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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
17 October 2019 (17.10.2019)

WIPO | PCT

(10) International Publication Number
WO 2019/200398 A1

(51) International Patent Classification:
C12Q 1/6827 (2018.01) *G16B 30/00* (2019.01)
C12Q 1/6886 (2018.01) *G16B 40/00* (2019.01)

(21) International Application Number:
PCT/US2019/027525

(22) International Filing Date:
15 April 2019 (15.04.2019)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

62/657,606	13 April 2018 (13.04.2018)	US
62/658,891	17 April 2018 (17.04.2018)	US
62/827,044	30 March 2019 (30.03.2019)	US

(71) Applicant: **DANA-FARBER CANCER INSTITUTE, INC.** [US/US]; 450 Brookline Avenue, Boston, MA 02215-5450 (US).

(72) Inventor; and

(71) Applicant: **CARTER, Scott L.** [US/US]; 141 Erie St. #2, Cambridge, MA 02139 (US).

(74) Agent: **HALSTEAD, David P** et al.; Foley Hoag LLP, 155 Seaport Boulevard, Boston, MA 02210-2600 (US).

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

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ,

(54) Title: ULTRA-SENSITIVE DETECTION OF CANCER BY ALGORITHMIC ANALYSIS

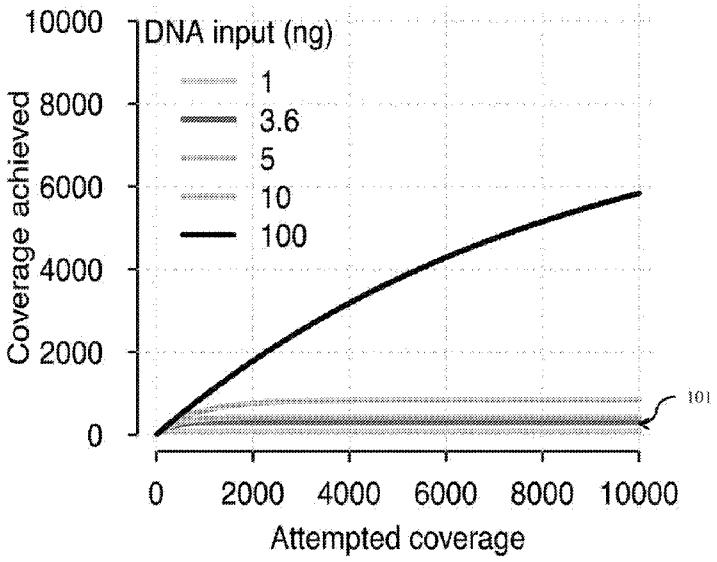

Fig. 1A

(57) Abstract: Ultra-sensitive detection of cancer by algorithmic analysis. In various embodiments, a sample comprising a plurality of polynucleotides is analyzed. A plurality of sequences of the plurality of polynucleotides is received. The plurality of sequences is provided to a trained classifier. The trained classifier is adapted to accept a plurality of sequences and output a class label indicative of the presence of a somatic variant within the plurality of sequences. A class label is received from the trained classifier indicative of the presence of a somatic clone within the plurality of sequences.

WO 2019/200398 A1

[Continued on next page]

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).

Declarations under Rule 4.17:

— *of inventorship (Rule 4.17(iv))*

Published:

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

ULTRA-SENSITIVE DETECTION OF CANCER BY ALGORITHMIC ANALYSIS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application No. 62/657,606, filed April 13, 2018, U.S. Provisional Application No. 62/658,891, filed April 17, 2018, and U.S. Provisional Application No. 62/827,044, filed March 30, 2019, which are hereby incorporated by reference in their entireties.

BACKGROUND

[0002] Embodiments of the present disclosure relate to whole-genome sequencing, and more specifically, to ultra-sensitive detection of cancer by algorithmic analysis.

BRIEF SUMMARY

[0003] According to embodiments of the present disclosure, methods of and computer program products for analyzing samples comprising a plurality of polynucleotides are provided. In various embodiments, a plurality of sequences of the plurality of polynucleotides is received. The plurality of sequences is provided to a trained classifier. The trained classifier is adapted to accept a plurality of sequences and output a class label indicative of the presence of a somatic variant within the plurality of sequences. A class label indicative of the presence of a somatic clone within the plurality of sequences is received from the trained classifier.

[0004] In some embodiments, based on the label, it is indicated whether a cancer clone is present in the sample. In some embodiments, based on the label, it is indicated whether a clonal expansion is present in the sample. In some embodiments, the label has an associated probability.

[0005] In some embodiments, clinical data are provided to the trained classifier, the clinical data being related to an originator of the sample. In some embodiments, the clinical data comprise an indication of smoking by the originator of the sample. In some embodiments, the clinical data comprise family history of the originator of the sample. In some embodiments, population data are provided to the trained classifier, the population data being related to an originator of the sample.

[0006] In some embodiments, the plurality of sequences comprise a plurality of somatic variants.

[0007] In some embodiments, trained classifier comprises an artificial neural network. In some embodiments, the trained classifier comprises a regression model. In some embodiments, the trained classifier comprises a random decision forest. In some embodiments, the trained classifier comprises an SVM. In some embodiments, the neural network is a convolutional neural network. In some embodiments, the artificial neural network is a recurrent neural network.

[0008] In some embodiments, the sample comprises blood. In some embodiments, the sample comprises cerebrospinal fluid. In some embodiments, the plurality of polynucleotides comprises DNA. In some embodiments, the plurality of polynucleotides comprises methylated DNA. In some embodiments, the plurality of polynucleotides comprises RNA.

[0009] In some embodiments, fragment lengths of the plurality of sequences are provided to the trained classifier.

[0010] In some embodiments, the sequencing is at a depth of 100x or less. In some embodiments, the sequencing is at a depth of 85x or less. In some embodiments, the sequencing is at a depth of about 20x to about 85x. In some embodiments, the sequencing is at a depth of about 20x to about 100x. In some embodiments, the plurality of polynucleotides is sequenced to obtain the plurality of sequences.

[0011] According to embodiments of the present disclosure, methods of and computer program products for analyzing samples comprising a plurality of polynucleotides are provided. In various embodiments, at least one prior sequence of a polynucleotide associated with a tumor genome is received. A generative model is fitted to the at least one prior sequence. A plurality of sequences of the plurality of polynucleotides is received. The generative model is applied to the plurality of sequences to determine a probability that a first somatic clone is present in the plurality of sequences. Based on the probability, a label indicative of the presence of the first somatic clone in the sample is determined.

[0012] In some embodiments, the label has an associated probability. In some embodiments, the generative model comprises a linear-Gaussian model. In some embodiments, the generative model comprises a linear-negative binomial model. In some embodiments, the generative model comprises a latent factor model. In some embodiments, the generative model comprises a factor analysis model.

[0013] In some embodiments, a phylogenetic tree is inferred from the at least one prior sequence. In some embodiments, the generative model is updated based on the plurality of sequences. In some embodiments, the at least one prior sequence comprises a second somatic clone, and based on the updated generative model, a probability that the sample contains the second somatic clone is determined.

[0014] In some embodiments, the at least one prior sequence comprises a second somatic clone, and, based on the updated generative model, a probability that the sample contains a descendent of the second somatic clone is determined. In some embodiments, the at least one prior sequence comprises a second somatic clone, and, based on the updated generative model, a probability that the sample contains a third somatic clone related to the second somatic clone within the phylogenetic tree is determined. In some embodiments, based on the updated generative model, a probability that the sample shares at least one somatic

mutation with the at least one prior sequence is determined. In some embodiments, based on the updated generative model, a probability that the sample shares at least one clonal expansion with the at least one prior sequence is determined.

[0015] In some embodiments, based on the label, it is indicated whether a cancer clone is present in the sample. In some embodiments, based on the label, it is indicated whether a clonal expansion is present in the sample. In some embodiments, the label has an associated probability. In some embodiments, the plurality of sequences comprise a plurality of somatic variants.

[0016] In some embodiments, the sample comprises blood. In some embodiments, the sample comprises cerebrospinal fluid. In some embodiments, the plurality of polynucleotides comprises DNA. In some embodiments, the plurality of polynucleotides comprises methylated DNA. In some embodiments, the plurality of polynucleotides comprises RNA.

[0017] In some embodiments, the sequencing is at a depth of 100x or less. In some embodiments, the sequencing is at a depth of 85x or less. In some embodiments, the sequencing is at a depth of about 20x to about 85x. In some embodiments, the sequencing is at a depth of about 20x to about 100x. In some embodiments, the plurality of polynucleotides is sequenced to obtain the plurality of sequences.

[0018] According to embodiments of the present disclosure, methods of and computer program products for analyzing samples comprising a plurality of polynucleotides are provided. In various embodiments, a plurality of sequences of the plurality of polynucleotides is received. One or more inherited variant and one or more somatic variant among the plurality of sequences are identified. The one or more inherited variant is provided to a first trained classifier. The one or more somatic variant is provided to a second trained classifier. The presence of aneuploidy in the plurality of polynucleotides is determined by the first and second trained classifier.

[0019] In various embodiments, the plurality of sequences comprise a plurality of somatic variants. In various embodiments, the first or second trained classifier comprises an artificial neural network. In various embodiments, the first or second trained classifier comprises a regression model. In various embodiments, the first or second trained classifier comprises a random decision forest. In various embodiments, the first or second trained classifier comprises an SVM. In various embodiments, the artificial neural network is a convolutional neural network. In various embodiments, the artificial neural network is a recurrent neural network.

[0020] In various embodiments, the sample comprises blood. In various embodiments, the sample comprises cerebrospinal fluid. In various embodiments, the plurality of polynucleotides comprises DNA. In various embodiments, the plurality of polynucleotides comprises methylated DNA. In various embodiments, the plurality of polynucleotides comprises RNA.

[0021] In various embodiments, the sequencing is at a depth of 100x or less. In various embodiments, the sequencing is at a depth of 85x or less. In various embodiments, the sequencing is at a depth of about 20x to about 85x. In various embodiments, the sequencing is at a depth of about 20x to about 100x. In various embodiments, the plurality of polynucleotides is sequenced to obtain the plurality of sequences.

BRIEF DESCRIPTION OF THE SEVERAL VIEWS OF THE DRAWINGS

[0022] Figs. 1A-B are plots of the expected number of unique reads as a function of raw sequencing depth for cfDNA samples.

[0023] Fig. 2 illustrates sensitive detection of cancer in cfDNA using reference-tumor somatic copy number alterations (SCNAs) according to embodiments of the present disclosure.

[0024] Figs. 3A-C depict power calculations for cancer detection using a typical lung-cancer genome as a tumor reference genome.

[0025] Figs. 4A-F illustrate improved detection of allelic imbalance in highly impure tumor DNA samples according to embodiments of the present disclosure.

[0026] Fig. 5 illustrates the power to detect *de novo* subclinical cancer using mutational signature analysis according to embodiments of the present disclosure.

[0027] Figs. 6A-C illustrate exemplary integrative analysis of genomic copy-number and point mutation according to embodiments of the present disclosure.

[0028] Fig. 7 illustrates a method of analyzing a sample comprising a plurality of polynucleotides according to embodiments of the present disclosure.

[0029] Fig. 8 illustrates a method of analyzing a sample comprising a plurality of polynucleotides according to embodiments of the present disclosure.

[0030] Figs. 9A-B provide data illustrating robust detection of subclinical lung cancer in a patient according to embodiments of the present disclosure.

[0031] Figs. 10A-E are 2D scatter plot of exemplary variant-allele fraction according to embodiments of the present disclosure.

[0032] Figs. 11A-C provide data illustrating robust detection of breast cancer in a patient according to embodiments of the present disclosure.

[0033] Figs. 12A-O are 2D scatter plots of exemplary variant-allele fraction according to embodiments of the present disclosure.

[0034] Figs. 13A-AD are 2D scatter plots of exemplary variant-allele fraction according to embodiments of the present disclosure.

[0035] Fig. 14A is a graph of the expected number of tracking variants detected against cancer fraction according to embodiments of the present disclosure.

[0036] Fig. 14B is a graph of detection power against cancer fraction according to embodiments of the present disclosure.

[0037] Fig. 15 depicts a computing node according to an embodiment of the present disclosure.

DETAILED DESCRIPTION

[0038] Early detection is one of the most effective ways to increase the cure rate for cancer, which remains the second leading cause of death in the United States. One approach to non-invasive early cancer detection is to perform genomic analysis of cell-free DNA (cfDNA), also known as liquid biopsy. Since genome sequencing of plasma cfDNA from routine blood draws can be used to detect the presence of DNA shed from tumors, this approach holds the potential to make early cancer detection a routine clinical procedure.

[0039] Genomic analysis of plasma cell free DNA (cfDNA) is useful due to its ability to non-invasively assay potential cancer-vulnerabilities in metastatic patients. Plasma cfDNA genotyping has a high positive predictive value (PPV) for detecting driver mutations in oncogenes like *EGFR* and *KRAS*, making it useful as a tool for detecting driver mutations and guiding use of targeted therapies. As discussed above, the fundamental limitation of plasma genotyping is that sensitivity for detecting known tumor-related mutations is imperfect. For example, CT-imaging can discern lung nodules as small as 0.034 cm^3, corresponding to an expected cancer-cfDNA fraction of 1.4×10^{-4} (95% CI: 6.4×10^{-6} – 3.1×10^{-3}) whereas the current limit of detection for targeted genotyping is approximately 1/1,000.

[0040] Clinically meaningful increases in detection sensitivity can be realized using the WGS approaches described herein. Among patients with advanced EGFR-mutant non small-cell lung cancer (NSCLC), rapid reductions in plasma mutations can be seen in the initial months on therapy, followed by re-emergence prior to radiographic disease progression. Some

metastatic lung cancers shed very little tumor DNA into the plasma, resulting in false negatives even with highly sensitive assays. This highlights the imperfect sensitivity of alternative PCR-based plasma genotyping assays, which are unable to detect circulating tumor DNA in patients with advanced NSCLC having a response to therapy.

[0041] Several alternative technical approaches may be taken to achieve improved sensitivity, including ultradeep sequencing and personalized (or bespoke) sequencing assays. Bespoke assays aim to increase sensitivity through PCR amplification of specific selector regions where mutations exist in a patient's cancer. Applicability of such an approach is limited by the need to sequence the tumor and then develop personalized PCR primers for detection of tumor DNA, a logistically complex process which might impair routine clinical use. Ultradeep sequencing assays aim to increase sensitivity through deep sequencing of key regions of the genome. Such an approach can be costly, is inherently limited in breadth, and may not improve sensitivity given the low levels of circulating DNA present in patients with a low burden of cancer.

[0042] The limited DNA in input material imposes a fundamental physical limitation on detection. Accordingly, targeted PCR or deep-coverage NGS assays are unlikely to achieve useful sensitivity for patients with subclinical cancer. Such patients will typically shed only a few nanograms of cfDNA per tube of blood – the mass equivalent of approximately 1,000 genomes (3.6ng) – so the resolution to quantify any given mutation cannot be greater than 1 in 1,000, even assuming error-free sequencing. In practice, this limitation is made even more severe by the limited efficiency (10-30%) of adaptor ligation prior to PCR amplification of DNA for NGS, resulting in loss of the majority of DNA fragments during library preparation.

[0043] Referring to **Figs. 1A-B**, plots are provided of the expected number of unique reads as a function of raw sequencing depth, illustrating the diminishing returns on increasing

attempted coverage when using limited-input cfDNA samples. The *y*-axis shows the expected number of unique reads (true coverage) as a function of raw sequencing depth attempted (*x*-axis). Highlighted curve **101** corresponds to an input DNA mass of 3.6ng, a typical quantity obtainable from plasma cfDNA in a patient with subclinical cancer. **Fig 1B** is a zoomed in view of **Fig. 1A**, showing that moderate-depth coverage is possible with limited DNA input. Assuming 30% adaptor-ligation efficiency, then 100X attempted coverage will yield 85X de-duplicated (true) coverage (dashed lines).

[0044] In general, coverage (or depth) in DNA sequencing is the number of reads that include a given nucleotide in the reconstructed sequence. Deep sequencing refers to the general concept of aiming for high number of replicate reads of each region of a sequence. Even though the sequencing accuracy for each individual nucleotide may be very high, the very large number of nucleotides in a genome means that if an individual genome is only sequenced once, there will be a significant number of sequencing errors. Many positions in a genome contain rare single-nucleotide polymorphisms (SNPs).

[0045] A further limitation of targeted-sequencing approaches is that that they are unable to distinguish benign from malignant clones, since many indolent clonal expansions harbor driver mutations identical to those found in deadly cancers. All somatic tissues result from an evolutionary process, whereby repeated rounds of mutation and selection result in mosaic genetic heterogeneity. Although cancer can be thought of as an extreme case involving strong positive selection and clonal expansion, such expansions have also been demonstrated in physiologically normal tissues, and are likely pervasive. Distinguishing benign from malignant clones requires methods that can incorporate genome-wide clonal co-variation to sensitively track clonal populations and match them to population-level knowledge banks.

[0046] To address these and other shortcomings of alternative approaches, various embodiments of the present disclosure leverage breadth of genomic coverage, rather than depth, to achieve ultra-sensitive detection of somatic clonal expansions consistent with cancer. Whole genome sequencing (WGS) to moderate depth (20-100x) can provide excellent sensitivity to detect the presence of somatic cancer clones by statistical integration of weak evidence from up to millions of somatic alterations, even though the sensitivity to detect any single cancer mutation is low. WGS approaches described herein address the challenge of clonal heterogeneity by statistical reconstruction of phylogenetic trees relating all somatic clones detected in a given patient. Longitudinal tracking of clones, rather than mutations, enables discrimination between benign (slow growing) and malignant (fast growing) clones; even if both harbor identical drivers, they may be distinguished by their unique passenger mutations, of which hundreds to millions may be revealed by WGS.

[0047] The calculations depicted in **Fig. 1** illustrate that various approaches described herein enable a transformational sensitivity increase of up to three orders of magnitude over alternative approaches, and thus, improve early detection of a range of cancers. The approaches described herein may be applied with existing whole-genome sequencing (WGS) of plasma-based cfDNA obtained from peripheral blood collected during the course of clinical care. It will be appreciated that, in addition, the present disclosure is applicable to any other polynucleotides that may be sequenced by WGS techniques, including DNA, RNA, methylated DNA, or isolated fragments thereof.

[0048] As set out herein, clonal phylogeny inference may be applied to low-purity longitudinal cfDNA sequencing data. In this way, improved signal-to-noise ratios is achieved in somatic DNA copy-number and mutation calling. Various embodiments leverage clonal deconvolution and statistical reconstruction of phylogenetic trees relating all somatic clones detected in a given patient.

[0049] In various embodiments, a flexible statistical analysis toolkit is provided to detect minute *de novo* somatic clones. Methods are provided for blinded detection of somatic clones consistent with cancer. Various embodiments apply Bayesian statistics to integrate evidence from signatures of aneuploidy and mutation that are learned from population-level analyses of TCGA datasets. These may be adapted to the risk profile of the patient being monitored.

[0050] The statistical approaches to WGS provided herein transform plasma-based genotyping liquid biopsy from a tool used for guiding palliative therapy of metastatic cancer into diagnostics that can be used to detect subclinical cancer and impact cure rates in early-stage, curable cancers. In various embodiments, the techniques described herein are useful to:

- Identify patients who would benefit from additional adjuvant therapy to prevent or delay clinical recurrence based on the presence of residual cancer-derived cfDNA following therapy with curative intent.
- Measure the extent of a patient's response to immune-checkpoint blockade therapy, where imaging is inherently ambiguous due to the pseudo-progression phenomenon, whereby inflammation due to therapeutic efficacy can be indistinguishable from tumor progression – a major problem in clinical decision-making.
- Enable precision interception of cancer in healthy individuals by ultrasensitive blood-based cfDNA screening and monitoring. Approaches are provided to tailor such screening to an individual's risk profile as determined by family history, germline risk alleles, and environmental exposures.

- Study the clonal dynamics of cancer during therapy using liquid biopsy to sample disease states at high temporal resolution – such studies can reveal the cancer subclones sensitive to a therapy and provide an early readout of efficacy as well as insights into genetic drivers of resistance and/or sensitivity. Ultrasensitive readout of clonal fractions at each timepoint allows these techniques to be applied in cases where the overall cancer-cfDNA shed is very low or when the change in clonal fraction between time-points is very small.

[0051] The statistical methods provided herein outperform alternative methods for accurate somatic variant calling from next-generation sequencing data, and thus have broad applicability to cancer genome characterization in virtually all contexts. The probabilistic methods provided herein can be directly applied to various sequencing strategies, including whole-exome sequencing, targeted clinical gene panels, or whole-genome sequencing. This confers enormous flexibility to design cost-effective studies and encourages widespread adoption.

[0052] The novel computational methods provided herein for fully automated clonal deconvolution and phylogenetic reconstruction are broadly applicable to the study of intra-tumor heterogeneity and evolution. This is particularly true for studies based on cfDNA and/or archival formalin-fixed paraffin-embedded (FFPE) blocks, for which single-cell sequencing will never be an option.

[0053] Whole-genome sequencing (WGS) to moderate depth (20-100x) can provide excellent sensitivity to detect the presence of somatic cancer clones, even though the sensitivity to detect any single cancer mutation is low. This is because the probabilities of various cancer-genome alterations being represented in a sample of cfDNA are statistically independent of one another. Therefore, sensitive detection can be achieved by statistical integration of weak evidence from up to millions of somatic alterations. This approach can realize a

transformational sensitivity increase of at least three orders of magnitude over alternative approaches.

[0054] In various embodiments, deep learning techniques are used to dramatically improve the sensitivity and specificity for distinguishing true point mutations from sequencing artifacts. Linear-Gaussian dimensionality reduction is used in some embodiments to remove noise from genomic copy-ratio profiles.

[0055] In various embodiments, joint estimation of allelic ratios at heterozygous SNP sites across all samples from a given patient is performed, which can be combined with panel-based SNP phasing using the Baum-Welch algorithm for hidden Markov models to achieve unprecedented sensitivity to identify allelic imbalance due to aneuploidy – a hallmark of cancer genomes.

[0056] In various embodiments, multi-sample inference of clonal phylogeny is provided. Methods are provided to statistically infer cancer phylogenies from mixed DNA sequencing data.

[0057] Ultrasensitive phylogenetic monitoring using reference tumor-genome is provided. Such embodiments enable optimal exploitation of all available information about a patient's cancer genome in order to increase sensitivity to detect recurrence.

[0058] Ultrasensitive detection of *de novo* somatic clones is provided without use of a reference tumor-genome, based on integrating signals from aneuploidy and mutational signatures associated with cancer. These can be tuned to a patient's personal risk profile based on family history and/or environmental exposures.

[0059] As set out below, in various embodiments, methods for clonal phylogeny inference are applied to low-purity longitudinal cfDNA sequencing data. All somatic tissues result from an evolutionary process, whereby repeated rounds of mutation and selection result in

mosaic genetic heterogeneity. Computational methods may be used to deconvolve complex DNA sequencing data obtained from genetically heterogeneous cancer tissue-samples in order to infer tumor content, DNA ploidy, whole-genome doubling, absolute copy-numbers, and subclonal alterations.

[0060] Statistical clonal-deconvolution is successful due to fact that aneuploidy is an early and pervasive event in many cancers. Most human tumor samples are parsimoniously explained by a single dominant clone discernable by aneuploidy, often containing smaller nested copy-number and sequence variants. This pattern has been confirmed by single-cell DNA sequencing, revealing nearly indistinguishable copy-number profiles among the 1,000s of primary-tumor cells sequenced. This observation is also highly consistent with classic models of tumorigenesis. Although human cancer tissues are genetically heterogeneous, discrete clonal phylogenies can be readily reconstructed using statistical analysis of bulk cancer-tissue DNA sequencing data. It follows that multiple samples of cancer-tissue from the same patient provide increased power to discern tumor clonal substructure and lessen the reliance on parsimony.

[0061] In various embodiments, robust and general Bayesian methods are provided. Such approaches intrinsically operate on all available samples from a given patient through all stages of analysis, including processing of raw read-count data to call somatic variants, resolution of clonal populations, and inference of phylogenetic relationships. Various approached herein cope with inherent ambiguity in the data when only a small number of patient-samples are available for inference, such as in TCGA or clinical sequencing. Many inferences of interest, such as whether a given point-mutation is homozygous, can still be made reliably even when other aspects of the sample(s) are not uniquely identified. In addition, these methods naturally gain more power for clonal reconstruction as more samples from a given cancer are sequenced.

[0062] Methods for robust multi-sample inference of somatic variation and clonal phylogeny can be extended to enable a transformative increase in the sensitivity at which subclinical cancer may be detected.

[0063] Referring to **Fig. 2**, sensitive detection of cancer in cfDNA using reference-tumor somatic copy number alterations (SCNAs) detected in WGS data is illustrated. Rectangles **201, 202** highlight a high-level focal amplicon visually apparent in both samples. cfDNA was collected from the cerebro-spinal fluid (CSF) and plasma at the same clinical visit. The patient had active lepto-meningeal disease, but no active extracranial disease.

[0064] Referring to **Figs. 3A-C**, power calculations for cancer detection using a typical lung-cancer genome as a tumor reference genome are provided. All calculations assume a sequencing error-rate of 1 in 1,000 and control the false positive rate at 1%. In **Fig. 3A**, SCNAs (50% of the genome aneuploid at 1-copy imbalance), Log copy-ratios were modeled at 1kb resolution as Gaussian with variance 0.15, similar to typical values obtained from WGS data. In **Fig. 3B**, Mutations (2 million SSNVs) are illustrated. In **Fig. 3C**, Allelic imbalance at SNPs is illustrated, an indicator of aneuploidy.

[0065] Anecdotal data with clinical specimens as illustrated in **Fig. 2**, as well as simulation data as illustrated in **Fig. 3** demonstrate that multi-sample inference of somatic variation and clonal phylogeny can be used with moderate-depth (20-85x) WGS using low-input DNA, and can detect low-level tumor DNA in settings where ultra-deep targeted sequencing would be inefficient. Alternative approaches to deal with cancer heterogeneity other than the statistical deconvolution methods are mostly inapplicable to cfDNA samples, since they utilize (i) single-cell sequencing data; (ii) complex structural genomic variants, which will not be well powered for detection in cfDNA samples due to fact that cfDNA is sheared into short

fragments of 166bp; or (iii) long-range haplotype-resolved reads, also not compatible with short cfDNA fragments.

[0066] In various embodiments, denoising total copy-ratio profiles with linear-Gaussian dimensionality reduction is provided. Denoising algorithms are applied for improved estimation of somatic copy-ratios from DNA sequencing depth. These methods extends probabilistic principal component analysis to model library size, using a generative linear-Gaussian model whereby each sample is projected onto a linear combination of latent bias factors, each having a mean and precision for each genomic region.

[0067] In various embodiments, multi-sample inference of allelic ratios is provided that operate on multiple patient-samples simultaneously. This is statistically more powerful the single sample approaches, since all samples from a given patient must share the same SNP genotypes. Thus, even DNA samples that contain almost no tumor may be assessed accurately for allelic imbalance, provided at least one high-purity tumor-sample is available to learn the SNP phases.

[0068] In various embodiments, multi-sample inference of clonal phylogeny is provided. A phylogenetic framework is provided that regularizes model complexity by leveraging biological assumptions regarding clonal evolution and phylogeny (*e.g.*, most mutations occur once during evolution and are clonally transmitted to all descendants, other than in cases of mutation loss by segmental deletion). These assumptions are highly consistent with plausible models of somatic clonal evolution and have demonstrated broad empirical support in numerous cancer datasets. This model thus has greater statistical power and confers robustness to sequencing artifacts, which are unlikely to be consistent with a well-resolved phylogeny. Furthermore, as a result of jointly modeling all samples from a given patient together, more power is gained as more patient-samples are sequenced, since more complex phylogenies can gain definitive support.

[0069] Inference of tumor purity and ploidy is provided that operates on multiple tissue samples simultaneously. Such methods infer the number of distinct somatic clones k represented in the s samples sequenced, as well as the clonal phylogeny relating them (represented as a $k \times k$ matrix $\mathbf{T}$, where $\mathbf{T}_{i,j} = 1$ if clone i is an ancestor of clone j, and 0 otherwise), and the clonal fractions in each sample (represented as an $s \times k$ matrix $\mathbf{P}$, where the rows sum to 1). In addition, the segmental somatic copy-numbers in each clone are represented as a $k \times n$ matrix $\mathbf{D}$, where n is the number of segments. The first row of $\mathbf{D}$, corresponding to the germline, is set to 2 in all segments other than regions of copy-number polymorphism or sex chromosomes, and subsequent rows (clones) contain the integer change in copy-number at that clone / segment.

[0070] The following matrix equation is obtained for total copy-ratio (tCR): $\mathrm{tCR} = \boldsymbol{\Psi}\mathbf{PTD}$. $\boldsymbol{\Psi}$ is an s × s diagonal matrix of nuisance factors: the inverse of the average ploidy of all clones in each sample, weighted by their clonal fractions. This factor, is necessary to convert mixed integer copy-numbers into copy-ratios, reflecting the fact that a constant mass of DNA library is input for sequencing, so sequencing depth reflects locus concentration, rather than absolute count. The above equation along with the copy-ratio error-model allows inference of $\mathbf{P}$, $\mathbf{T}$, and $\mathbf{D}$. Furthermore, this equation can be extended to also incorporate allelic-fractions at each segment (using allelic depth at heterozygous SNPs), as well as the multiplicity of somatic mutations in each clone. Because each of these data-sources are independent and measured with defined error, a combined likelihood can be calculated, allowing integration of evidence from each source of data available for a given cancer. Algorithms are provided to fit this model using a hybrid approach incorporating greedy search and importance sampling.

[0071] In various embodiments, ultrasensitive phylogenetic monitoring is provided using reference tumor-genome. In order to optimally leverage prior knowledge of a patient's tumor genome to enable ultrasensitive detection, the model above is fit using all available patient samples with an appreciable amount of cancer-derived cfDNA (*e.g.*, pre-operative high-shed cfDNA samples, or else sampled cancer tissue). A low-shed cfDNA sample may then be analyzed using this phylogenetic model of the patient's cancer simply by adding a new row to **P** for the new sample. This leaves only k-1 free parameters to optimize (the fraction of each somatic clone present in the new sample), and optimally utilizes all of the copy-ratio, allelic ratio, and mutational profiles of all known clones. The probability of cancer being present in the sample is equivalent to the probability than any of the k-1 somatic clonal fractions is > 0.

[0072] Because cancers are likely to continue evolving during monitoring, this can be extended to increase the weight of evidence from variants at internal branches of the cancer's clonal phylogeny. This could increase sensitivity since those variants are more likely to be present at relapse. One advantageous property of this approach is that, as more samples are sequenced, the more resolved the phylogeny will become, and detection power will increase monotonically. This approach can deliver clinically transformative increases in sensitivity, as illustrated in **Fig. 3**.

[0073] As set out above, optimal statistical methods are provided for multi-sample somatic variant calling and clonal phylogenetic reconstruction. Many samples undetectable using ddPCR (~1/1,000) sensitivity will be detectible using the WGS approaches provided herein. Because WGS of cfDNA alone may not be adequately powered to detect all cryptic cancer clones of potential interest, aberrant methylation patterns may be incorporated by performing whole-genome bisulfide sequencing (WGBS).

[0074] In various embodiments, a flexible statistical analysis toolkit to detect minute *de novo* somatic clones is provided. For cases where no reference tumor-genome exists at appreciable fraction, such as cancer screening, techniques are provided for maximally sensitive detection of *de novo* somatic clones consistent with the presence of cancer. These techniques leverage the genomic properties of cancer – such as common mutational processes and/or pervasive aneuploidy – in order to achieve sensitive discrimination of cancer clones from sequencing artifacts in the largest number of samples. As in the monitoring case discussed above, the fact that millions of somatic variants may be independently observed using WGS is leveraged to achieve sensitivity. Several statistical approaches are provided to this problem, that can achieve high sensitivity.

[0075] Referring to **Figs. 4A-F**, improved detection of allelic imbalance in highly impure tumor DNA samples by combining panel-based statistical phasing with cfDNA-based phasing is illustrated. Curves show inferred posterior probability density functions over the quantity f, the fraction of DNA derived from the minor homologous chromosome. Results obtained using only observed imbalances in allelic depth at heterozygous SNP sites to infer segmental allelic imbalance are compared with those obtained by combining panel-based haplotype phasing with tumor allelic depth. The combined approach yields densities substantially more concentrated at the true f values (dashed vertical lines). This effect is especially pronounced as f approaches 0.5 (perfect allelic balance). These results were obtained by simulating 100X coverage of a genomic segment containing 10,000 heterozygous SNP sites and a probability of panel-based switch-error of 1 in 1,000 SNPs.

[0076] Referring to **Fig. 5**, power to detect *de novo* subclinical cancer using mutational signature analysis is illustrated. Even without a paired tumor tissue or high-shed cfDNA sample, low-level cancer clones can still be identified at clinically relevant sensitivity using Bayesian methods for mutational signature analysis. This will allow them to be applied in a

cancer-screening setting, or when no reference tumor-genome is available. Results are based on simulated 100X whole-genome sequencing with an error-rate of 1 in 1,000 sequenced bases (with 300M total spurious variants due to sequencing-errors expected in the data). The indicated number of mutations generated by the smoking signature were simulated. Each point on the curves was generated by repeating the simulation 1,000 times and computing the fraction of runs where cancer was detected. The detection threshold was selected by holding the false-positive rate fixed at 0.01. Dashed lines around each curve represent 99% credible intervals.

[0077] As set out herein, even without a reference tumor genome, sensitive detection of cancer in cfDNA is possible using statistical approaches. A flexible and robust statistical analysis toolkit is provided to detect minute somatic clones consistent with subclinical cancer in cfDNA samples using moderate depth whole-genome sequencing. In order to optimally integrate evidence from multiple classes of genomic alterations so that sensitivity is maximized, statistical methods are provided integrating evidence from multiple cancer-associated mutational processes using a hierarchical Bayesian model.

