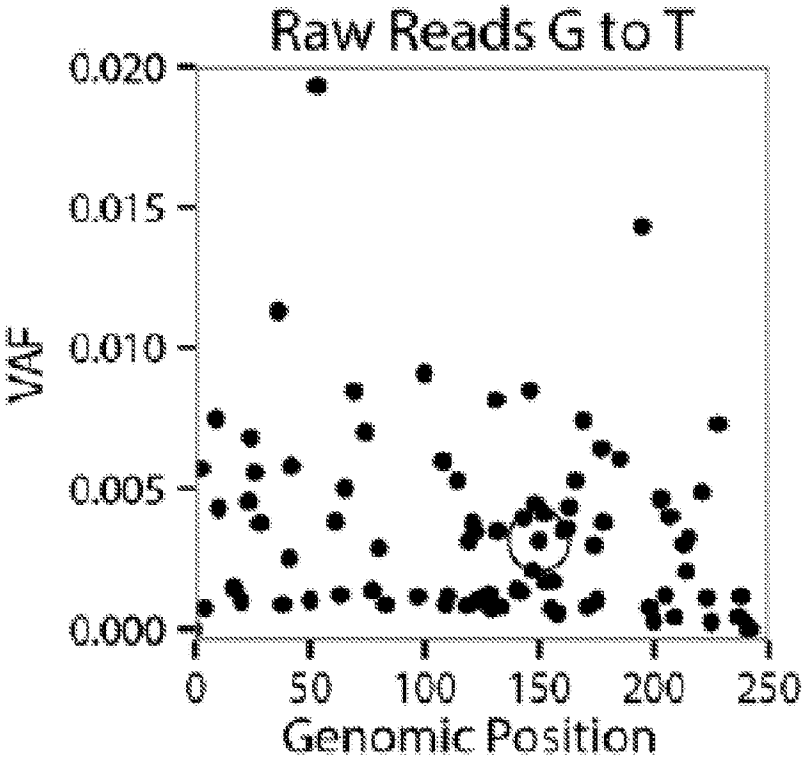
Raw Reads G to T
0.020
0.015
0.010
0.005
0.000
VAF
0
50
100
150
200
250
Genomic Position

FIG. 6A

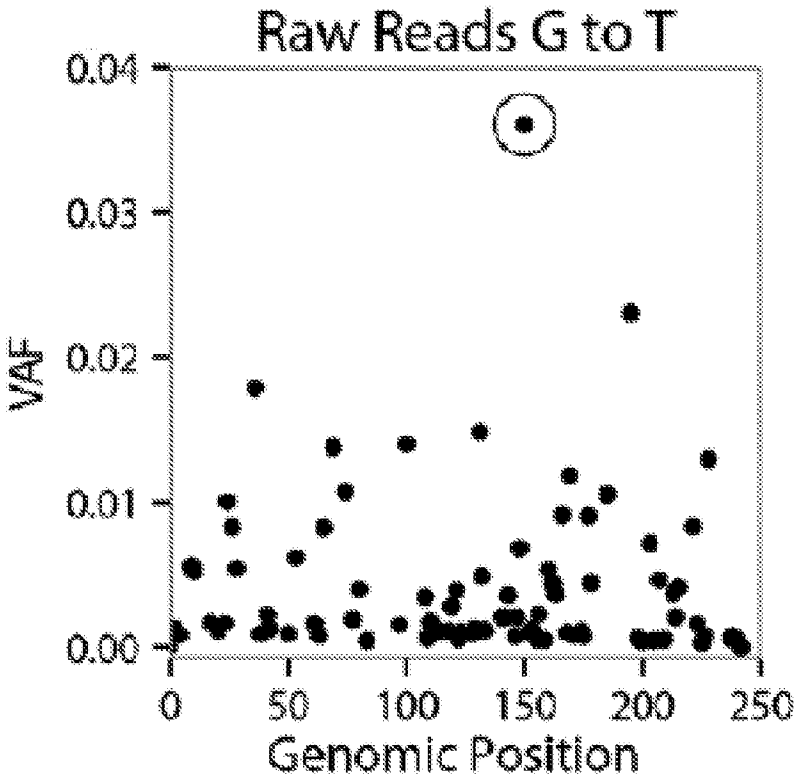
Raw Reads G to T
0.04
0.03
0.02
0.01
0.00
VAF
0
50
100
150
200
250
Genomic Position

FIG. 6B

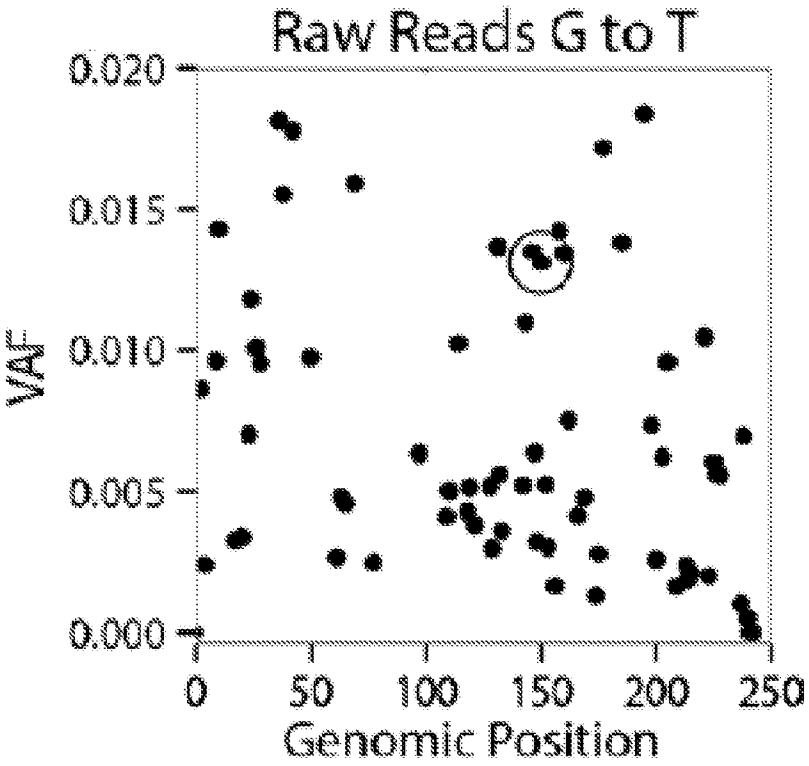
Raw Reads G to T
0.020
0.015
0.010
0.005
0.000
VAF
0
50
100
150
200
250
Genomic Position

FIG. 6C

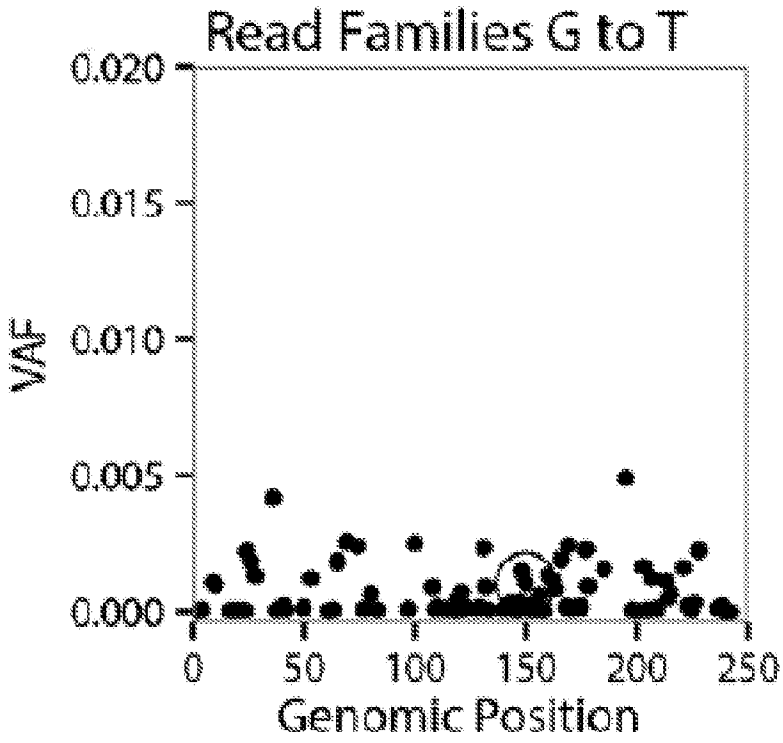
Read Families G to T
0.020
0.015
0.010
0.005
0.000
VAF
0
50
100
150
200
250
Genomic Position

FIG. 6D

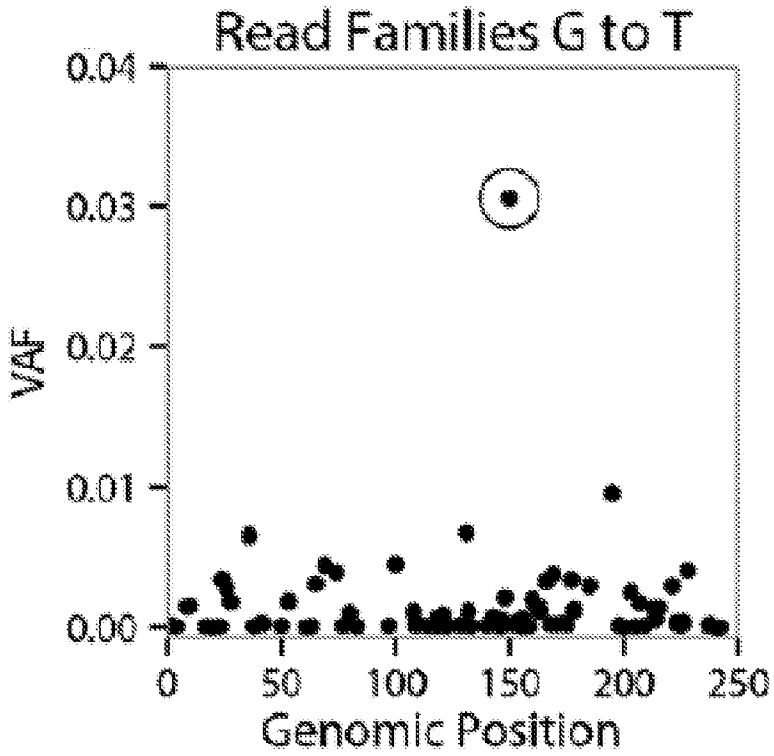
Read Families G to T
0.04
0.03
0.02
0.01
0.00
VAF
0
50
100
150
200
250
Genomic Position

FIG. 6E


Read Families G to T
0.020
0.015
0.010
0.005
0.000
VAF
0
50
100
150
200
250
Genomic Position

FIG. 6F

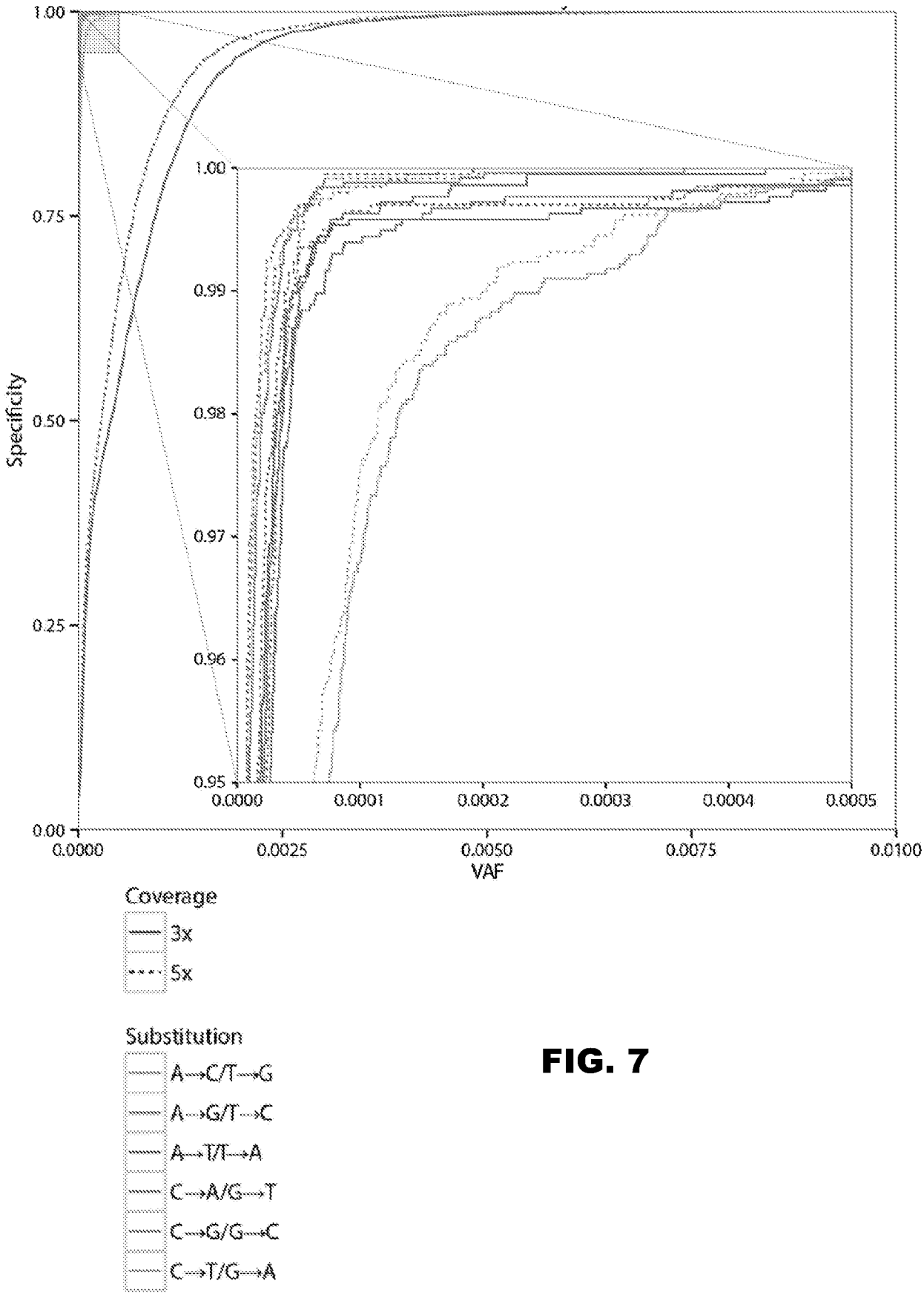
1.00
0.75
0.50
0.25
0.00
Specificity
0.0000
0.0025
0.0050
0.0075
0.0100
VAF
1.00
0.99
0.98
0.97
0.96
0.95
0.0000
0.0001
0.0002
0.0003
0.0004
0.0005
Coverage
3x
5x
Substitution
A→C/T→G
A→G/T→C
A→T/T→A
C→A/G→T
C→G/G→C
C→T/G→A