[0078] In various embodiments, ultrasensitive detection of de novo somatic clones using aneuploidy is provided. Aneuploidy is a hallmark of nearly all cancers, and is pervasive in epithelial cancers. Statistical frameworks are provided to use somatic copy number alterations (SCNAs) in two ways to detect *de novo* clonal expansions: (a) by the resulting imbalance in the allelic depth – the number of reads supporting the alternate (variant) and reference allele at heterozygous SNPs (hets), and (b) by the effect on the total number of reads at each locus, quantified in terms of total copy-ratio (tCR).

[0079] The allelic imbalance model may be described as follows. At het h, the number of alt reads is distributed as $a_h \sim \text{Binomial}(N_h, f_h)$, where N_h is the total read count (alt+ref) and

f_h is the fraction of copies in the sample that contain the alt allele. In particular, in copy neutral regions, $a_h \sim \text{Binomial}(N_h, 1/2)$. More generally, for all hets within a region of constant copy number, f_h is either ϕ or $1-\phi$ depending on the phase, where ϕ or $1-\phi$ are the fractions of copies of that region coming from each homologous chromosome. To model this, a time-inhomogeneous hidden Markov model (HMM) with a Binomial emission distribution is used, with the hidden state representing both the phase and the value of ϕ. For each cancer type t, an HMM is defined with a transition matrix sequence informed by previous data on commonly amplified/deleted regions, and informed by a patient-specific estimate of phase using a panel-based haplotype-phasing method such as SHAPEIT2. Maximum likelihood estimates of the set of ϕ values are computed using the Baum-Welch algorithm. For a patient with alt read counts $a = (a_1, \ldots, a_H)$ at hets $h = 1, \ldots, H$, the marginal likelihood $p(a|t)$ of cancer type t is computed using a specialized Laplace approximation, summing over all possible sequences of hidden states and integrating over the ϕ parameters. Total log copy-ratios $r = (r_1, \ldots, r_B)$ for bins $b = 1, \ldots, B$ can be incorporated into the HMM using a Gaussian emission model.

[0080] Combining tumor allelic-imbalance with panel-based phasing dramatically improves the concentration of the posterior distribution over f as illustrated in **Fig. 4**, demonstrating that this technique improves sensitivity to confidently identify aneuploid genomic segments in cryptic somatic clones. We will further extend these methods to incorporate prior information from genetic recombination maps tailored to the inferred ancestry of the input sample. Such information can be incorporated into our HMM by increasing the state-transition probability corresponding to switch-errors in regions of high meiotic recombination.

[0081] In various embodiments, ultrasensitive detection of de novo somatic clones using mutational signatures is provided. The idea behind this technique is to examine all possible genomic variants, even if they have only a single supporting sequencing read. In the case of mutations, this is expected to produce an enormous number of false positives: 3 billion bases × 100x coverage × 1/1,000 error-rate = 300M errors expected. Detection leverages the fact that many lung cancer genomes associated with tobacco smoking will be heavily mutagenized by a well-established mutational process that mutates DNA bases in a highly non-random manner easily appreciated by considering the base substitution within its 5' and 3' sequence context.

[0082] Specifically, each candidate somatic single-nucleotide substitution (SSNV) can be classified according to the bases involved in the substitution and the bases immediately adjacent on the 5' and 3' sides, *e.g.*, $_{5'}$ACG$_{3'}$ > $_{5'}$AGG$_{3'}$, resulting in 96 such classes of SSNV. Different types of cancer exhibit different relative proportions of these SSNV classes. Further, these proportions can be decomposed into characteristic signatures or frequency profiles $(w_{1k}, w_{2k}, \ldots, w_{96,k})$ corresponding to common mutational processes $k = 1, \ldots, K$. The number of SSNVs of each class is well modeled by a Poisson distribution – specifically, $x_i \sim \text{Poisson}(\sum_{k=1}^{K} w_{ik}\theta_k)$ where x_i denotes the number of SSNVs of class i in the sample, and θ_k represents the patient's exposure to mutational process k.

[0083] Novel statistical methods are provided to leverage this information to detect cancerous clonal expansions using the presence of signatures that are highly consistent with cancer. More precisely, a Bayesian approach is taken where for each cancer type t a targeted model is defined specifying a subset of active mutational processes and a prior distribution $p(\theta|t)$ on the patient's exposure to each active process, which may be informed by covariates such as known environmental exposure or familial history. The marginal likelihood of cancer

type t is then $p(x|t) = \int p(x|\theta)p(\theta|t)d\theta$ for a patient with SSNV substitution class counts $x = (x_1, \ldots, x_{96})$. Although exact computation of these integrals is prohibitively complex, a fast algorithm using the Laplace approximation may be applied. This approximation may be validated using Markov chain Monte Carlo (MCMC) sampling.

[0084] Power calculations as shown in **Fig. 5** demonstrate the utility of this approach for early detection of lung cancer. Variable numbers of mutations from the smoking process are simulated, as well as 300M errors. For realistic numbers of smoking-associated mutations (between 100K and 1M), this technique alone is powered to detect cancer clones as small as 1/1,000 shedding cells as shown in **Fig. 5**. This technique may be applied to mutational signatures (or combinations thereof) other than tobacco exposure. This technique may be combined with other genomic analyses (for, *e.g.*, aneuploidy) to improve sensitivity.

[0085] To evaluate the evidence for clonal expansions, various embodiments use a Bayesian approach to integrate information from all of these models. A patient-specific prior probability of each cancer type t is defined based on covariates and any previous test data that may be available for the patient. The posterior probability is computed of each cancer type t – as well as the null model of no cancer ($t = 0$) – using Bayes' formula: $p(t|x, a, r) \propto p(x|t)p(a, r|t)p(t)$, modeling the SSNV class counts x as conditionally independent of the het alt read counts a and the total copy ratios r, given t. From this, the posterior probability of the presence of a cancer clone can easily be computed as $\sum_{t\neq 0} p(t|x, a, r)$.

[0086] In various embodiments, filtering of sequencing artifacts is provided using deep learning. One of the challenges with somatic mutation analysis in the ultrasensitive detection paradigm is that many somatic variants appear at low allele fractions (even a single supporting read), making it difficult to distinguish them from sequencing errors. In practice, this is made even worse because the lack of adequate base-specific error-models from cfDNA

WGS data requires an assumption that the per-site error-rate is the genome-wide average of 1/1,000, even though many sites have lower error, and some have more. Furthermore, various properties of the sequencing reads can inform error-probabilities, meaning that even genome-wide base-level models will not be optimal for error suppression.

[0087] A bank of filters may be used for read-covariates and a site-level panel-of-normals (PoN). In such approaches, a filter removes potential calls based on a single attribute the filter is designed to detect. However, such filters are hard and do not consider other covariates. If one filter detects something mildly suspicious, but the potential variant passes all other filters easily, it is still removed. Additionally, such filters are generally developed off of general human observations, which may not identify more subtle tells. Another issue is that all the filters are tweaked to work well with a specific kind of sequencer and other sequencers may have a completely different error space. A deep learning model addresses these problems of filter-based approaches.

[0088] In various embodiments, a neural network classifier is applies. A neural network learns useful covariates, provided the model is deep enough, allowing the network to detect patterns that may be indicative of sequence errors that humans aren't going to notice. Other benefits include a reduced reliance on a panel of normal genome sequences, which serves to detect rare kinds of errors that aren't covered by filters in the old model. If a neural network is trained well it has no need for such a panel, since it should have learned ways to identify these kinds of errors.

[0089] Suitable artificial neural networks include but are not limited to a feedforward neural network, a radial basis function network, a self-organizing map, learning vector quantization, a recurrent neural network, a Hopfield network, a Boltzmann machine, an echo state network, long short term memory, a bi-directional recurrent neural network, a hierarchical recurrent neural network, a stochastic neural network, a modular neural network, an associative neural

network, a deep neural network, a deep belief network, a convolutional neural networks, a convolutional deep belief network, a large memory storage and retrieval neural network, a deep Boltzmann machine, a deep stacking network, a tensor deep stacking network, a spike and slab restricted Boltzmann machine, a compound hierarchical-deep model, a deep coding network, a multilayer kernel machine, or a deep Q-network.

[0090] Various examples provided herein use a neural network classifier. However, it will be appreciated that a variety of other classifiers are suitable for use according to the present disclosure, including random decision forest, linear classifiers, support vector machines (SVM), or neural networks such as recurrent neural networks (RNN).

[0091] In an exemplary embodiment, a deep neural network model learns on the sequence context surrounding a potential somatic variant, as well as some upstream MuTect annotations and summary statistics obtained from the reads at the site of a potential variant. The network is designed to initially process the positional information of the sequence context separately from the read annotations. The sequence context surrounding the potential variant is converted to a 1-hot encoding and then fed through a series of 1-dimensional convolution layers. The goal is to have this section learn specific sequence motifs that can provide information on how to classify a potential variant. The read annotations are fed in as a normalized vector and processed through a series of dense layers. The goal is to have this section learn relationships between annotations that can be useful in classifying a potential variant. The information from these upstream sections of the network is merged and processed through more dense layers. This allows the network to learn relationships between annotations and sequence motifs. The information is then passed through a softmax function producing a vector of length two with elements that sum to one. The first element of the vector is the probability the variant provided is real and the second element is the probability that it is a sequencing artifact. When training the network this output is then given to a

binary cross-entropy loss function along with the actual label of the variant, which allows the network to be trained through back-propagation. After the network is trained the greater of the two values identifies how the potential variant is classified.

[0092] In an exemplary implementation, gold-standard variant calls from DREAM challenge batches 1 and 3 are used for training and its performance is tested on batches 2 and 4. An accuracy of 97.12% is achieved with a true-positive rate (TPR) of 96.02% and a false-positive rate of 1.79%. This approach therefore has great promise in increasing our ability to sensitively identify cryptic *de novo* cancer clones using mutational signatures.

[0093] The methods described herein may be evaluated by performing *in-silico* mixing of WGS data from tumor-normal pairs generated, for example, from the Cancer Genome Atlas (TCGA). Aligned reads from tumor-normal BAM pairs may be mixed in variable defined fractions. The resulting mixtures may be analyzed as set forth herein to infer the fraction of tumor DNA present – both using the pure tumor genome as a reference, and using the *de novo* detection techniques. These mixtures allow generation of validation data for these methods over mixing fractions that might be difficult or expensive to achieve in the laboratory.

[0094] Referring now to **Fig. 6**, integrative analysis of genomic copy-number and point mutation data across 8 samples is illustrated. In this example CSF, plasma, brain metastasis, and primary tumor samples are drawn from the same patient and subjected to whole-exome sequencing. In the 1st column of **Fig. 6A**, inference of genome-wide copy-ratios from sequencing data is shown. The black horizontal line indicates segmented copy-ratio values. Colored points show the raw data, with color alternating between orange and blue for adjacent segments. The 2nd column shows inference of allelic copy-ratios from sequencing data. Variant allele fraction (VAF; y-axis) is plotted at all heterozygous SNP sites. Points

are colored according to whether they fall on the major (red) or minor (allele). Purple points are in regions of equal allelic copy-numbers. The 3rd column shows inference of absolute allelic copy-number from sequencing data. Horizontal bars indicate absolute copy-numbers (adjusted for sample purity and ploidy). Black corresponds to total copy-number. Pink corresponds to minor-allele copy number. The 4th column shows rescaling of VAF of somatic point mutations to units of multiplicity: the number of alleles per cell. This rescaling adjusts for sample's purity and ploidy and allows subclonal mutations to be identified and quantitated. Green corresponds to mutations present in 100% of the cancer cells in the sequenced sample. Pink corresponds to subclonal mutations present in <100% of cancer cells. In **Fig. 6B** a phylogenetic tree is depicted derived from these same samples showing that the cfDNA from CSF samples was more related to the brain metastasis, and was divergent from plasma samples and the primary tumor. **Fig. 6C** shows a matrix of SSNVs (columns) by samples (rows) showing the evidence for the inferred tree.

[0095] As described herein, cfDNA is used as a biomarker for residual disease after definitive therapy, or as a pre-diagnostic test to screen for early-stage lung cancers. However, the imperfect sensitivity of existing assays represents a huge roadblock for these pursuits. If 20% of metastatic NSCLC are undetectable in plasma cfDNA (particularly those involving the lungs only), assay sensitivity will likely be inadequate as a screening tool for early stage curable lung cancers.

[0096] The use of liquid biopsies as a window to the genetic evolution of central nervous system (CNS) metastases has significant implications for the diagnosis and monitoring of these patients. The genetic drivers of central nervous system (CNS) metastases differ from their matched primary tumors and extracranial metastases in ways that impact clinical decision-making. The results of whole-exome sequencing was conducted in order to construct a global phylogenetic tree relating all cancer subclones detected in each patient.

These analyses revealed stark divergence between inter- and extra-cranial disease. This finding presents a major barrier for the use of plasma-based liquid biopsy to effectively inform treatment decisions in patients with CNS metastases, since the cancer clones represented in plasma may by highly divergent from those in the CNS. Furthermore, in patients without extra-cranial disease, plasma may contain only very small amounts of cfDNA, only a small fraction of which is likely to be cancer-derived. Unfortunately, lumbar puncture to collect CSF is an invasive procedure and may not be possible for most patients with CNS metastases. Thus, the capability to access CNS tumor genetics via a peripheral blood (or some other readily accessed biospecimen) would extend the potential benefit of precision targeting of therapy to CNS cancer.

[0097] The ultra-sensitive detection methods described herein may be applied to WGS data from paired plasma-CSF cfDNA samples in order to better understand anatomic partitioning of cancer subclones in CNS metastasis.

[0098] In the below discussion of an exemplary Allelic CapSeg model, the following terminology is used. For observed data (per segment), a denotes the number of reads supporting alternate base at each SNP and n denotes the total number of reads at each SNP. Model parameters (segment level) includes $f \in [0 - 0.5]$, the fraction of minor homologous chromosome and $\tau \geq 0$, the total copy-ratio. The minor homologous chromosome copy-ratio is denoted $f\tau$. The major homologous chromosome copy-ratio is denoted $(1 - f)\tau$. f and τ remain independent, even conditionally on the data, unlike SNP arrays. This simplifies inference. The simplest model for the probability of a_i alternate reads at SNP i given f is formed from a mixture of 2 binomial distributions:

$$\Pr(a_i|f,n_i) = \frac{1}{2}\left(\text{Binom}(a_i|f,n_i) + \text{Binom}(a_i|(1-f),n_i)\right)$$

Equation 1

[0099] The two binomial mixtures represent two possible phases to observe heterozygous SNPs: in phase 1 (alt minor) and phase 2 (ref minor). Using Bayes' rule, this can be inverted to obtain an expression for the likelihood of f given data for all the SNPs on the segment:

$$\Pr(f|\boldsymbol{a},\boldsymbol{n}) \propto \prod_i \left(\frac{1}{2}\left(\text{Binom}(a_i|f,n_i) + \text{Binom}(a_i|(1-f),n_i)\right)\right)\Pr(f|\mathcal{H})$$

Equation 2

[0100] Outliers may also be observed, which can be modeled using a third mixture distribution (uniform on unit interval), having weight ε:

$$\Pr(f|\boldsymbol{a},\boldsymbol{n}) \propto \prod_i \left(\frac{1-\varepsilon}{2}\left(\text{Binom}(a_i|f,n_i) + \text{Binom}(a_i|(1-f),n_i)\right) + \varepsilon\right)\Pr(f|\mathcal{H})$$

Equation 3

[0101] Adding an outlier component is important to make inferences on f robust to outliers (*e.g.*, genotyping errors). In an exemplary embodiments, $\varepsilon = 0.005$.

[0102] The quality of this model may be evaluated using normal (germline) samples, where $f = 0.5$.

[0103] The simple Binomial sampling model may be further improved by introducing a **sample-level** model parameter, s_f, The average value by which SNPs are skewed in this sample. For example, $observed\ AF = s_f \times true\ AF$.

[0104] This model may be further improved, as can be seen by a quantile-quantile plot of P-values for the SNPs against their expected quantities under the null distribution. The model is inflated, having an excess of highly significant SNPs (lambda =1.34).

[0105] The model may be improved by adding an additional sample level parameter v. This parameter allows specification of additional overdispersion using a Beta-Binomial model. Lower values imply more weight in the tails of the distribution, meaning that it expects to see more SNPs with skewed allelic fractions even when $f = 0.5$. This parameter can be interpreted as the number of prior pseudo-observations in the sense implied by the conjugate relationship between the binomial density and with a Beta prior distribution over the true value of p. Adding this parameter allows a much better fit on normal samples, with many values of lambda near 1.0.

[0106] In general, binomial sampling of X successes in n trials with probability of success p is given as:

$$X \sim \mathrm{Bin}(n, p)$$, then

$$P(X = k|p, n) = L(p|k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Equation 4

[0107] In Bayesian modeling, p is considered as a random variable with a prior distribution:

$$\pi(p|\alpha, \beta) = \mathrm{Beta}(\alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}$$

Equation 5

[0108] This leads to the analytically tractable composition integral to obtain the predictive distribution:

$$f(k|n,\alpha,\beta) = \int_0^1 L(p|k)\pi(p|\alpha,\beta)\,dp$$
$$= \binom{n}{k}\frac{1}{\mathrm{B}(\alpha,\beta)}\int_0^1 p^{k+\alpha-1}(1-p)^{n-k+\beta-1}\,dp$$
$$= \binom{n}{k}\frac{\mathrm{B}(k+\alpha,n-k+\beta)}{\mathrm{B}(\alpha,\beta)}.$$

Equation 6

[0109] Under maximum entropy assumptions, this distribution represents the optimal integration of information relevant to predict the next $(n+1)$th trial after observing n prior trials with α successes and β failures. Notice that the above quantity may be calculated directly, without the need for grid or numerical optimization methods – this makes it very easy to use.

[0110] To make the optimal inferences regarding the true allelic fraction f in the context of expected bias favoring observation of the reference allele, this distribution is adapted to represent uncertainty about observed $\boldsymbol{a}$ and $\boldsymbol{n}$ values equivalent to having information about f from v observations with α_1 alternate and β_1 reference allele counts in the case of SNPs in phase 1 (alt minor):

$$\alpha_1 = s_f f v,$$
$$\beta_1 = (1 - s_f f)v.$$

Equation 7

and analogous counts α_2 and β_2 for phase 2 (ref minor):

$$\alpha_2 = (1 - s_f^{-1} f)v,$$
$$\beta_2 = s_f^{-1} f v.$$

Equation 8

[0111] We then obtain:

$$\Pr(f|\boldsymbol{a},\boldsymbol{n},s_f,\nu) \propto \mathcal{L}(f=0.5|\boldsymbol{a},\boldsymbol{n},\hat{s}_f,\hat{\nu})\Pr(f|\mathcal{H}) = \prod_i \left(\frac{1-\varepsilon}{2}\left(\text{Beta-Binom}(\boldsymbol{a}_i|\alpha_1,\beta_1,\boldsymbol{n}_i)+\text{Beta-Binom}(\boldsymbol{a}_i|\alpha_2,\beta_2,\boldsymbol{n}_i)\right)+\varepsilon\right)\Pr(f|\mathcal{H})$$

Equation 9

[0112] If $s_f = 1$ then in the limit of $\nu \to \infty$ this is equivalent to the above Binomial model.

[0113] Before jointly fitting all segment level parameters τ and f, along with the two sample level parameters s_f and ν, the simpler case of normal samples may be considered, where it may be assumed that that $\tau = 1$ and $f = 0.5$ for all segments. s_f and ν are therefore jointly optimized using coordinate ascent with a numerical optimizer to obtain $\hat{s}_f$ and $\hat{\nu}$.

[0114] Fitting this model in ~100 normal exomes, there is a strong correlation between $\hat{s}_f$ and $\log\hat{\nu}$. The quality of this fit indicates that model fitting could be simplified by estimating only $\hat{s}_f$ from the data and then estimating $\log\hat{\nu}$ by inserting $\hat{s}_f$ into a linear model: $\log\hat{\nu} = \hat{a}\hat{s}_f + \hat{b}$, with $\hat{a}$ and $\hat{b}$ learned in a panel of normal exomes.

[0115] This simplification facilitates model fitting in cancer samples, where segment level parameters τ and f must also be fit. Because f and the remaining sample level parameter s_f are conditionally independent of τ, they may be optimized separately using coordinate ascent with a numerical optimization.

[0116] Even in an aneuploid tumor sample, many segments will be present at allelic balance, that is, both homologous chromosomes have the same copy-ratio. In these cases, $f = 0.5$. However, random fluctuations in the observed data may cause ML estimates of f to be < 0.5. From a Bayesian perspective, for a given segment, the free parameter f may not be a justified addition to the simpler model where $f = 0.5$.

[0117] We write the *evidence* for the two models, this quantity represents the extent to which the model fits the data, averaged over all configurations of free parameters weighted by their probability given the data (posterior) multiplied by their prior distributions. For the allelic balance model, there is no free parameter f at the segment level, and the evidence is equivalent to the above data likelihood with $f = 0.5$:

$$\Pr(\boldsymbol{a}, \boldsymbol{n}|\mathcal{H}_0) = \mathcal{L}\left(f = 0.5|\boldsymbol{a}, \boldsymbol{n}, \hat{s}_f, \hat{v}\right)$$

Equation 10

[0118] In the case of the more complex model $\mathcal{H}_1$, this evidence calculation requires integration over the posterior distribution of the free parameter f:

$$\Pr(\boldsymbol{a}, \boldsymbol{n}|\mathcal{H}_1) = \int_0^{0.5} \mathcal{L}\left(f|\boldsymbol{a}, \boldsymbol{n}, \hat{s}_f, \hat{v}\right)\Pr(f|\mathcal{H}_1)df$$

Equation 11

[0119] A uniform prior density is specified on f over 0 to 0.5, resulting in $Pr(f|\mathcal{H}_1) = 2$. Considering $\mathcal{H}_0$ and $\mathcal{H}_1$ to be the only possible models, Bayes' rule may be applied to obtain:

$$\Pr(\mathcal{H}_0|\boldsymbol{a}, \boldsymbol{n}) = \frac{\Pr(\boldsymbol{a}, \boldsymbol{n}|\mathcal{H}_0)\Pr(\mathcal{H}_0)}{\Pr(\boldsymbol{a}, \boldsymbol{n}|\mathcal{H}_0)\Pr(\mathcal{H}_0) + \Pr(\boldsymbol{a}, \boldsymbol{n}|\mathcal{H}_1)\Pr(\mathcal{H}_1)},$$

Equation 12

and $Pr(\mathcal{H}_1|a, n) = 1 - Pr(\mathcal{H}_0|a, n)$. Equal prior probabilities for the two models are specified: $Pr(\mathcal{H}_0) = Pr(\mathcal{H}_1) = \frac{1}{2}$.

[0120] In DNA, there are four bases: cytosine (C), thymine (T), adenine (A), and guanine (G). Cytosine always pairs with guanine, and thymine always pairs with adenine. Thus,

distinguishing one of the two strands of a given DNA molecule, there are four possible base pairs at each point: C-G, G-C, T-A, and A-T.

[0121] When considering base pair substitutions at a given point, the convention is to distinguish the strand containing the pyrimidine (C or T) before the substitution has been made. (Cytosine and thymine are pyrimidines, whereas adenine and guanine are purines.) With this convention, there are six possible types of substitutions at any given point:

	before	after
1	C-G	A-T
2	C-G	G-C
3	C-G	T-A
4	T-A	A-T
5	T-A	C-G
6	T-A	G-C

Table 1

[0122] Sometimes, these are abbreviated as C>A, C>G, C>T, T>A, T>C, and T>G, denoting only the pre-substitution pyrimidine and what it changes to.

[0123] These six classes can be further divided by considering the left-right context, that is, the bases adjacent to the point of substitution. The convention is to label the context in terms of the bases (C, T, A, or G) on the 5' and 3' sides on the strand containing the pre-substitution pyrimidine. For instance, in a substitution C>A, the C may be flanked by a T on the 5' side and a G on the 3' side:

before		after	
TCG		TAG	
5'	3'	5'	3'

Table 2

[0124] There are $4 \times 4 = 16$ different contexts for each of the original six substitution types. Therefore, there are $16 \times 6 = 96$ substitution types when context is taken into account.

[0125] At each position in the genome (except for a negligible fraction at chromosome ends), one of the two strands contains a pyrimidine C or T, flanked by bases on the 5' and 3' sides,

say, X and Y, respectively: *i.e.*, XCY or XTY. This defines $2 \times 4 \times 4 = 32$ possible states at each point. For each such state, a mutation causing a substitution at this point can result in one of three possibilities:

before	1	2	3
XCY	XAY	XGY	XTY
XTY	XAY	XCY	XGY

Table 3

[0126] Note that, as before, there are $32 \times 3 = 96$ types of substitution.

[0127] Focusing still on one position ℓ in the genome, assume mutations at ℓ occur as a time-homogeneous continuous-time Markov process, holding the left-right context fixed (*i.e.*, holding the neighboring base pairs constant). More precisely, when the current state is a, it remains a for an $Exp(|\Lambda_{aa}|)$ amount of time and then transitions to $b \neq a$ with probability $\Lambda_{ab}/|\Lambda_{aa}|$, where Λ is a 32×32 matrix such that (i) $\Lambda_{ab} \geq 0$ for a $a \neq b$, and (ii) $\sum_b \Lambda_{ab} = 0$. This is equivalent to saying that transitions from a to b occur with rate Λ_{ab}; thus, Λ is called the transition rate matrix.

[0128] Let S_ℓ^t denote the state at locus ℓ at time t, and let s_ℓ^0 be the state at ℓ for the normal (germline) genome of the individual under consideration. Let $P_{ab}^t = \mathbb{P}(S_\ell^t = b \mid S_\ell^0 = a)$ be the probability that the state is b at time t given that the state is a at time 0. From the theory of continuous-time Markov chains, we have that

$$P^t = \exp(t\Lambda) = \sum_{k=0}^{\infty} \frac{(t\Lambda)^k}{k!}$$

Equation 13

where $exp(\cdot)$ denotes the matrix exponential. Since the mutation rates Λ_{ab} are very small, it is reasonable to use a first-order Taylor approximation, $P^t \approx I + t\Lambda$.

[0129] Let a_i and b_i denote the starting and ending states, respectively, for each of the substitution types $i = 1, \ldots, 96$. Let $\lambda_i = \lambda_{a_i b_i}$, and define $X_i^t = \#\{\ell : S_\ell^0 = a_i, S_\ell^t = b_i\}$, *i.e.*, X_i^t is the number of positions that undergo substitution i.

[0130] Consider all of the positions ℓ that are in state a at time 0, and to simplify the math, let us assume that no two of these positions are adjacent. Of the 32 states, only four of them can be reached from a: the state can remain at a, or one of three substitutions can occur. Suppose these three substitutions are $i = 1,2,3$, so that the starting states are $a_1 = a_2 = a_3 = a$ and the ending states are b_1, b_2, b_3, respectively. Letting $n = \#\{\ell : s_\ell^0 = a\}$, $x_0 = x_1 + x_2 + x_3$, and $\lambda_0 = \lambda_1 + \lambda_2 + \lambda_3$, we have

$$\mathbb{P}(X_{1:3}^t = x_{1:3} \mid S^0 = s^0) = \frac{n!}{(n-x_0)!x_1!x_2!x_3!}(P_{aa}^t)^{n-x_0}\prod_{i=1}^{3}(P_{a_i b_i}^t)^{x_i}$$

$$\approx \frac{n!}{(n-x_0)!x_1!x_2!x_3!}(1-t\lambda_0)^{n-x_0}(t\lambda_1)^{x_1}(t\lambda_2)^{x_2}(t\lambda_3)^{x_3}$$

Equation 14

[0131] Since n is large and $t\lambda_0$ is on the order of $1/n$, then letting $c = nt\lambda_0$ we have $(1-t\lambda_0)^n = (1-c/n)^n \approx e^{-c} = \exp(-nt\lambda_0)$ and $(1-t\lambda_0)^{-x_0} = (1-c/n)^{-x_0} \approx 1$ when $x_0 \ll n$, which is the case with high probability. Plugging these into Equation 14 yields

$$\approx \frac{n!}{(n-x_0)!x_1!x_2!x_3!}\exp(-nt\lambda_0)(t\lambda_1)^{x_1}(t\lambda_2)^{x_2}(t\lambda_3)^{x_3}$$

$$= \frac{n!\,n^{-x_0}}{(n-x_0)!}\prod_{i=1}^{3}\exp(-nt\lambda_i)\frac{(nt\lambda_i)^{x_i}}{x_i!}$$

Equation 15

[0132] By Stirling's approximation,

$$\frac{n!\,n^{-x_0}}{(n-x_0)!} \sim \frac{\sqrt{2\pi n}\,(n/e)^n n^{-x_0}}{\sqrt{2\pi(n-x_0)}\,((n-x_0)/e)^{n-x_0}} = \sqrt{\frac{n}{n-x_0}}\,\frac{e^{-x_0}n^n}{(n-x_0)^n}\,\frac{(n-x_0)^{x_0}}{n^{x_0}} \longrightarrow 1$$

Equation 16

as $n \to \infty$. Hence, we have

$$\mathbb{P}(X^t_{1:3} = x_{1:3} \mid S^0 = s^0) \approx \prod_{i=1}^{3} \text{Poisson}(x_i \mid nt\lambda_i)$$

Equation 17

when n is large, $x_0 \ll n$, and $t\lambda_0$ is on the order of $1/n$.

[0133] For each of the 32 states a, the same approximation applies to the set of positions starting in state a. Modeling these 32 sets independently, we have

$$\mathbb{P}(X^t_{1:96} = x_{1:96} \mid S^0 = s^0) \approx \prod_{i=1}^{96} \text{Poisson}(x_i \mid n_i t\lambda_i)$$

Equation 18

where $n_i = \#(\ell : s^0_\ell = a_i)$. In other words, the counts of the 96 substitution types are approximately distributed as independent Poisson random variables with rates $n_i t\lambda_i$.

[0134] The preceding derivation ignores the fact that a substitution at one position changes the context of the two adjacent positions. However, since it is rare for mutations to occur at two adjacent positions, this should be negligible.

[0135] Suppose counts x_{ij} are obtained for subjects $j = 1, \ldots, J$, for substitution types $i = 1, \ldots, I$,

where $I = 96$. The derivation above suggests using the following model:

$$\text{Poisson}(x_{ij} \mid n_{ij}\lambda_{ij}t_j)$$

Equation 19

where t_j is the age (or exposure time) of subject j, and n_{ij} is the number of positions that are in state a_i in the normal genome of subject j, out of all positions that were measured, which maybe a subset of the genome due to low sequencing depth or exome sequencing for example.

[0136] Each subject is exposed to multiple mutational processes, $e.g.$, UV radiation, smoking, replication errors, $etc.$, each of which has a different rate profile across the 96 substitution types. Since rates are additive in a continuous-time Markov process, it is natural to model the subject-specific mutation rates λ_{ij} as linear combinations of these mutational process rate profiles, with nonnegative weights depending on the exposure of the subject to each process. Thus, we assume a representation of the form

$$\lambda_{ij} t_j = \sum_{k=1}^{K} r_{ik} \theta_{jk}$$

Equation 20

where the weight $\theta_{jk} > 0$ is the exposure of subject j to process k, and $(r_{1k}, \ldots, r_{1k})$ is the rate profile for mutational process k.

[0137] The exposures θ_{jk} are measured in units of time, and the rates $r_{jk}$are measured in mutations per time unit for a single position. If we measure time in years, then the parameters have the following interpretation:

$n_{ij} r_{ik}$ is the expected number of substitutions of type i in subject j due to one year of exposure to process k.

θ_{jk} is the number of years that subject j has been exposed to process k.

[0138] Since the rates for single position are exceedingly small, it is convenient to put them on a more interpretable scale by moving a fixed constant from n_{ij} to r_{ik}. Specifically, by

moving a factor of 10^6, one can measure n_{ij} in Mbp (millions of basepair positions) and measure r_{ij} in mutations/Mbp/year.

[0139] Due to sequencing errors and other technical artifacts, occasionally a position will be erroneously counted as mutated. To account for this in the model, we consider the $k = 1$ process to represent technical errors, and we set the corresponding profile to a fixed vector of rates $(r_{11}, \ldots, r_{I1})$ based on historical data.

[0140] A statistical issue with this particular representation is that there is a non-identifiability between the r_{ik}'s and θ_{jk}'s, since arbitrary constants c_k can be moved between them. This makes it problematic to consistently estimate r and θ from x data alone. One way of dealing with this is to normalize the r's by enforcing the constraint $\sum_{i=1}^{I} r_{ik} = 1$ for all k.

[0141] However, from a Bayesian perspective, this non-identifiability is not a fundamental problem since the posterior distribution is still well defined. Additional data on process-specific mutation rates (such as experimental data) can be used to help estimate r_{ik}, resolving the non-identifiability. A practical disadvantage of enforcing the normalization constraint is that it complicates inference somewhat when the start counts n_{ij} vary across subjects j. We explore both normalized and unnormalized approaches.

[0142] Often, covariates such as age, sex, smoking habits, family history, UV exposure, and so on, are available on each subject. Later, we construct a regression-based prior on the θ's that takes into account subject-specific covariates. Further, we present an algorithm for inferring the regression coefficients.

[0143] First, we present a simplified model that is interesting in its own right, and illustrates some key aspects of the inference algorithm we propose. Later, we will augment the model by (i) introducing indicator variables to allow θ_{jk} values to be exactly 0 (enabling inference

for K as well as inferring which processes are active in which subjects), (ii) incorporating the start counts n_{ij}, and (iii) using a regression model for the prior on θ_{jk} in order to handle covariates.

[0144] We describe two versions of the model: a normalized version with $\sum_i r_{ik} = 1$, and an unnormalized version. Suppose

$$\begin{aligned} &X_{ij}|r,\theta \sim \text{Poisson}\left(\sum_{k=1}^{K} r_{ik}\theta_{jk}\right) \text{ independently for } i = 1,\ldots,I,\ j = 1,\ldots,J, \\ &\theta_{jk} \sim \text{Gamma}(a,b) \text{ independently for } j = 1,\ldots,J,\ k = 1,\ldots,K, \\ &(r_{1k},\ldots,r_{Ik}) \sim \text{Dirichlet}(\alpha_1,\ldots,\alpha_I) \text{ independently for } k = 1,\ldots,K, \end{aligned}$$

Equation 21

for the normalized version, or

$$r_{ik} \sim \text{Gamma}(\alpha,\beta) \text{ independently for } i = 1,\ldots,I,\ k = 1,\ldots,K,$$

Equation 22

for the unnormalized version. The likelihood for this model is

$$p(x|r,\theta) = \prod_{i=1}^{I}\prod_{j=1}^{J} p(x_{ij}|r,\theta) = \prod_{i,j} \exp\left(-\sum_k r_{ik}\theta_{jk}\right) \frac{\left(\sum_k r_{ik}\theta_{jk}\right)^{x_{ij}}}{x_{ij}!}$$

Equation 23

[0145] Due to the sum over k in the factor $\left(\sum_k r_{ik}\theta_{jk}\right)^{x_{ij}}$, it seems that there do not exist conditionally conjugate priors for the r's and θ's. Interestingly, however, by introducing a well-chosen auxiliary variable, we can find conditionally conjugate priors.

[0146] Let us introduce

$$Y_{ij} = (Y_{ij1},\ldots,Y_{ijK})|x,r,\theta \sim \text{Multinomial}\left(x_{ij},(q_{ij1},\ldots,q_{ijK})\right)$$

Equation 24

for $i = 1, \ldots, I$ and $j = 1, \ldots, J$, where $q_{ijk} = r_{ik}\theta_{jk} / \sum_{k'=1}^{K} r_{ik'}\theta_{jk'}$. Then

$$p(y_{ij}|x,r,\theta) = \frac{x_{ij}!}{\prod_k y_{ijk}!}\prod_{k=1}^{K} q_{ijk}^{y_{ijk}} = \frac{x_{ij}!}{(\sum_k r_{ik}\theta_{jk})^{x_{ij}}}\prod_{k=1}^{K}\frac{(r_{ik}\theta_{jk})^{y_{ijk}}}{y_{ijk}!}$$

Equation 25

Since $\sum_k y_{ijk} = x_{ij}$ for all $(y_{ij1},\ldots,y_{ijK})$ in the support of this distribution. Therefore,

$$p(x|r,\theta)p(y|x,r,\theta) = \prod_{ij} p(x_{ij}|r,\theta)p(y_{ij}|x,r,\theta) = \prod_{ijk}\exp(-r_{ik}\theta_{jk})\frac{(r_{ik}\theta_{jk})^{y_{ijk}}}{y_{ijk}!},$$

Equation 26

which is, incidentally, equal to $\prod_{ijk}\text{Poisson}(y_{ijk}\mid r_{ik}\theta_{jk})$. Thus, the auxiliary variable Y has the effect of breaking up the sum over k, and as a result it is easy to do inference using Gibbs sampling on the augmented joint distribution.

[0147] The augmented joint distribution is $p(x,y,r,\theta) = p(x|r,\theta)p(y|x,r,\theta)p(r)p(\theta)$. The full conditionals are derived as follows.

Updating y. Sample from the multinomial distribution in Equation 24.

Updating θ. For each j and k,

$$\begin{aligned} p(\theta_{jk}|\cdots) &\propto p(x|r,\theta)p(y|x,r,\theta)p(\theta_{jk}) \propto \left(\prod_i e^{-r_{ik}\theta_{jk}}(r_{ik}\theta_{jk})^{y_{ijk}}\right)\theta_{jk}^{a-1}e^{-b\theta_{jk}} \\ &\propto \theta_{jk}^{A-1}e^{-B\theta_{jk}} \propto \text{Gamma}(\theta_{jk}|A,B) \end{aligned}$$

Equation 27

where $A = a + \sum_{i=1}^{I} y_{ijk}$ and $B = b + \sum_{i=1}^{I} r_{ik}$.

Updating r (normalized case).

$$\begin{aligned} p(r_{ik},\ldots,r_{ik}\mid\cdots) &\propto p(x|r,\theta)p(y|x,r,\theta)p(r_{ik},\ldots,r_{ik}) \propto \left(\prod_{ij} e^{-r_{ik}\theta_{jk}}(r_{ik}\theta_{jk})^{y_{ijk}}\right)\prod_j r_{ik}^{a_j-1} \\ &\overset{(a)}{\propto} \left(\prod_{ij} r_{ik}^{y_{ijk}}\right)\prod_j r_{ik}^{a_j-1} = \prod_j r_{ik}^{A_j-1} \propto \text{Dirichlet}(r_{ik},\ldots,r_{ik}\mid A_1,\ldots,A_J) \end{aligned}$$

Equation 28

where $A_k = \alpha_k + \sum_{j=1}^{J} y_{jk}$, and step (a) holds since $\sum_i r_{ik} = 1$ and thus $\prod_i e^{-r_{ik}\theta_{jk}} = e^{-\theta_{jk}\sum_i r_{ik}} = e^{-\theta_{jk}}$, which does not depend on r.

Updating r (unnormalized case).

$$p(r_{ik}|\cdots) \propto p(x|r,\theta)p(y|x,r,\theta)p(r_{ik}) \propto \left(\prod_j e^{-r_{ik}\theta_{jk}}(r_{ik}\theta_{jk})^{y_{ik}}\right) r_{ik}^{\alpha-1} e^{-\beta r_{ik}}$$
$$\propto r_{ik}^{A-1} e^{-B r_{ik}} \propto \text{Gamma}(r_{ik} \mid A, B)$$

Equation 29

where $A = \alpha + \sum_{j=1}^{J} y_{jk}$ and $B = \beta + \sum_{j=1}^{J} \theta_{jk}$.

[0148] To accurately recover the true values of r and θ, it is essential to infer K (the number of mutational processes). A principled Bayesian approach would be to place a prior on K and simply consider the posterior distribution on K. However, this tends to be complicated and computationally burdensome, involving dimension-changing moves or using indicator variables to include or exclude a subset of processes.

[0149] Automatic relevance determination (ARD) is an expedient alternative that effectively selects a subset of processes by driving the coefficients of any unneeded processes to be close to zero. For some models, ARD can be implemented by maximization of the marginal likelihood (also known as type-II maximum likelihood). However, for the NMF models we are interested in, the marginal likelihood does not have a closed form. Nonetheless, ARD can still be performed via a particular choice of data-dependent hyperprior that we describe here. The basic idea is to choose a hyperprior that favors having few active processes, and the key is to calibrate the strength of the hyperprior to match the corresponding likelihood, so that it is not too strong and not too weak.

[0150] First, as usual in ARD, we parameterize the model in a way that gives each process k a common scale. In the r-normalized version of the model, we modify the prior on θ:

$$\theta_{jk} \sim \text{Gamma}(a, a/\mu_k)$$

Equation 30

and in the unnormalized version, we also modify the prior on r:

$$\theta_{jk} \sim \text{Gamma}(a, a/\mu_k)$$
$$r_{ik} \sim \text{Gamma}(a, a/\mu_k),$$

Equation 31

with a common mean $\mu_k = \mathbb{E}(\theta_{jk}|\mu_k) = \mathbb{E}(r_{ik}|\mu_k)$ for all parameters involved in process k.

Then, in both versions we use a hyperprior on the means: $\mu_k \sim \text{InverseGamma}(a_0, b_0)$ for $k = 1, \ldots, K$. The full conditional distribution for μ_k is then, for the normalized version,

$$p(\mu_k|\cdots) \propto \left(\prod_j p(\theta_{jk}|\mu_k)\right) p(\mu_k)$$
$$\propto \left(\prod_j \mu_k^{-a} e^{-a\theta_{jk}/\mu_k}\right) \mu_k^{-a_0-1} e^{-b_0/\mu_k} \propto \text{InverseGamma}(\mu_k|A, B)$$

Equation 32

where $A = a_0 + J_a$ and $B = b_0 + a\sum_j \theta_{jk}$, and for the unnormalized version,

$$p(\mu_k|\cdots) \propto \left(\prod_i p(r_{ik}|\mu_k)\right)\left(\prod_j p(\theta_{jk}|\mu_k)\right) p(\mu_k)$$
$$\propto \left(\prod_i \mu_k^{-a} e^{-ar_{ik}/\mu_k}\right)\left(\prod_j \mu_k^{-a} e^{-a\theta_{jk}/\mu_k}\right) \mu_k^{-a_0-1} e^{-b_0/\mu_k} \propto \text{InverseGamma}(\mu_k|A, B)$$

Equation 33

where $A = a_0 + Ia + Ja$ and $B = b_0 + a\sum_i r_{ik} + a\sum_j \theta_{jk}$. The full conditionals for θ_{jk} and r_{ik} are modified to use a/μ_k or a/μ_k instead of b or β.

[0151] The key is to choose a_0 and b_0 appropriately, so as (i) to favor small values of μk and (ii) to have appropriately calibrated strength so that the hyperprior will be just strong enough to drive down μk for any unneeded processes. To see how to do this, we will focus on the

normalized version; the unnormalized version is similar. Since the prior mean is $\mathbb{E}(\mu_k) = b_0/(a_0 - 1)$, in order to accomplish (i), we can choose $b_0 = \varepsilon(a_0 - 1)$ for some relatively small ε, so that $\mathbb{E}(\mu_k) = \varepsilon$. For (ii), note that the full conditional for μk has mean

$$\frac{B}{A-1} = \frac{b_0 + a\sum_j \theta_{jk}}{a_0 + Ja - 1} = \frac{a_0 - 1}{a_0 - 1 + Ja}\left(\frac{b_0}{a_0 - 1}\right) + \frac{Ja}{a_0 - 1 + Ja}\left(\frac{1}{J}\sum_{j=1}^{J}\theta_{jk}\right)$$

Equation 34

a convex combination of the prior mean $b_0/(a_0 - 1)$ and the average $\frac{1}{J}\sum_{j=1}^{J}\theta_{jk}$. Thus, in order to make the prior have commensurate strength, we need to choose $a_0 - 1$ to be the same order of magnitude as J_a. Empirically, just choosing $a_0 = Ja + 1$ seems to work fine.

[0152] For processes k that are unneeded to fit the data, the parameters μk will be driven to 0, or more precisely, driven to $\approx \varepsilon$. Consequently, ε should be chosen to be significantly smaller than the values of θ_{jk} that one expects to encounter.

[0153] Here, we put together the pieces from the previous sections, to provide the Markov chain Monte Carlo (MCMC) algorithm for this basic model.

Input: $x \in \mathbb{Z}_{\geq 0}^{I \times J}$, K, ε, a, and α

Output: Samples of $\theta \in \mathbb{R}_{>0}^{J \times K}$, $r \in \mathbb{R}_{>0}^{I \times K}$, and $\mu \in \mathbb{R}_{>0}^{K}$

Set $a_0 = Ja + 1$ (normalized) or $a_0 = Ia + Ja + 1$ (unnormalized), and set $b_0 = \varepsilon(a_0 - 1)$

.

Initialize θ, r, and μ randomly (*e.g.*, by sampling from the prior).

Iteratively repeat the following steps to obtain MCMC samples:

1. Sample $y_{ij} \sim \text{Multinomial}(x_{ij}, (q_{ij1}, \ldots, q_{ijK}))$ where

$q_{ijk} = r_{ik}\theta_{jk} / \sum_{k'=1}^{K} r_{ik'}\theta_{jk'}$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

2. Sample $\theta_{jk} \sim \text{Gamma}(A, B)$ where $A = a + \sum_{i=1}^{I} y_{ijk}$ and $B = a/\mu_k + \sum_{i=1}^{I} r_{ik}$ for $j = 1, \ldots, J$ and $k = 1, \ldots, K$.

3. (Normalized version) Sample $(r_{1k}, \ldots, r_{Ik}) \sim \text{Dirichlet}(A_1, \ldots, A_I)$, where $A_i = a + \sum_{j=1}^{J} y_{ijk}$, for $k = 1, \ldots, K$.

(Unnormalized version) Sample $r_{ik} \sim \text{Gamma}(A, B)$ where $A = a + \sum_{j=1}^{J} y_{ijk}$ and $B = a/\mu_k + \sum_{j=1}^{J} \theta_{jk}$, for $i = 1, \ldots, I$ and $k = 1, \ldots, K$.

4. (Normalized version) Sample $\mu_k \sim \text{InverseGamma}(A, B)$ where $A = a_0 + Ja$ and $B = b_0 + a\sum_{j=1}^{J} \theta_{jk}$, for $k = 1, \ldots, K$.

(Unnormalized version) Sample $\mu_k \sim \text{InverseGamma}(A, B)$ where $A = a_0 + Ia + Ja$ and $B = b_0 + a\sum_{i=1}^{I} r_{ik} + a\sum_{j=1}^{J} \theta_{jk}$, for $k = 1, \ldots, K$.

[0154] To briefly demonstrate this model, we apply it to simulated data. Data for three different scenarios are generated: (a) $(I, J, K_0) = (96,100,5)$, (b) $(I, J, K_0) = (96,200,10)$, and (c) $(I, J, K_0) = (96,400,20)$. For each scenario, we generate 10 simulated data sets by sampling true parameters $\theta_{jk}^0 \sim \text{Gamma}(1, 0.5)$ and $r_{ik}^0 \sim \text{Gamma}(0.5, 0.5)$ independently, and then sampling data $x_{ij} \sim \text{Poisson}(\sum_{k=1}^{K_0} r_{ik}^0 \theta_{jk}^0)$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$ independently given the parameters.

[0155] For the model, we choose $a = 1$, $\alpha = 0.5$, and set K to be (a) 10, (b) 20, and (c) 30 for

the three scenarios, respectively. We use $\varepsilon = 0.01$ for the ARD tolerance parameter. For illustration purposes, we only use the normalized version of the model; the unnormalized version yields similar results. On each data set, we run the MCMC chain for 2000 iterations total, discarding the first 1000 as burn-in.

[0156] The model is symmetric up to permutations of the labels k, so label switching is a potential issue complicating estimation of the parameters. However, empirically we do not observe any label switching, so averaging across MCMC samples provides interpretable estimates of the parameters, even though in theory one would need to account for the possibility of label switching.

[0157] To estimate K_0, we apply a threshold to the estimated hyperparameters $\hat{\mu}k$, keeping only those processes k such that $\hat{\mu}k > 4\varepsilon$.

[0158] To facilitate comparison with the true parameters (which are not identifiable due to the possibility of moving a constant c_k between r_{ik}^0 and θ_{jk}^0), we normalize the true parameters by setting $\theta_{jk}^N = c_k\theta_{jk}^0$ and $r_{ik}^N = r_{ik}^0/c_k$ where $\tilde{c}_k = \sum_i r_{ik}^0$.

[0159] As a benchmark, we compare results with the ARD NMF algorithm of Tan and Févotte (2013) (specifically, their ℓ^1 algorithm for Poisson data), with their parameters set to $a = 5$ and their suggested choice of b. We use the same values of K as for our algorithm. The Tan-Févotte algorithm seeks a maximum a posteriori (MAP) estimate. ARD NMF results for scenario (a) are given in Table 4.

	$\chi^2(\hat{r}, r^N)$	$\chi^2(r^N, \hat{r})$	$\text{rmse}(\hat{r}, r^N)$	$\text{rmse}(\hat{\theta}, \theta^N)$	time (sec)
Our algorithm	15.85	0.849	0.00155	25.27	50.26
Tan–Fevotte	984.98	∞	0.00186	35.77	2.55

Table 4

[0160] ARD NMF results for scenario (b) are given in Table 5.

	$\chi^2(\hat{r}, r^N)$	$\chi^2(r^N, \hat{r})$	$\text{rmse}(\hat{r}, r^N)$	$\text{rmse}(\hat{\theta}, \theta^N)$	time (sec)
Our algorithm	961.3	7.01	0.00191	40.7	162.9
Tan–Fevotte	25809.7	∞	0.00303	120.0	16.08

Table 5

[0161] ARD NMF results for scenario (c) are given in Table 6

	$\chi^2(\hat{r}, r^N)$	$\chi^2(r^N, \hat{r})$	$\mathrm{rmse}(\hat{r}, r^N)$	$\mathrm{rmse}(\hat{\theta}, \theta^N)$	time (sec)
Our algorithm	96.89	6.17	0.00471	101.5	503.6
Tan–Fevotte	18353.1	∞	0.00745	225.1	35.26

Table 6

[0162] The results shown are averaged over the 10 randomly generated data sets for each scenario. In scenarios (a) and (b), both algorithms estimated the number of processes correctly in all 10 data sets, *i.e.*, (a) $\widehat{K} = K_0 = 5$ and (b) $\widehat{K} = K_0 = 10$. In scenario (c), the estimates in the 10 runs were $\widehat{K} = 20,19,21,20,20,20,19,20,19,19$ for our algorithm, and $\widehat{K} = 18,19,19,20,19,19,18,19,18,18$ for the Tan-Févotte algorithm.

[0163] The distance metrics are defined as follows: $\chi^2(p, q)$ is the chi-squared distance averaged across the columns, *i.e.*, $\chi^2(p, q) = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{I}(p_{ik} - q_{ik})^2/q_{ik}$. Similarly, $rmse(x, y)$ is the root-mean-squared error averaged across the columns, $i.e.$, $rmse(xy) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{I}\sum_{i=1}^{I}(x_{ik} - y_{ik})^2\right)^{1/2}$. Since the order of the columns is not identifiable, to compute the distances we first permute the columns to minimize the distance, using a greedy algorithm to avoid searching over all $K!$ permutations. In cases where $\widehat{K} \neq K_0$, we pad the smaller matrix with dummy columns filled with $1/I$ for r and 0 for θ.

[0164] Our MCMC algorithm is consistently much more accurate in estimating r and θ than the Tan-Févotte MAP algorithm. For estimating K, the two methods seem to be comparable in accuracy. The reason why the Tan-Févotte estimates perform so poorly with respect to $\chi^2(r^N, \hat{r})$ is that a number of its $\hat{r}_{ik}$ values are very small or equal to 0, causing the chi-squared distance to blow up.

[0165] A further advantage of our MCMC algorithm is that it provides uncertainty quantification rather than simply point estimates. For example, the MCMC samples can be used to construct posterior credible intervals for quantities of interest.

[0166] A copy ratio is the relative concentration of copies of a given locus in a given sample. In other words, it is the proportion of copies of a given locus out of all copies of observed loci in a sample, divided by the corresponding proportion in a normal sample. Total copy ratio refers to the copy ratio including copies from both homologous chromosomes, as opposed to haplotype-specific copy ratio.

[0167] In order to estimate copy ratios from sequencing data, it is necessary to account for the fact that different loci have different propensities of being observed, due to the physics of the measurement technology. Further, a significant reduction in copy ratio estimation error can be obtained by measuring typical noise profiles from a large panel of normal samples, and subtracting off the projection onto this set of noise profiles.

[0168] Let $n_{s\ell} :=$ number of reads of locus ℓ in sample s. This is the observed data. Later, in the point mutation model, we will also consider the number of reference versus alternate reads. (Note: We use the notation $a := b$ to mean that a is defined to be b.)

[0169] Let $C_s :=$ number of cells in sample s, for $s = 1, \ldots, S$. (Unobserved)

[0170] Let $p_{sk} :=$ proportion of cells in sample s from population k, for $s = 1, \ldots, S$, $k = 1, \ldots, K$. (Unobserved)

[0171] Let $q_{km} :=$ number of copies of segment m in each cell from population k, for $k = 1, \ldots, K$, $m = 1, \ldots, M$. (Unobserved for cancer populations, 2 for normal diploid populations)

[0172] Let $t_{ki} := 1$ if population i is an ancestor of population k, otherwise $t_{ki} := 0$ (for $k = 1, \ldots, K$, $i = 1, \ldots, K$). (Unobserved) We take the convention that the set of ancestors of k includes k itself, so $t_{kk} = 1$. Note that this defines a phylogenetic tree with each population as a node.

[0173] Let $m_\ell :=$ index of the segment containing locus ℓ, for $\ell = 1, \ldots, L$ (Unobserved)

[0174] Let us define the normal population to be indexed by $k = 1$.

[0175] We would like to infer the matrices $P = (p_{sk})$, $Q = (q_{km})$, and $T = (t_{ki})$, and the segmentation assignments m_ℓ. Estimating the copy ratios is the first step toward doing that.

[0176] One approach is to construct a generative model, and perform posterior inference given the $n_{s\ell}$'s. However, standard inference techniques such as MCMC and variational Bayes do not seem to work well on this problem, since the space of Q's and T's is large and is complicated by strong dependencies and constraints, making the space difficult to explore.

[0177] The alternative explored here is to incrementally infer what we can from what we know, chipping away at the problem until we arrive at sets of plausible values of P, Q, and T. The plan is then to use these inferred values to construct an importance sampling distribution that will yield fully Bayesian inferences.

[0178] $C_s p_{sk}$ = number of cells in sample s from population k.

[0179] $C_s p_{sk} q_{km_\ell}$ = number of copies of segment m from population k, in sample s.

[0180] Then the true concentration of locus ℓ in sample s is

$$\frac{\sum_{k=1}^{K} C_s p_{sk} q_{km_\ell}}{\sum_{\ell'=1}^{L} \sum_{k=1}^{K} C_s p_{sk} q_{km_{\ell'}}} = \frac{\sum_{k=1}^{K} p_{sk} q_{km_\ell}}{\sum_{\ell'=1}^{L} \sum_{k=1}^{K} p_{sk} q_{km_{\ell'}}}$$

Equation 35

[0181] Dividing this by the true concentration of locus ℓ for a normal sample, namely $1/L$, yields the true copy ratio,

$$r_{s\ell} := \frac{\sum_{k=1}^{K} p_{sk} q_{km_\ell}}{\frac{1}{L} \sum_{\ell'=1}^{L} \sum_{k=1}^{K} p_{sk} q_{km_{\ell'}}}$$

Equation 36

[0182] However, we do not get to observe the true copy ratio directly, since only a fraction

of copies are observed as reads. Note that, by definition, $\frac{1}{L}\sum_{\ell=1}^{L} r_{s\ell} = 1$, so we only need to estimate the $r_{s\ell}$'s up to a multiplicative constant, say ψ_s, such that $r_{s\ell} = \psi_s \sum_{k=1}^{K} p_{sk} q_{km_\ell}$. In particular, we have

$$r_{s\ell} \propto_\ell \sum_{k=1}^{K} p_{sk} q_{km_\ell}$$

Equation 37

where $\propto_\ell$ denotes proportionality with respect to ℓ.

[0183] Let $\gamma_{s\ell} :=$ proportion of copies of locus ℓ that are read (*i.e.*, observed) in sample s.

[0184] Then $n_{s\ell} = \sum_{k=1}^{K} C_s p_{sk} q_{km_\ell} \gamma_{s\ell} =$ number of copies of locus ℓ read in sample s. This is our observed number of reads, from above.

[0185] $\sum_{\ell=1}^{L} n_{s\ell} = \sum_{\ell=1}^{L} \sum_{k=1}^{K} C_s p_{sk} q_{km_\ell} \gamma_{s\ell} =$ total number of reads in sample s, assuming each read contains only one locus.

[0186] So, from our observed data, we can compute the empirical copy ratio,

$$\hat{r}_{s\ell}^{(1)} := \frac{n_{s\ell}}{\frac{1}{L}\sum_{\ell=1}^{L} n_{s\ell}} = \frac{\gamma_{s\ell} \sum_{k=1}^{K} p_{sk} q_{km_\ell}}{\frac{1}{L}\sum_{\ell=1}^{L} \gamma_{s\ell} \sum_{k=1}^{K} p_{sk} q_{km_\ell}} \propto_\ell \gamma_{s\ell} \sum_{k=1}^{K} p_{sk} q_{km_\ell}.$$

Equation 38

[0187] Note that this eliminates C_s, one of our unknowns.

[0188] However, comparing this with the true copy ratio, we see that we are still left with the issue of $\gamma_{s\ell}$ varying across different loci ℓ. As a first step toward mitigating this, we could divide by the empirical copy ratio for a normal sample s', in which everything but the γ's cancel,

$$r_{s\ell}^{(1)} := \frac{n_{s\ell}}{\frac{1}{L}\sum_{\ell'=1}^{L} n_{s\ell'}} = \frac{\gamma_{s\ell}}{\frac{1}{L}\sum_{\ell'=1}^{L} \gamma_{s\ell'}} \propto_{\ell} \gamma_{s\ell}.$$

Equation 39

[0189] However, there is still the issue that $\gamma_{s\ell} \neq \gamma_{s'\ell}$ due to randomness. This noise can be reduced somewhat by averaging across a panel of normal samples $s' \in S_N$,

$$\frac{1}{|S_N|}\sum_{s'\in S_N} r_{s'\ell}^{(1)} = \frac{1}{|S_N|}\sum_{s'\in S_N} \frac{\gamma_{s'\ell}}{\frac{1}{L}\sum_{\ell'=1}^{L}\gamma_{s'\ell'}} \approx \frac{1}{|S_N|}\sum_{s'\in S_N} \psi\gamma_{s'\ell} \approx \psi\mathbb{E}(\gamma_{s'\ell}) \propto_{\ell} \mathbb{E}(\gamma_{s'\ell})$$

Equation 40

[0190] Here, the first approximation is by applying the law of large numbers to the denominators: $\frac{1}{L}\sum_{\ell'=1}^{L}\gamma_{s\ell'} \to 1/\psi$ as $L \to \infty$, where $1/\psi$ is defined to be this limit, and it is assumed that $\gamma_{s\ell} \sim G_{\ell}$ independently for some distributions G_{ℓ}. If $L \gg |S_N|$ then this is approximation is reasonable. The second approximation is by applying the law of large numbers as $|S_N| \to \infty$.

[0191] Thus, we could estimate $r_{s\ell}$ by

$$r_{s\ell}^{(2)} := \frac{r_{s\ell}^{(1)}}{\frac{1}{|S_N|}\sum_{s'\in S_N} r_{s'\ell}^{(1)}} \overset{\propto}{\approx} \frac{\gamma_{s\ell}}{\mathbb{E}(\gamma_{s\ell})} \sum_{k=1}^{K} p_{sk} q_{k\ell} m_{\ell},$$

Equation 41

where $\overset{\propto}{\approx}$ means approximately proportional to, as a function of ℓ.

[0192] Still, there is quite a bit of noise in $\gamma_{s\ell}/\mathbb{E}(\gamma_{s\ell})$. One more step is used to remove some of this noise. Note that if s' is a normal sample, then $r_{s'\ell}^{(2)} \approx \gamma_{s'\ell}/\mathbb{E}(\gamma_{s'\ell})$.

[0193] Thus, the vectors $r_{s'}^{(2)} := (r_{s'1}^{(2)}, \ldots, r_{s'L}^{(2)})^{\mathrm{T}} \in \mathbb{R}^L$, for $s' \in S_N$, represent directions (or, noise profiles) in L-dimensional space that are typical for samples of the random vector $(\gamma_{s1}/\mathbb{E}(\gamma_{s1}), \ldots, \gamma_{sL}/\mathbb{E}(\gamma_{sL}))^{\mathrm{T}} \in \mathbb{R}^L$.

[0194] The idea is to remove any component of our cancer copy ratio profile estimate $r_s^{(2)} = (r_{s1}^{(2)}, \ldots, r_{sL}^{(2)})^{\mathrm{T}}$ that can be explained by the noise profiles. Specifically, what is done is to center the data at zero (subtract off the mean), take the projection of our centered estimate onto the space spanned by the noise profiles, and then subtract this projection from the centered estimate. Another way of viewing this is that we are computing the residuals after regressing the estimate onto the noise profiles.

[0195] The mathematical details are as follows. For notational convenience, let's reindex the samples so that the normal samples are numbered $s' = 1, \ldots, |S_N|$ (and the cancer sample s is not one of these). Let $a_{\ell s'} = r_{s'\ell}^{(2)} - 1$ (*i.e.*, subtract off the sample mean over s'), and form the matrix $A = (a_{\ell s'}) \in \mathbb{R}^{L \times |S_N|}$.

[0196] Likewise, let $y_{s\ell} = r_{s\ell}^{(2)} - 1$, where s is our cancer sample, and denote $y_s = (y_{s1}, \ldots, y_{sL})^{\mathrm{T}}$.

[0197] Let $A^+ := (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$ (*i.e.*, A^+ is the pseudoinverse of A). Then AA^+y is the projection of y onto the column space of A (*i.e.*, onto the span of the mean-centered noise profiles). The residuals, then, are

$$r_s^{(3)} := y_s - AA^+y_s = (I - AA^+)y_s$$

Equation 42

[0198] The vector $r_s^{(3)} = (r_{s1}^{(3)}, \ldots, r_{sL}^{(3)})^T$ is an improved estimate of total copy ratios for cancer sample s. This is repeated for each cancer sample s, using the same matrix A from the panel of normals.

[0199] The total copy ratio estimates $r_{s\ell}^{(3)}$ are then used (along with the alt/ref read counts at heterozygous loci) to obtain haplotype-specific copy ratio estimates. The haplotype-specific copy ratio estimates are, in turn, used to choose segment assignments m_ℓ. This allelic CapSeg method is described further above.

[0200] From the segment assignments m_ℓ and the locus-level total copy ratio estimates $r_{s\ell}^{(3)}$, we can improve our estimates of total copy ratio by combining the estimates within each segment, since the true total copy ratio is constant within each segment (assuming the segmentation is valid). Thus, we can estimate total copy ratio for all loci in segment m as

$$\hat{r}_{sm} \propto \frac{1}{|L_m|} \sum_{\ell \in L_m} r_{s\ell}^{(3)}$$

Equation 43

where $L_m := \{\ell : m_\ell = m\}$ is the set of loci in segment m, and the constant of proportionality is determined by the constraint that $\frac{1}{L} \sum_{\ell=1}^{L} \hat{r}_{sm_\ell} = 1$.

[0201] The standard error of this average (adjusted for multiplying the constant of proportionality), say $\hat{\sigma}_{sm}$, can be used as a rough estimate of the standard deviation of the estimator $\hat{r}_{sm}$.

[0202] Using a probabilistic model, or at least some form of shrinkage, can provide substantial improvement in copy ratio estimation accuracy. In high-dimensional estimation problems like this, empirical averages are often not very accurate (*e.g.*, the classic example is

Stein's paradox). A natural choice would be a Poisson model with sample-specific effects and locus-specific effects, along with some form of noise reduction using the normal panel.

[0203] Also, subtracting off the noise projection doesn't make complete sense to me. Since the noise $\gamma_{sc}/\mathbb{E}(\gamma_{sc})$ is multiplicative rather than additive, it seems like it would make more sense to divide by something rather than subtract something.

[0204] In order to separate the populations, as a next step toward recovering P, Q, and T, we will consider the estimated total copy ratios $\hat{r}_{sm}$ as our input data, along with estimates of the standard deviation $\hat{\sigma}_{sm}$ of the $\hat{r}_{sm}$ estimators.

[0205] Indexed by segment m, the true total copy ratios are $r_{sm} := \psi_s \sum_{k=1}^{K} p_{sk} q_{km}$, where ψ_s is such that $\frac{1}{L}\sum_{\ell=1}^{L} r_{sm_\ell} = 1$.

[0206] Thus, defining $\varepsilon_{sm} := \hat{r}_{sm} - r_{sm}$, we have

$$\hat{r}_{sm} = \psi_s \sum_{k=1}^{K} p_{sk} q_{km} + \varepsilon_{sm}$$

Equation 44

where the error ε_{sm} should be on the order of $\hat{\sigma}_{sm}$. Equivalently, in matrix notation,

$$\hat{R} = \Psi P Q + \varepsilon,$$

Equation 45

letting $\hat{R} := (\hat{r}_{sm})$ and $\Psi := \mathrm{diag}(\psi)$.

[0207] Let $d_{1m} := 2$, and for $k = 2, \ldots, K$, let $d_{km} := q_{km} - q_{pa_k m}$ where pa_k is the parent of population k in the tree T. ($k = 1$ is the normal population.) In other words, d_{km} is the change in segment m's total copy number that occurred when population k arose from its parent population. We refer to d_{km} as the copy delta at population k, for segment m.

[0208] Then $q_{km} = \sum_{i=1}^{K} t_{ki} d_{im}$. Therefore,

$$\sum_{k=1}^{K} p_{sk} q_{km} = \sum_{k=1}^{K} p_{sk} \sum_{i=1}^{K} t_{ki} d_{im} = \sum_{i=1}^{K} d_{im} \sum_{k=1}^{K} p_{sk} t_{ki}.$$

Equation 46

Defining $b_{si} = \sum_{k=1}^{K} p_{sk} t_{ki}$, we therefore have

$$\sum_{k=1}^{K} p_{sk} q_{km} = \sum_{k=1}^{K} b_{sk} d_{km}$$

Equation 47

[0209] Equivalently, in matrix notation, $Q = TD$, where $D = (d_{km})$. Therefore,

$$PQ = PTD = BD$$

Equation 48

where $B := PT$.

[0210] Plugging this into the equation for $\hat{r}_{sm}$ above, we get

$$\hat{r}_{sm} = \sum_{k=1}^{K} \psi_s b_{sk} d_{km} + \varepsilon_{sm}.$$

Equation 49

Or, in matrix form, $\hat{R} = \Psi BD + \varepsilon$.

[0211] The B matrix has a natural interpretation, namely, b_{sk} is the proportion of cells in sample s that are descended from population k (*i.e.*, that belong to a population descended from k). We refer to the b's as descended cell fractions (DCFs).

[0212] For any given segment m, the copy deltas $d_{1m}, \ldots, d_{Km}$ should be approximately independent across the different populations. This is not exactly true, since the copy number cannot be negative, and if it reaches zero then it cannot become positive after that. However,

the idea is that $d_{1m}, \ldots, d_{Km}$ are close enough to independent to be exploited to obtain a reasonable estimate in practice.

[0213] If $d_{1m}, \ldots, d_{Km}$ are independent, and are not Gaussian distributed, then the problem of

recovering ΨB and D given $\hat{R}$ is precisely the independent components analysis (ICA) problem.

[0214] Note that since $b_{s1} = 1$ for all s, we can solve for Ψ and B given the product ΨB. Also, by definition, the first row of D is $d_1 = (2,2, \ldots,2,2)$, so we only need to infer the remaining rows. To simplify notation a bit, let's define $A := \Psi B$.