FIG. 7

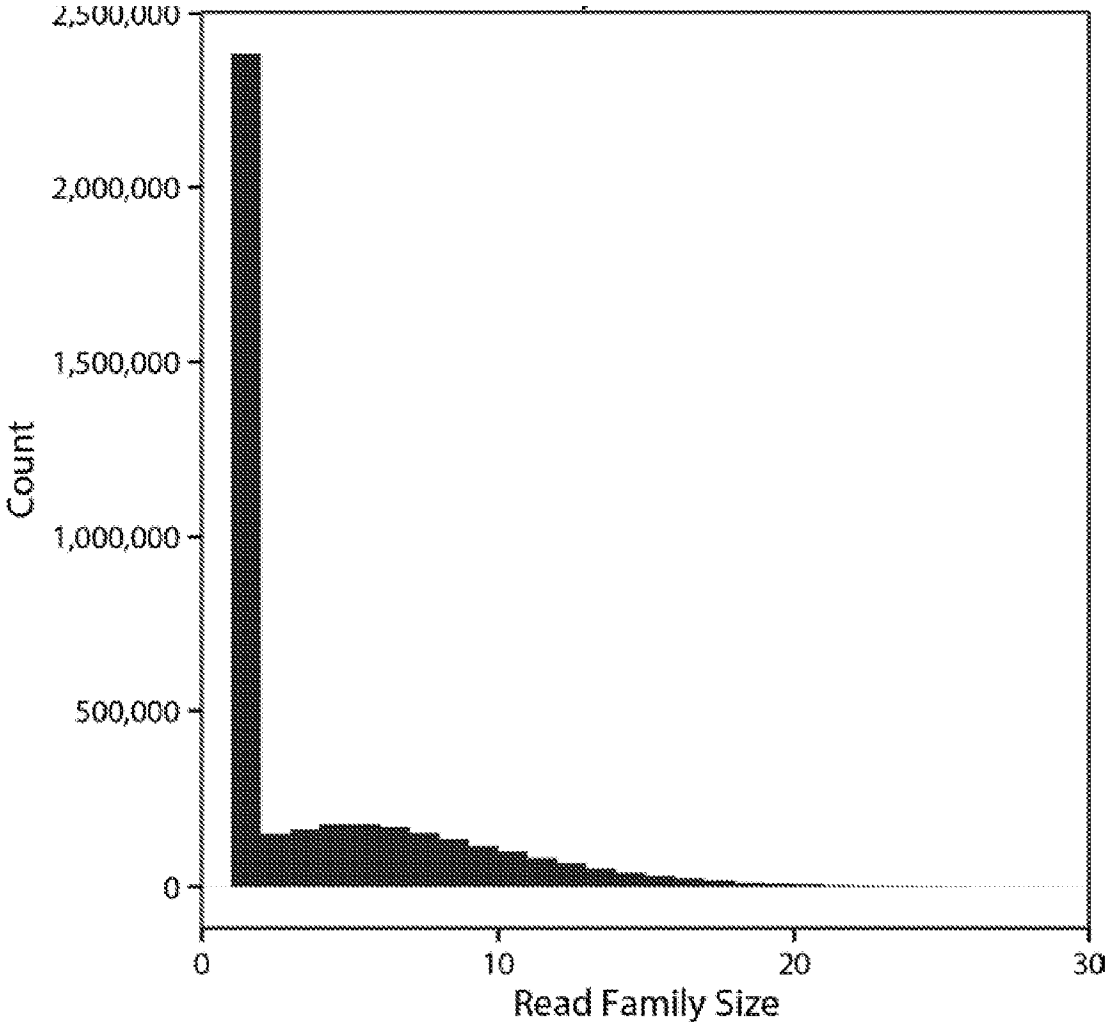
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
Count
0
10
20
30
Read Family Size

FIG. 8

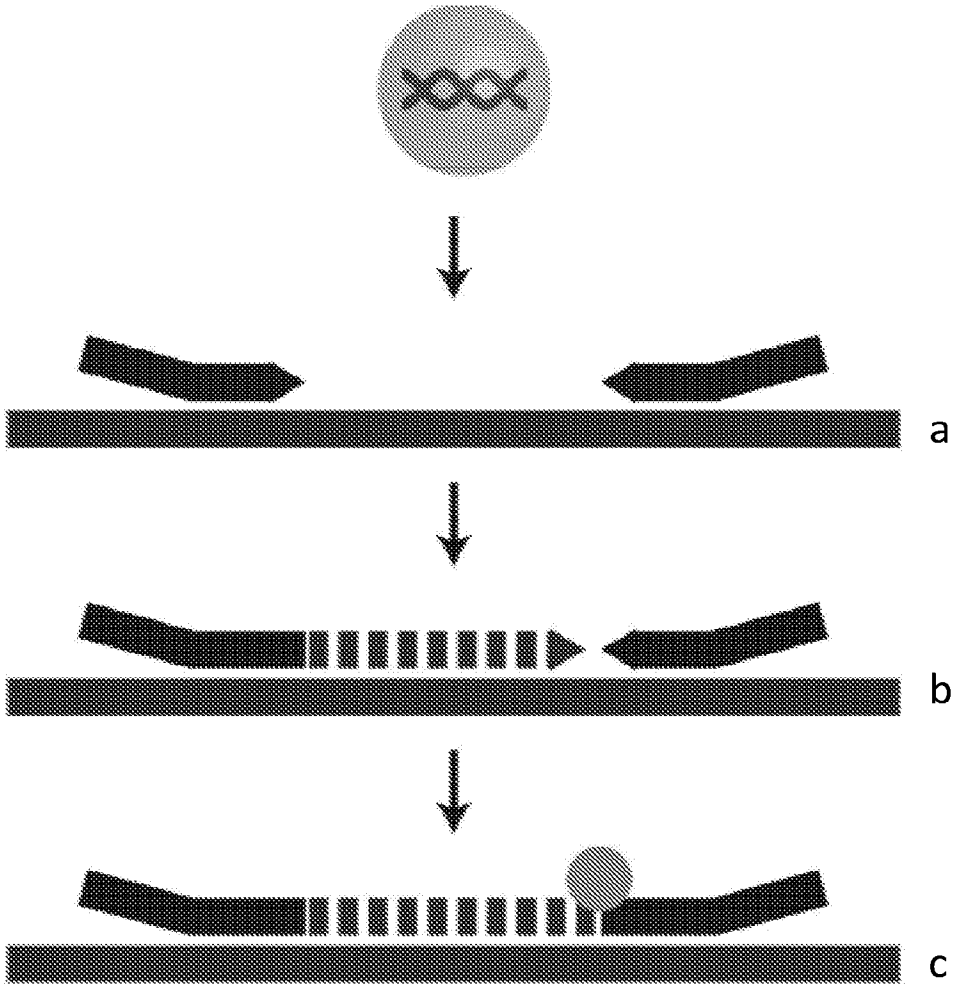

FIG. 9A

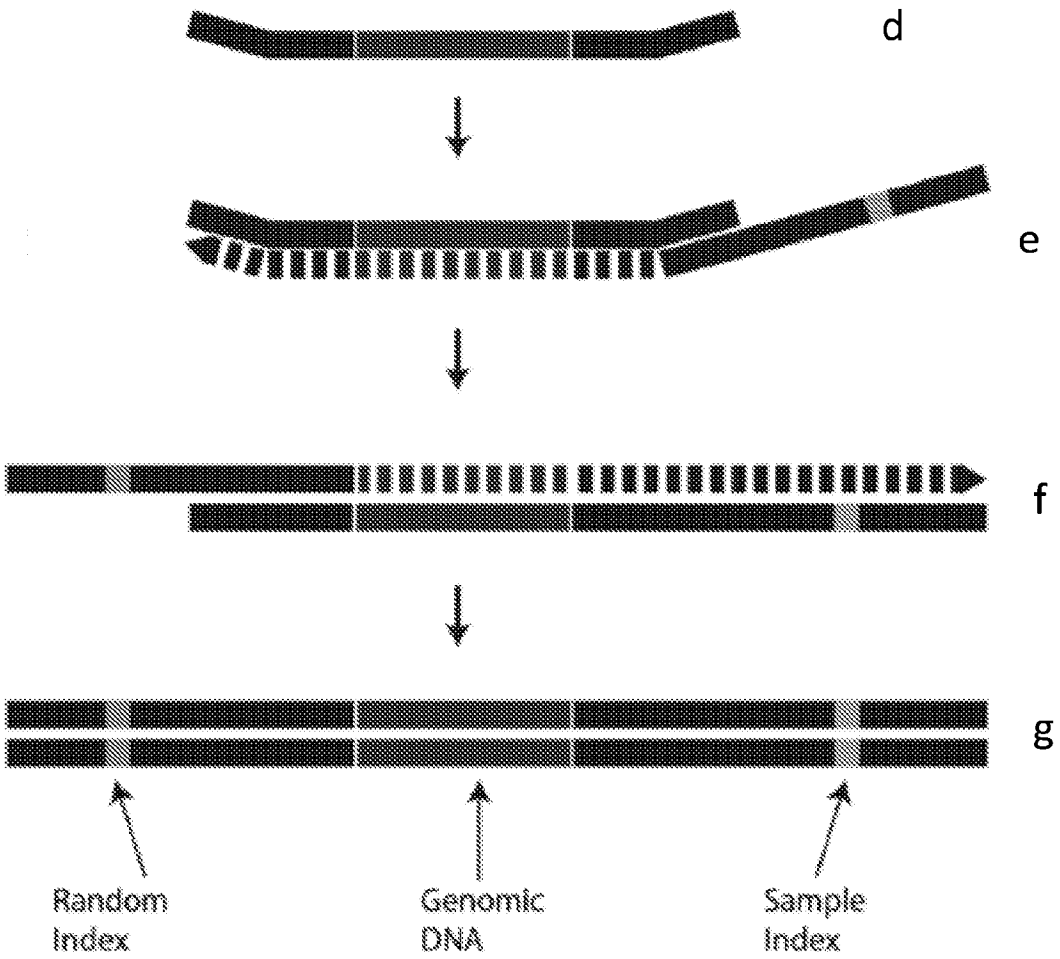

FIG. 9B

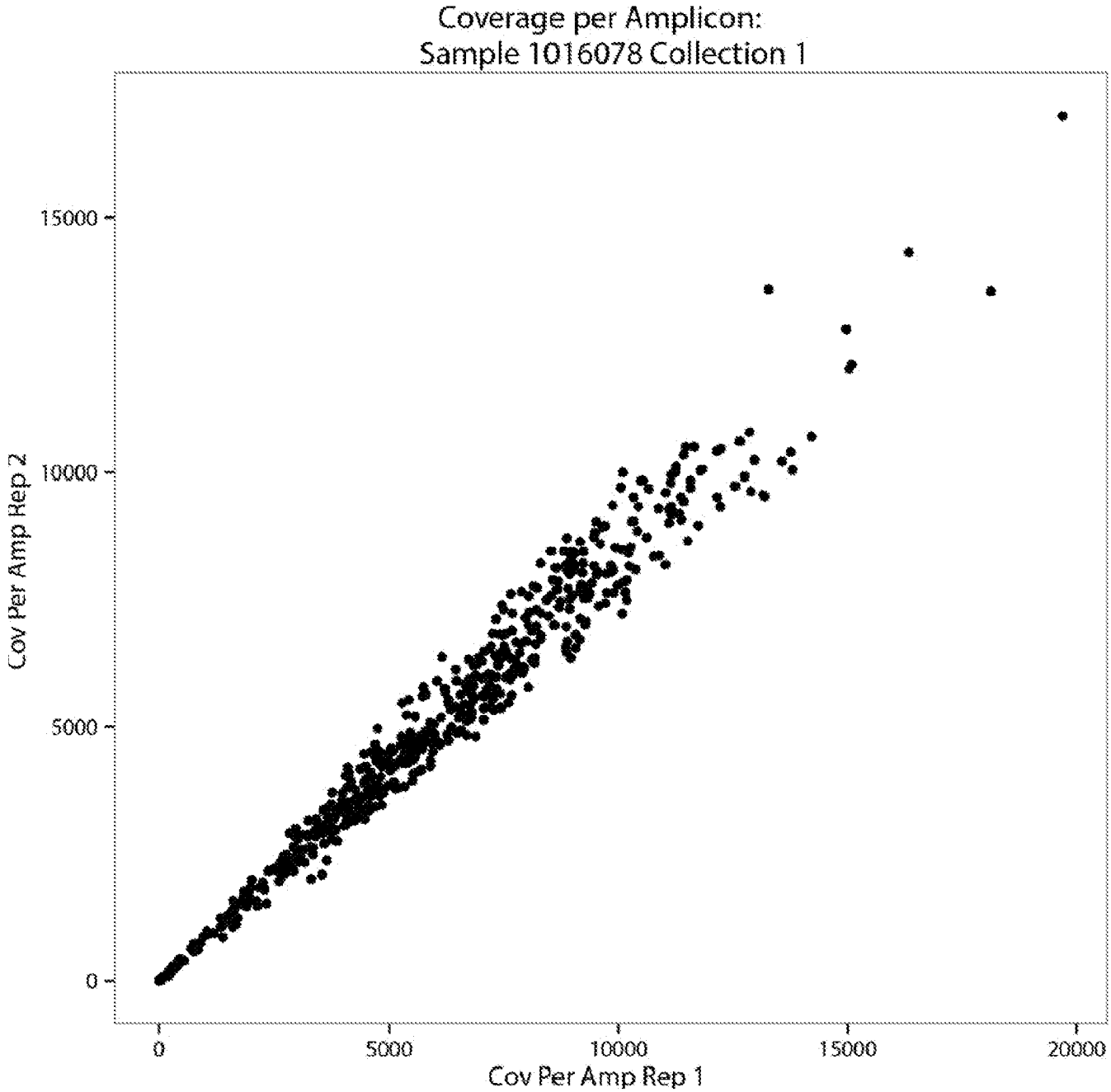
Coverage per Amplicon:
Sample 1016078 Collection 1
Cov Per Amp Rep 2
0
5000
10000
15000
Cov Per Amp Rep 1
0
5000
10000
15000
20000

FIG. 10A

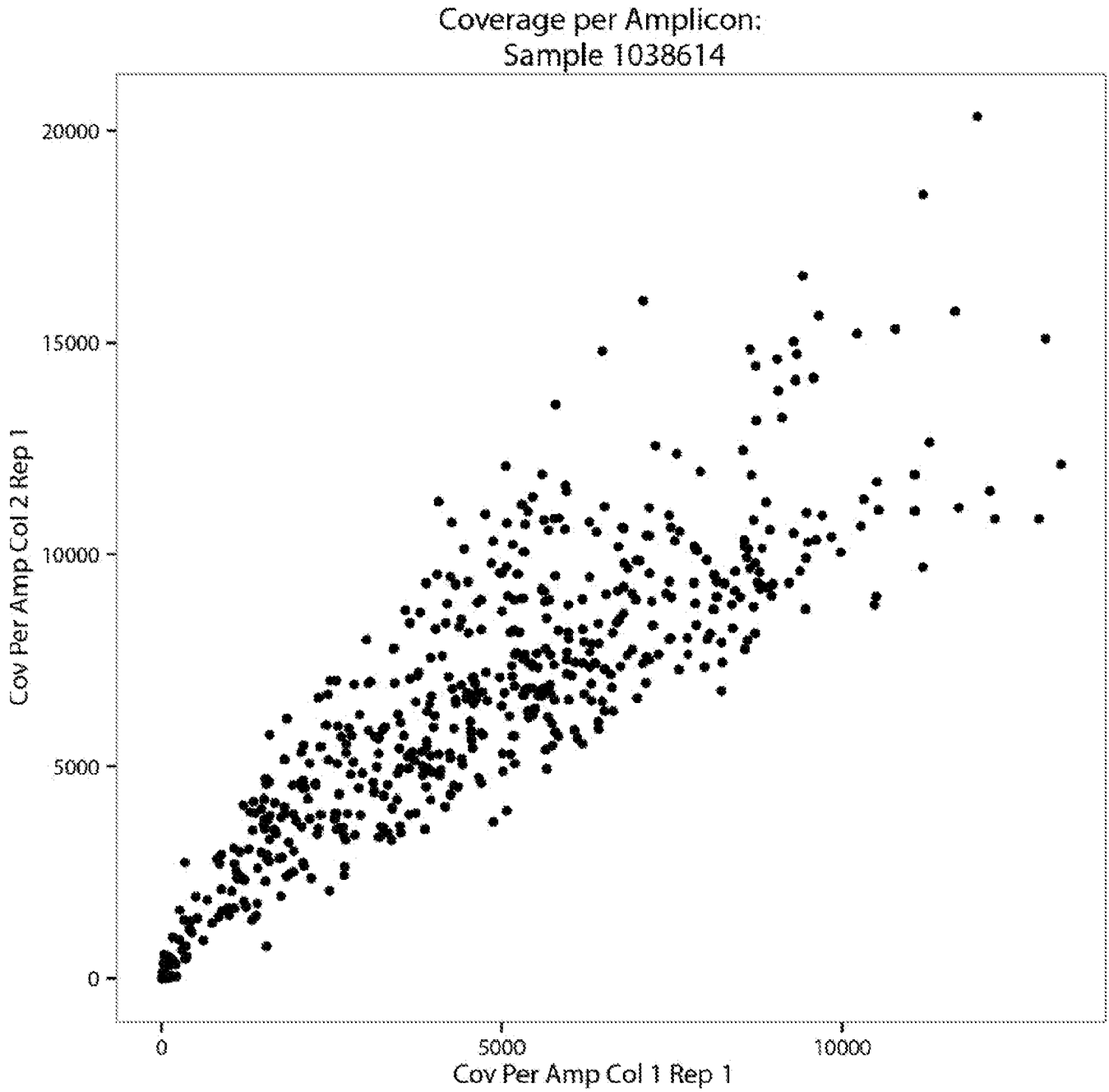
Coverage per Amplicon:
Sample 1038614
Cov Per Amp Col 2 Rep 1
20000
15000
10000
5000
0
0
5000
10000
Cov Per Amp Col 1 Rep 1

FIG. 10B

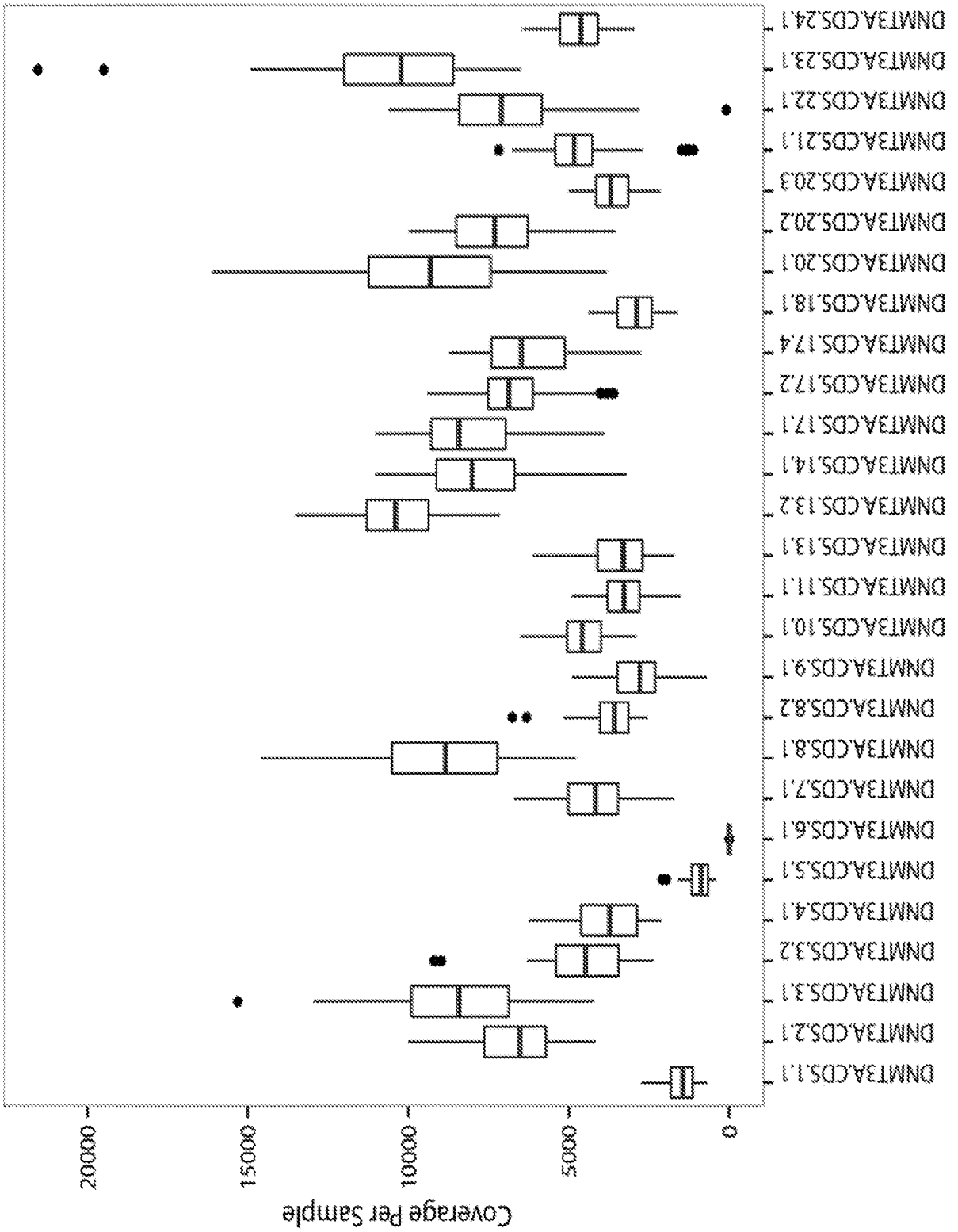
Coverage Per Sample
20000
15000
10000
5000
0
Amplicon
DNMT3A.CDS.1.1
DNMT3A.CDS.2.1
DNMT3A.CDS.3.1
DNMT3A.CDS.3.2
DNMT3A.CDS.4.1
DNMT3A.CDS.5.1
DNMT3A.CDS.6.1
DNMT3A.CDS.7.1
DNMT3A.CDS.8.1
DNMT3A.CDS.8.2
DNMT3A.CDS.9.1
DNMT3A.CDS.10.1
DNMT3A.CDS.11.1
DNMT3A.CDS.13.1
DNMT3A.CDS.13.2
DNMT3A.CDS.14.1
DNMT3A.CDS.17.1
DNMT3A.CDS.17.2
DNMT3A.CDS.17.4
DNMT3A.CDS.18.1
DNMT3A.CDS.20.1
DNMT3A.CDS.20.2
DNMT3A.CDS.20.3
DNMT3A.CDS.21.1
DNMT3A.CDS.22.1
DNMT3A.CDS.23.1
DNMT3A.CDS.24.1

FIG. 11

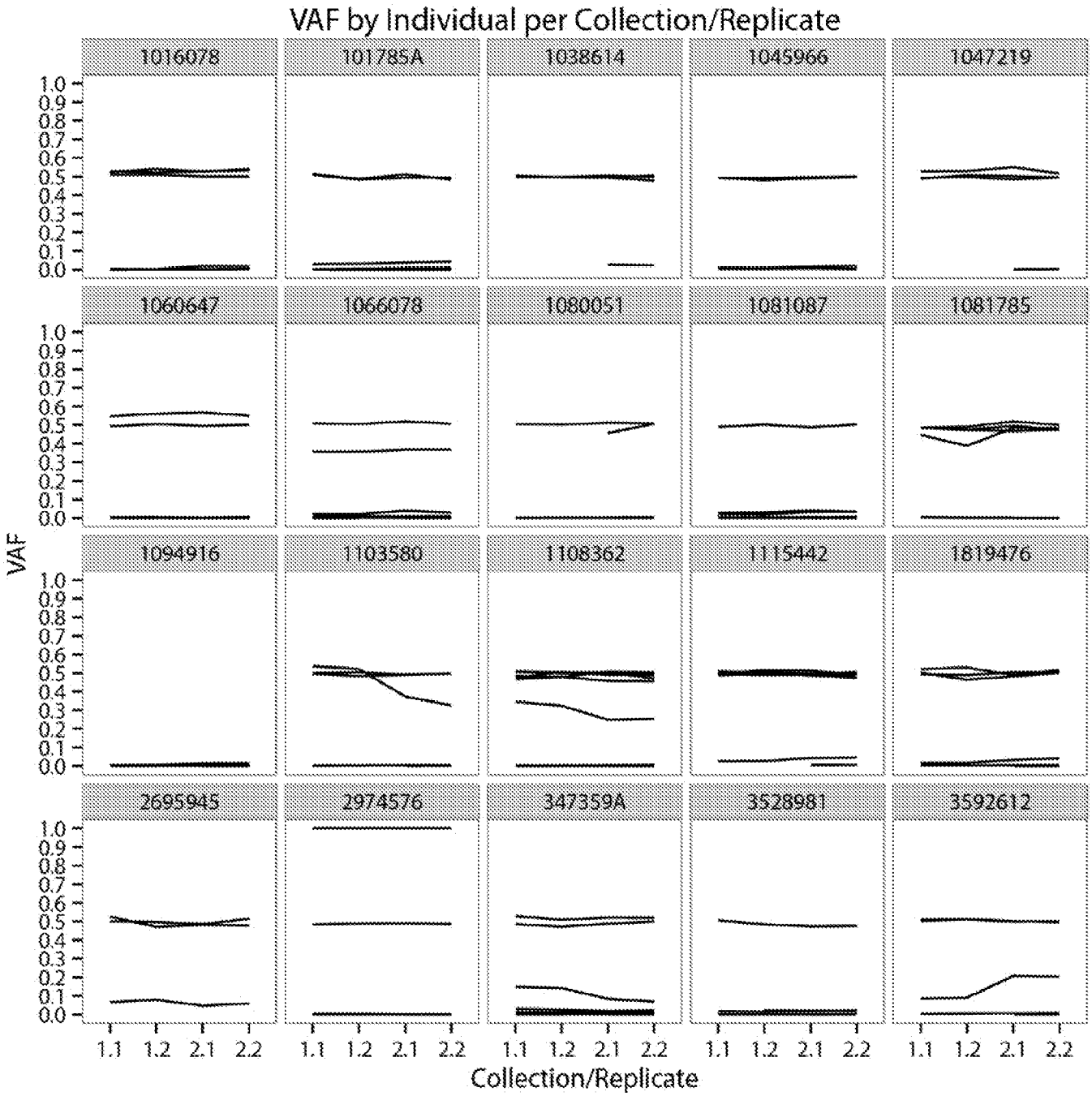
VAF by Individual per Collection/Replicate
1016078
101785A
1038614
1045966
1047219
1060647
1066078
1080051
1081087
1081785
1094916
1103580
1108362
1115442
1819476
2695945
2974576
347359A
3528981
3592612
VAF
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1.1
1.2
2.1
2.2
Collection/Replicate

FIG. 12

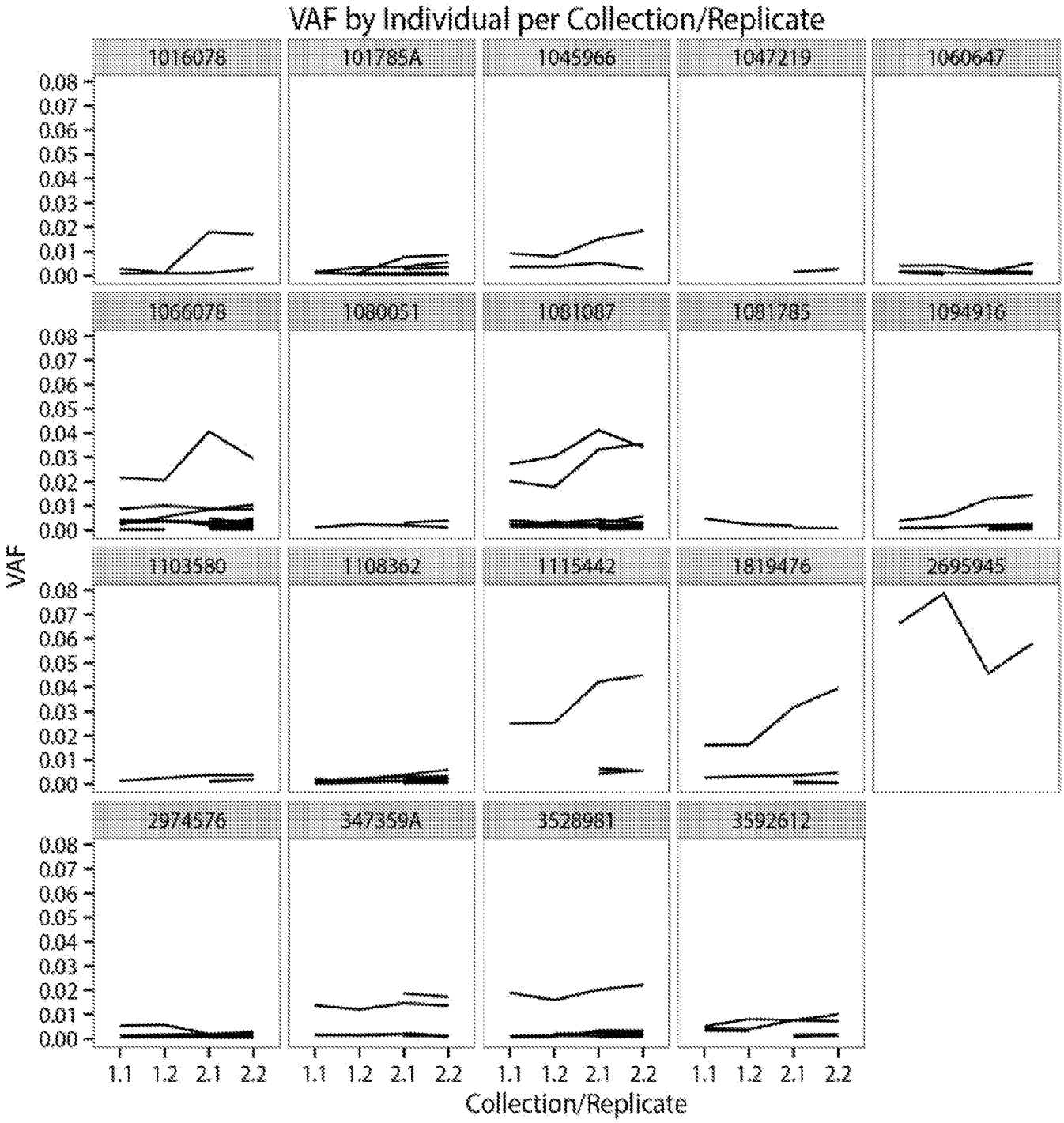
VAF by Individual per Collection/Replicate
1016078
101785A
1045966
1047219
1060647
1066078
1080051
1081087
1081785
1094916
1103580
1108362
1115442
1819476
2695945
2974576
347359A
3528981
3592612
VAF
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
1.1
1.2
2.1
2.2
Collection/Replicate