[0215] For this problem, ICA algorithms do not work well for recovering the full matrices A and D. In part, this is because the inverse transformations (the w's below) are often not orthogonal, as assumed in standard ICA, and in part it is because we have additional information about the distribution of the copy deltas d_{km}.

[0216] Single-unit ICAmay be applied to recover one row of D at a time. An estimate of D then directly yields an estimate of A via a pseudoinverse calculation.

[0217] First we look at how to estimate a single row of D using single-unit ICA. Then we show how to estimate A given an estimate of D. Finally, we describe how to find the remaining rows of D by using residual $\hat{R}$ matrices.

[0218] The first step is to standardize the data matrix $\hat{R}$ to have zero mean and covariance matrix equal to the identity, considering each column as an input vector. Specifically, let $z_m := (\hat{r}_{1m}, \ldots, \hat{r}_{Sm})^T$ be the mth column of $\hat{R}$. Let C be the empirical covariance matrix of the z's, *i.e.*, $C := \frac{1}{M}\sum_{m=1}^{M}(z_m - \bar{z})(z_m - \bar{z})^T$, where $\bar{z} = \frac{1}{M}\sum_{m=1}^{M} z_m$. Factorize C

as $C = UVU^T$, where the columns of U are the eigenvectors of C, and V is the diagonal matrix with the corresponding eigenvalues on the diagonal. Then, letting

$$x_m := UV^{-1/2}U^T(z_m - \bar{z})$$

Equation 50

we have that $x_1, \ldots, x_M \in \mathbb{R}^s$ are linearly transformed versions of the columns of $\hat{R}$ with zero empirical mean and empirical covariance equal to the identity.

[0219] After this standardization step, single-unit ICA searches for a vector $w \in \mathbb{R}^s$ with $\|w\| = 1$ such that the empirical distribution of $w^T x_1, \ldots, w^T x_M \in \mathbb{R}$ is as non-Gaussian as possible. This is done using a rapidly converging approximate Newton's method optimization initialized at a random w. Non-Gaussianity is measured by an approximation to negentropy.

[0220] A linear combination of independent variables is more Gaussian than each of the uncombined variables. As a result, Gaussianity should be minimized when w extracts one of the original rows of D (recall that $d_{1m}, \ldots, d_{Km}$ are assumed to be roughly independent). More precisely, when Gaussianity is minimized, $(w^T x_1, \ldots, w^T x_M)$ should be approximately equal to one of the rows of D, up to a multiplicative factor of undetermined scale and sign.

[0221] To infer the scale σ of the multiplicative factor, we use the fact that the entries of D are integers, mostly zeros or near zero. So in order for $(\sigma w^T x_1, \ldots, \sigma w^T x_M)$ to approximate a row of D, its entries should look like samples from a Gaussian mixture with components centered at the integers, with the mixture weights favoring the components near

0. Thus, we use the likelihood under such a Gaussian mixture model to infer the scale σ and also to score the quality of w.

[0222] The single-unit ICA procedure does not always yield a high-quality w, since there are usually multiple local optima in the optimization. Thus, we generate a number of candidate w's by running the procedure with different randomly chosen initializations.

[0223] The candidate w's are then ranked by the mixture model-based score. The highest ranking w is selected, and an estimate of a row of D is formed by setting $\hat{d} := (\lfloor \sigma w^T x_1 \rceil, \ldots, \lfloor \sigma w^T x_M \rceil)$, where σ is the estimated scale factor and $[y]$ denotes $y \in \mathbb{R}$ rounded off to the nearest integer. (This may be improved by assigning points to components using the mixture model.)

[0224] Given an estimate of the copy delta matrix $\hat{D}$, even if only for a subset of populations, we can obtain an estimate $\hat{A}$ of the columns of ΨB for the corresponding populations.

[0225] This is done by first setting $\hat{A} := R\hat{D}^+$, where $\hat{D}^+$ is the pseudoinverse of $\hat{D}$.

[0226] Since the correct sign for each row of $\hat{D}$ is not determined by ICA, we need to infer the appropriate sign for each column of $\hat{A}$. The entries of A must all be nonnegative. Thus, we can simply multiply each column of $\hat{A}$ by the sign of its average in order to correct the sign for that column. We then force all entries of $\hat{A}$ to be nonnegative by setting any remaining negative values to 0. The same sign flips performed on columns of $\hat{A}$ are also applied to the corresponding rows of $\hat{D}$ to correct its signs as well.

[0227] To estimate the full matrices A and D, we use the procedures described above in a stepwise fashion, adding one population at a time, as follows.

[0228] Initialize with only the normal population: $\hat{A} = (1, \ldots, 1)^T \in \mathbb{R}^{S \times 1}$ and $\hat{D} = (2, \ldots, 2) \in \mathbb{Z}^{1 \times M}$.

[0229] Then, repeat the following steps to add populations:

1. Compute the residual copy ratio matrix $\hat{R}^{res} = \hat{R} - \hat{A}\hat{D}$

2. Apply the single-unit ICA procedure described above to $\hat{R}^{res}$ (instead of $\hat{R}$), to obtain a new estimated row $\hat{d}$.

3. Append $\hat{d}$ to $\hat{D}$, and update $\hat{A}$ using the procedure above. Also correct the signs of $\hat{D}$ as described there.

4. Stop adding populations if the newly added column of $\hat{A}$ has very small entries and the score (or some other test statistic for fit) is not very good. Otherwise, go to step 1.

[0230] Discard any rows of $\hat{D}$ that are low quality (using a criterion based on the score or some other test statistic for fit), and recompute $\hat{A}$ as above.

[0231] Lastly, solve for the unique $\hat{\Psi}$ and $\hat{B}$ such that $\hat{\Psi}\hat{B} = \hat{A}$, $\hat{b}_{s1} = 1$ for all s, and $\hat{\Psi}$ is diagonal.

[0232] Occasionally it happens that at a given step, the newly added column of $\hat{A}$ is nearly identical to an existing column. If this occurs (*e.g.*, if the correlation is greater than 0.995), then discard this column and instead use the next highest scoring candidate $\hat{d}$.

[0233] With a minor adjustment, the same algorithm can be applied to haplotype-specific copy ratio (HSCR) data that is phased within each segment (but not necessarily phased across segments).

[0234] Let q^h_{km} be the copy number of segment m of haplotype h, for a cell in population k (for $h \in \{1,2\}$). Here, it is assumed that within a given segment m, the label h refers to the same haplotype in all populations k, but importantly, the labeling does not have to be consistent from segment to segment.

[0235] Then $r^h_{sm} := \psi_s \sum_{k=1}^{K} p_{sk} q^h_{km}$ is the haplotype-specific copy ratio of haplotype h for sample s and segment m. Here, ψ_s and p_{sk} are the same values as in the case of total copy ratio.

[0236] By definition, $q_{km} = q^1_{km} + q^2_{km}$, and therefore, $r_{sm} = r^1_{sm} + r^2_{sm}$, *i.e.*, the total copy ratio is the sum of the HSCRs.

[0237] An estimate $\hat{r}^h_{sm}$ of the HSCR r^h_{sm} can be obtained from an estimate $\hat{r}_{sm}$ of the total copy ratio along with alt/ref read counts at heterozygous (non-mutated) loci within the segment. Here, we will assume that estimates $\hat{r}^h_{sm}$ are given, which we will treat as input data, along with estimates $\hat{\sigma}^h_{sm}$ of the standard deviations of these estimators.

[0238] The main adjustment that needs to be made from the case of total copy ratio is that for the normal population $k = 1$, the haplotype-specific copy numbers q^h_{1m} are all equal to 1 rather than 2. So, to define the haplotype-specific copy deltas, we let $d^h_{1m} := 1$ and for $k = 2, \ldots, K$, let $d^h_{km} := q^h_{km} - q^h_{\mathrm{pa}_k m}$ where pa_k is the parent of k in the tree T.

[0239] Then, just in the case of total copy ratio, in matrix form we have $Q^h = TD^h$ and thus $R^h = \Psi P Q^h = \Psi B D^h$. So, letting $\varepsilon^h = \hat{R}^h - R^h$, we have

$$\hat{R}^h = \Psi PQ^h + \varepsilon^h = \Psi BD^h + \varepsilon^h.$$

Equation 51

[0240] Therefore, in block matrix notation, letting $\hat{R}^{HS} := [\hat{R}^1\ \hat{R}^2]$, $D^{HS} := [D^1\ D^2]$, and $\varepsilon^{HS} := [\varepsilon^1\ \varepsilon^2]$

$$\hat{R}^{HS} = \Psi BD^{HS} + \varepsilon^{HS}$$

Equation 52

[0241] So, to obtain estimates $\hat{B}$ and $\hat{D}^{HS}$, we can use the same algorithm as in the case of total copy ratio, but with $\hat{R}^{HS}$ as input instead of $\hat{R}$, and with the first row of $\hat{D}^{HS}$ set to 1 instead of 2.

[0242] Another adjustment is to modify the model distribution on haplotype-specific copy deltas to favor somewhat smaller values compared to the model on total copy deltas (since a total copy delta is the sum of the two haplotype-specific copy deltas).

[0243] The estimated haplotype-specific copy delta matrix $\hat{D}^{HS}$ is phased across populations but not across segments (unless the input HSCR matrix $\hat{R}^{HS}$ is phased across segments). In other words, $\hat{d}_{km}$ and $\hat{d}_{k'm}$ both refer to changes in the copy number of segment m for the same haplotype, however, $\hat{d}_{km}$ and $\hat{d}_{km'}$ may refer to changes in the copy number of segment m for different haplotypes.

[0244] The DCF matrix B provides a substantial amount of information about the tree T. In this section, we describe how to find the set of all trees that are consistent with an estimated DCF matrix $\hat{B}$, assuming estimates of the uncertainty in $\hat{B}$ are also given. This is done using a recursive tree construction procedure combined with a sequence of hypothesis tests for the plausibility of parent-children relationships. A threshold for rejecting implausible trees can

be adjusted to make the procedure more permissive (find more trees) or less permissive (find fewer trees).

[0245] Information about T can be recovered from B do to an inequality between a parent's DCF and the DCFs of its children.

[0246] $B = PT$, *i.e.*, $b_{sk} = \sum_{i=1}^{K} p_{si} t_{ik}$. Thus, b_{sk} is the proportion of cells in sample s that are descended from population k (*i.e.*, that belong to a population descended from k).

[0247] Suppose the children of k are $j_1, \ldots, j_N$. Then the sets of cells descended from $j_1, \ldots, j_N$ are disjoint, and all of these cells are descended from k. Thus,

$$b_{sj_1} + \cdots + b_{sj_N} \leq b_{sk}$$

Equation 53

for all $s = 1, \ldots, S$. In other words, the sum of the children is less or equal to their parent, in terms of DCFs. We refer to this as the child-sum inequality.

[0248] The child-sum inequality can also be derived as follows. $t_{ik} = 1$ if population i is a descendant of population k (including k itself), otherwise $t_{ik} = 0$. Thus, $t_{ik} = \mathbb{1}(i = k) + t_{ij_1} + \cdots + t_{ij_N}$, since (a) i is a descendant of k if either $i = k$ or i is a descendant of a child of k, and (b) the descendants of the children do not overlap. (We use $\mathbb{1}$ for the indicator function, *i.e.*, $\mathbb{1}(E)$ is 1 if E is true, and is 0 otherwise.) Hence, plugging this into the formula for b_{sk}, we have

$$b_{sk} = \sum_{i=1}^{K} p_{si}\left(\mathbb{1}(i = k) + \sum_n t_{ij_n}\right) = p_{sk} + \sum_{n=1}^{N}\sum_{i=1}^{K} p_{si} t_{ij_n} = p_{sk} + \sum_{n=1}^{N} b_{sj_n}$$

Equation 54

[0249] Note that in particular, the DCFs are non-increasing as we go down the tree, *i.e.*, if i is a descendant of k, then $b_{si} \leq b_{sk}$ for all s.

[0250] It turns out that the child-sum inequality is a necessary and sufficient condition for a tree to be consistent with B, when P is unknown. Suppose $B \in [0,1]^{S \times K}$ with $b_{s1} = 1$ for all s, and $T \in \{0,1\}^{K \times K}$ is a tree matrix with root 1. Then B and T satisfy the child-sum inequality for all s and k if and only if there exists a matrix $P \in [0,1]^{S \times K}$ such that $B = PT$. $B = PT$ implies the child-sum inequality. Suppose the child-sum inequality is satisfied for all s and k. For each s and k, define $p_{sk} := b_{sk} - \sum_{n=1}^{N} b_{sj_n}$ where $j_1, \ldots, j_N$ are the children of k.

Then $p_{sk} \geq 0$ for all s, k, and $B = PT$.

[0251] Here is described an algorithm that produces the set of trees that are consistent with B, when P is unknown. To convey the basic structure of the algorithm, we first consider the case in which B is known exactly.

[0252] By definition, we know that node 1 is the root of the true tree T, since it represents the normal population.

[0253] Given $J \subseteq \{1, \ldots, K\}$, we say that a tree T' with nodes J is valid if it has root 1 and all of the child-sum inequalities are satisfied with respect to the corresponding columns of B.

[0254] Given a rooted tree T', an extension of T' is a rooted tree that contains all of the nodes of T' and exhibits the same ancestor-descendant relationships among these nodes as T' (but not necessarily the same parent-child relationships).

[0255] The problem of finding all valid trees on $\{1, \ldots, K\}$ reduces to the following subproblem: given a tree $T^{(k-1)}$ with nodes $1, \ldots, k-1$, find all valid extensions to $1, \ldots, k$ (*i.e.*, all valid ways of adding node k to the tree in a way that preserves the existing ancestor-descendant relationships).

[0256] This subproblem can be solved using the following recursive procedure, with input $T^{(k-1)}$. For each $i = 1, \ldots, k-1$, consider making k a child of i. Let J_i be the children of i in $T^{(k-1)}$. For each subset $J \subseteq J_i$, consider making J the children of k (and k a child of i), while preserving all other parent-child relationships.

1. If the child-sum inequality would be violated for k (with J being the children of k) or violated for i (with $(J_i \setminus J) \cup \{k\}$ being the children of i), then reject this extension.

2. Otherwise, accept the extension as $T^{(k)}$. If $k = K$ then we have a complete valid tree, so add $T^{(k)}$ to a running list of complete valid trees. If $k < K$ then recurse to find all valid extensions of $T^{(k)}$ to $1, \ldots, k+1$.

[0257] Several optimizations may be applied to this procedure:

If $b_{sk} > b_{si}$ for some s, then we can reject all extensions in which k is a descendant of i. To use this efficiently, we can visit nodes i of $T^{(k-1)}$ in a depth-first traversal, and if $b_{sk} > b_{si}$ for some s, then skip to the first non-descendant of i in $T^{(k-1)}$.

The search over subsets $J \subseteq J_i$ can be sped up by using the fact that if J is not a valid set of children for k, then no superset $J' \supseteq J$ is valid either.

[0258] This algorithm is guaranteed to generate all valid trees on $\{1, \ldots, K\}$. Any tree with root 1 can be constructed by starting with 1 as root and incrementally adding nodes $k = 2, \ldots, K$ by making k a child of some $i \in \{1, \ldots, k-1\}$ and moving a subset of the current children of this i to instead be children of k, while preserving all other parent-child

relationships. If there is a valid extension of a tree, then that tree is valid. By construction, any tree generated by the algorithm is valid. Consider any valid tree T on $1, \ldots, K$. As above, T can be constructed via a sequence of node insertions as in the algorithm, and each of the intermediate trees $T^{(1)}, \ldots, T^{(K-1)}$ in this sequence is guaranteed to be valid (since T is a valid extension of each of them), so it will not be rejected. Therefore, the algorithm will generate T.

[0259] Referring now to **Fig. 7**, a method of analyzing a sample comprising a plurality of polynucleotides is illustrated. At **701**, the plurality of polynucleotides is sequenced to obtain a plurality of sequences. At **702**, the plurality of sequences is provided to a trained classifier. The trained classifier is adapted to accept a plurality of sequences and output a class label indicative of the presence of a somatic variant within the plurality of sequences. At **703**, a class label is received from the trained classifier indicative of the presence of a somatic variant within the plurality of sequences.

[0260] Referring now to **Fig. 8**, a method of analyzing a sample comprising a plurality of polynucleotides is illustrated. At **801**, at least one prior sequence of a polynucleotide associated with a tumor genome is read. At **802**, a generative model is fit to the at least one prior sequence. At **803**, a plurality of sequences of the plurality of polynucleotides is read. At **804**, the generative model is applied to the plurality of sequences to determine a probability that a first somatic clone is present in the plurality of sequences. At **805**, based on the probability, a label indicative of the presence of the first somatic clone in the sample is determined.

[0261] Referring now to **Figs. 9A-B**, data illustrating robust detection of subclinical lung cancer in a patient according to embodiments of the present disclosure is provided. In **Fig. 10A**, each column represents the comparison of one sample to 36 cfDNA and gDNA WGS samples. In each column, the 2nd and 4th rows are relevant. The 4th (elongated) row depicts the number of alternate reads in samples (x-axis) vs. sites (y-axis). Shade indicates the number of reads supporting the site, from 0 (white) to 5 or more (black). The sites shown in each comparison must have been detected as somatic variants in both the test sample and in the test patient. In each comparison, sites are sorted by the probability that they represent somatic variants (sorted to the top of the y-axis of each plot), vs. sites w. high error rates (sorted to the bottom). The test sample increments by column, ranging from 1 to 36 (the total number of samples sequenced). The first two samples are from a single test patient.

[0262] **Fig. 9B** is a detail view of the inset of **Fig. 9A**. The inset shows the posterior distribution over the number of variant sites that are due to cancer detected in the test sample (second row). Only samples from the test patient show appreciable non-zero probability of cancer detection. Note that data from the test sample is withheld from this calculation if it is from the test patient.

[0263] Referring to **Figs. 10A-E**, each plot shows a 2D scatter plot of the variant-allele fraction (the fraction of total reads reporting the alternate allele at a given locus; VAF). Sample 2 represents a post-relapse high-tumor-shed state, whereas sample 1 was obtained at a timepoint in which no cancer was detectible in the patient, either by imaging, deep DNA sequencing (ddPCR), or other biomarker assays. Both samples are plasma-derived cfDNA.

[0264] Each column corresponds to classification of variants as inherited (het, AA, or BB), or potentially somatic (outlier). The variants labeled as AA may also comprise error-prone sites identified using a model trained on 30 other cfDNA WGS samples. The last column shows all variants together will most-likely class indicated by color. The presence of a large number

of somatic mutations (blue points > 0) in all both samples demonstrates the robust detection somatic clones. These variants are heavily enriched in smoking-associated mutations (not shown).

[0265] Referring now to **Figs. 11A-C**, data illustrating robust detection of breast cancer in a patient according to embodiments of the present disclosure is provided. In **Fig. 11A**, each column represents the comparison of one sample to 36 cfDNA and gDNA WGS samples. In each column, the 2nd and 4th rows are relevant. The 4th (elongated) row depicts the number of alternate reads in samples (x-axis) vs. sites (y-axis). Shade indicates the number of reads supporting the site, from 0 (white) to 5 or more (black). The sites shown in each comparison must have been detected as somatic variants in both the test sample and in the test patient. In each comparison, sites are sorted by the probability that they represent somatic variants (sorted to the top of the y-axis of each plot), vs. sites w. high error rates (sorted to the bottom). The test sample increments by column, ranging from 1 to 36 (the total number of samples sequenced). The first three samples are from a single test patient.

[0266] **Figs. 11B-C** are detail views of **Fig. 11A**, with **Fig. 11B** covering columns 1-3 and **Fig. 11C** covering columns 4-7. The inset shows the posterior distribution over the number of variant sites that are due to cancer detected in the test sample (2nd row). Only samples from the test patient show appreciable non-zero probability of cancer detection. Note that data from the test sample is withheld from this calculation if it is from the test patient.

[0267] Here, the three samples from the test patient represent cfDNA samples from plasma and cerebrospinal fluid of a patient with brain metastasis and leptomeningeal disease from breast cancer.

[0268] Referring to Fig. **12A-O**, each plot shows a 2D scatter plot of the variant-allele fraction (the fraction of total reads reporting the alternate allele at a given locus; VAF). Each row (**Figs. 12A-E**; **F-J**; and **K-O**) corresponds to a comparison between two samples

sequenced from a melanoma test patient. Sample SM-61EX7 is from tissue obtained from surgical resection. The other two samples are from plasma-derived cfDNA.

[0269] Each column corresponds to classification of variants as inherited (het, AA, or BB), or potentially somatic (outlier). The last column (**Figs. 12E, J, O**) shows all variants together will most-likely class indicated by color. The variants labeled as AA may also comprise error-prone sites identified using a model trained on 30 other cfDNA WGS samples. The presence of a large number of somatic mutations (blue points > 0) in all samples demonstrates the robust detection somatic clones. These variants are heavily enriched in UV-associated mutations (not shown).

[0270] Referring to **Figs. 13A-AD**, each plot shows a 2D scatter plot of the variant-allele fraction (the fraction of total reads reporting the alternate allele at a given locus; VAF). Each row (**Figs. 13A-E**; **F-J**; **K-O**; **P-T**, **U-Y**; **Z-AD**) corresponds to a comparison between two samples sequenced from a melanoma test patient. Sample SM-9ZE13 is from tissue obtained from surgical resection. The other two samples are from plasma-derived cfDNA.

[0271] Each column corresponds to classification of variants as inherited (het, AA, or BB), or potentially somatic (outlier). The variants labeled as AA may also comprise error-prone sites identified using a model trained on 30 other cfDNA WGS samples. The last column (**Figs. 13E, J, O, T, Y, AD**) shows all variants together will most-likely class indicated by color. The presence of a large number of somatic mutations (blue points > 0) in all samples demonstrates the robust detection somatic clones. These variants are heavily enriched in UV-associated mutations (not shown).

[0272] In various embodiments, fingerprint mutation detection in a subclinical cancer sample is distinguished from sequencing error using mutational signatures and phylogenetic analysis.

[0273] In various embodiments, the mutational signature model is extended beyond 96 single-base-substitution classes to include insertions and deletions, as well as di-nucleotide and multi-nucleotide frequencies associated with mutational processes active in human cancers.

[0274] In various embodiments, a statistical model is constructed to estimate the sequencing error rate at any location in the genome, based on local sequence composition and additional covariates. This model also utilizes ground-truth sequencing errors identified by overlapping paired-end reads.

[0275] Various embodiments do not rely on a germline gDNA reference sample for a given patient – the multi-sample phylogenetic classification of variants as germline vs. somatic based on VAF combined with the mutational signatures classifier based on sequence context obviates the need for such a reference sample. This is advantageous, as such reference samples are often difficult to obtain, and even when they are available, may be misleading as cancer-derived cfDNA may contaminate the gDNA.

[0276] In various embodiments, the strategy of paired whole-genome sequencing is applied where DNA sequencing libraries are constructed using dual-index barcodes that uniquely identify DNA-fragments in the input sample. Detection of cancer-fingerprint mutations in the whole-genome sequencing of a test sample may subsequently be followed-up by ultradeep sequencing of the same library, with targeted hybrid-capture of genomic regions harboring detected fingerprint mutations. This provides another level of validation that the detected fingerprint mutation was not due to sequencing error, and can substantially increase the sensitivity of detection, while not compromising specificity.

[0277] **Fig. 14A** shows the expected number of fingerprint somatic single-nucleotide substitutions (SSNVs) detected in a theoretical MRD liquid biopsy cfDNA sample, as a function of cancer fraction. The total number of tracking fingerprint substitutions was

100,000. 3.6 ng input DNA was modeled, resulting in 85x unique coverage, with base-level WGS sequencing error rate was assumed to be 1/100. Detected tracking mutations are designated as true-positive (TPs) if they were derived from cancer in the simulation, or false-positives (FPs) if they were detected due to sequencing error.

[0278] Fig. 14B shows detection power as a function of cancer fraction using the paired WGS approach, or with additional validation sequencing of all detected mutations using deep targeted UMI sequencing in the original library. Power was calculated using a binomial model and first considering the scenario where all detected fingerprint mutations were false positives in the test WGS sample. In this case, the number of these errors that also produce errors in the followup UMI validation was modeled using a binomial distribution with N equal to the expected number of false-positives and p equal to the deep UMI-sequencing error rate modeled as 1e-4, which accounts for both PCR errors in the 1st cycle of library construction and read-mapping error. The quantile k of this distribution was computed corresponding to 1 – the FDR rate – here 1e-4. The number of correctly validated mutations was modeled as a binomial distribution with N equal to the number of expected true positives detected in the paired WGS, and p equal to 1 – the UMI-error rate. Power was computed as the probability of detecting at least k fingerprint mutations, which would cause rejection of the null-hypothesis that all deep-UMI-validated mutations were due to errors at the desired FDR.

[0279] Referring now to **Fig. 15**, a schematic of an example of a computing node is shown. Computing node **10** is only one example of a suitable computing node and is not intended to suggest any limitation as to the scope of use or functionality of embodiments described

herein. Regardless, computing node **10** is capable of being implemented and/or performing any of the functionality set forth hereinabove.

[0280] In computing node **10** there is a computer system/server **12**, which is operational with numerous other general purpose or special purpose computing system environments or configurations. Examples of well-known computing systems, environments, and/or configurations that may be suitable for use with computer system/server **12** include, but are not limited to, personal computer systems, server computer systems, thin clients, thick clients, handheld or laptop devices, multiprocessor systems, microprocessor-based systems, set top boxes, programmable consumer electronics, network PCs, minicomputer systems, mainframe computer systems, and distributed cloud computing environments that include any of the above systems or devices, and the like.

[0281] Computer system/server **12** may be described in the general context of computer system-executable instructions, such as program modules, being executed by a computer system. Generally, program modules may include routines, programs, objects, components, logic, data structures, and so on that perform particular tasks or implement particular abstract data types. Computer system/server **12** may be practiced in distributed cloud computing environments where tasks are performed by remote processing devices that are linked through a communications network. In a distributed cloud computing environment, program modules may be located in both local and remote computer system storage media including memory storage devices.

[0282] As shown in **Fig. 15**, computer system/server **12** in computing node **10** is shown in the form of a general-purpose computing device. The components of computer system/server **12** may include, but are not limited to, one or more processors or processing units **16**, a system memory **28**, and a bus **18** that couples various system components including system memory **28** to processor **16**.

[0283] Bus **18** represents one or more of any of several types of bus structures, including a memory bus or memory controller, a peripheral bus, an accelerated graphics port, and a processor or local bus using any of a variety of bus architectures. By way of example, and not limitation, such architectures include Industry Standard Architecture (ISA) bus, Micro Channel Architecture (MCA) bus, Enhanced ISA (EISA) bus, Video Electronics Standards Association (VESA) local bus, Peripheral Component Interconnect (PCI) bus, Peripheral Component Interconnect Express (PCIe), and Advanced Microcontroller Bus Architecture (AMBA).

[0284] Computer system/server **12** typically includes a variety of computer system readable media. Such media may be any available media that is accessible by computer system/server **12**, and it includes both volatile and non-volatile media, removable and non-removable media.

[0285] System memory **28** can include computer system readable media in the form of volatile memory, such as random access memory (RAM) **30** and/or cache memory **32**. Computer system/server **12** may further include other removable/non-removable, volatile/non-volatile computer system storage media. By way of example only, storage system **34** can be provided for reading from and writing to a non-removable, non-volatile magnetic media (not shown and typically called a "hard drive"). Although not shown, a magnetic disk drive for reading from and writing to a removable, non-volatile magnetic disk (*e.g.*, a "floppy disk"), and an optical disk drive for reading from or writing to a removable, non-volatile optical disk such as a CD-ROM, DVD-ROM or other optical media can be provided. In such instances, each can be connected to bus **18** by one or more data media interfaces. As will be further depicted and described below, memory **28** may include at least one program product having a set (*e.g.*, at least one) of program modules that are configured to carry out the functions of embodiments of the disclosure.

[0286] Program/utility **40**, having a set (at least one) of program modules **42**, may be stored in memory **28** by way of example, and not limitation, as well as an operating system, one or more application programs, other program modules, and program data. Each of the operating system, one or more application programs, other program modules, and program data or some combination thereof, may include an implementation of a networking environment. Program modules **42** generally carry out the functions and/or methodologies of embodiments as described herein.

[0287] Computer system/server **12** may also communicate with one or more external devices **14** such as a keyboard, a pointing device, a display **24**, etc.; one or more devices that enable a user to interact with computer system/server **12**; and/or any devices (*e.g.*, network card, modem, etc.) that enable computer system/server **12** to communicate with one or more other computing devices. Such communication can occur via Input/Output (I/O) interfaces **22**. Still yet, computer system/server **12** can communicate with one or more networks such as a local area network (LAN), a general wide area network (WAN), and/or a public network (*e.g.*, the Internet) via network adapter **20**. As depicted, network adapter **20** communicates with the other components of computer system/server **12** via bus **18**. It should be understood that although not shown, other hardware and/or software components could be used in conjunction with computer system/server **12**. Examples, include, but are not limited to: microcode, device drivers, redundant processing units, external disk drive arrays, RAID systems, tape drives, and data archival storage systems, etc.

[0288] The present disclosure may be embodied as a system, a method, and/or a computer program product. The computer program product may include a computer readable storage medium (or media) having computer readable program instructions thereon for causing a processor to carry out aspects of the present disclosure.

[0289] The computer readable storage medium can be a tangible device that can retain and store instructions for use by an instruction execution device. The computer readable storage medium may be, for example, but is not limited to, an electronic storage device, a magnetic storage device, an optical storage device, an electromagnetic storage device, a semiconductor storage device, or any suitable combination of the foregoing. A non-exhaustive list of more specific examples of the computer readable storage medium includes the following: a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), a static random access memory (SRAM), a portable compact disc read-only memory (CD-ROM), a digital versatile disk (DVD), a memory stick, a floppy disk, a mechanically encoded device such as punch-cards or raised structures in a groove having instructions recorded thereon, and any suitable combination of the foregoing. A computer readable storage medium, as used herein, is not to be construed as being transitory signals per se, such as radio waves or other freely propagating electromagnetic waves, electromagnetic waves propagating through a waveguide or other transmission media (*e.g.*, light pulses passing through a fiber-optic cable), or electrical signals transmitted through a wire.

[0290] Computer readable program instructions described herein can be downloaded to respective computing/processing devices from a computer readable storage medium or to an external computer or external storage device via a network, for example, the Internet, a local area network, a wide area network and/or a wireless network. The network may comprise copper transmission cables, optical transmission fibers, wireless transmission, routers, firewalls, switches, gateway computers and/or edge servers. A network adapter card or network interface in each computing/processing device receives computer readable program instructions from the network and forwards the computer readable program instructions for

storage in a computer readable storage medium within the respective computing/processing device.

[0291] Computer readable program instructions for carrying out operations of the present disclosure may be assembler instructions, instruction-set-architecture (ISA) instructions, machine instructions, machine dependent instructions, microcode, firmware instructions, state-setting data, or either source code or object code written in any combination of one or more programming languages, including an object oriented programming language such as Smalltalk, C++ or the like, and conventional procedural programming languages, such as the "C" programming language or similar programming languages. The computer readable program instructions may execute entirely on the user's computer, partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider). In some embodiments, electronic circuitry including, for example, programmable logic circuitry, field-programmable gate arrays (FPGA), or programmable logic arrays (PLA) may execute the computer readable program instructions by utilizing state information of the computer readable program instructions to personalize the electronic circuitry, in order to perform aspects of the present disclosure.

[0292] Aspects of the present disclosure are described herein with reference to flowchart illustrations and/or block diagrams of methods, apparatus (systems), and computer program products according to embodiments of the disclosure. It will be understood that each block of the flowchart illustrations and/or block diagrams, and combinations of blocks in the flowchart

illustrations and/or block diagrams, can be implemented by computer readable program instructions.

[0293] These computer readable program instructions may be provided to a processor of a general purpose computer, special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the computer or other programmable data processing apparatus, create means for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks. These computer readable program instructions may also be stored in a computer readable storage medium that can direct a computer, a programmable data processing apparatus, and/or other devices to function in a particular manner, such that the computer readable storage medium having instructions stored therein comprises an article of manufacture including instructions which implement aspects of the function/act specified in the flowchart and/or block diagram block or blocks.

[0294] The computer readable program instructions may also be loaded onto a computer, other programmable data processing apparatus, or other device to cause a series of operational steps to be performed on the computer, other programmable apparatus or other device to produce a computer implemented process, such that the instructions which execute on the computer, other programmable apparatus, or other device implement the functions/acts specified in the flowchart and/or block diagram block or blocks.

[0295] The flowchart and block diagrams in the Figures illustrate the architecture, functionality, and operation of possible implementations of systems, methods, and computer program products according to various embodiments of the present disclosure. In this regard, each block in the flowchart or block diagrams may represent a module, segment, or portion of instructions, which comprises one or more executable instructions for implementing the specified logical function(s). In some alternative implementations, the functions noted in the

block may occur out of the order noted in the figures. For example, two blocks shown in succession may, in fact, be executed substantially concurrently, or the blocks may sometimes be executed in the reverse order, depending upon the functionality involved. It will also be noted that each block of the block diagrams and/or flowchart illustration, and combinations of blocks in the block diagrams and/or flowchart illustration, can be implemented by special purpose hardware-based systems that perform the specified functions or acts or carry out combinations of special purpose hardware and computer instructions.

[0296] The descriptions of the various embodiments of the present disclosure have been presented for purposes of illustration, but are not intended to be exhaustive or limited to the embodiments disclosed. Many modifications and variations will be apparent to those of ordinary skill in the art without departing from the scope and spirit of the described embodiments. The terminology used herein was chosen to best explain the principles of the embodiments, the practical application or technical improvement over technologies found in the marketplace, or to enable others of ordinary skill in the art to understand the embodiments disclosed herein.

CLAIMS

What is claimed is:

1. A method of analyzing a sample comprising a plurality of polynucleotides, the method comprising:

receiving a plurality of sequences of the plurality of polynucleotides;

providing the plurality of sequences to a trained classifier, the trained classifier adapted to accept a plurality of sequences and output a class label indicative of the presence of a somatic variant within the plurality of sequences;

receiving from the trained classifier a class label indicative of the presence of a somatic clone within the plurality of sequences.