FIG. 13

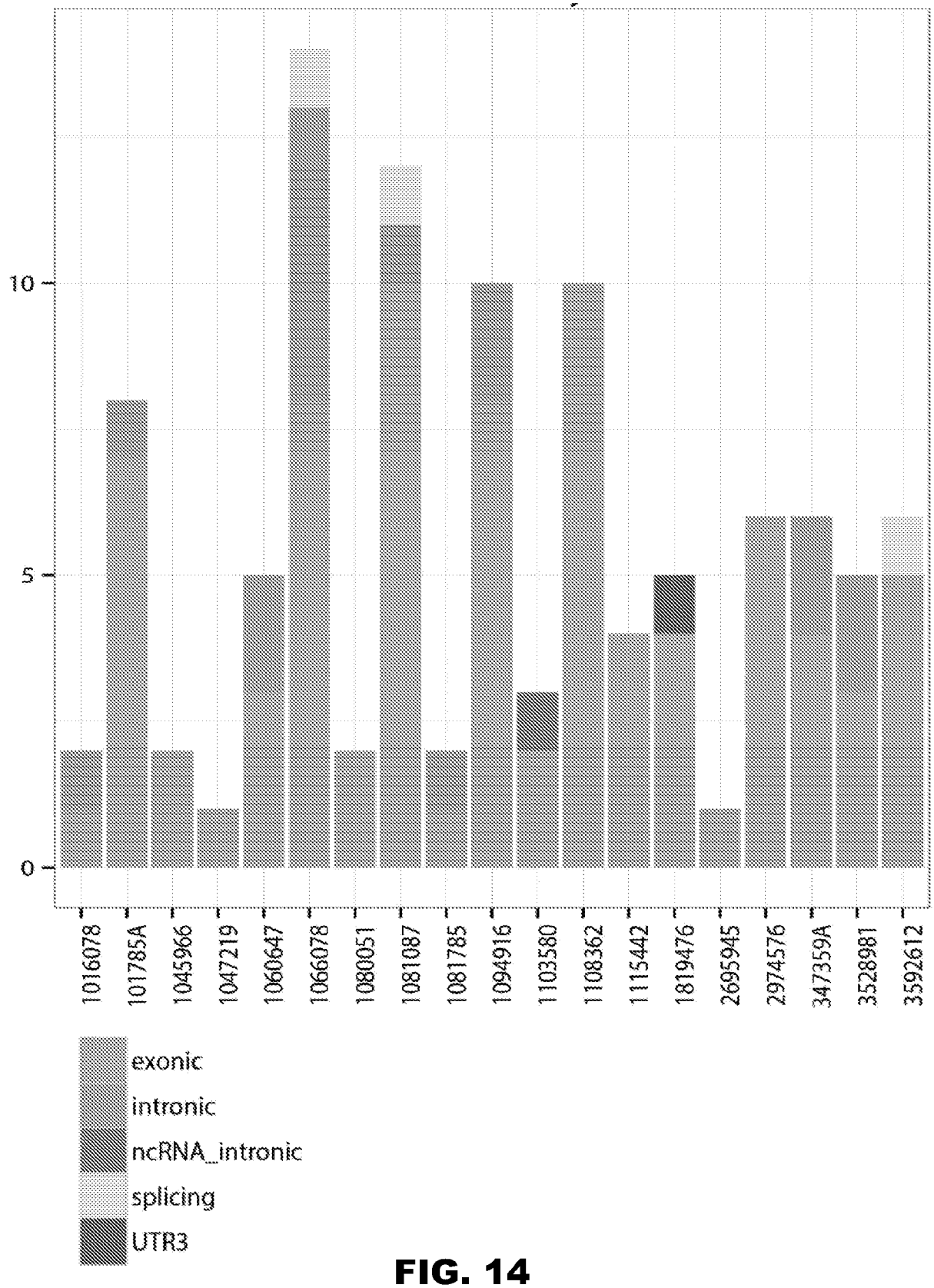
10
5
0
1016078
101785A
1045966
1047219
1060647
1066078
1080051
1081087
1081785
1094916
1103580
1108362
1115442
1819476
2695945
2974576
347359A
3528981
3592612
exonic
intronic
ncRNA_intronic
splicing
UTR3

FIG. 14

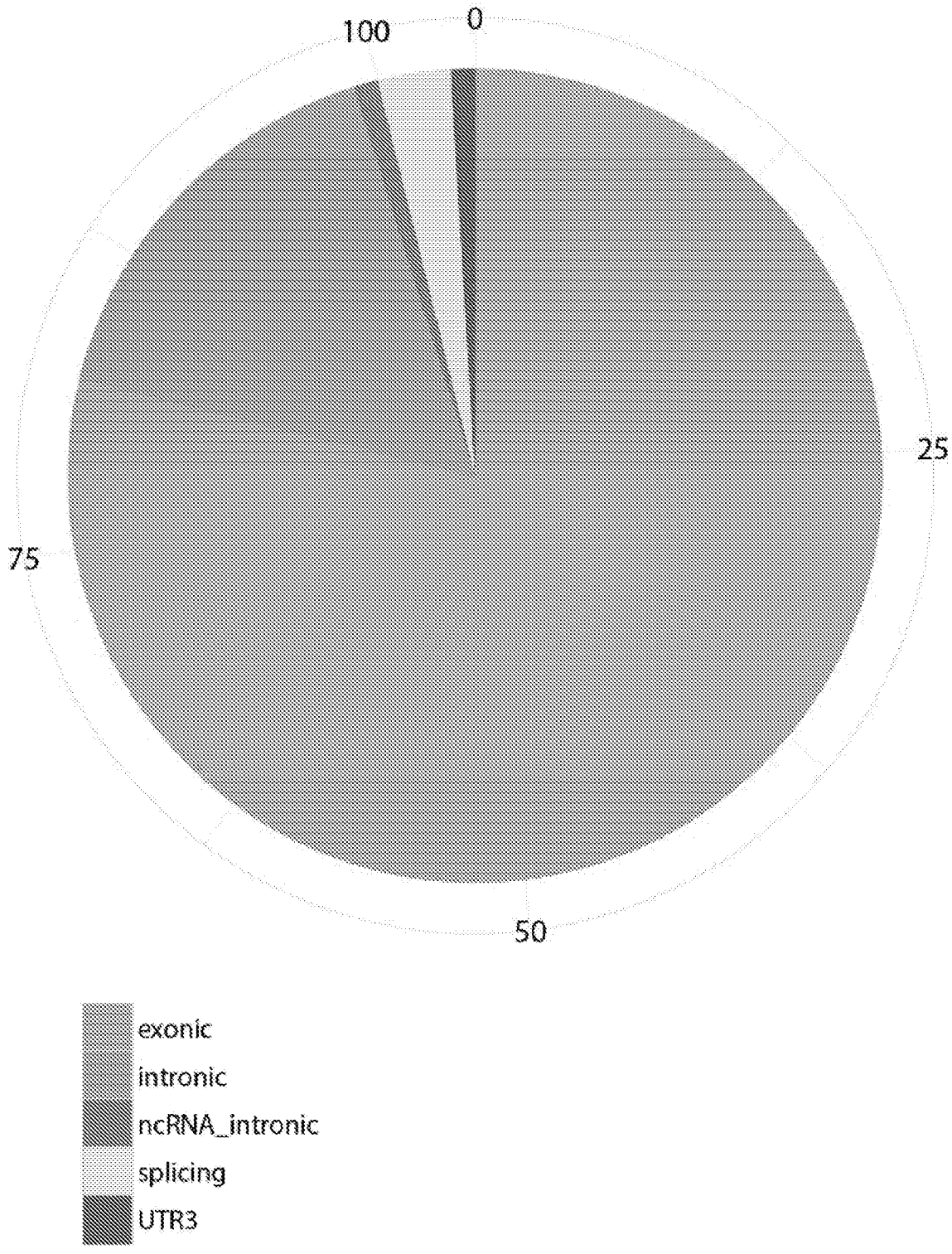

FIG. 15A

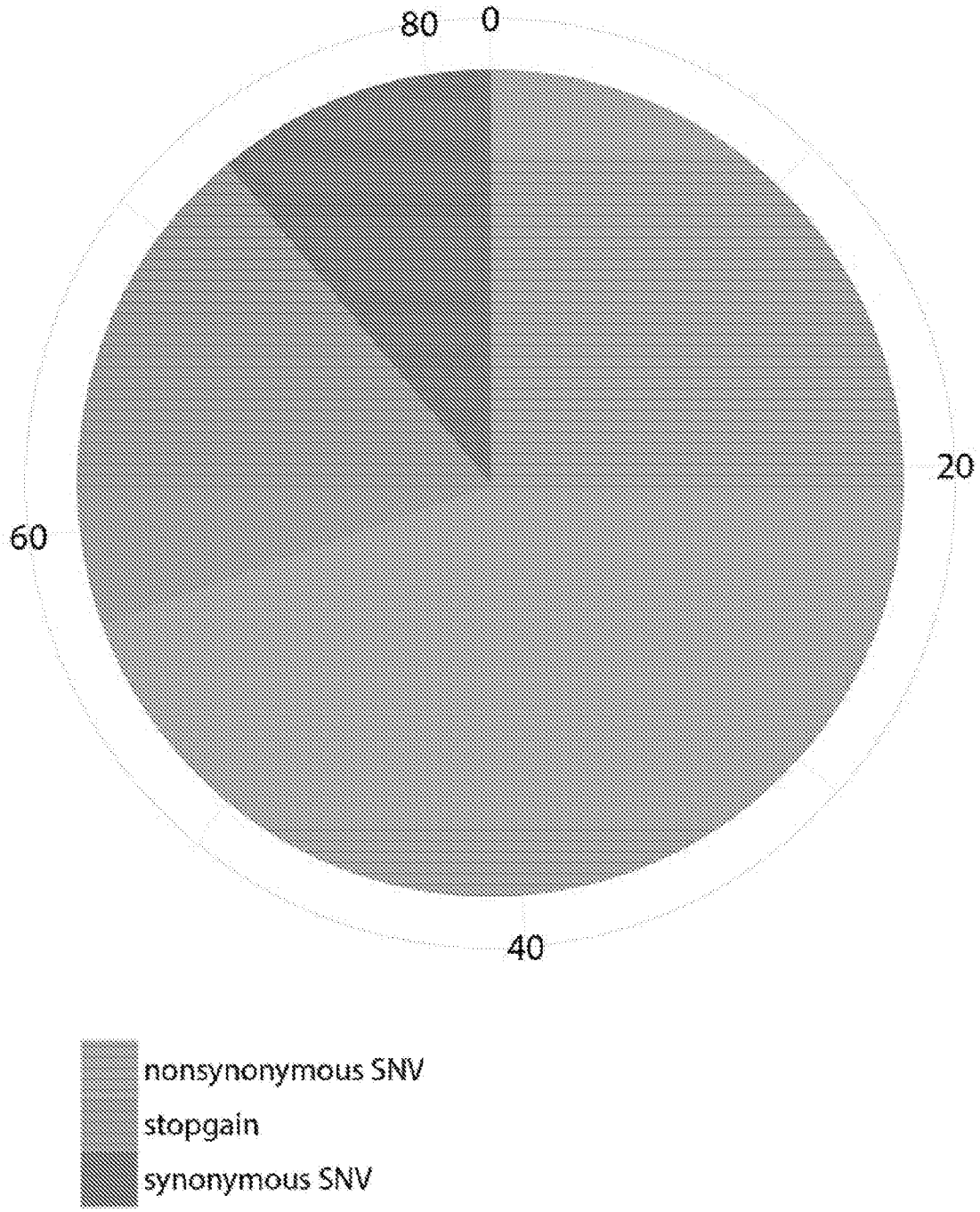

FIG. 15B

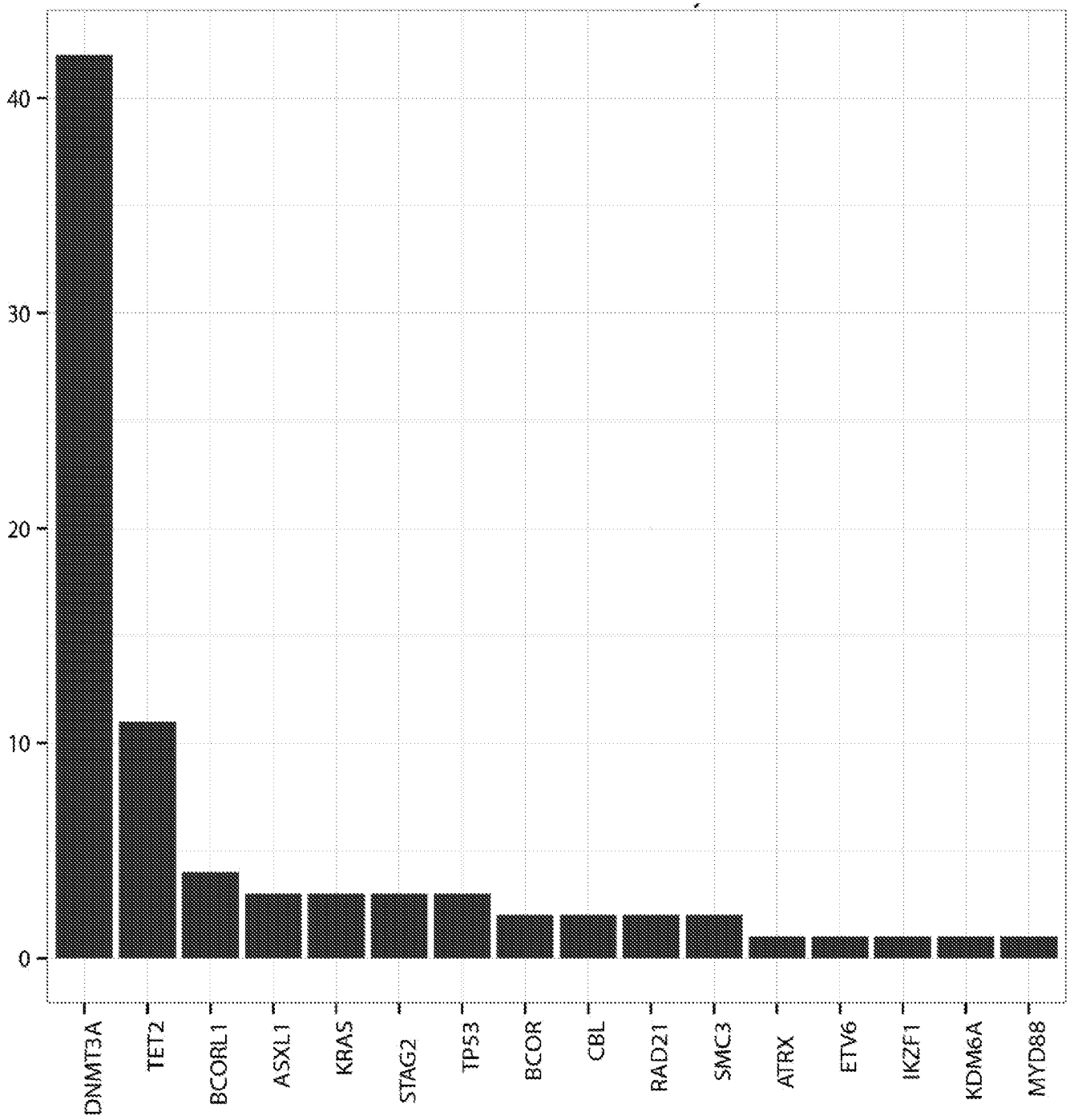
40
30
20
10
0
DNMT3A
TET2
BCORL1
ASXL1
KRAS
STAG2
TP53
BCOR
CBL
RAD21
SMC3
ATRX
ETV6
IKZF1
KDM6A
MYD88