2. The method of claim 1, further comprising:

based on the label, indicating whether a cancer clone is present in the sample.

3. The method of claim 1, further comprising:

based on the label, indicating whether a clonal expansion is present in the sample.

4. The method of claim 1, wherein the label has an associated probability.

5. The method of claim 1, further comprising:

providing clinical data to the trained classifier, the clinical data being related to an originator of the sample.

6. The method of claim 1, wherein the clinical data comprises an indication of smoking by the originator of the sample.

7. The method of claim 1, wherein the clinical data comprises family history of the originator of the sample.

8. The method of claim 1, further comprising:

providing population data to the trained classifier, the population data being related to an originator of the sample.

9. The method of claim 1, wherein the plurality of sequences comprise a plurality of somatic variants.

10. The method of claim 1, wherein the trained classifier comprises an artificial neural network.

11. The method of claim 1, wherein the trained classifier comprises a regression model.

12. The method of claim 1, wherein the trained classifier comprises a random decision forest.

13. The method of claim 1, wherein the trained classifier comprises an SVM.

14. The method of claim 10, wherein the artificial neural network is a convolutional neural network.

15. The method of claim 10, wherein the artificial neural network is a recurrent neural network.

16. The method of claim 1, wherein the sample comprises blood.

17. The method of claim 1, wherein the sample comprises cerebrospinal fluid.

18. The method of claim 1, wherein the plurality of polynucleotides comprises DNA.

19. The method of claim 1, wherein the plurality of polynucleotides comprises methylated DNA.

20. The method of claim 1, further comprising:

providing fragment lengths of the plurality of sequences to the trained classifier.

21. The method of claim 1, wherein the plurality of polynucleotides comprises RNA.

22. The method of claim 1, wherein the sequencing is at a depth of 100x or less.

23. The method of claim 1, wherein the sequencing is at a depth of 85x or less.

24. The method of claim 1, wherein the sequencing is at a depth of about 20x to about 85x.

25. The method of claim 1, wherein the sequencing is at a depth of about 20x to about 100x.

26. The method of claim 1, further comprising sequencing the plurality of polynucleotides to obtain the plurality of sequences.

27. A system comprising:

a computing node comprising a computer readable storage medium having program instructions embodied therewith, the program instructions executable by a processor of the computing node to cause the processor to perform a method comprising:

sequencing the plurality of polynucleotides to obtain a plurality of sequences;

providing the plurality of sequences to a trained classifier, the trained classifier adapted to accept a plurality of sequences and output a class label indicative of the presence of a somatic variant within the plurality of sequences;

receiving from the trained classifier a class label indicative of the presence of a somatic variant within the plurality of sequences.

28. The system of claim 27, the method further comprising:

based on the label, indicating whether a cancer clone is present in the sample.

29. The system of claim 27, the method further comprising:

based on the label, indicating whether a clonal expansion is present in the sample.

30. The system of claim 27, wherein the label has an associated probability.

31. The system of claim 27, the method further comprising:

providing clinical data to the trained classifier, the clinical data being related to an originator of the sample.

32. The system of claim 27, wherein the clinical data comprises an indication of smoking by the originator of the sample.

33. The system of claim 27, wherein the clinical data comprises family history of the originator of the sample.

34. The system of claim 27, the method further comprising:

providing population data to the trained classifier, the population data being related to an originator of the sample.

35. The system of claim 27, wherein the plurality of sequences comprise a plurality of somatic variants.

36. The system of claim 27, wherein the trained classifier comprises an artificial neural network.

37. The system of claim 27, wherein the trained classifier comprises a regression model.

38. The system of claim 27, wherein the trained classifier comprises a random decision forest.

39. The system of claim 27, wherein the trained classifier comprises an SVM.

40. The system of claim 36, wherein the artificial neural network is a convolutional neural network.

41. The system of claim 36, wherein the artificial neural network is a recurrent neural network.

42. The system of claim 27, wherein the sample comprises blood.

43. The system of claim 27, wherein the sample comprises cerebrospinal fluid.

44. The system of claim 27, wherein the plurality of polynucleotides comprises DNA.

45. The system of claim 27, wherein the plurality of polynucleotides comprises methylated DNA.

46. The system of claim 27, the method further comprising:

providing fragment lengths of the plurality of sequences to the trained classifier.

47. The system of claim 27, wherein the plurality of polynucleotides comprises RNA.

48. The system of claim 27, wherein the sequencing is at a depth of 100x or less.

49. The system of claim 27, wherein the sequencing is at a depth of 85x or less.

50. The system of claim 27, wherein the sequencing is at a depth of about 20x to about 85x.

51. The system of claim 27, wherein the sequencing is at a depth of about 20x to about 100x.

52. The system of claim 27, further comprising sequencing the plurality of polynucleotides to obtain the plurality of sequences.

53. A computer program product for analyzing a sample comprising a plurality of polynucleotides, the computer program product comprising a computer readable storage medium having program instructions embodied therewith, the program instructions executable by a processor to cause the processor to perform a method comprising:

sequencing the plurality of polynucleotides to obtain a plurality of sequences;

providing the plurality of sequences to a trained classifier, the trained classifier adapted to accept a plurality of sequences and output a class label indicative of the presence of a somatic variant within the plurality of sequences;

receiving from the trained classifier a class label indicative of the presence of a somatic variant within the plurality of sequences.

54. The computer program product of claim 53, the method further comprising:

based on the label, indicating whether a cancer clone is present in the sample.

55. The computer program product of claim 53, the method further comprising:

based on the label, indicating whether a clonal expansion is present in the sample.

56. The computer program product of claim 53, wherein the label has an associated probability.

57. The computer program product of claim 53, the method further comprising:

providing clinical data to the trained classifier, the clinical data being related to an originator of the sample.

58. The computer program product of claim 53, wherein the clinical data comprises an indication of smoking by the originator of the sample.

59. The computer program product of claim 53, wherein the clinical data comprises family history of the originator of the sample.

60. The computer program product of claim 53, the method further comprising:

providing population data to the trained classifier, the population data being related to an originator of the sample.

61. The computer program product of claim 53, wherein the plurality of sequences comprise a plurality of somatic variants.

62. The computer program product of claim 53, wherein the trained classifier comprises an artificial neural network.

63. The computer program product of claim 53, wherein the trained classifier comprises a regression model.

64. The computer program product of claim 53, wherein the trained classifier comprises a random decision forest.

65. The computer program product of claim 53, wherein the trained classifier comprises an SVM.

66. The computer program product of claim 62, wherein the artificial neural network is a convolutional neural network.

67. The computer program product of claim 62, wherein the artificial neural network is a recurrent neural network.

68. The computer program product of claim 53, wherein the sample comprises blood.

69. The computer program product of claim 53, wherein the sample comprises cerebrospinal fluid.

70. The computer program product of claim 53, wherein the plurality of polynucleotides comprises DNA.

71. The computer program product of claim 53, wherein the plurality of polynucleotides comprises methylated DNA.

72. The computer program product of claim 53, the method further comprising:

providing fragment lengths of the plurality of sequences to the trained classifier.

73. The computer program product of claim 53, wherein the plurality of polynucleotides comprises RNA.

74. The computer program product of claim 53, wherein the sequencing is at a depth of 100x or less.

75. The computer program product of claim 53, wherein the sequencing is at a depth of 85x or less.

76. The computer program product of claim 53, wherein the sequencing is at a depth of about 20x to about 85x.

77. The computer program product of claim 53, wherein the sequencing is at a depth of about 20x to about 100x.

78. The computer program product of claim 53, further comprising sequencing the plurality of polynucleotides to obtain the plurality of sequences.

79. A method of analyzing a sample comprising a plurality of polynucleotides, the method comprising:

reading at least one prior sequence of a polynucleotide associated with a tumor genome;

fitting a generative model to the at least one prior sequence;

receiving a plurality of sequences of the plurality of polynucleotides;

applying the generative model to the plurality of sequences to determine a probability that a first somatic clone is present in the plurality of sequences;

based on the probability, determining a label indicative of the presence of the first somatic clone in the sample.

80. The method of claim 79, wherein the label has an associated probability.

81. The method of claim 79, wherein the generative model comprises a linear-Gaussian model.

82. The method of claim 79, wherein the generative model comprises a linear-negative binomial model.

83. The method of claim 79, wherein the generative model comprises a latent factor model.

84. The method of claim 79, wherein the generative model comprises a factor analysis model.

85. The method of claim 79, further comprising

inferring a phylogenetic tree from the at least one prior sequence.

86. The method of claim 79, further comprising:

updating the generative model based on the plurality of sequences.

87. The method of claim 86, wherein the at least one prior sequence comprises a second somatic clone, the method further comprising:

based on the updated generative model, determining a probability that the sample contains the second somatic clone.

88. The method of claim 86, wherein the at least one prior sequence comprises a second somatic clone, the method further comprising:

based on the updated generative model, determining a probability that the sample contains a descendent of the second somatic clone.

89. The method of claim 86, wherein the at least one prior sequence comprises a second somatic clone, the method further comprising:

based on the updated generative model, determining a probability that the sample contains a third somatic clone related to the second somatic clone within the phylogenetic tree.

90. The method of claim 86, further comprising:

based on the updated generative model, determining a probability that the sample shares at least one somatic mutation with the at least one prior sequence.

91. The method of claim 86, further comprising:

based on the updated generative model, determining a probability that the sample shares at least one clonal expansion with the at least one prior sequence.

92. The method of claim 79, further comprising:

based on the label, indicating whether a cancer clone is present in the sample.

93. The method of claim 79, further comprising:

based on the label, indicating whether a clonal expansion is present in the sample.

94. The method of claim 79, wherein the label has an associated probability.

95. The method of claim 79, wherein the plurality of sequences comprise a plurality of somatic variants.

96. The method of claim 79, wherein the sample comprises blood.

97. The method of claim 79, wherein the sample comprises cerebrospinal fluid.

98. The method of claim 79, wherein the plurality of polynucleotides comprises DNA.

99. The method of claim 79, wherein the plurality of polynucleotides comprises methylated DNA.

100. The method of claim 79, wherein the plurality of polynucleotides comprises RNA.

101. The method of claim 79, wherein the sequencing is at a depth of 100x or less.

102. The method of claim 79, wherein the sequencing is at a depth of 85x or less.

103. The method of claim 79, wherein the sequencing is at a depth of about 20x to about 85x.

104. The method of claim 79, wherein the sequencing is at a depth of about 20x to about 100x.

105. The method of claim 79, further comprising sequencing the plurality of polynucleotides to obtain the plurality of sequences.

106. A system comprising:

a computing node comprising a computer readable storage medium having program instructions embodied therewith, the program instructions executable by a processor of the computing node to cause the processor to perform a method comprising:

reading at least one prior sequence of a polynucleotide associated with a tumor genome;

fitting a generative model to the at least one prior sequence;

receiving a plurality of sequences of the plurality of polynucleotides;

applying the generative model to the plurality of sequences to determine a probability that a first somatic clone is present in the plurality of sequences;

based on the probability, determining a label indicative of the presence of the first somatic clone in the sample.

107. The system of claim 106, wherein the label has an associated probability.

108. The system of claim 106, wherein the generative model comprises a linear-Gaussian model.

109. The system of claim 106, wherein the generative model comprises a linear-negative binomial model.

110. The system of claim 106, wherein the generative model comprises a latent factor model.

111. The system of claim 106, wherein the generative model comprises a factor analysis model.

112. The system of claim 106, the method further comprising

inferring a phylogenetic tree from the at least one prior sequence.

113. The system of claim 106, the method further comprising:

updating the generative model based on the plurality of sequences.

114. The system of claim 113, wherein the at least one prior sequence comprises a second somatic clone, the method further comprising:

based on the updated generative model, determining a probability that the sample contains the second somatic clone.

115. The system of claim 113, wherein the at least one prior sequence comprises a second somatic clone, the method further comprising:

based on the updated generative model, determining a probability that the sample contains a descendent of the second somatic clone.

116. The system of claim 113, wherein the at least one prior sequence comprises a second somatic clone, the method further comprising:

based on the updated generative model, determining a probability that the sample contains a third somatic clone related to the second somatic clone within the phylogenetic tree.

117. The system of claim 113, further comprising:

based on the updated generative model, determining a probability that the sample shares at least one somatic mutation with the at least one prior sequence.

118. The system of claim 113, further comprising:

based on the updated generative model, determining a probability that the sample shares at least one clonal expansion with the at least one prior sequence.

119. The system of claim 106, the method further comprising:

based on the label, indicating whether a cancer clone is present in the sample.

120. The system of claim 106, the method further comprising:

based on the label, indicating whether a clonal expansion is present in the sample.

121. The system of claim 106, wherein the label has an associated probability.

122. The system of claim 106, wherein the plurality of sequences comprise a plurality of somatic variants.

123. The system of claim 106, wherein the sample comprises blood.

124. The system of claim 106, wherein the sample comprises cerebrospinal fluid.

125. The system of claim 106, wherein the plurality of polynucleotides comprises DNA.

126. The system of claim 106, wherein the plurality of polynucleotides comprises methylated DNA.

127. The system of claim 106, wherein the plurality of polynucleotides comprises RNA.

128. The system of claim 106, wherein the sequencing is at a depth of 100x or less.

129. The system of claim 106, wherein the sequencing is at a depth of 85x or less.

130. The system of claim 106, wherein the sequencing is at a depth of about 20x to about 85x.

131. The system of claim 106, wherein the sequencing is at a depth of about 20x to about 100x.

132. The system of claim 106, further comprising sequencing the plurality of polynucleotides to obtain the plurality of sequences.

133. A computer program product for analyzing a sample comprising a plurality of polynucleotides, the computer program product comprising a computer readable storage medium having program instructions embodied therewith, the program instructions executable by a processor to cause the processor to perform a method comprising:

reading at least one prior sequence of a polynucleotide associated with a tumor genome;

fitting a generative model to the at least one prior sequence;

receiving a plurality of sequences of the plurality of polynucleotides;

applying the generative model to the plurality of sequences to determine a probability that a first somatic clone is present in the plurality of sequences;

based on the probability, determining a label indicative of the presence of the first somatic clone in the sample.

134. The computer program product of claim 133, wherein the label has an associated probability.

135. The computer program product of claim 133, wherein the generative model comprises a linear-Gaussian model.

136. The computer program product of claim 133, wherein the generative model comprises a linear-negative binomial model.

137. The computer program product of claim 133, wherein the generative model comprises a latent factor model.

138. The computer program product of claim 133, wherein the generative model comprises a factor analysis model.

139. The computer program product of claim 133, the method further comprising

inferring a phylogenetic tree from the at least one prior sequence.

140. The computer program product of claim 133, the method further comprising:

updating the generative model based on the plurality of sequences.

141. The computer program product of claim 140, wherein the at least one prior sequence comprises a second somatic clone, the method further comprising:

based on the updated generative model, determining a probability that the sample contains the second somatic clone.

142. The computer program product of claim 140, wherein the at least one prior sequence comprises a second somatic clone, the method further comprising:

based on the updated generative model, determining a probability that the sample contains a descendent of the second somatic clone.

143. The computer program product of claim 140, wherein the at least one prior sequence comprises a second somatic clone, the method further comprising:

based on the updated generative model, determining a probability that the sample contains a third somatic clone related to the second somatic clone within the phylogenetic tree.

144. The computer program product of claim 140, further comprising:

based on the updated generative model, determining a probability that the sample shares at least one somatic mutation with the at least one prior sequence.

145. The computer program product of claim 140, further comprising:

based on the updated generative model, determining a probability that the sample shares at least one clonal expansion with the at least one prior sequence.

146. The computer program product of claim 133, the method further comprising:

based on the label, indicating whether a cancer clone is present in the sample.

147. The computer program product of claim 133, the method further comprising:

based on the label, indicating whether a clonal expansion is present in the sample.

148. The computer program product of claim 133, wherein the label has an associated probability.

149. The computer program product of claim 133, wherein the plurality of sequences comprise a plurality of somatic variants.

150. The computer program product of claim 133, wherein the sample comprises blood.

151. The computer program product of claim 133, wherein the sample comprises cerebrospinal fluid.

152. The computer program product of claim 133, wherein the plurality of polynucleotides comprises DNA.

153. The computer program product of claim 133, wherein the plurality of polynucleotides comprises methylated DNA.

154. The computer program product of claim 133, wherein the plurality of polynucleotides comprises RNA.

155. The computer program product of claim 133, wherein the sequencing is at a depth of 100x or less.

156. The computer program product of claim 133, wherein the sequencing is at a depth of 85x or less.

157. The computer program product of claim 133, wherein the sequencing is at a depth of about 20x to about 85x.

158. The computer program product of claim 133, wherein the sequencing is at a depth of about 20x to about 100x.

159. The computer program product of claim 133, further comprising sequencing the plurality of polynucleotides to obtain the plurality of sequences.

160. A method of analyzing a sample comprising a plurality of polynucleotides, the method comprising:

receiving a plurality of sequences of the plurality of polynucleotides;

identifying one or more inherited variant and one or more somatic variant among the plurality of sequences;

providing the one or more inherited variant to a first trained classifier;

providing the one or more somatic variant to a second trained classifier;

determining by the first and second trained classifier the presence of aneuploidy in the plurality of polynucleotides.

161. The method of claim 160, wherein the plurality of sequences comprise a plurality of somatic variants.

162. The method of claim 160, wherein the first or second trained classifier comprises an artificial neural network.

163. The method of claim 160, wherein the first or second trained classifier comprises a regression model.

164. The method of claim 160, wherein the first or second trained classifier comprises a random decision forest.

165. The method of claim 160, wherein the first or second trained classifier comprises an SVM.

166. The method of claim 162, wherein the artificial neural network is a convolutional neural network.

167. The method of claim 162, wherein the artificial neural network is a recurrent neural network.

168. The method of claim 160, wherein the sample comprises blood.

169. The method of claim 160, wherein the sample comprises cerebrospinal fluid.

170. The method of claim 160, wherein the plurality of polynucleotides comprises DNA.

171. The method of claim 160, wherein the plurality of polynucleotides comprises methylated DNA.

172. The method of claim 160, wherein the plurality of polynucleotides comprises RNA.

173. The method of claim 160, wherein the sequencing is at a depth of 100x or less.

174. The method of claim 160, wherein the sequencing is at a depth of 85x or less.

175. The method of claim 160, wherein the sequencing is at a depth of about 20x to about 85x.

176. The method of claim 160, wherein the sequencing is at a depth of about 20x to about 100x.

177. The method of claim 160, further comprising sequencing the plurality of polynucleotides to obtain the plurality of sequences.

178. A system comprising:

a computing node comprising a computer readable storage medium having program instructions embodied therewith, the program instructions executable by a processor of the computing node to cause the processor to perform a method comprising:

receiving a plurality of sequences of the plurality of polynucleotides;

identifying one or more inherited variant and one or more somatic variant among the plurality of sequences;

providing the one or more inherited variant to a first trained classifier;

providing the one or more somatic variant to a second trained classifier;

determining by the first and second trained classifier the presence of aneuploidy in the plurality of polynucleotides.

179. The system of claim 178, wherein the plurality of sequences comprise a plurality of somatic variants.

180. The system of claim 178, wherein the first or second trained classifier comprises an artificial neural network.

181. The system of claim 178, wherein the first or second trained classifier comprises a regression model.

182. The system of claim 178, wherein the first or second trained classifier comprises a random decision forest.

183. The system of claim 178, wherein the first or second trained classifier comprises an SVM.

184. The system of claim 180, wherein the artificial neural network is a convolutional neural network.

185. The system of claim 180, wherein the artificial neural network is a recurrent neural network.

186. The system of claim 178, wherein the sample comprises blood.

187. The system of claim 178, wherein the sample comprises cerebrospinal fluid.

188. The system of claim 178, wherein the plurality of polynucleotides comprises DNA.

189. The system of claim 178, wherein the plurality of polynucleotides comprises methylated DNA.

190. The system of claim 178, wherein the plurality of polynucleotides comprises RNA.

191. The system of claim 178, wherein the sequencing is at a depth of 100x or less.

192. The system of claim 178, wherein the sequencing is at a depth of 85x or less.

193. The system of claim 178, wherein the sequencing is at a depth of about 20x to about 85x.

194. The system of claim 178, wherein the sequencing is at a depth of about 20x to about 100x.

195. The system of claim 178, further comprising sequencing the plurality of polynucleotides to obtain the plurality of sequences.

196. A computer program product for analyzing a sample comprising a plurality of polynucleotides, the computer program product comprising a computer readable storage medium having program instructions embodied therewith, the program instructions executable by a processor to cause the processor to perform a method comprising:

receiving a plurality of sequences of the plurality of polynucleotides;

identifying one or more inherited variant and one or more somatic variant among the plurality of sequences;

providing the one or more inherited variant to a first trained classifier;

providing the one or more somatic variant to a second trained classifier;

determining by the first and second trained classifier the presence of aneuploidy in the plurality of polynucleotides.

197. The computer program product of claim 196, wherein the plurality of sequences comprise a plurality of somatic variants.

198. The computer program product of claim 196, wherein the trained classifier comprises an artificial neural network.

199. The computer program product of claim 196, wherein the trained classifier comprises a regression model.

200. The computer program product of claim 196, wherein the trained classifier comprises a random decision forest.

201. The computer program product of claim 196, wherein the trained classifier comprises an SVM.

202. The computer program product of claim 198, wherein the artificial neural network is a convolutional neural network.

203. The computer program product of claim 198, wherein the artificial neural network is a recurrent neural network.

204. The computer program product of claim 196, wherein the sample comprises blood.

205. The computer program product of claim 196, wherein the sample comprises cerebrospinal fluid.

206. The computer program product of claim 196, wherein the plurality of polynucleotides comprises DNA.

207. The computer program product of claim 196, wherein the plurality of polynucleotides comprises methylated DNA.

208. The computer program product of claim 196, wherein the plurality of polynucleotides comprises RNA.

209. The computer program product of claim 196, wherein the sequencing is at a depth of 100x or less.

210. The computer program product of claim 196, wherein the sequencing is at a depth of 85x or less.

211. The computer program product of claim 196, wherein the sequencing is at a depth of about 20x to about 85x.

212. The computer program product of claim 196, wherein the sequencing is at a depth of about 20x to about 100x.

213. The computer program product of claim 196, further comprising sequencing the plurality of polynucleotides to obtain the plurality of sequences.

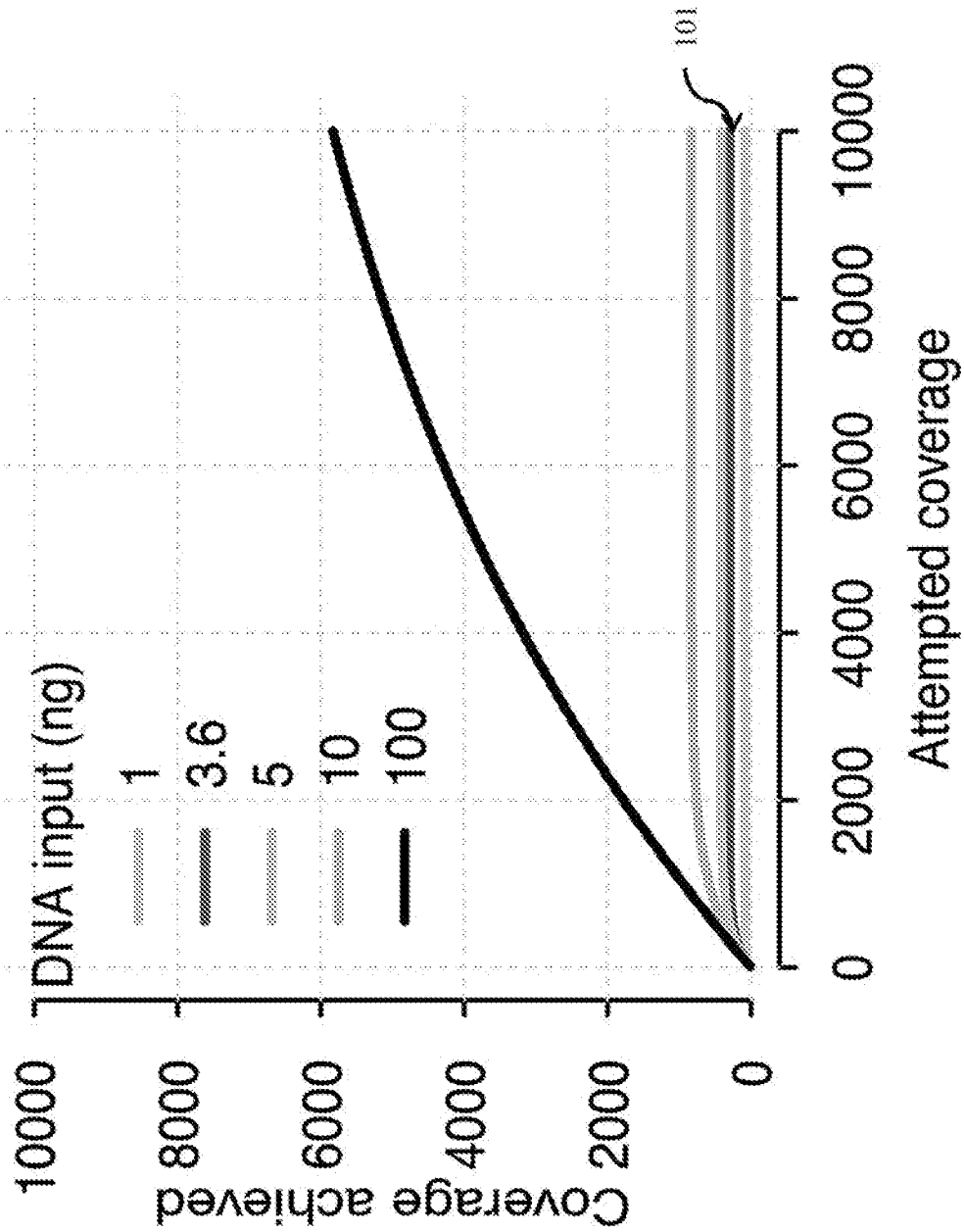
DNA input (ng)
1
3.6
5
10
100
Coverage achieved
10000
8000
6000
4000
2000
0
0
2000
4000
6000
8000
10000
Attempted coverage
101

Fig. 1A

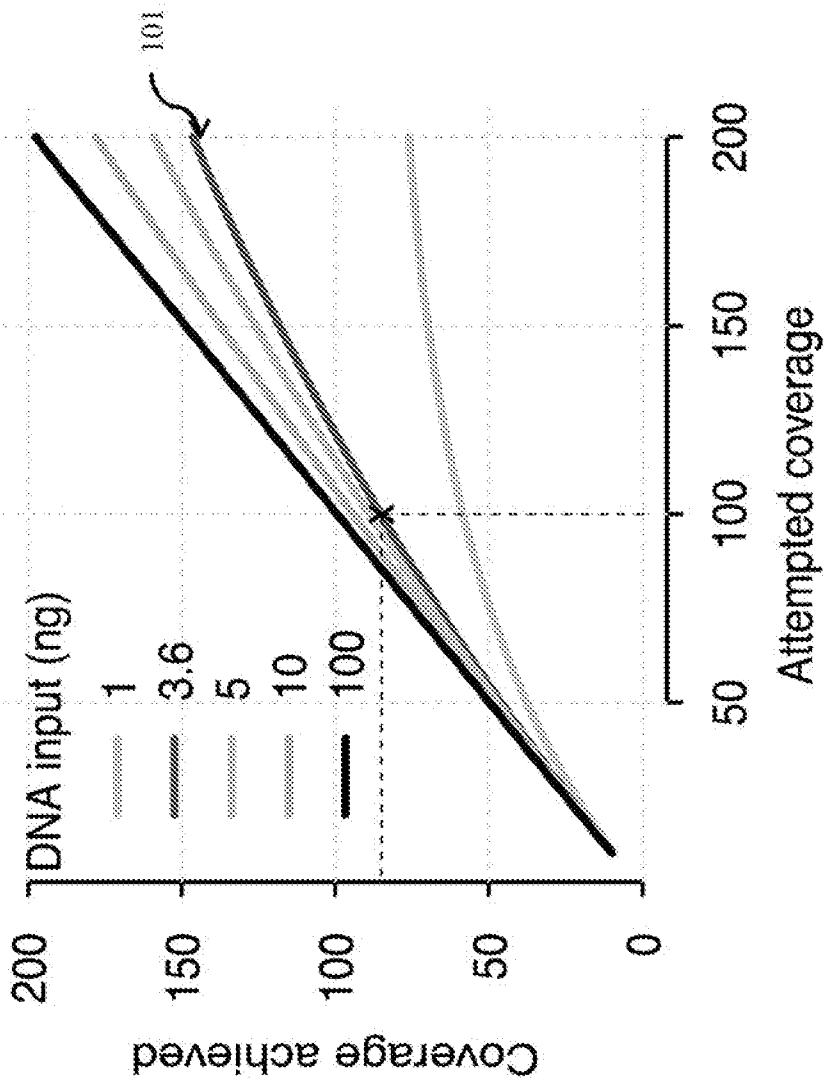
101
DNA input (ng)
1
3.6
5
10
100
Coverage achieved
200
150
100
50
0
50
100
150
200
Attempted coverage

Fig. 1B

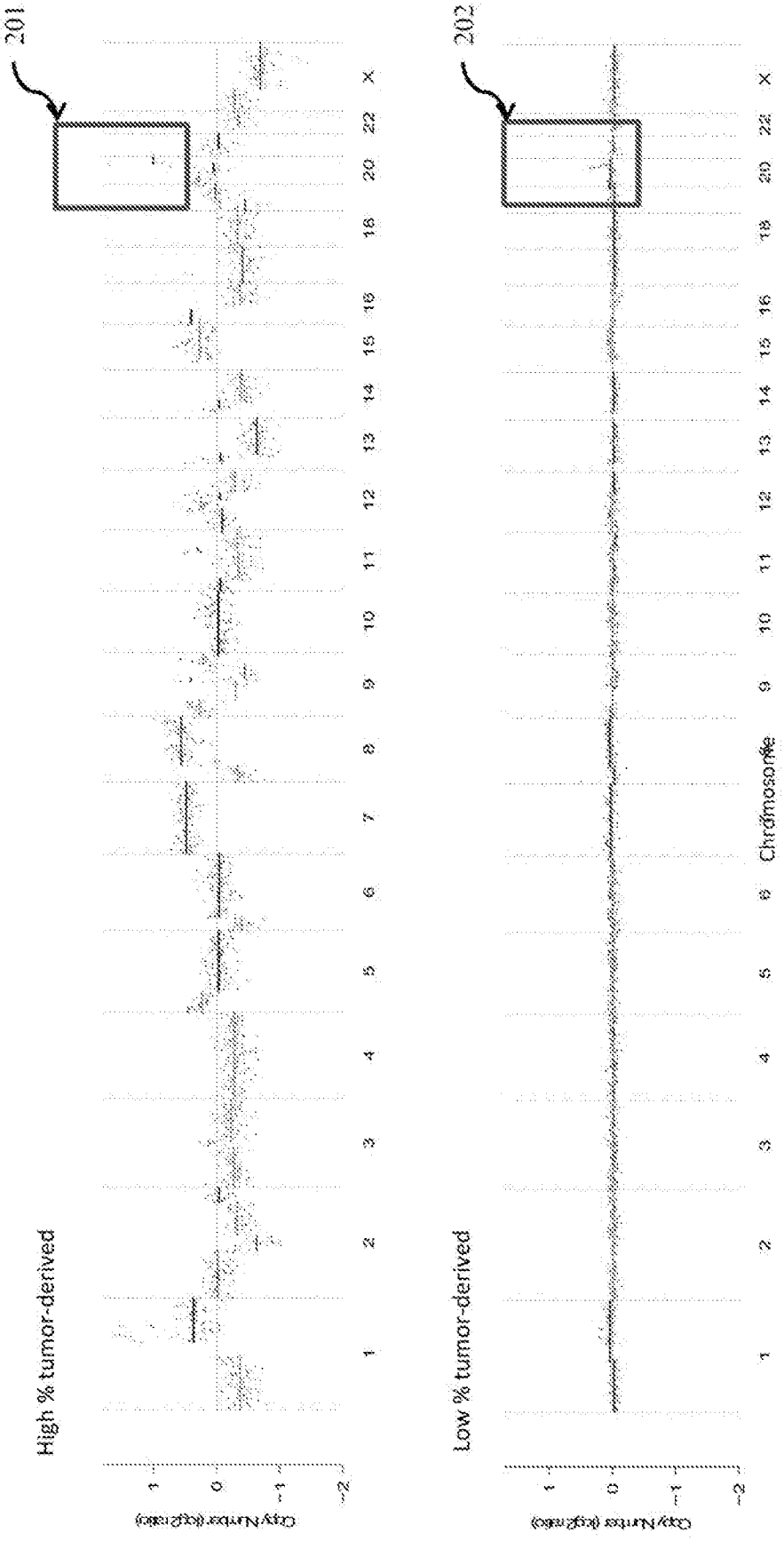
201
202
High % tumor-derived
Low % tumor-derived
Copy Number (log2 ratio)
1
0
-1
-2
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
18
20
22
X
Chromosome

Fig. 2

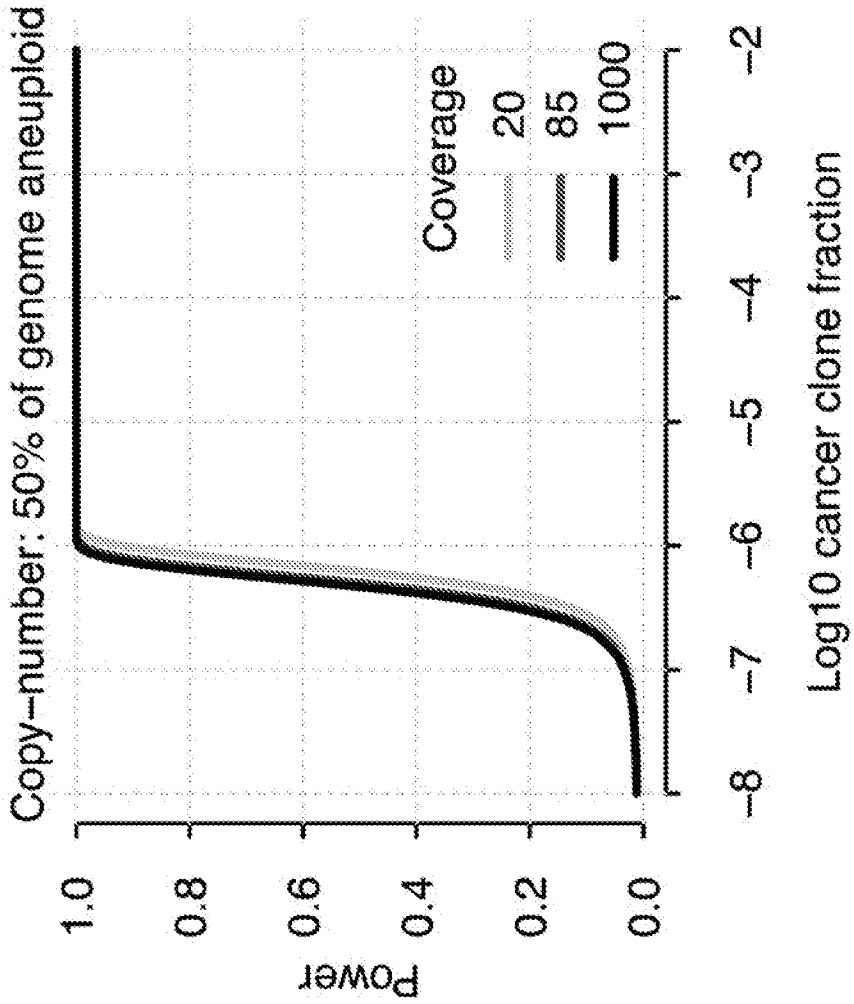
Copy−number: 50% of genome aneuploid
Power
1.0
0.8
0.6
0.4
0.2
0.0
Coverage
20
85
1000
−8
−7
−6
−5
−4
−3
−2
Log10 cancer clone fraction

Fig. 3A

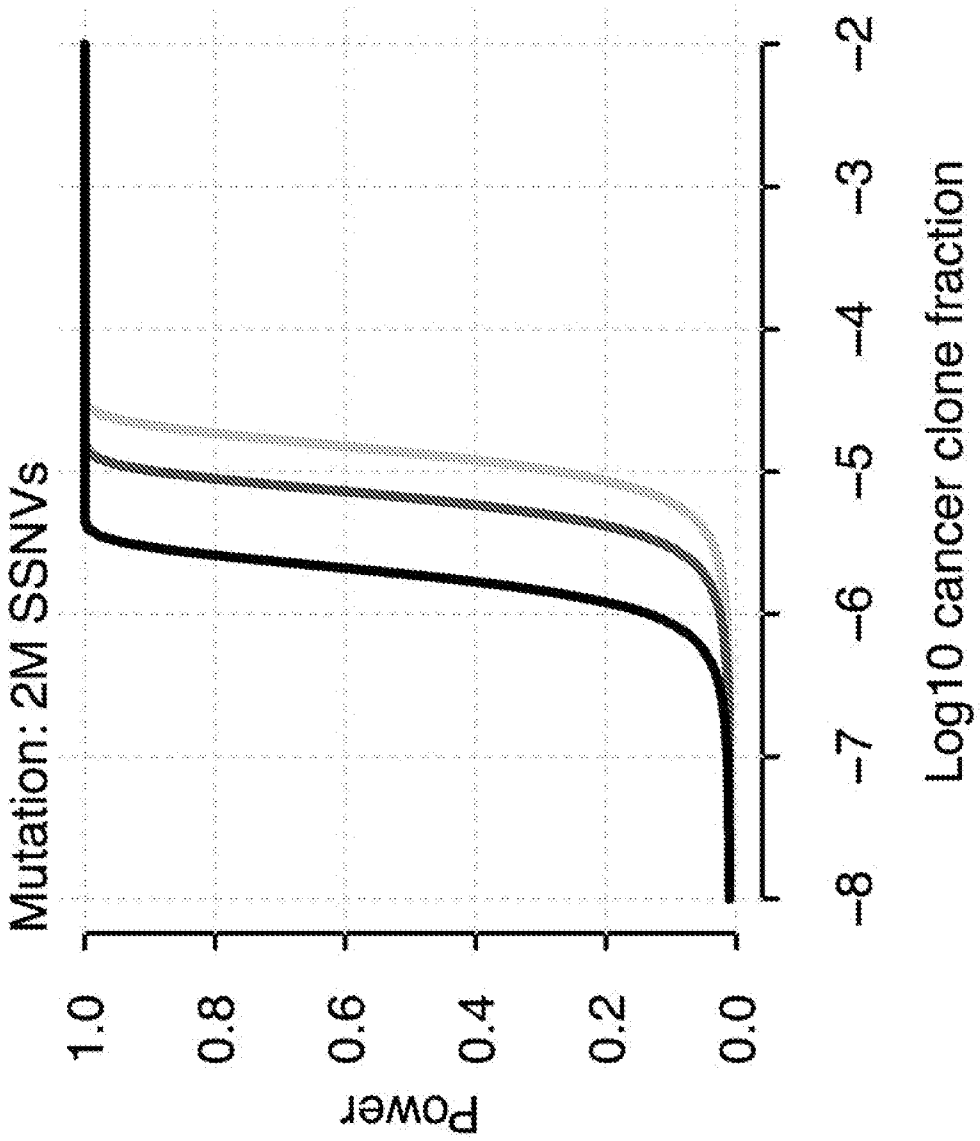
Mutation: 2M SSNVs
Power
1.0
0.8
0.6
0.4
0.2
0.0
-8
-7
-6
-5
-4
-3
-2
Log10 cancer clone fraction

Fig. 3B

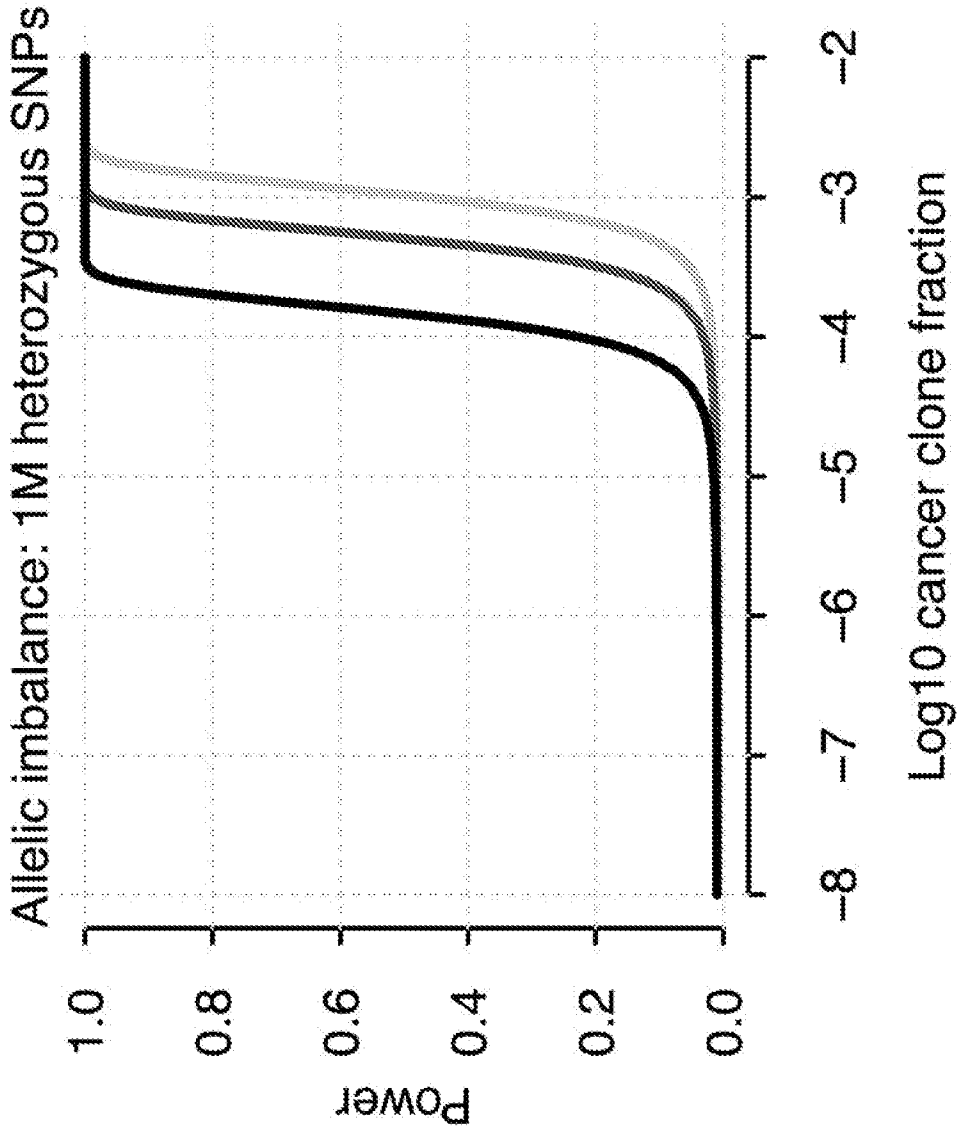
Allelic imbalance: 1M heterozygous SNPs
Power
1.0
0.8
0.6
0.4
0.2
0.0
−8
−7
−6
−5
−4
−3
−2
Log10 cancer clone fraction

Fig. 3C

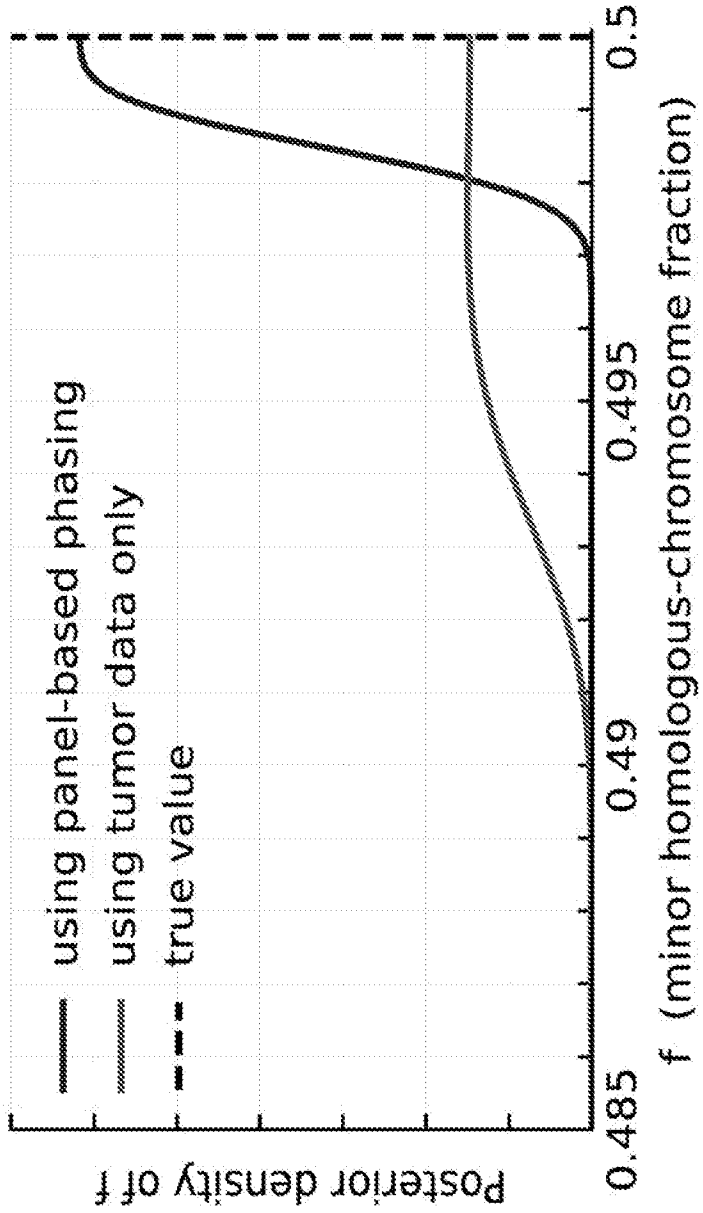
using panel-based phasing
using tumor data only
true value
Posterior density of f
0.485
0.49
0.495
0.5
f (minor homologous-chromosome fraction)

Fig. 4A

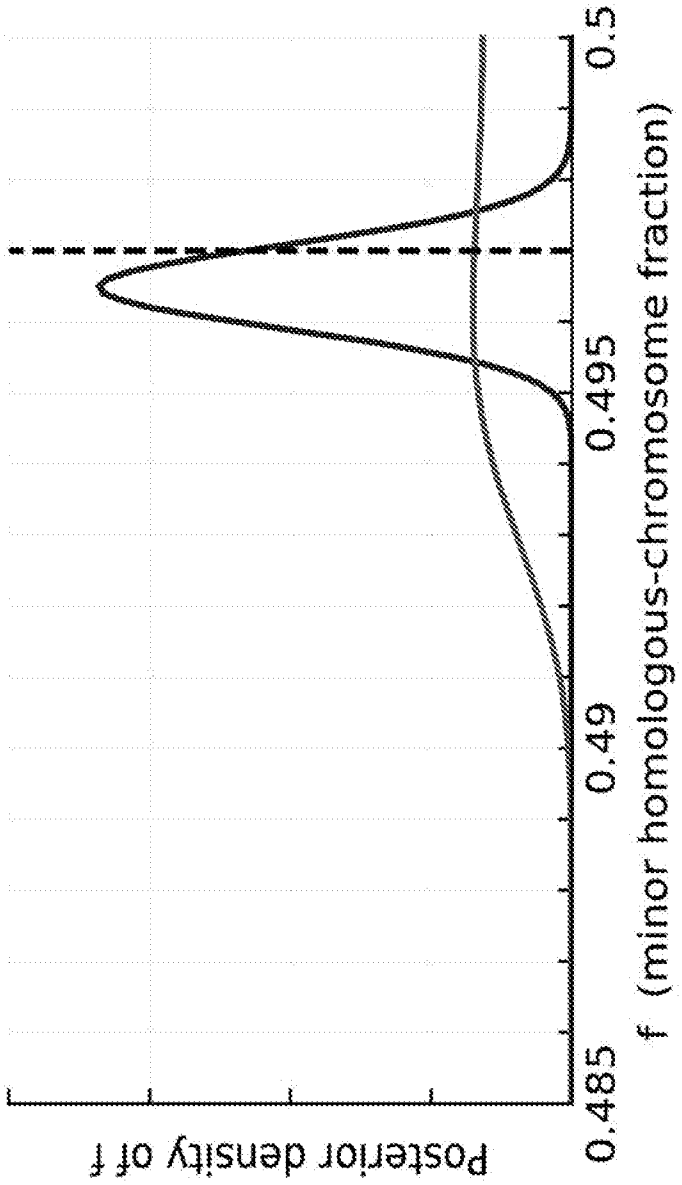
Posterior density of f
0.485
0.49
0.495
0.5
f (minor homologous-chromosome fraction)

Fig. 4B

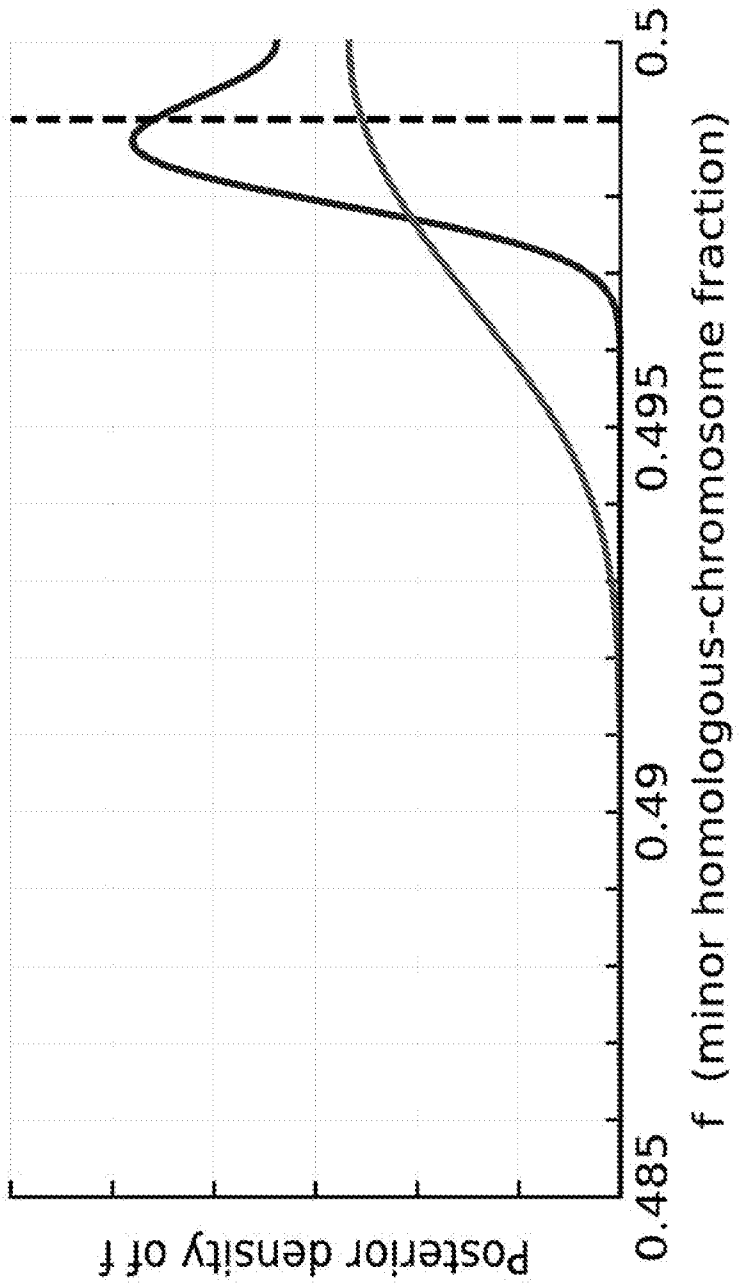
Posterior density of f
0.485
0.49
0.495
0.5
f (minor homologous-chromosome fraction)

Fig. 4C

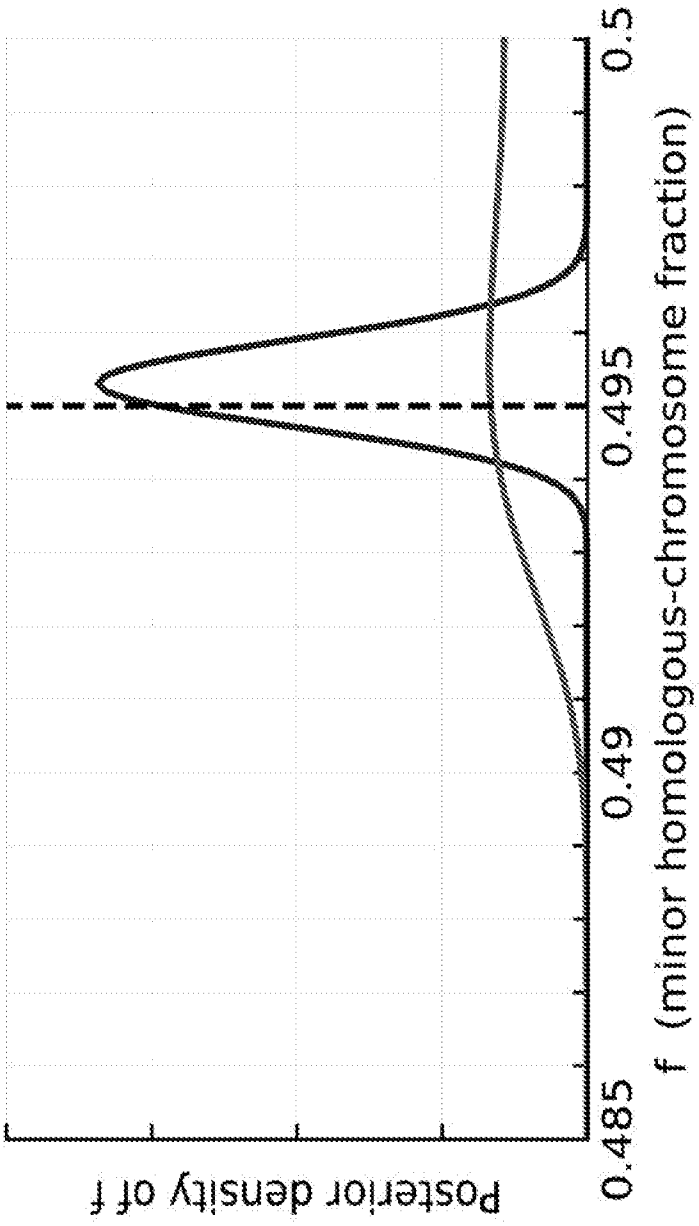
Posterior density of f
0.485
0.49
0.495
0.5
f (minor homologous-chromosome fraction)

Fig. 4D

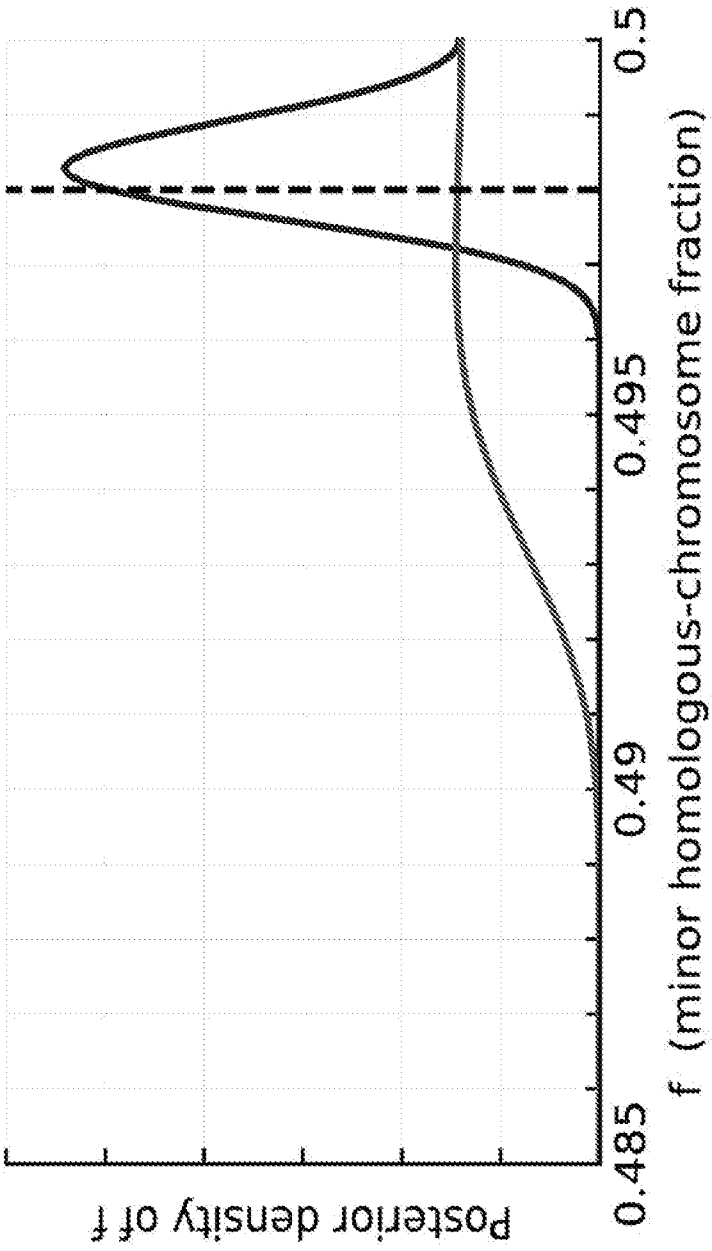
Posterior density of f
0.485
0.49
0.495
0.5
f (minor homologous-chromosome fraction)

Fig. 4E

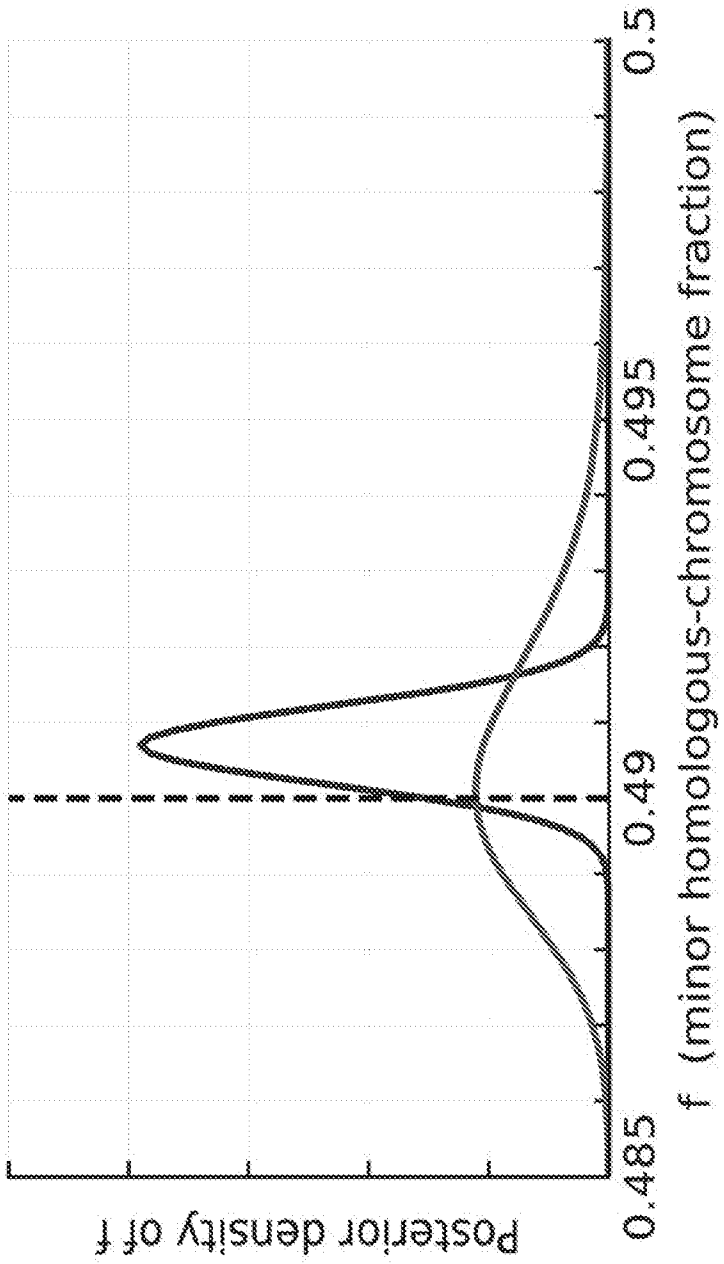
Posterior density of f
0.485
0.49
0.495
0.5
f (minor homologous-chromosome fraction)

Fig. 4F

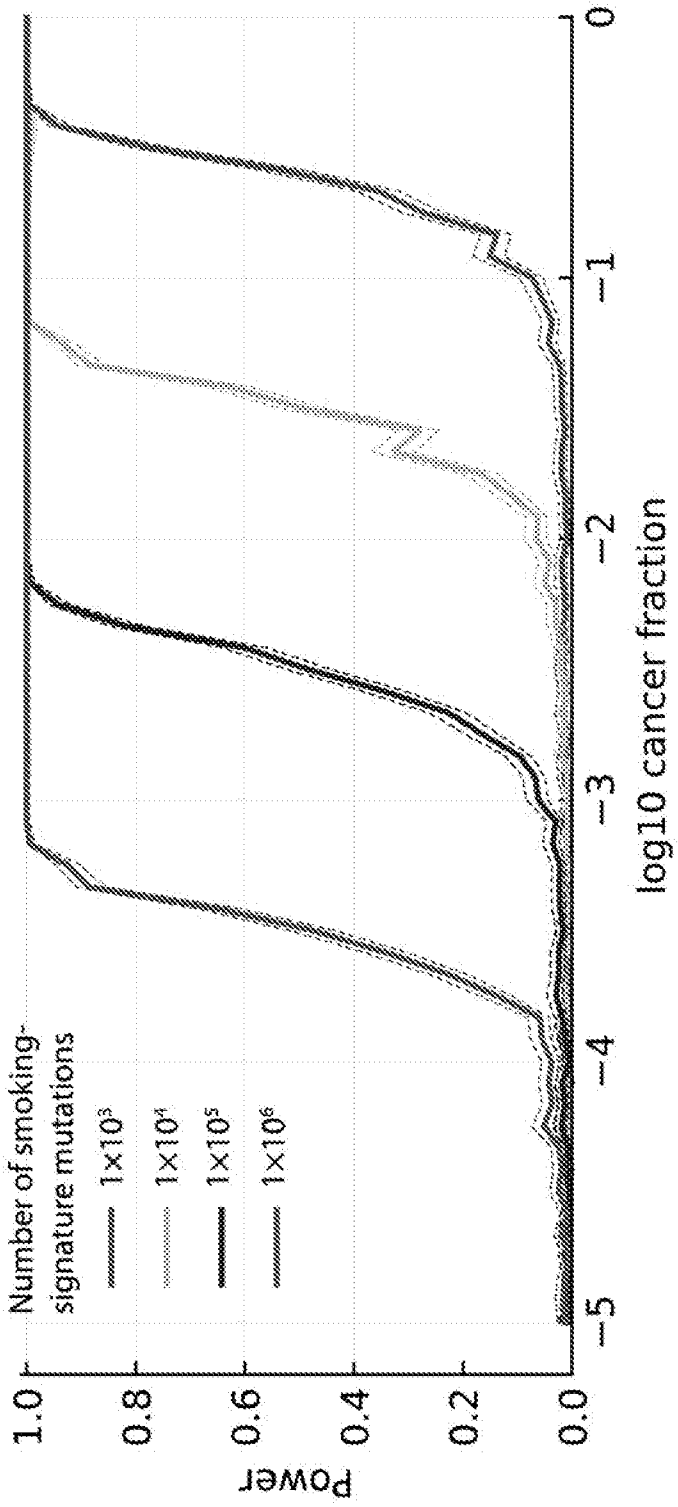
Power
1.0
0.8
0.6
0.4
0.2
0.0
−5
−4
−3
−2
−1
0
log10 cancer fraction
Number of smoking-signature mutations
1×10³
1×10⁴
1×10⁵
1×10⁶

Fig. 5

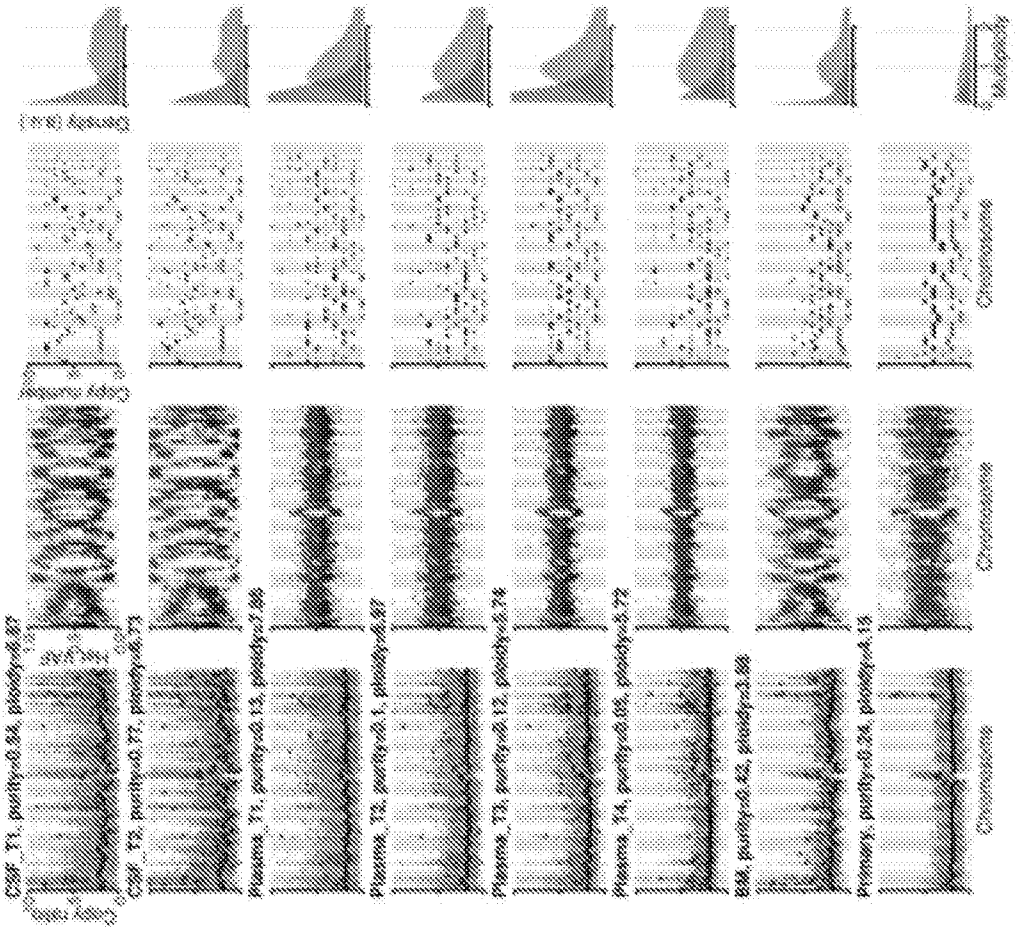

Fig. 6A

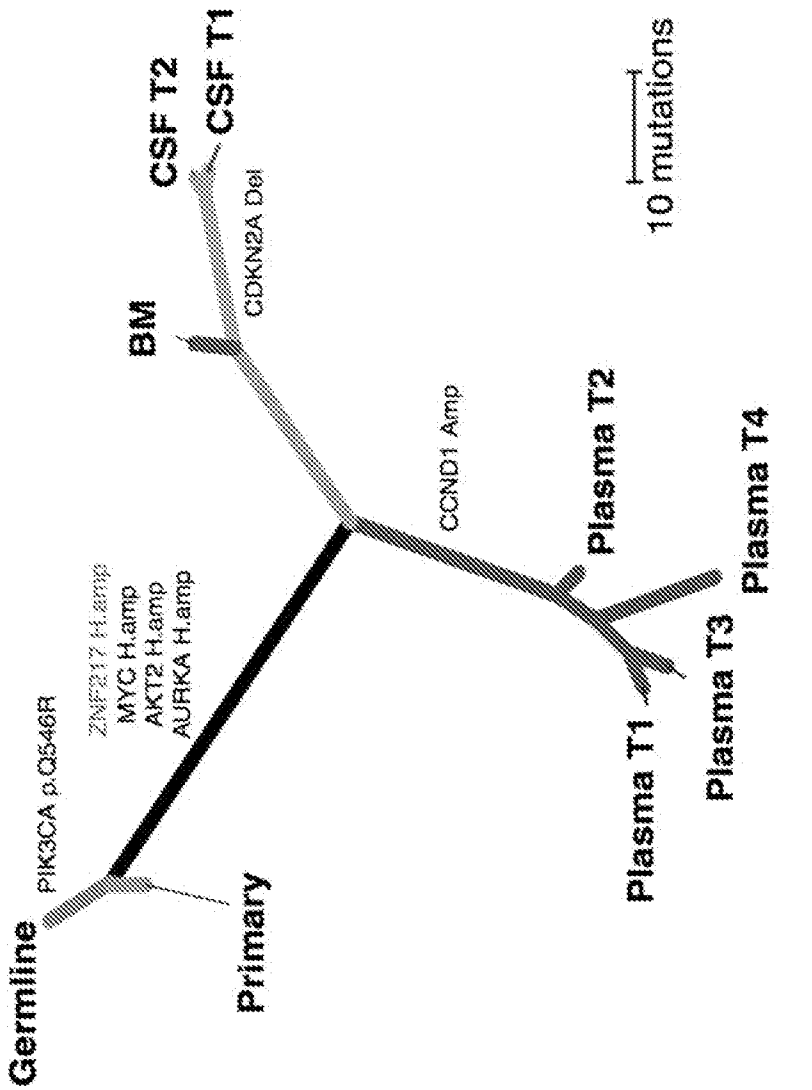
Germline
PIK3CA p.Q546R
Primary
ZNF217 H.amp
MYC H.amp
AKT2 H.amp
AURKA H.amp
BM
CSF T2
CSF T1
CDKN2A Del
CCND1 Amp
Plasma T2
Plasma T4
Plasma T3
Plasma T1
10 mutations