FIG. 16

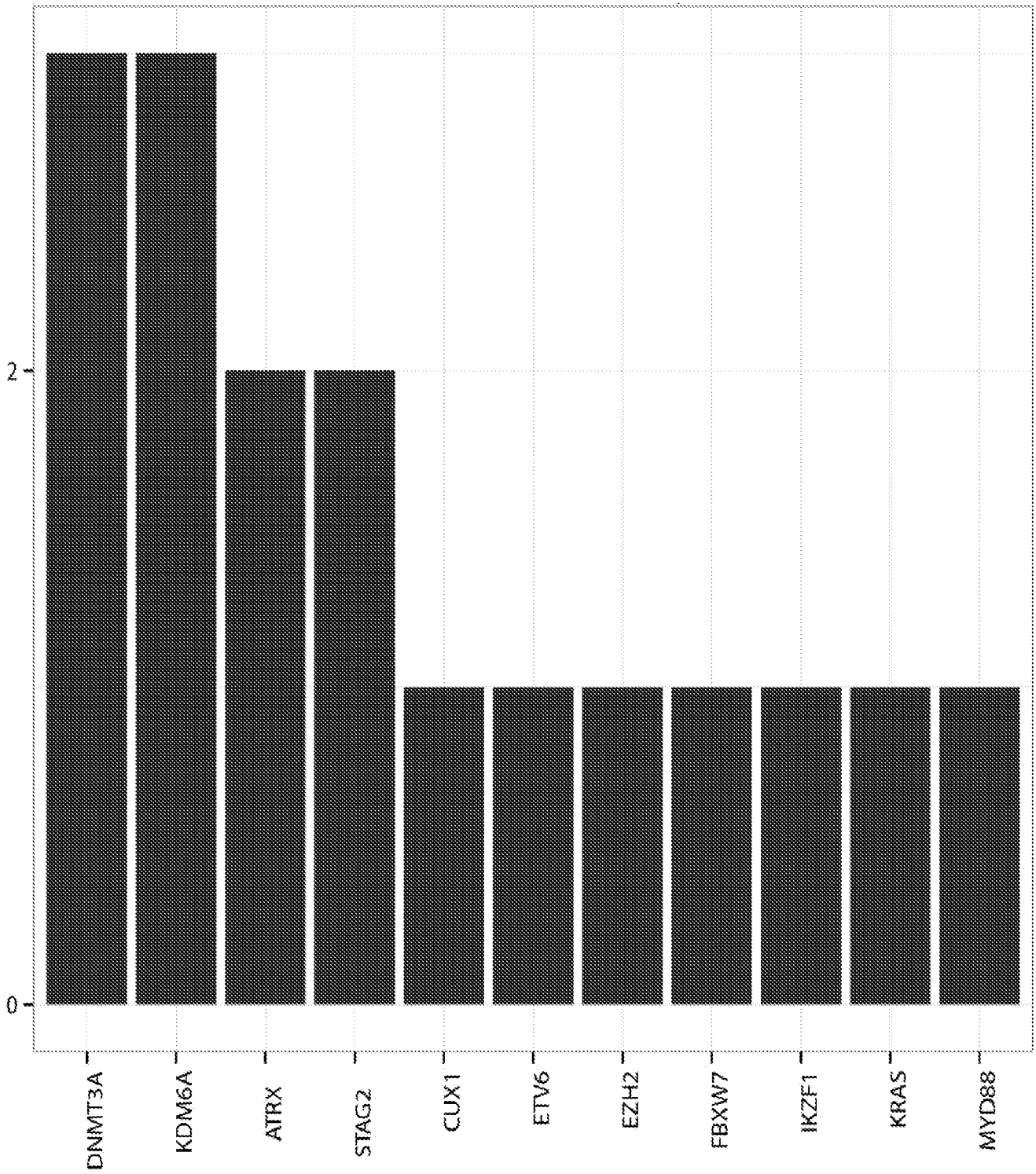
2
0
DNMT3A
KDM6A
ATRX
STAG2
CUX1
ETV6
EZH2
FBXW7
IKZF1
KRAS
MYD88

FIG. 17

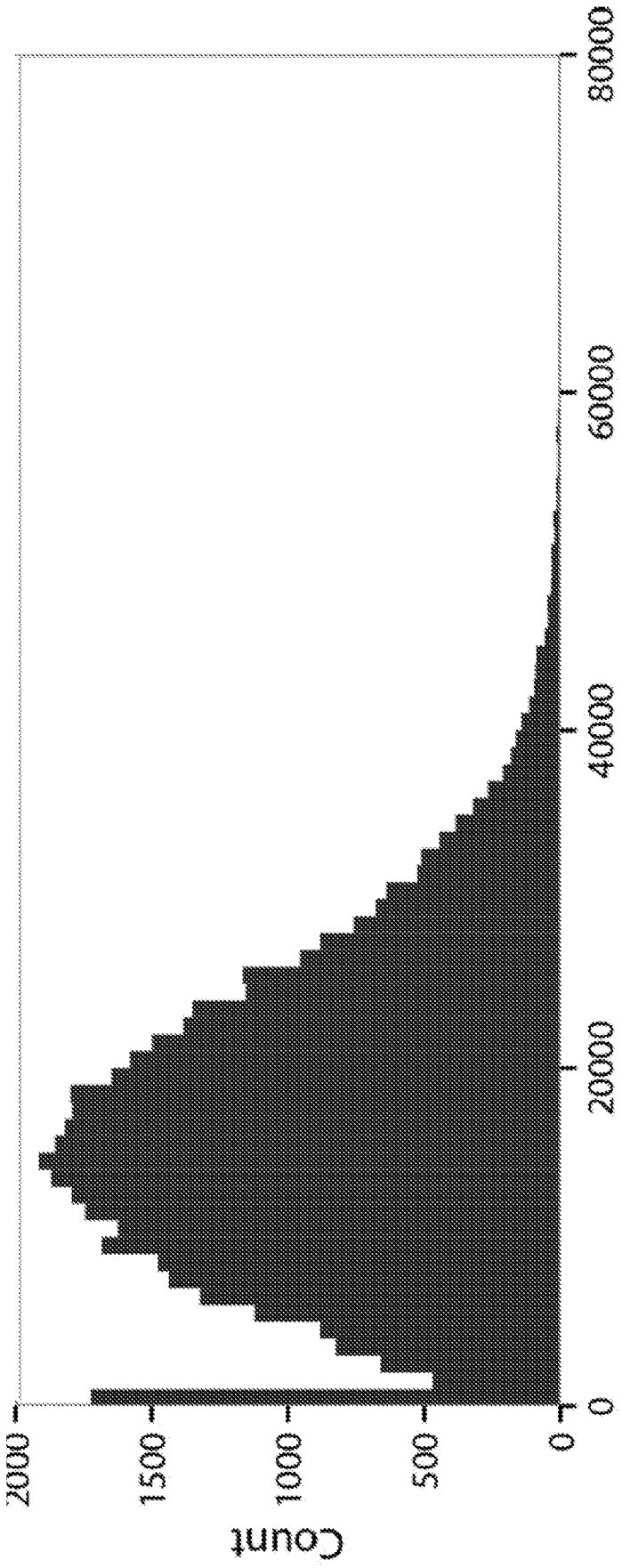
Count
2000
1500
1000
500
0
0
20000
40000
60000
80000
Amplicon Coverage

FIG. 18A

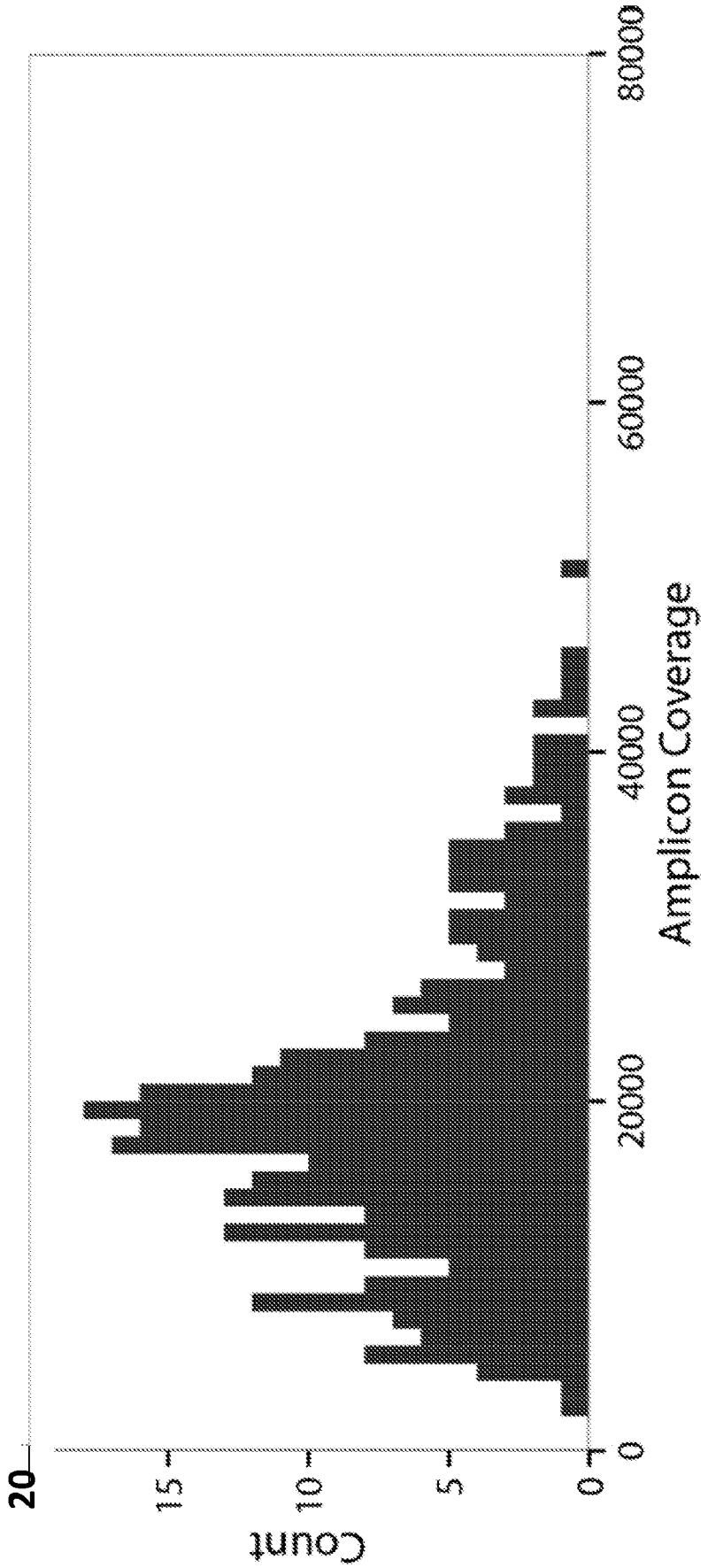
Count
0
5
10
15
20
0
20000
40000
60000
80000
Amplicon Coverage

FIG. 18B

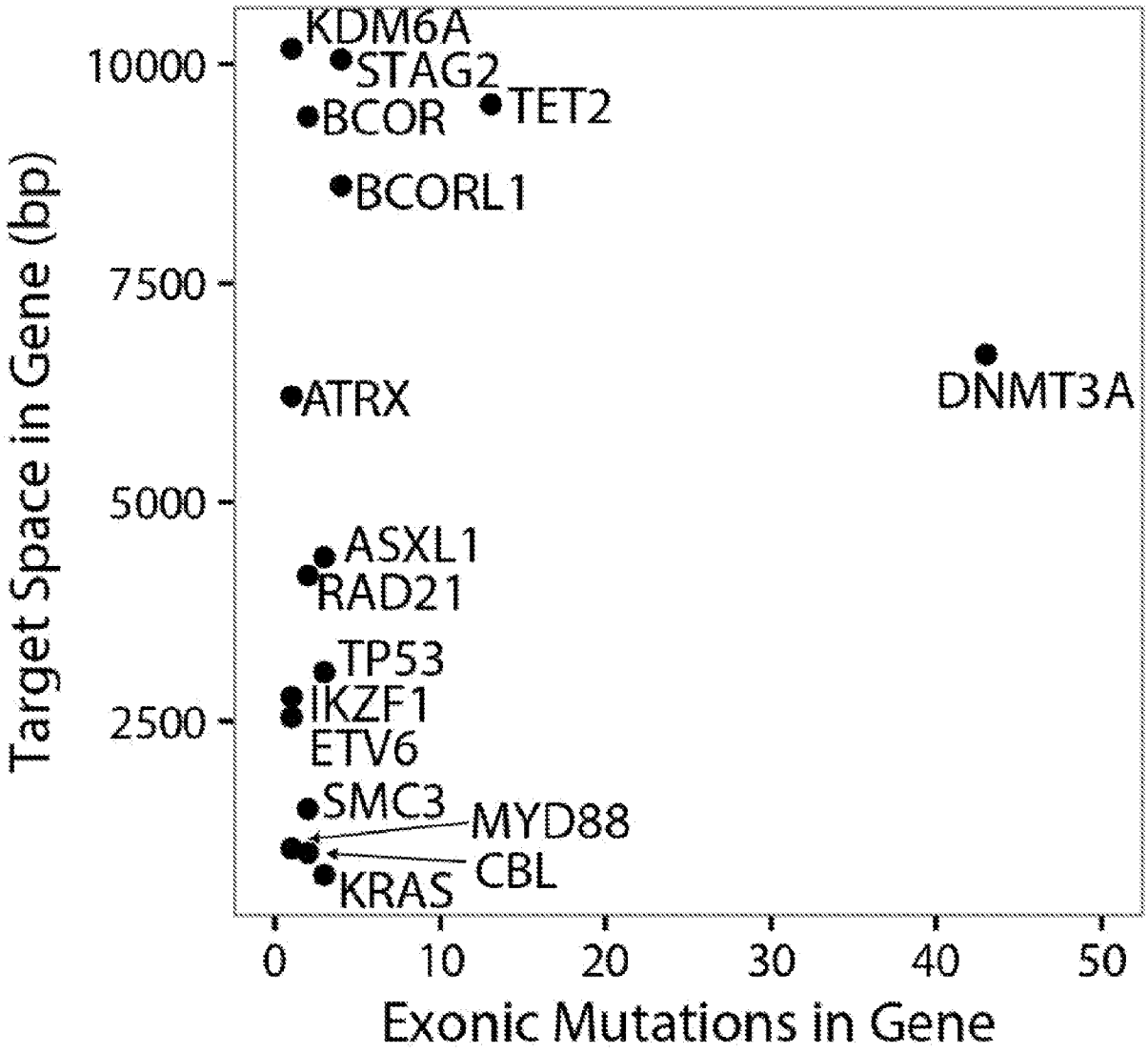
KDM6A
STAG2
TET2
BCOR
BCORL1
ATRX
DNMT3A
ASXL1
RAD21
TP53
IKZF1
ETV6
SMC3
MYD88
CBL
KRAS
10000
7500
5000
2500
0
10
20
30
40
50
Target Space in Gene (bp)
Exonic Mutations in Gene

FIG. 19A

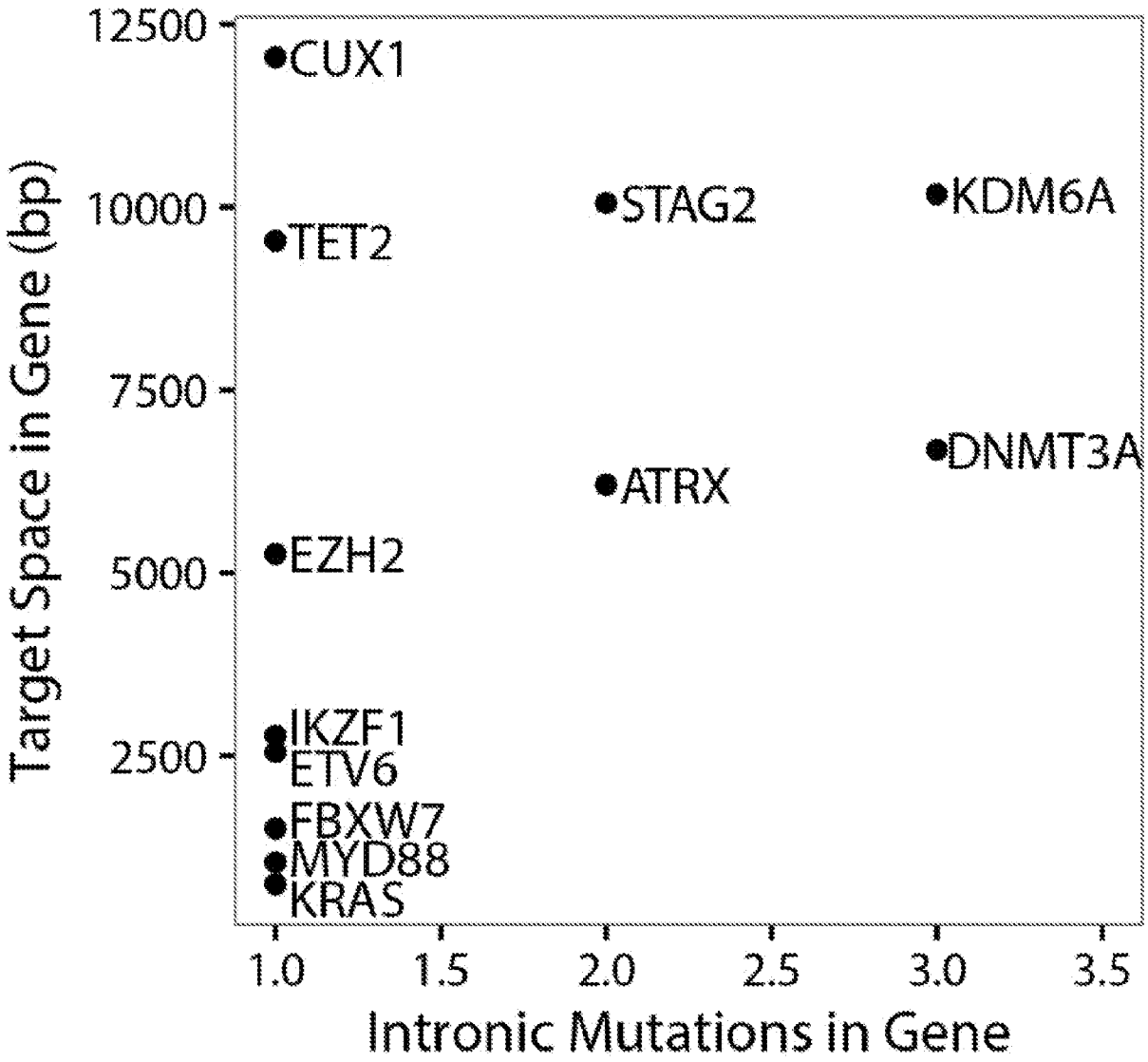
12500
10000
7500
5000
2500
Target Space in Gene (bp)
CUX1
TET2
STAG2
KDM6A
DNMT3A
ATRX
EZH2
IKZF1
ETV6
FBXW7
MYD88
KRAS
1.0
1.5
2.0
2.5
3.0
3.5
Intronic Mutations in Gene

FIG. 19B

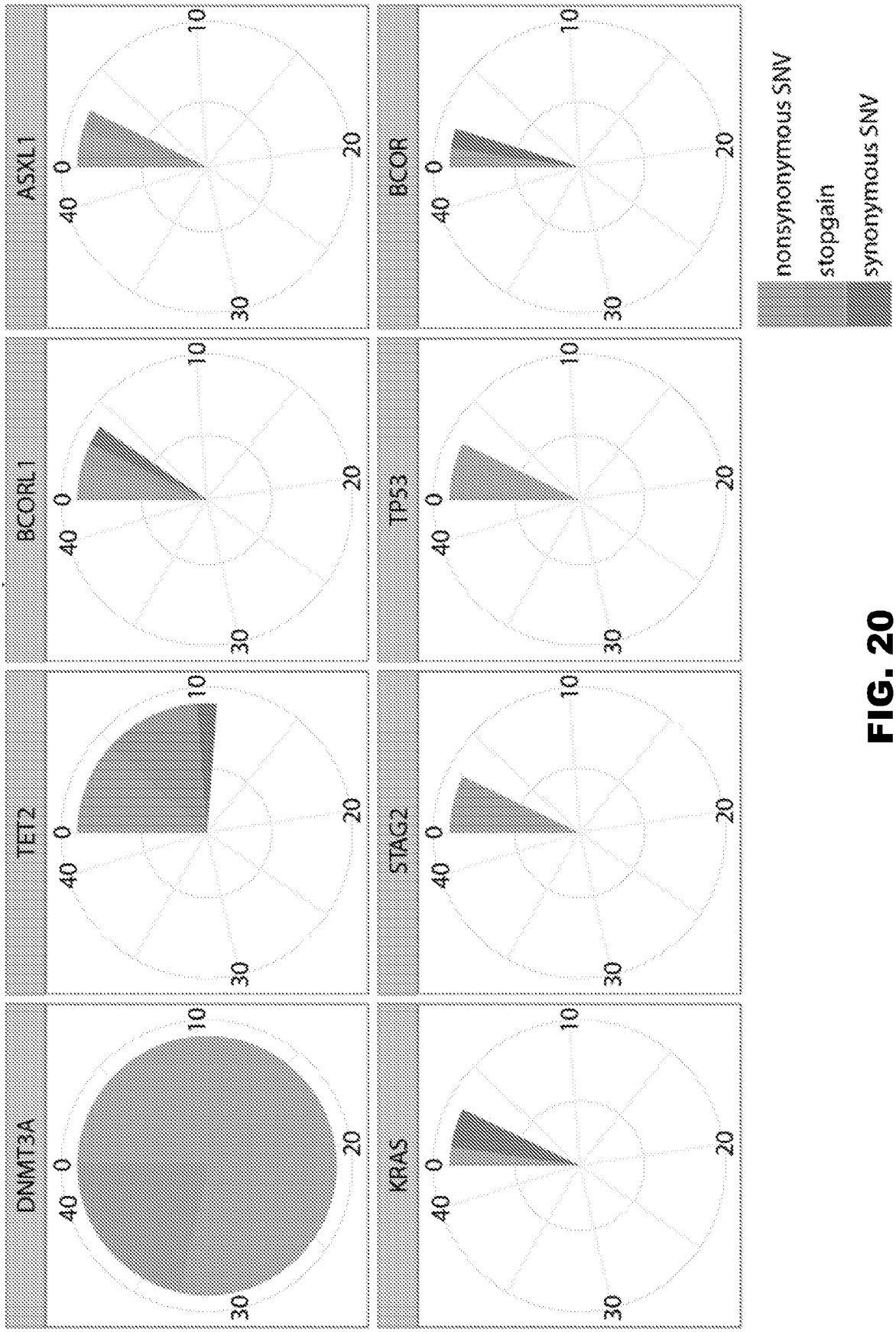
DNMT3A
TET2
BCORL1
ASXL1
KRAS
STAG2
TP53
BCOR
0
10
20
30
40
nonsynonymous SNV
stopgain
synonymous SNV

FIG. 20

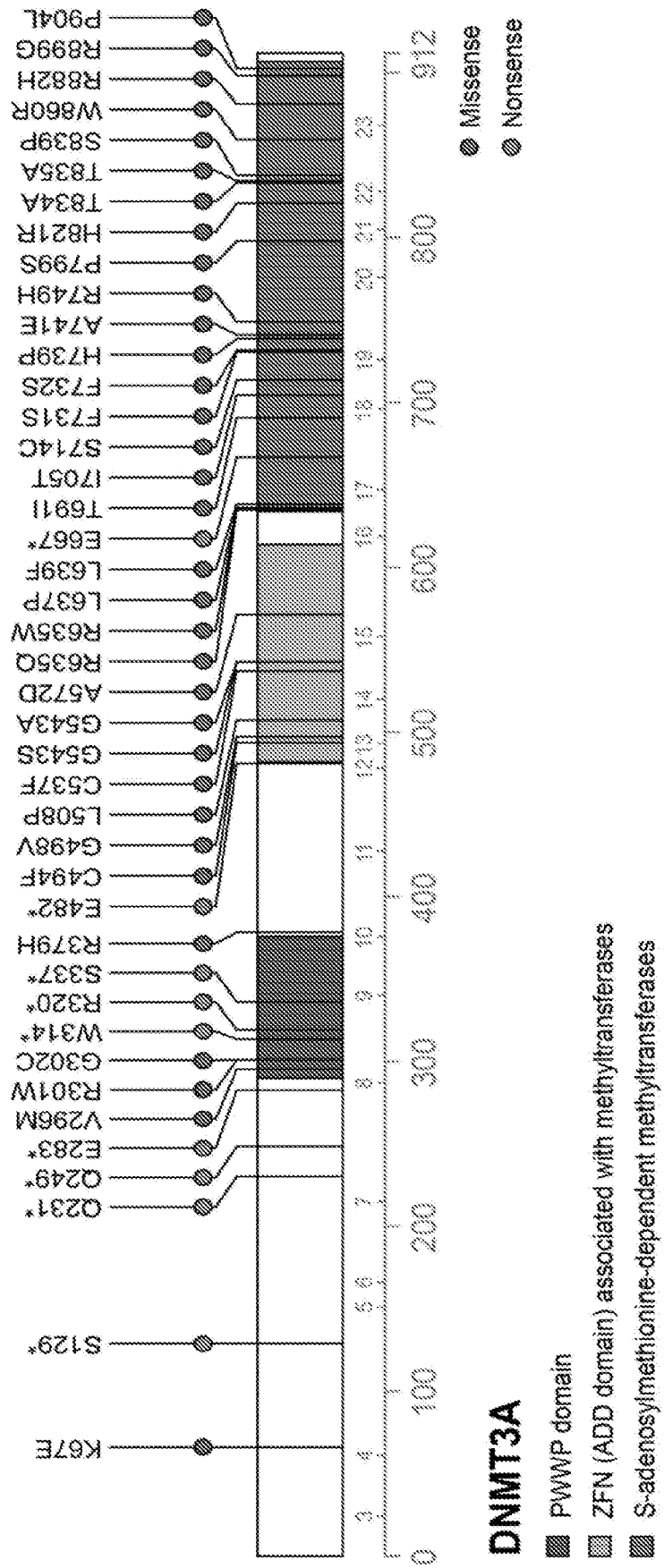
DNMT3A
PWWP domain
ZFN (ADD domain) associated with methyltransferases
S-adenosylmethionine-dependent methyltransferases
Missense
Nonsense
0
100
200
300
400
500
600
700
800
912
K67E
S129*
Q231*
Q249*
E283*
V296M
R301W
G302C
W314*
R320*
S337*
R379H
E482*
C494F
G498V
L508P
C537F
G543S
G543A
A572D
R635Q
R635W
L637P
L639F
E667*
T691I
I705T
S714C
F731S
F732S
H739P
A741E
R749H
P799S
H821R
T834A
T835A
S839P
W860R
R882H
R899G
P904L