Fig. 6B

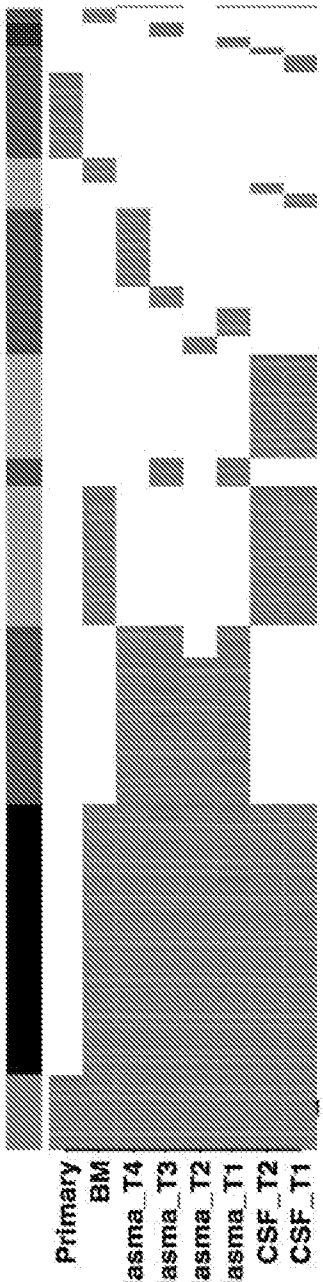
Primary
BM
Plasma_T4
Plasma_T3
Plasma_T2
Plasma_T1
CSF_T2
CSF_T1

Fig. 6C

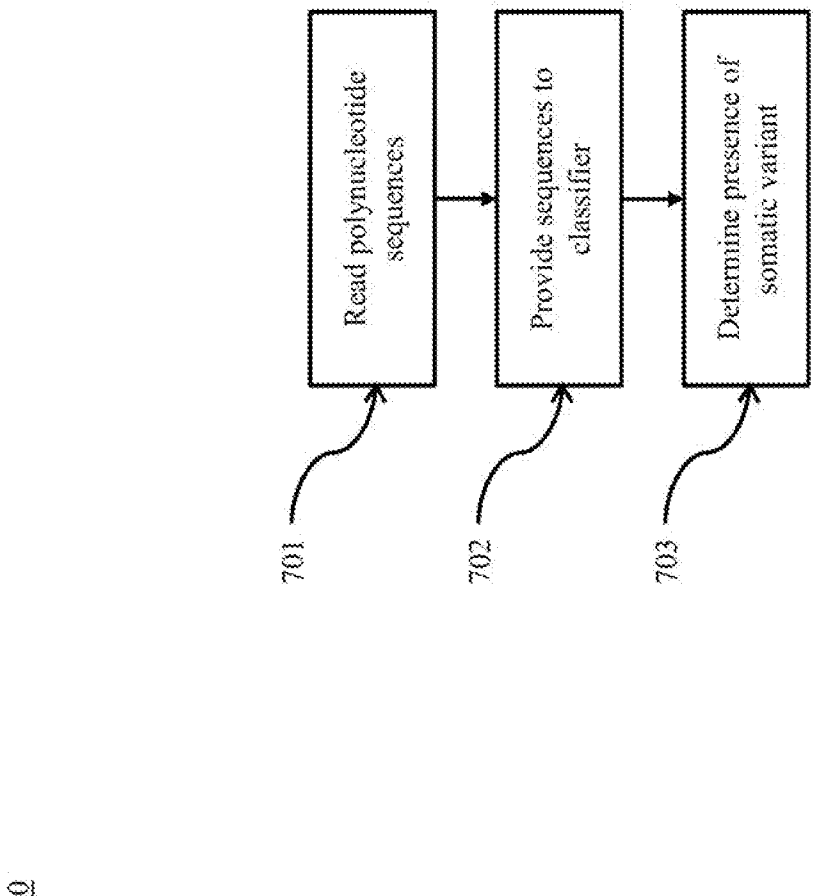
700
Read polynucleotide sequences
Provide sequences to classifier
Determine presence of somatic variant
701
702
703

Fig. 7

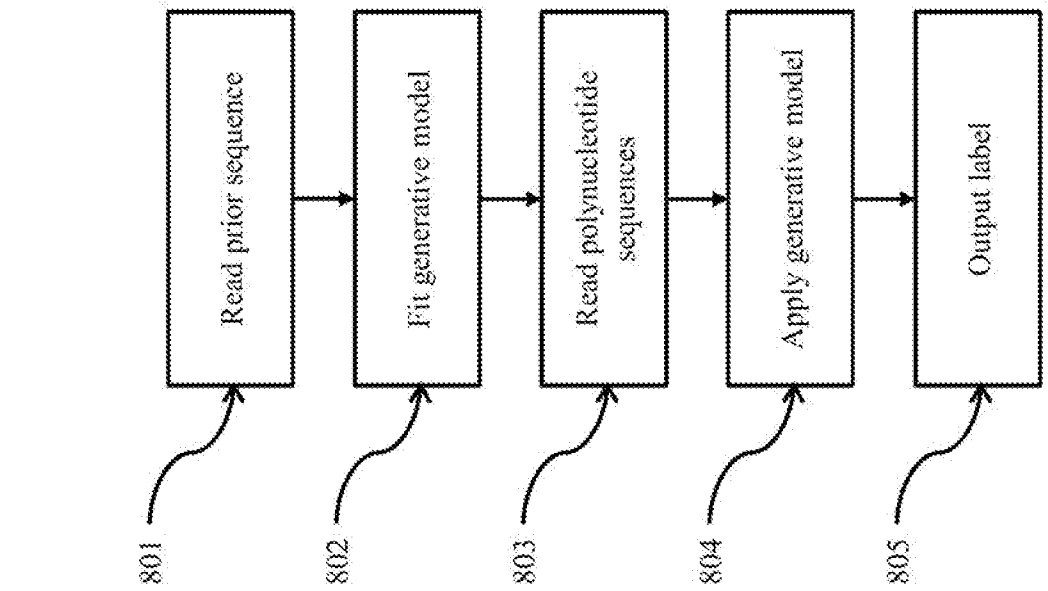
800
801
Read prior sequence
802
Fit generative model
803
Read polynucleotide sequences
804
Apply generative model
805
Output label