FIG. 21

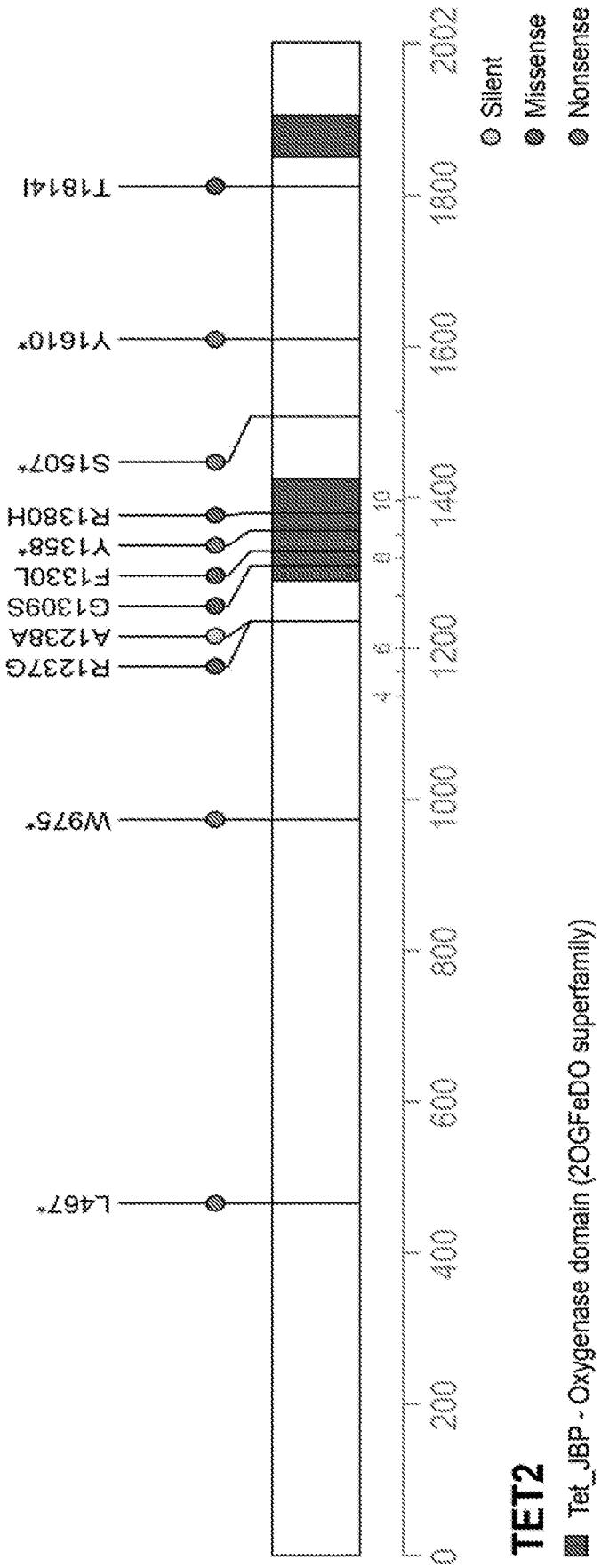
TET2
Tet_JBP - Oxygenase domain (2OGFeDO superfamily)
Silent
Missense
Nonsense
0
200
400
600
800
1000
1200
1400
1600
1800
2002
T1814I
Y1610*
S1507*
R1380H
Y1358*
F1330L
G1309S
A1238A
R1237G
W975*
L467*

FIG. 22

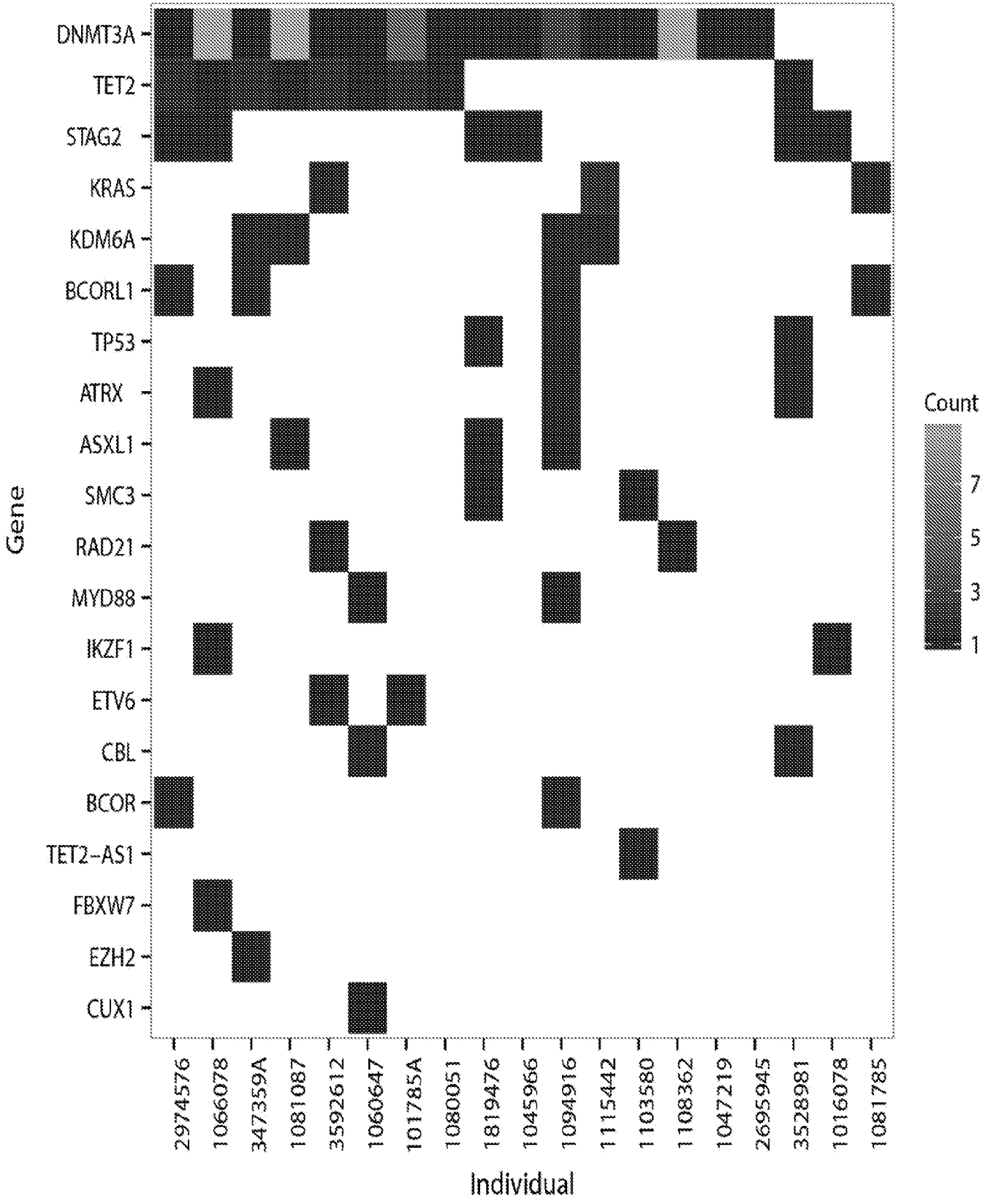
DNMT3A
TET2
STAG2
KRAS
KDM6A
BCORL1
TP53
ATRX
ASXL1
SMC3
RAD21
MYD88
IKZF1
ETV6
CBL
BCOR
TET2-AS1
FBXW7
EZH2
CUX1
Gene
2974576
1066078
347359A
1081087
3592612
1060647
101785A
1080051
1819476
1045966
1094916
1115442
1103580
1108362
1047219
2695945
3528981
1016078
1081785
Individual
Count
7
5
3
1

FIG. 23

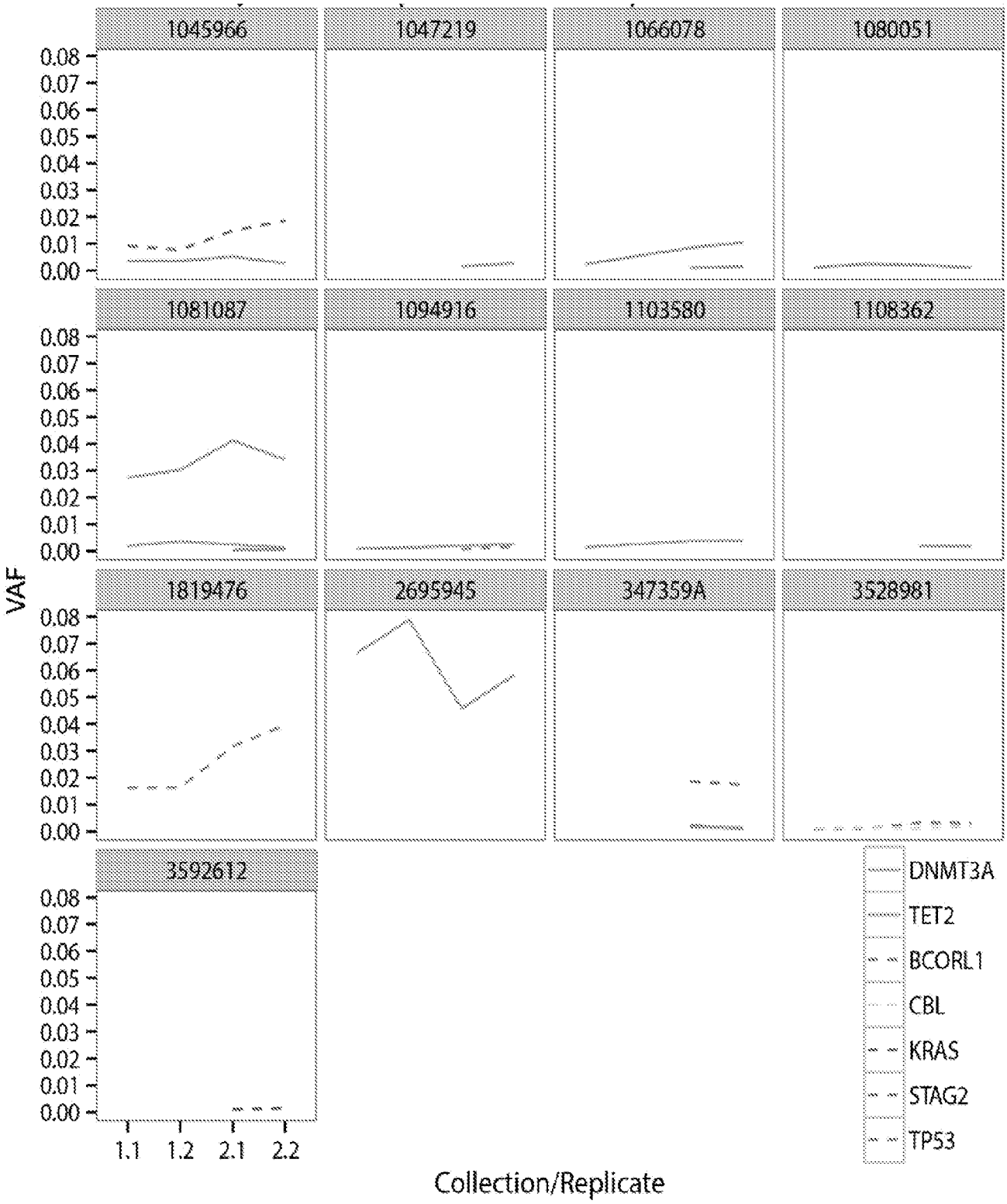
1045966
1047219
1066078
1080051
1081087
1094916
1103580
1108362
1819476
2695945
347359A
3528981
3592612
VAF
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
1.1
1.2
2.1
2.2
Collection/Replicate
DNMT3A
TET2
BCORL1
CBL
KRAS
STAG2
TP53

FIG. 24

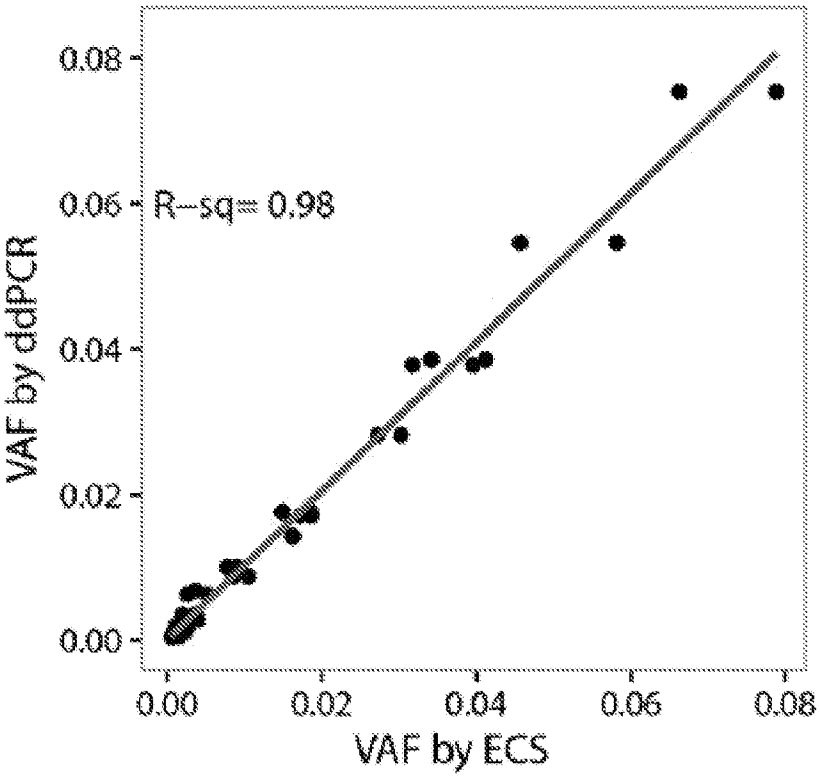

FIG. 25A

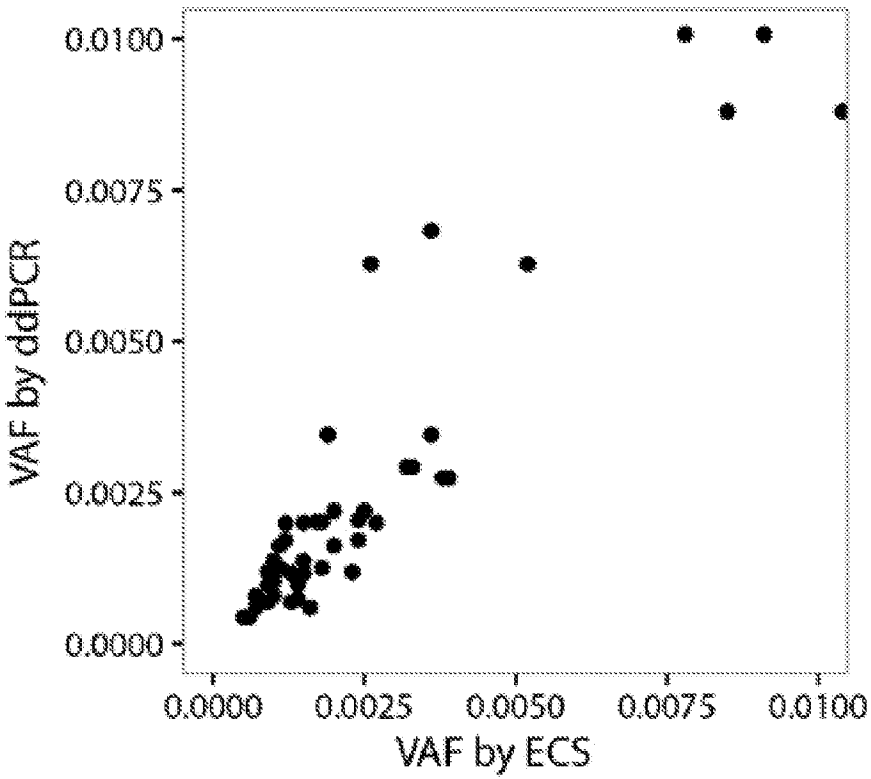

FIG. 25B

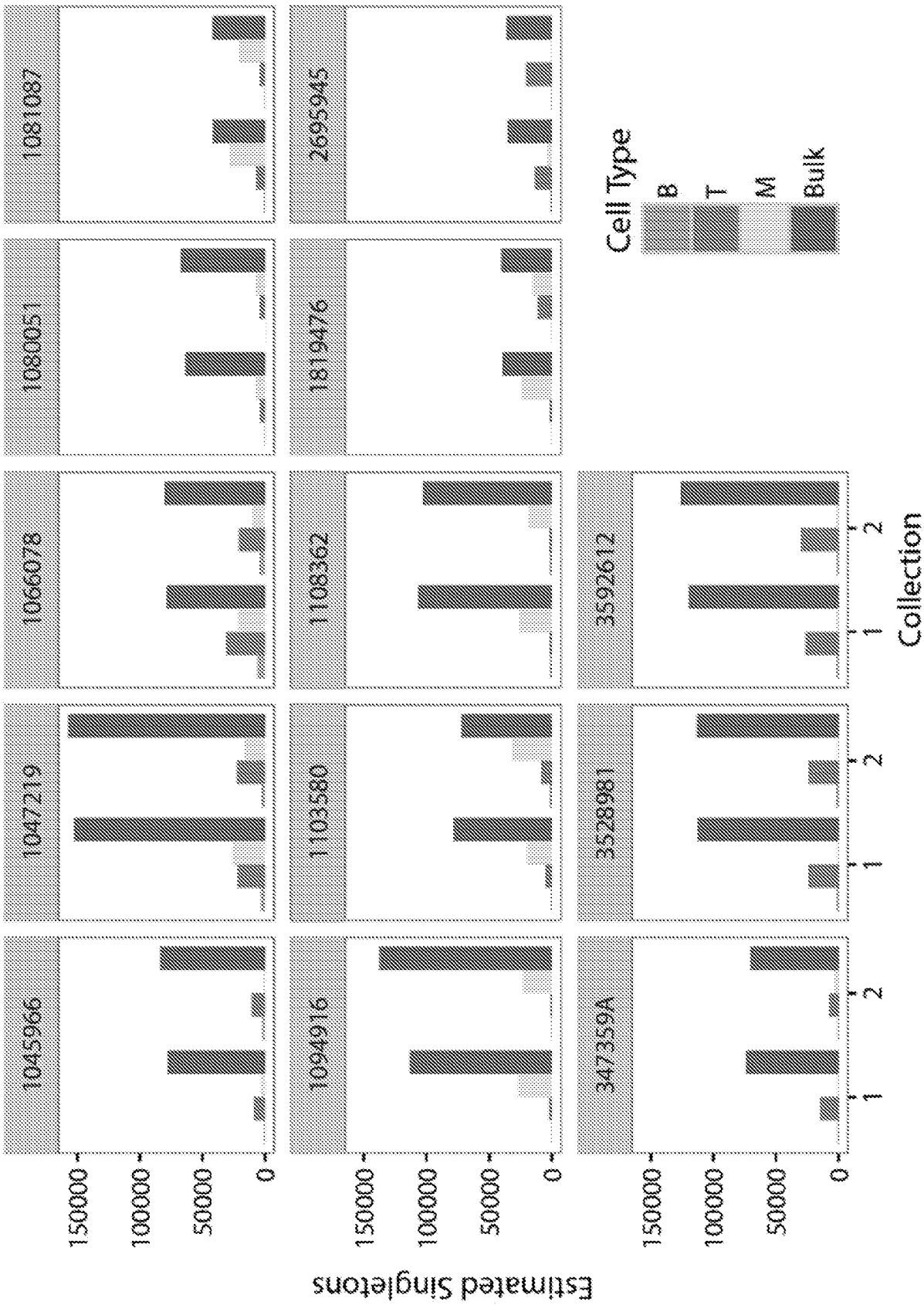
1045966
1047219
1066078
1080051
1081087
1094916
1103580
1108362
1819476
2695945
347359A
3528981
3592612
Estimated Singletons
150000
100000
50000
0
1
2
Collection
Cell Type
B
T
M
Bulk

FIG. 26

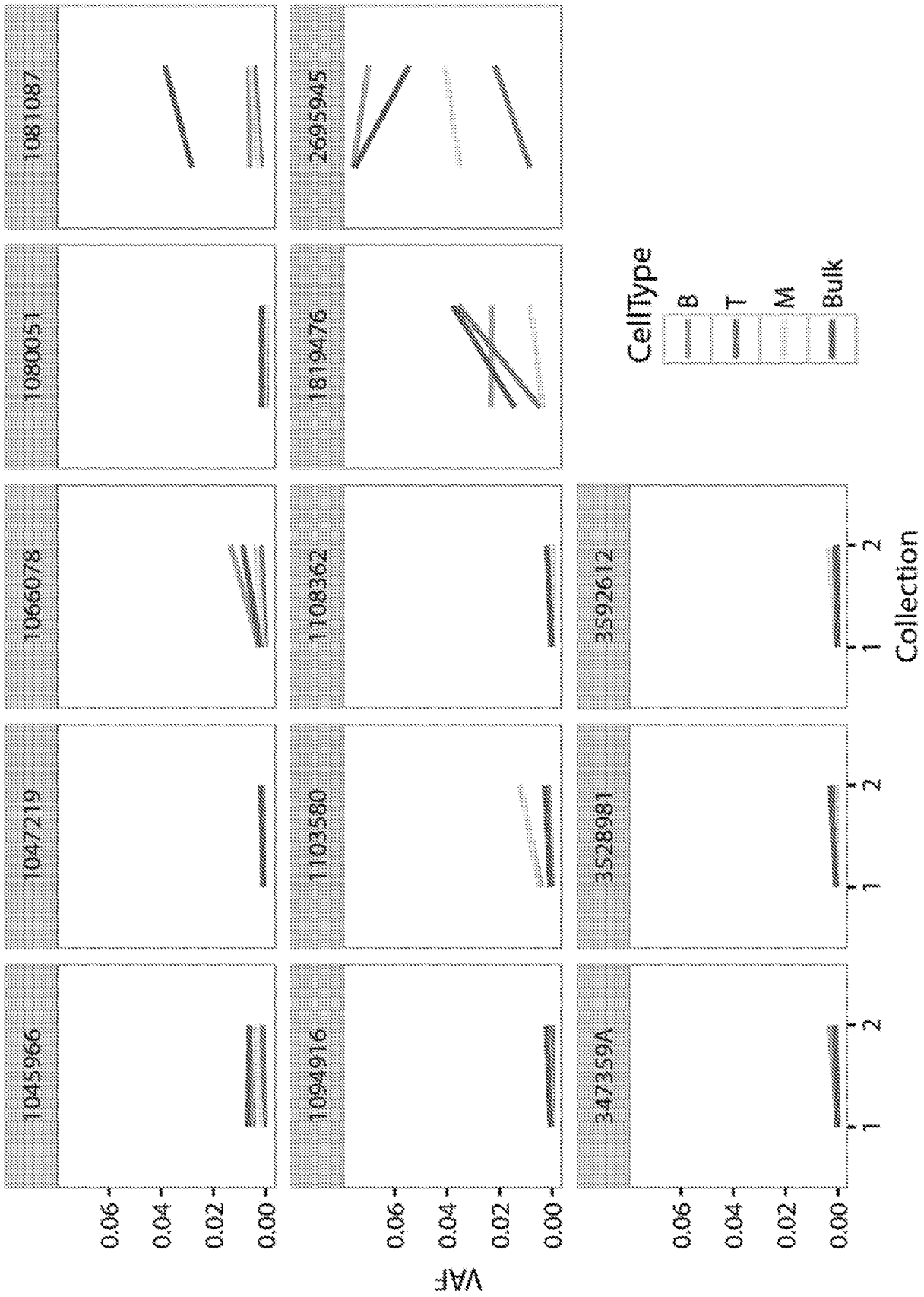
1045966
1047219
1066078
1080051
1081087
1094916
1103580
1108362
1819476
2695945
347359A
3528981
3592612
VAF
0.00
0.02
0.04
0.06
1
2
Collection
CellType
B
T
M
Bulk

FIG. 27

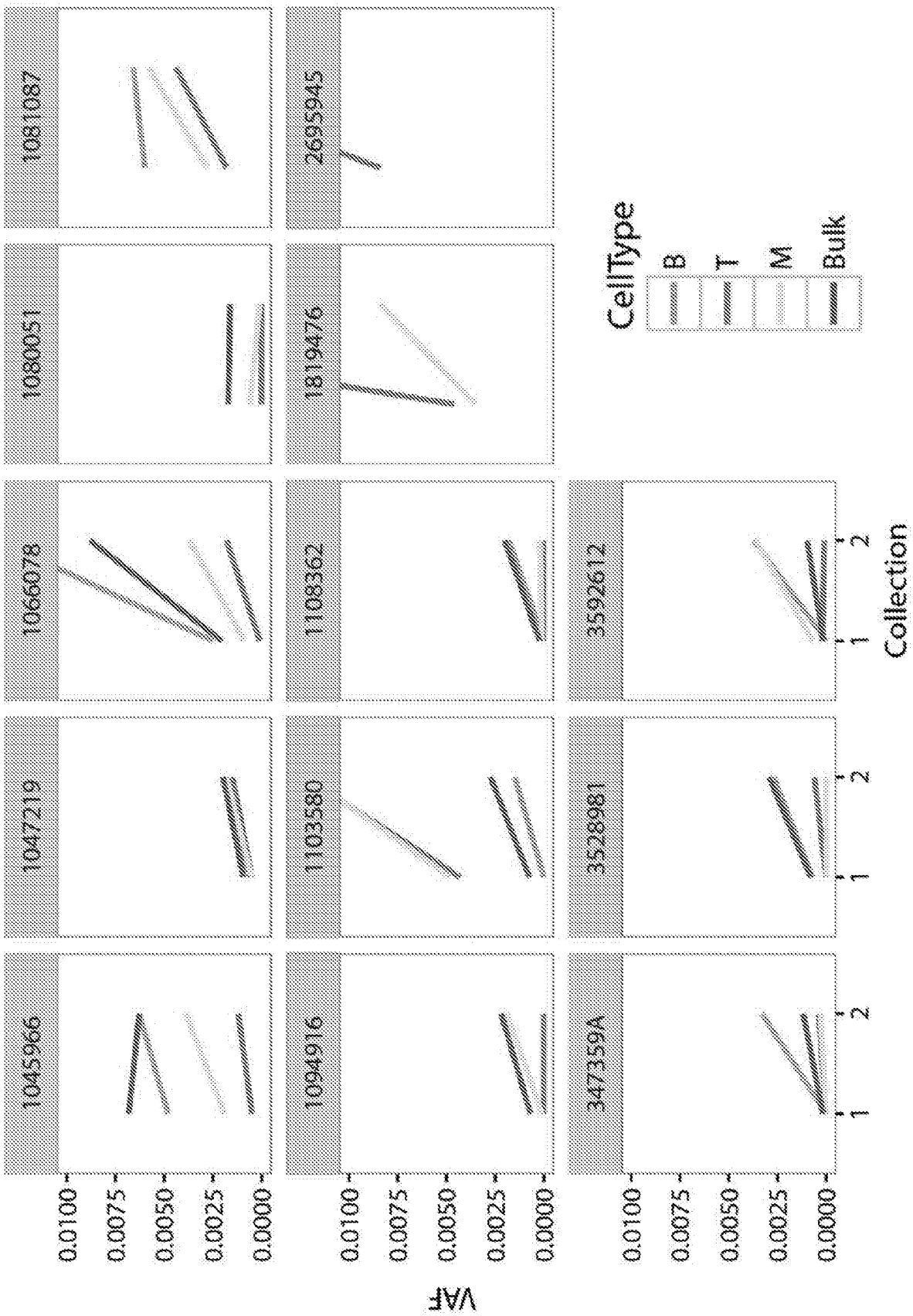
1045966
1047219
1066078
1080051
1081087
1094916
1103580
1108362
1819476
2695945
347359A
3528981
3592612
VAF
0.0100
0.0075
0.0050
0.0025
0.0000
1
2
Collection
CellType
B
T
M
Bulk

FIG. 28

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US16/14559

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - C12Q 1/68; C12N 15/10; G06F 19/22 (2016.01)
CPC - C12Q 1/6869; G06F 19/22; C12N 15/1065
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)
IPC(8): C12Q 1/68; C12N 15/10; G06F 19/22 (2016.01)
CPC: C12Q 1/6869, 1/6876; G06F 19/22; C12N 15/1065

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PatSeer (US, EP, WO, JP, DE, GB, CN, FR, KR, ES, AU, IN, CA, INPADOC Data); Google; Google Scholar; PubMed; EBSCO Discovery Service; 'amplifying', 'Nucleic acid', 'Attach', "ligation', 'random', 'Adapter', 'adaptor', 'sequencing', 'Genetic mutation', 'Error corrected sequencing', 'Next generation sequencing'

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	(WONG, TN et al.) The Role of TP53 Mutations in the Origin and Evolution of Therapy-Related AML. Nature. 08 December 2014, Vol. 518, pages 552-555; page 2, first paragraph; page 3, fourth paragraph – page 4, first paragraph; page 7, third paragraph; figure 4a; Supplementary Tables 3, 11. DOI: 10.1038/nature13968	1, 3-4, 6-7, 9-14 ------ 2, 5, 8, 15-27
Y	(QUAIL, MA et al.) SASI-Seq: Sample Assurance Spike-Ins, and Highly Differentiating 384 Barcoding for Illumina Sequencing. BMC Genomics. 2014, Vol. 15, No. 110, pages 1-12; page 3, second – third paragraphs; page 5, second column, second paragraph; page 9, second column, first paragraph; figure 1. DOI: doi:10.1186/1471-2164-15-110	2, 5, 16
Y	(SCHMITT, MW et al.) Detection of Ultra-Rare Mutations by Next-Generation Sequencing. PNAS. 04 September 2012, Vol. 109, No. 36, pages 14508-14513; page 14509, first column, first paragraph; figure 1. DOI: 10.1073/pnas.1208715109	8, 19
Y	TruSight Tumor Sample Preparation Guide [online]. (ILLUMINA., INC.) May 2013 [retrieved on 2016-03-24]. Retrieved from the Internet: <URL: http://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_trusight/trusight_tumor_sample_prep_guide_15042911_a.pdf>; pages 15-24, 39	15-27
Y	(SALIPANTE, SJ et al.) Detection of Minimal Residual Disease in NPM1-Mutated Acute Myeloid Leukemia by Next-Generation Sequencing. Modern Pathology. 2014, Vol. 27, pages 1438-1446; abstract; DOI: 0.1038/modpathol.2014.57	26-27

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
08 April 2016 (08.04.2016)	28 APR 2016
Name and mailing address of the ISA/ Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Authorized officer Shane Thomas PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (January 2015)