Fig. 8

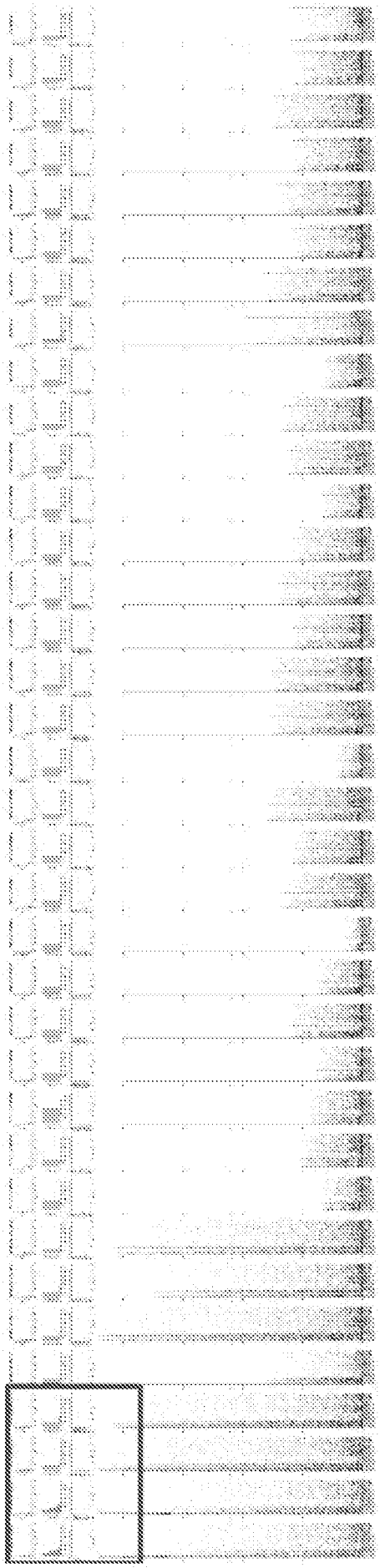

Fig. 9A

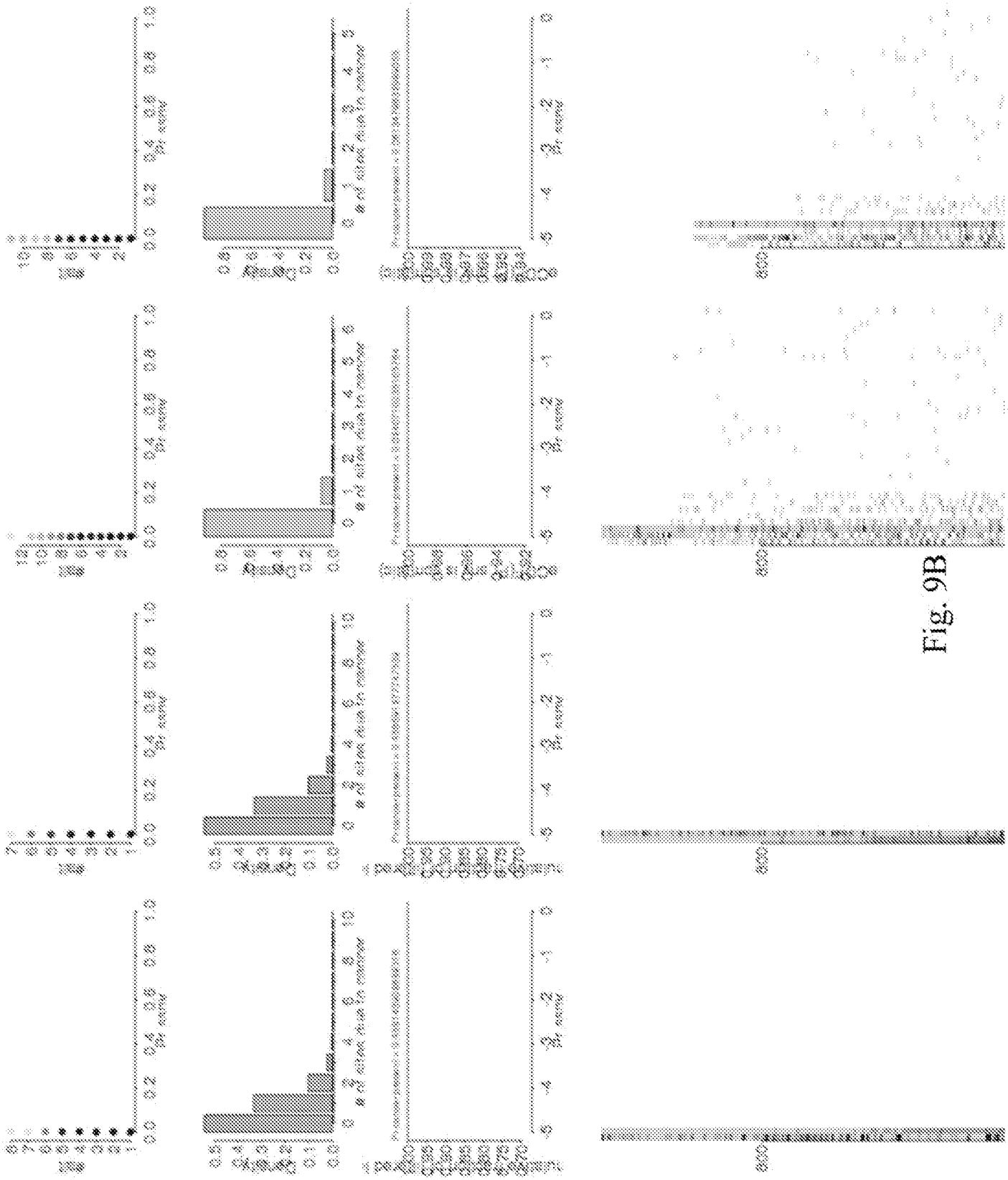

Fig. 9B

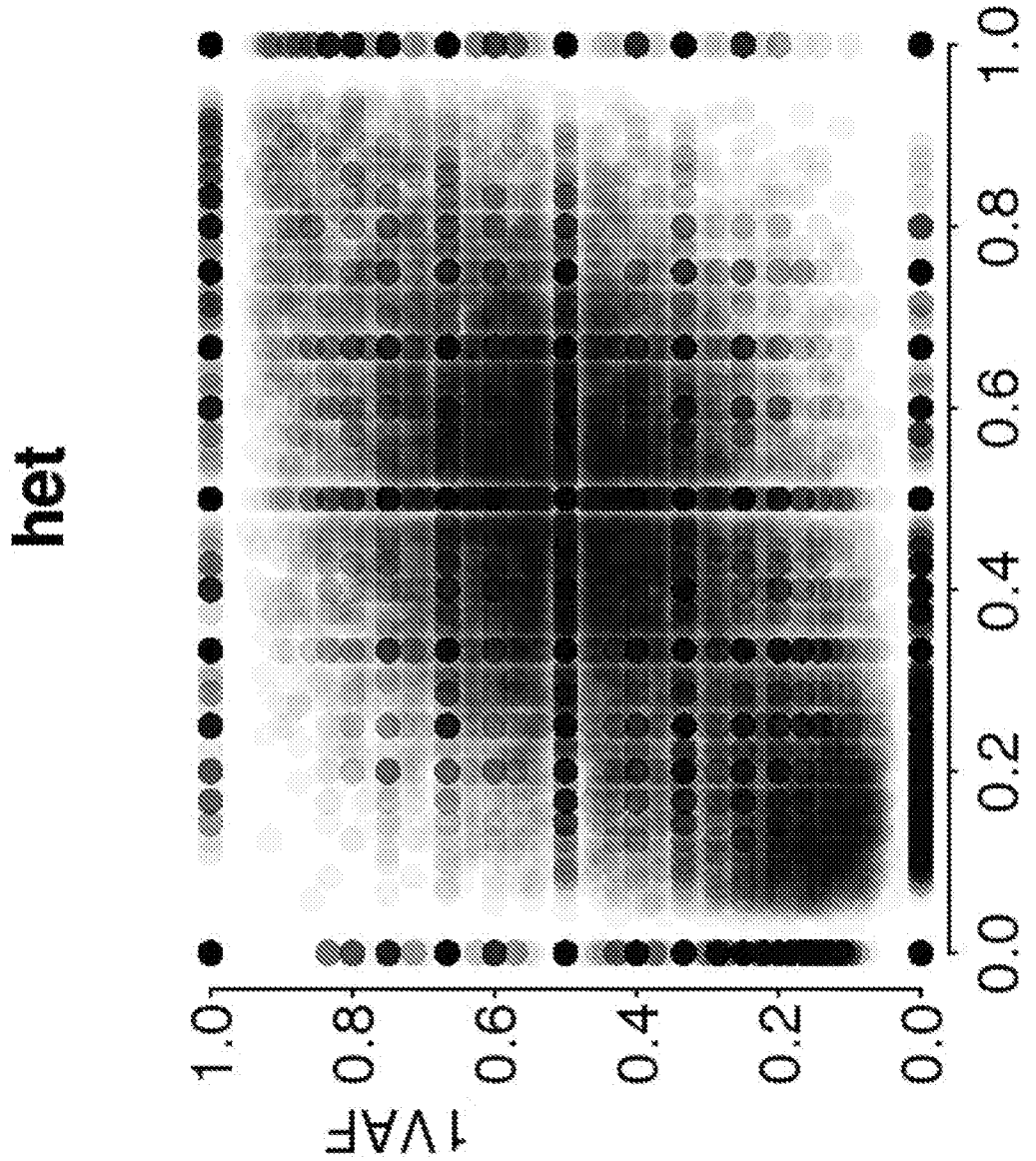

Fig. 10A

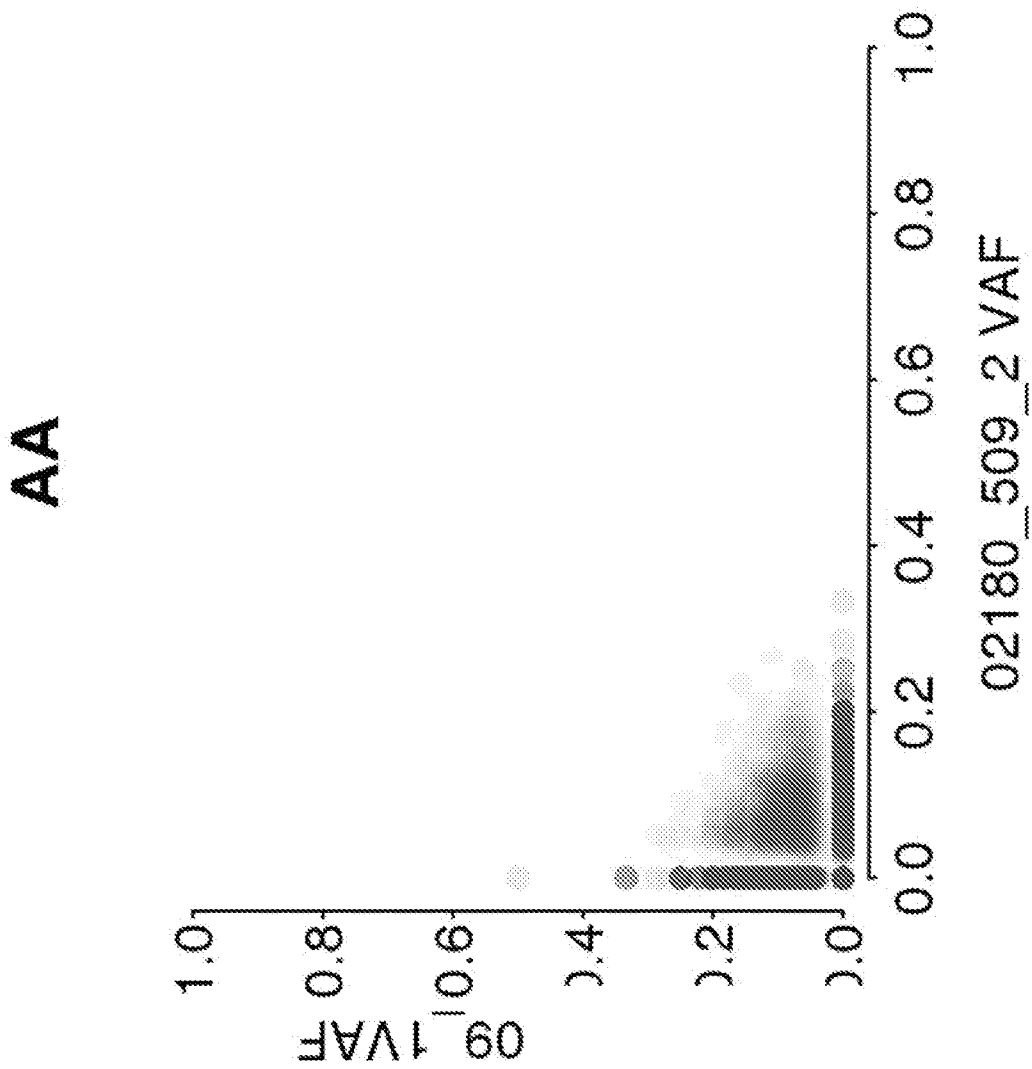

Fig. 10B

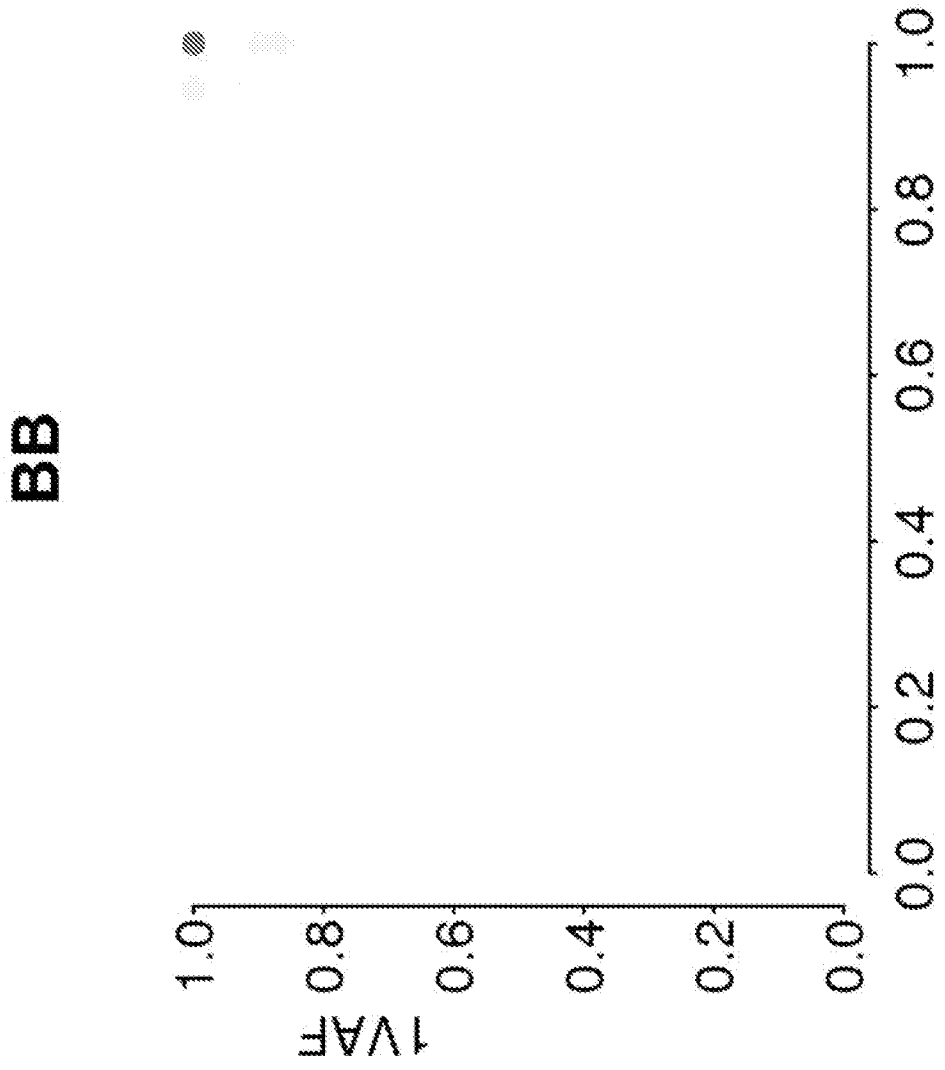

Fig. 10C

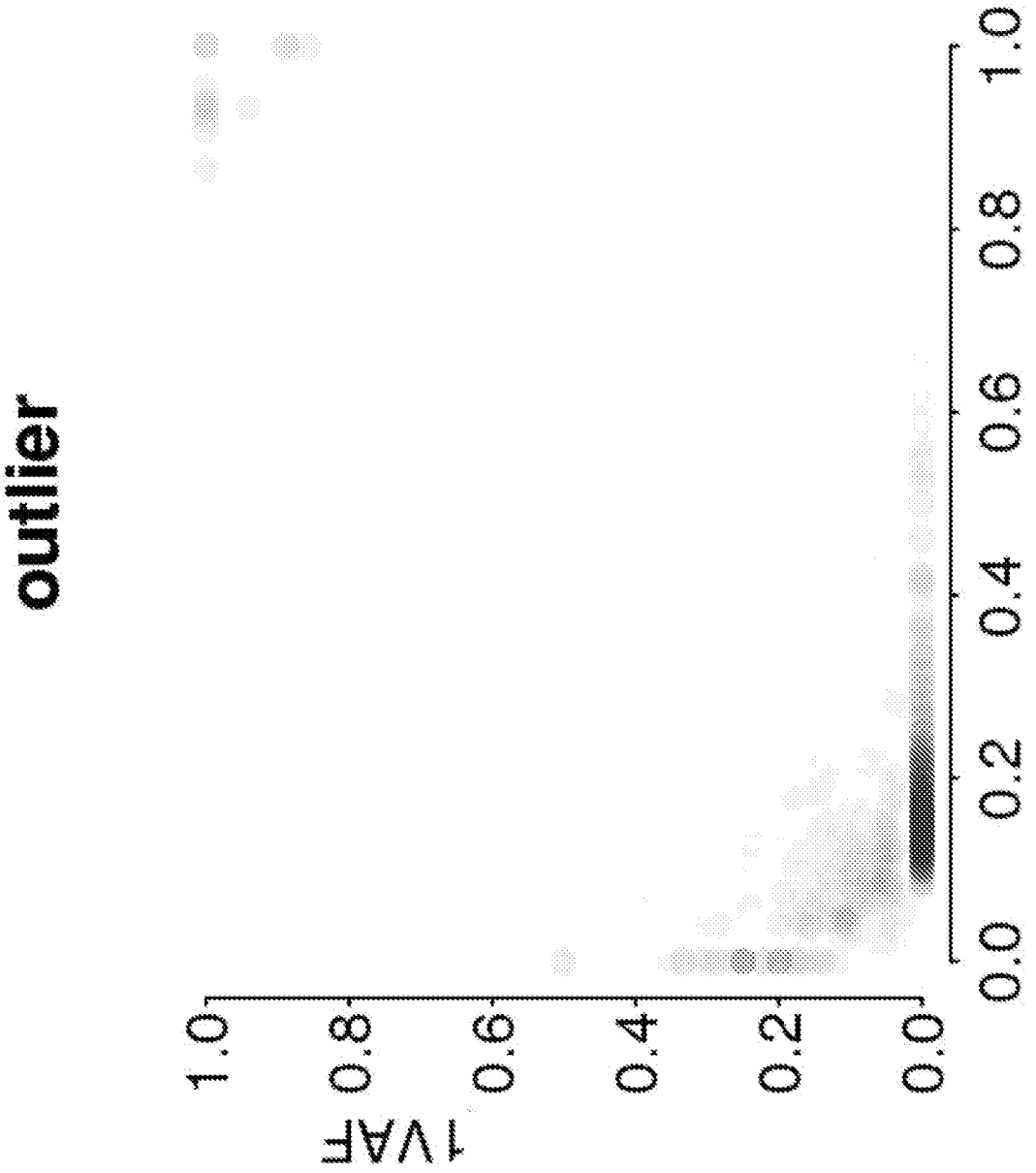

Fig. 10D

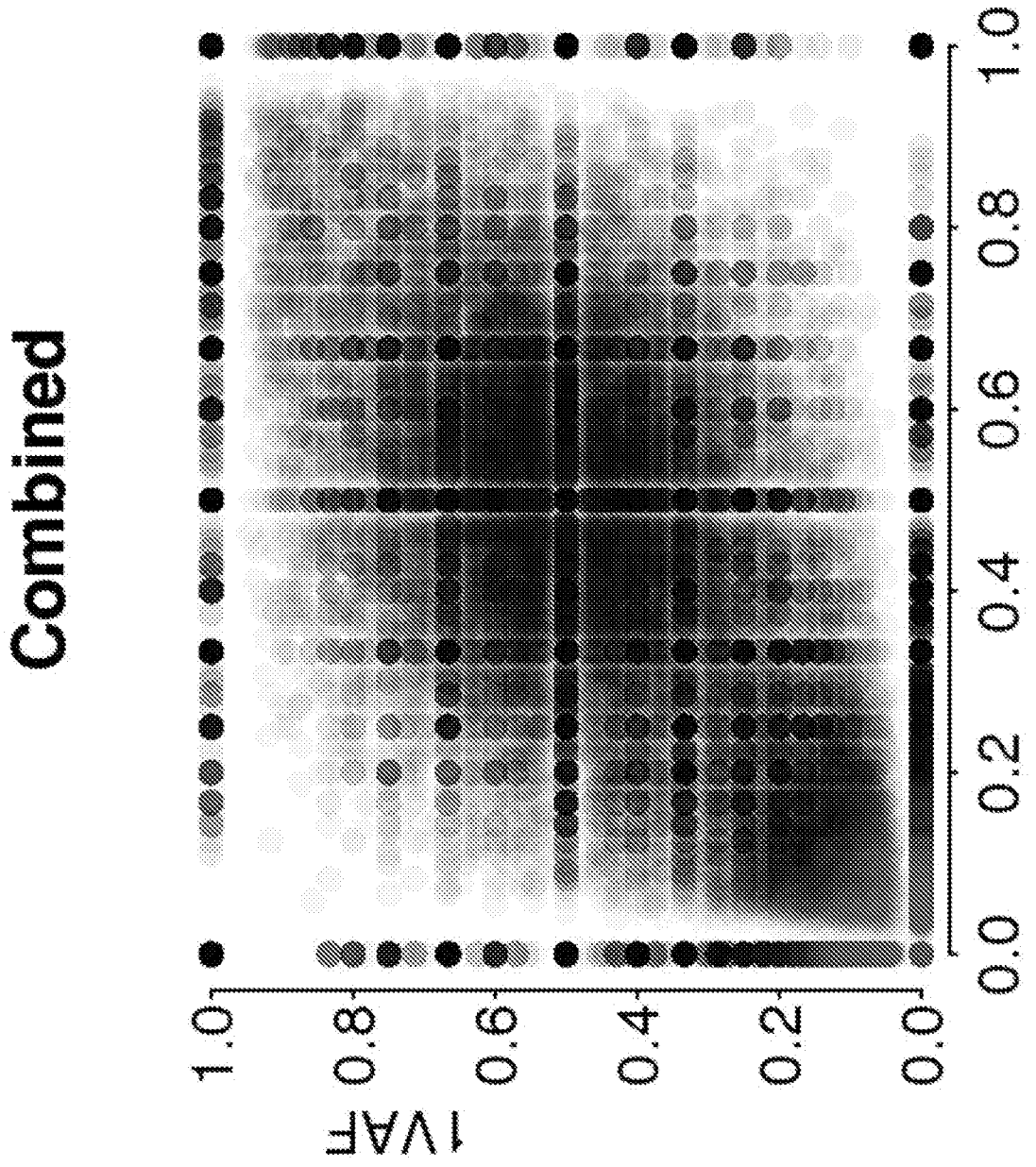

Fig. 10E

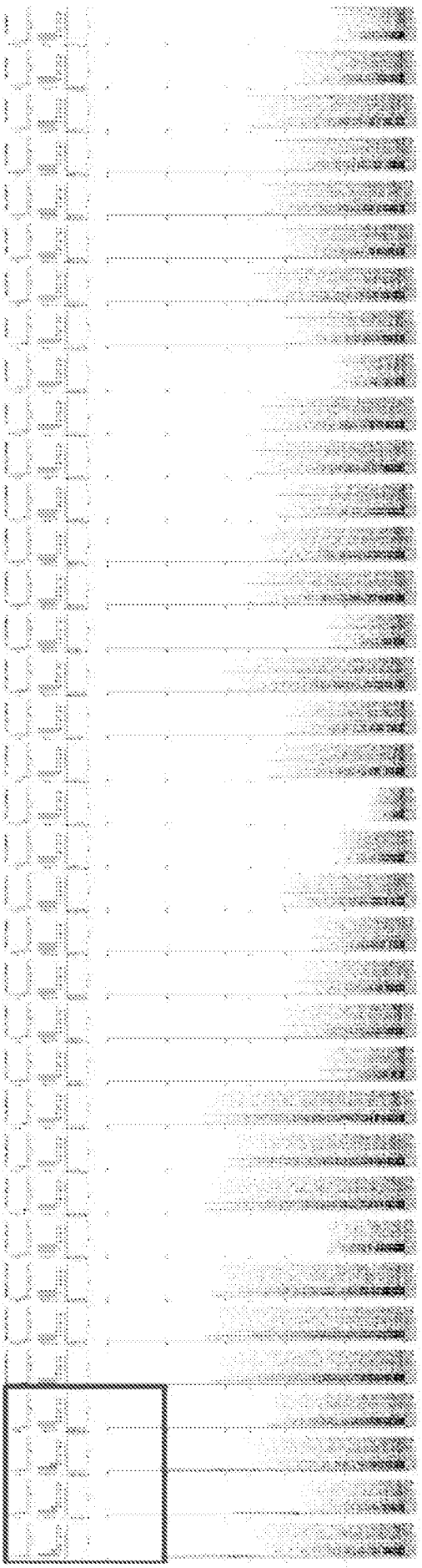

Fig. 11A

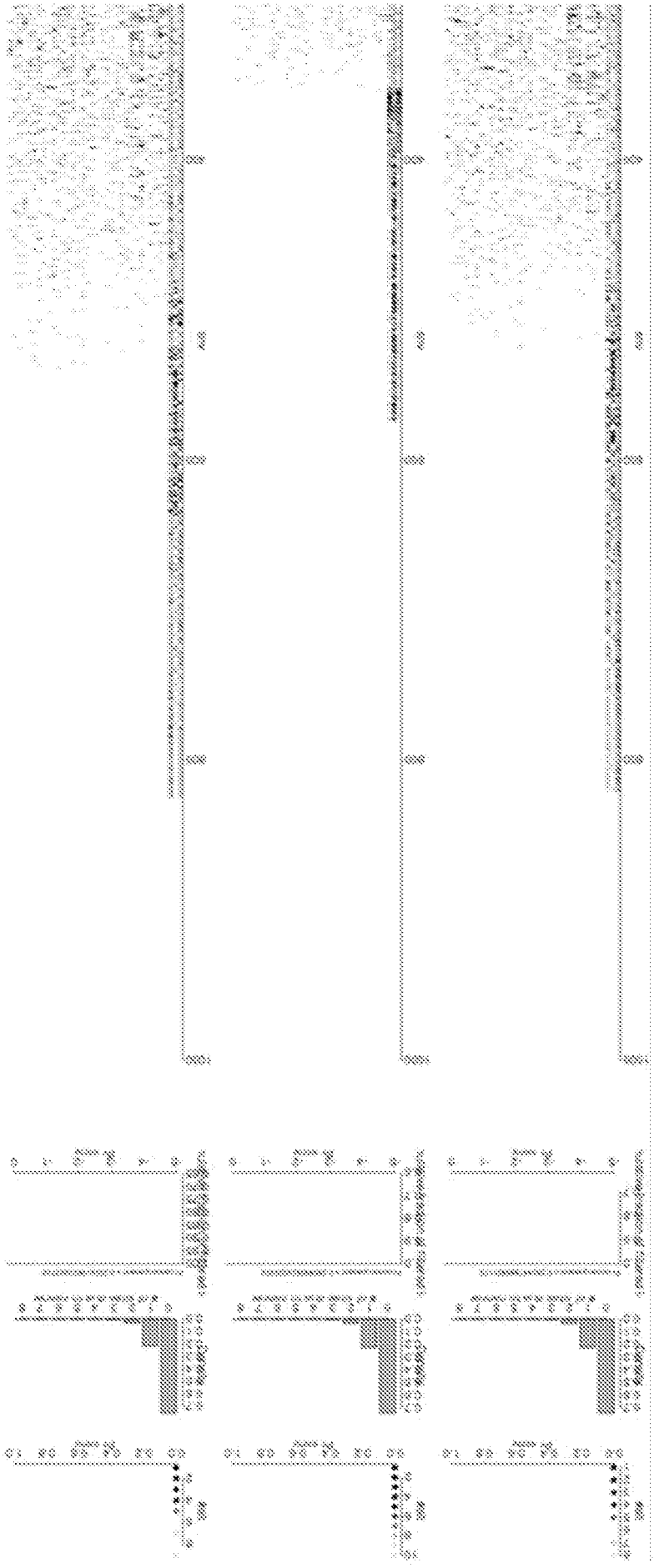

Fig. 11B

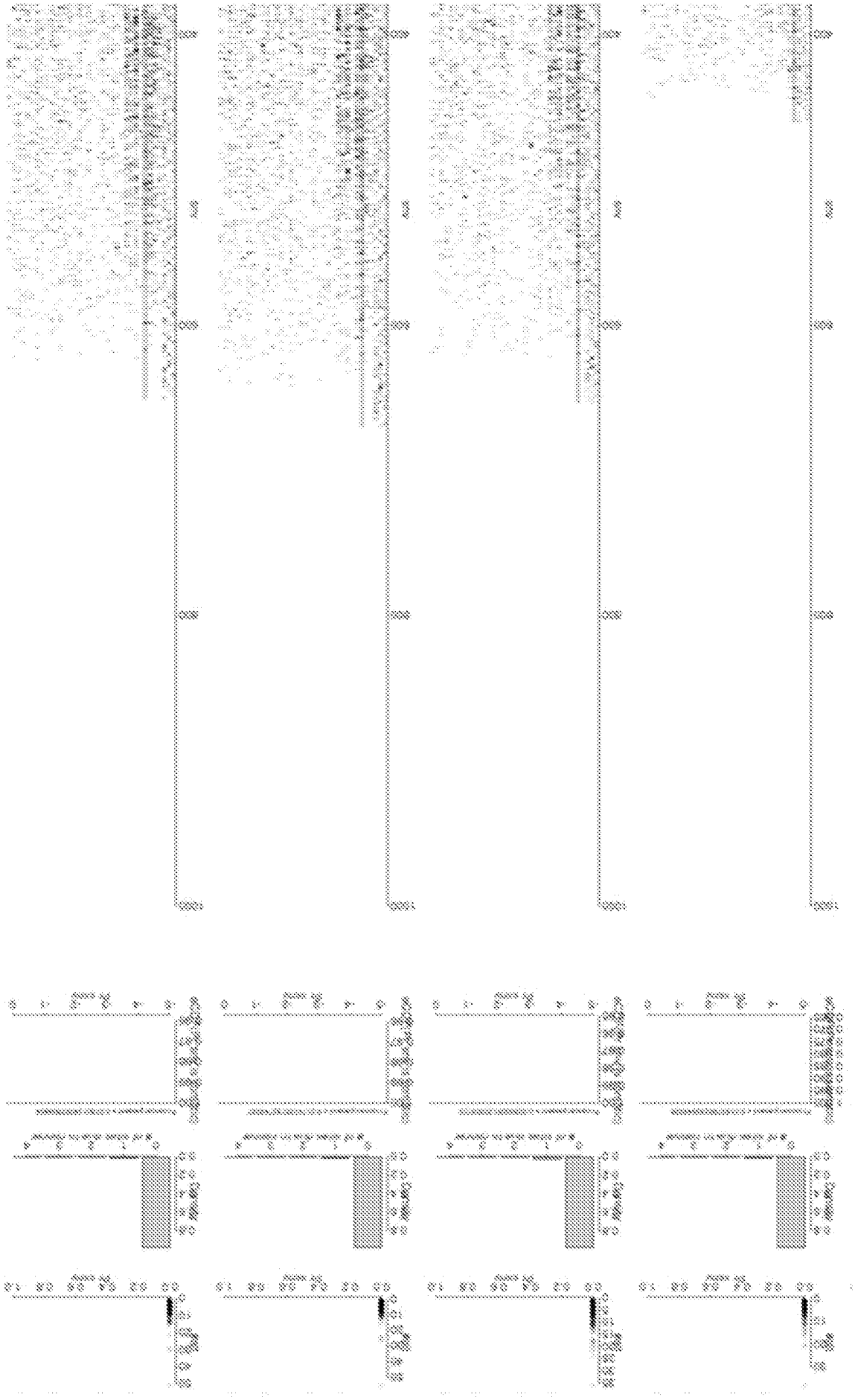

Fig. 11C

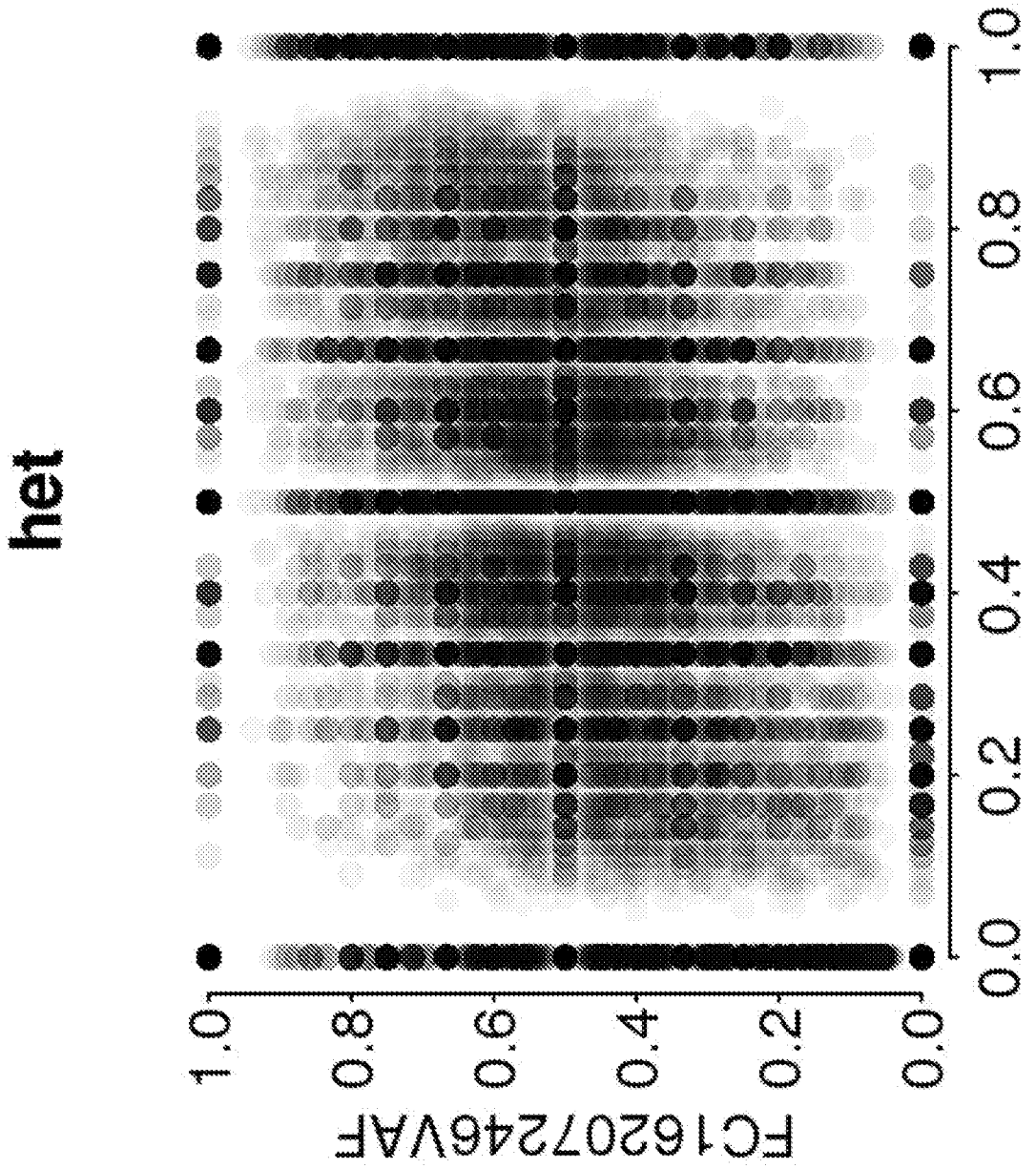
het
FC16207246VAF
1.0
0.8
0.6
0.4
0.2
0.0
FC16207248 VAF
0.0
0.2
0.4
0.6
0.8
1.0

Fig. 12A

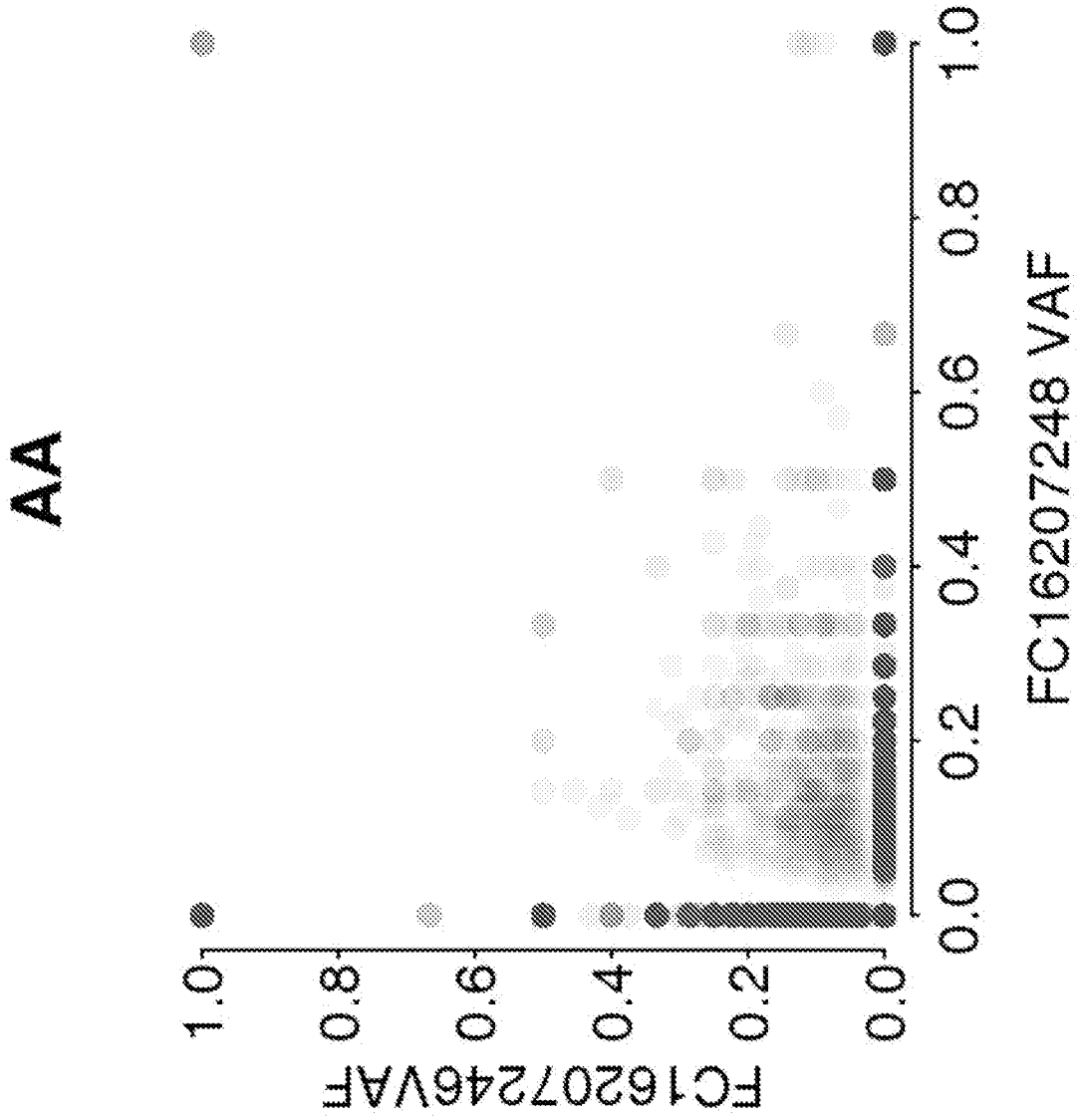
AA
FC16207246VAF
1.0
0.8
0.6
0.4
0.2
0.0
FC16207248 VAF
0.0
0.2
0.4
0.6
0.8
1.0

Fig. 12B

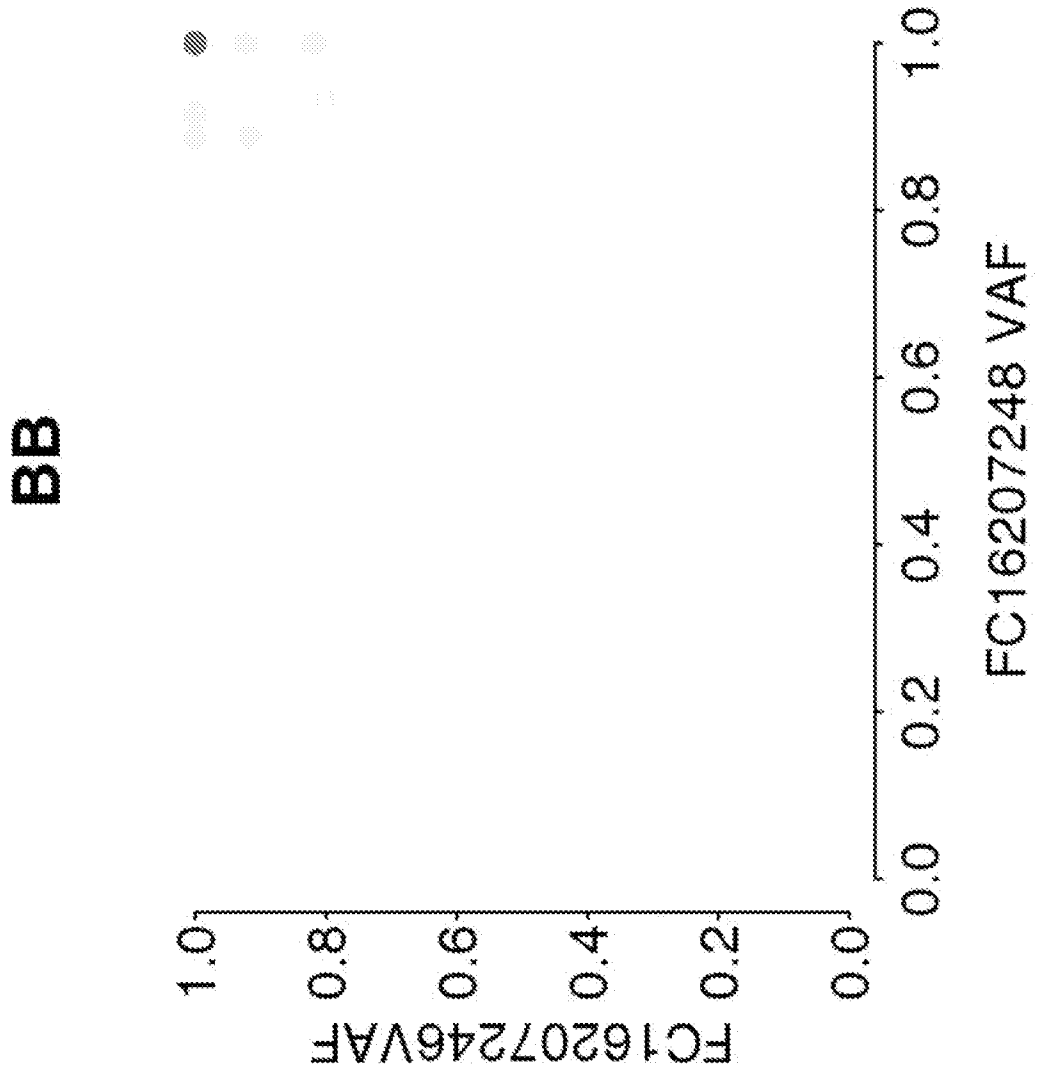
BB
FC16207246VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FC16207248 VAF

Fig. 12C

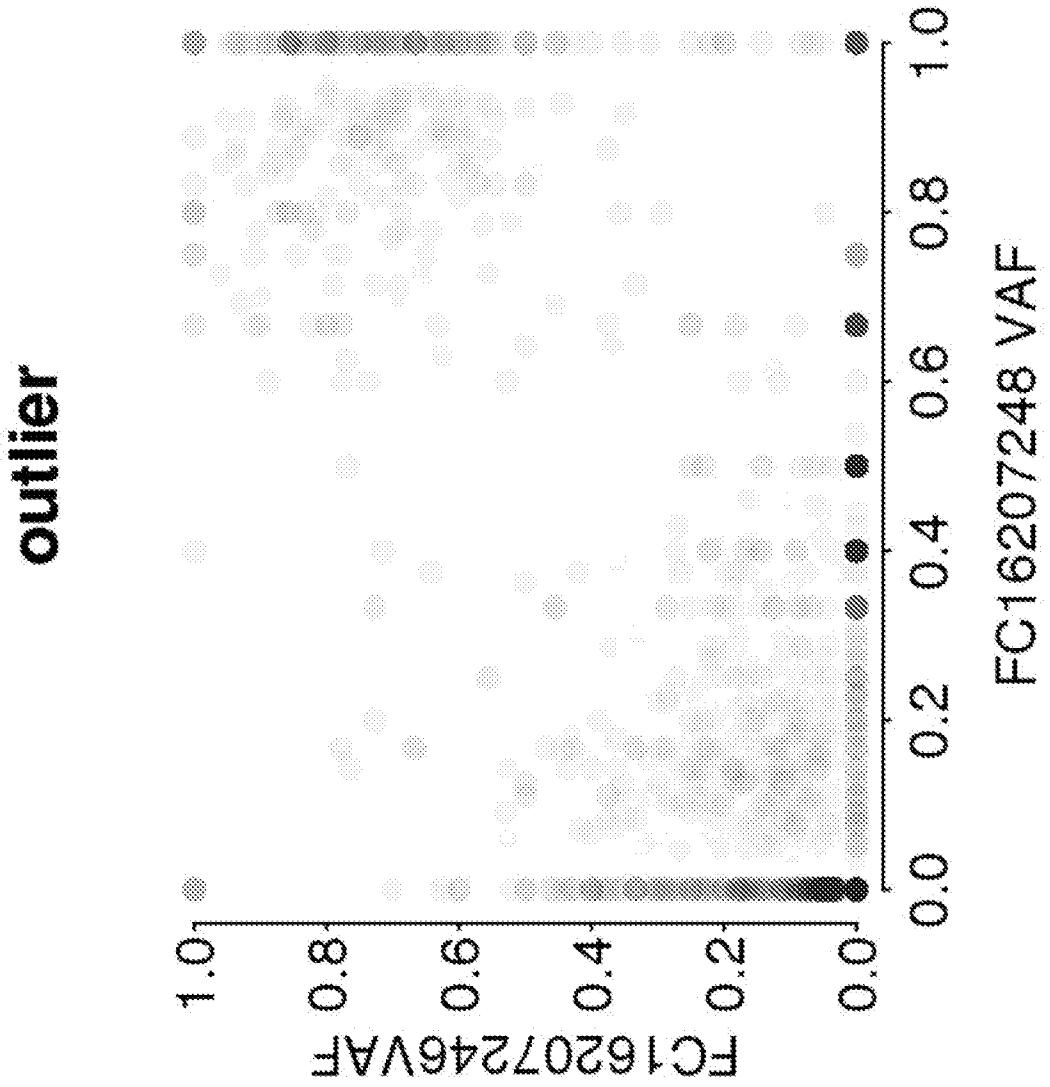
outlier
FC16207246VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FC16207248 VAF

Fig. 12D

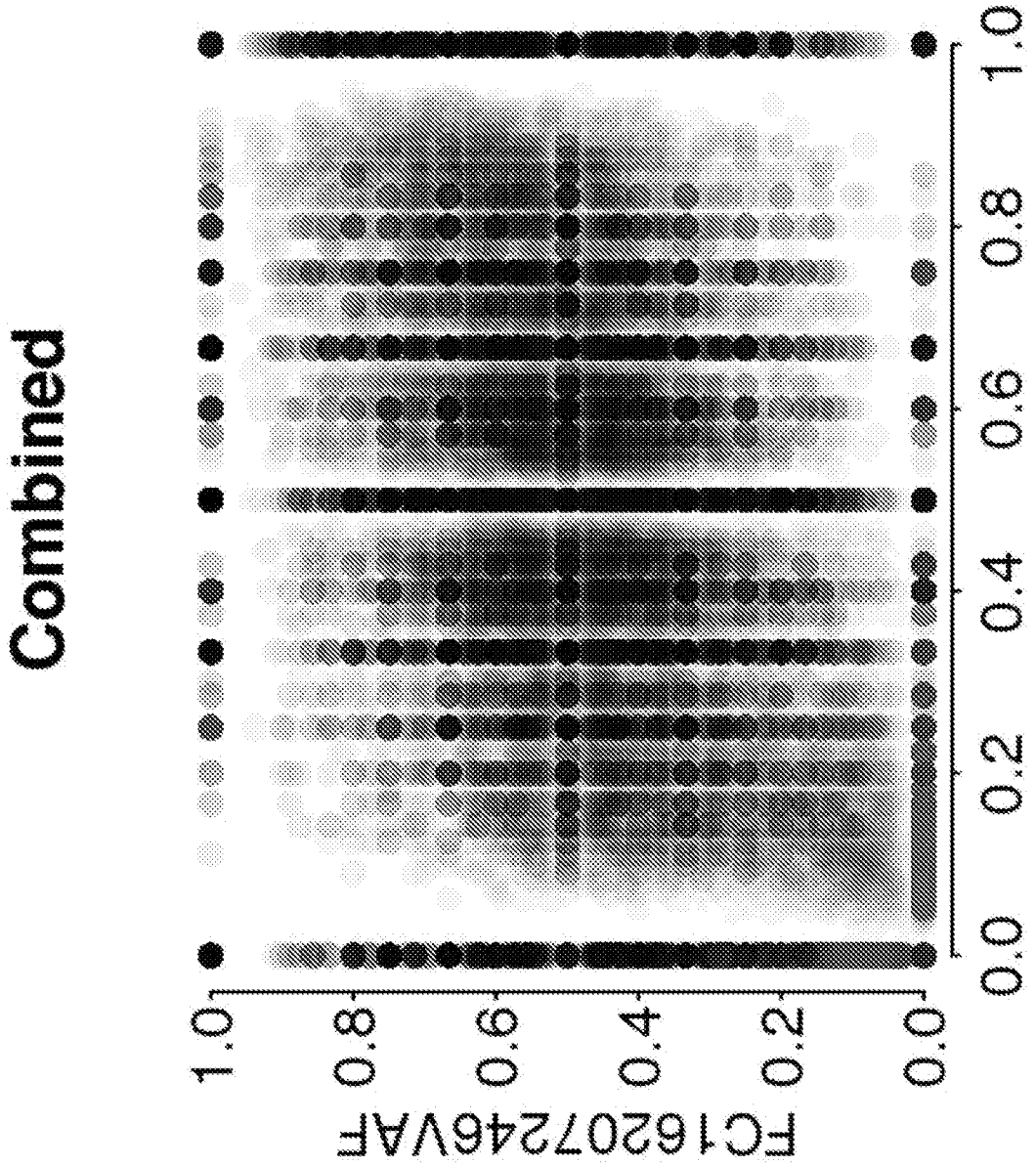
Combined
FC16207246VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FC16207248 VAF

Fig. 12E

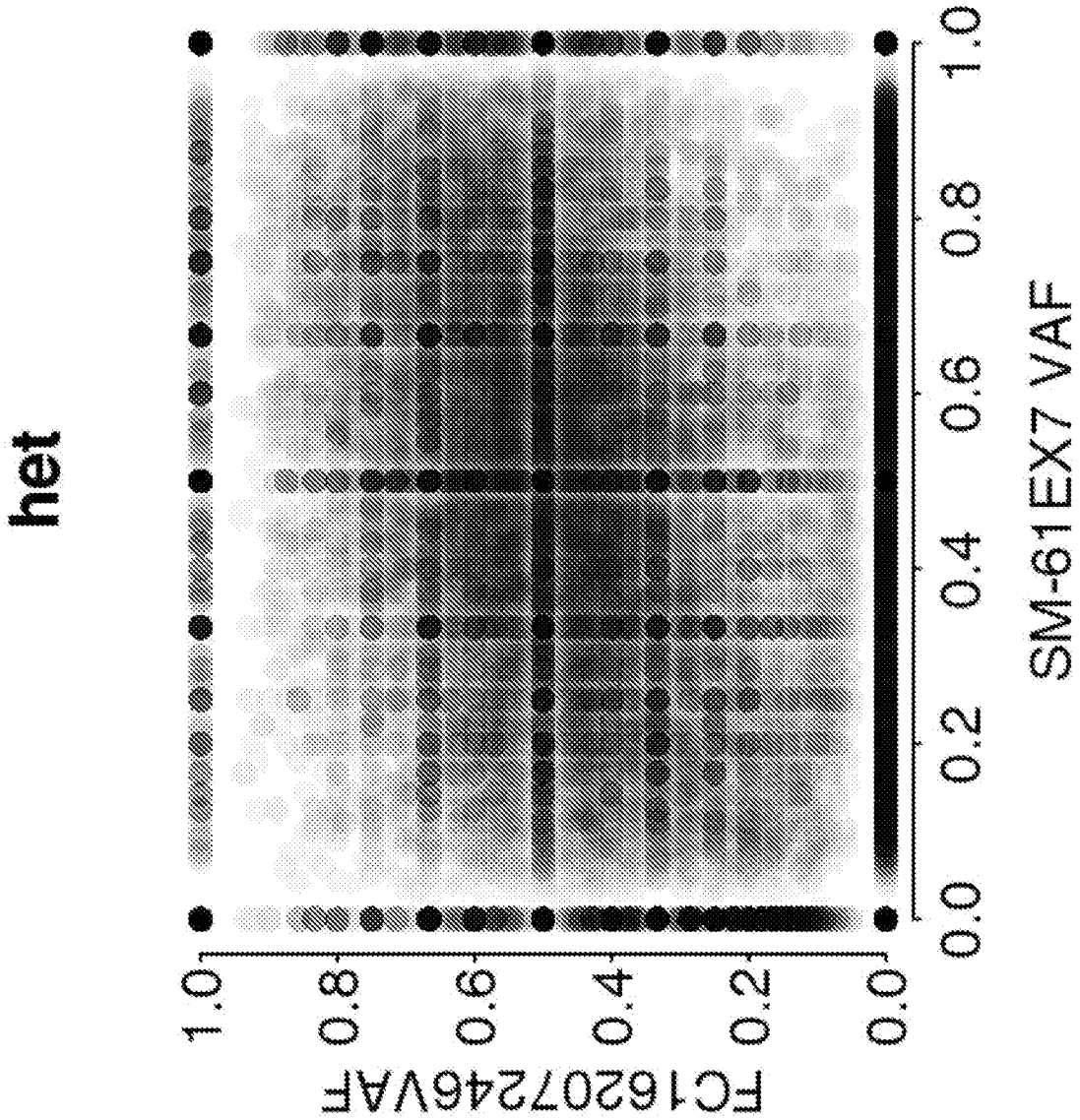
het
FC16207246VAF
1.0
0.8
0.6
0.4
0.2
0.0
SM-61EX7 VAF
0.0
0.2
0.4
0.6
0.8
1.0

Fig. 12F

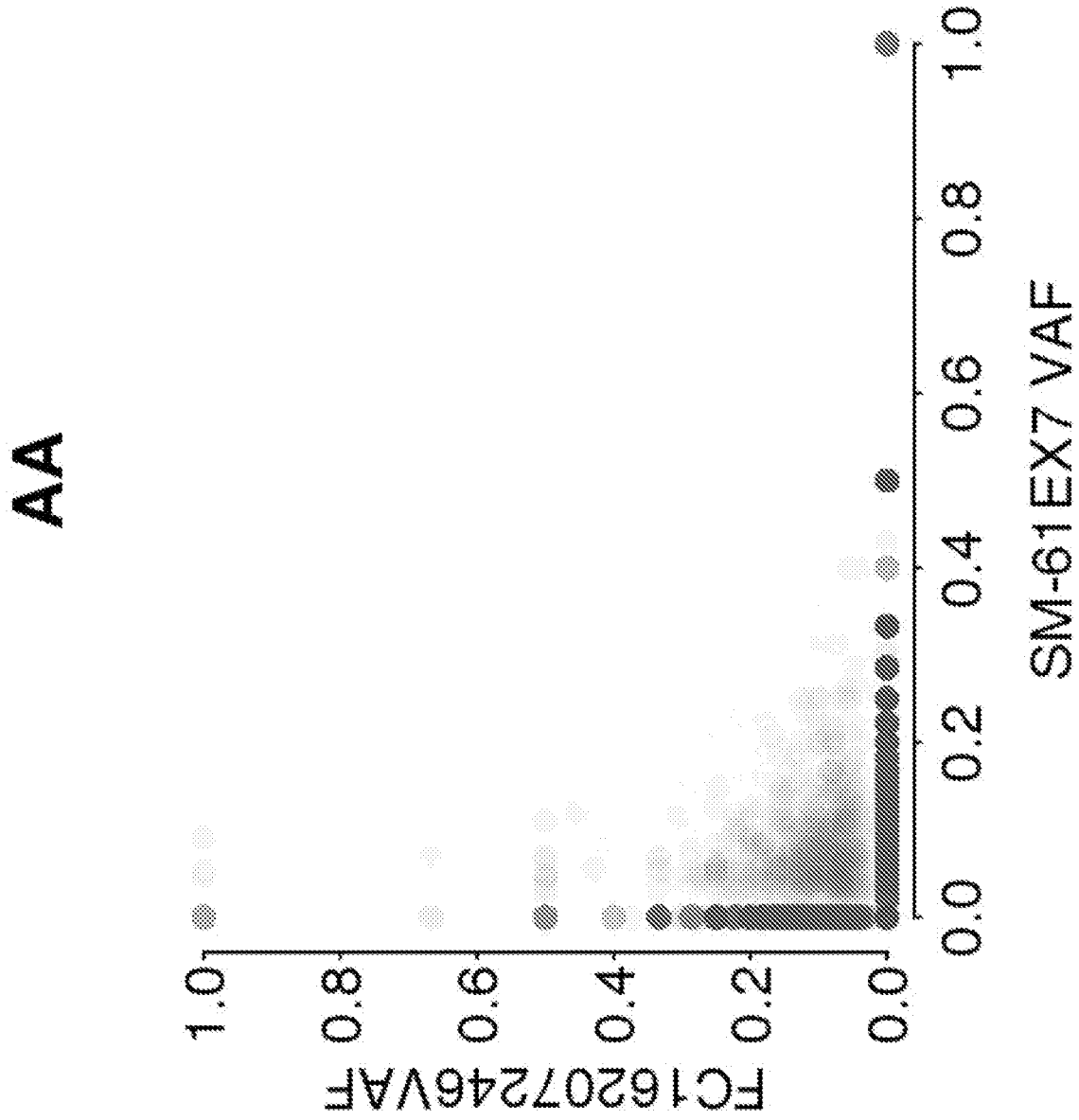

Fig. 12G

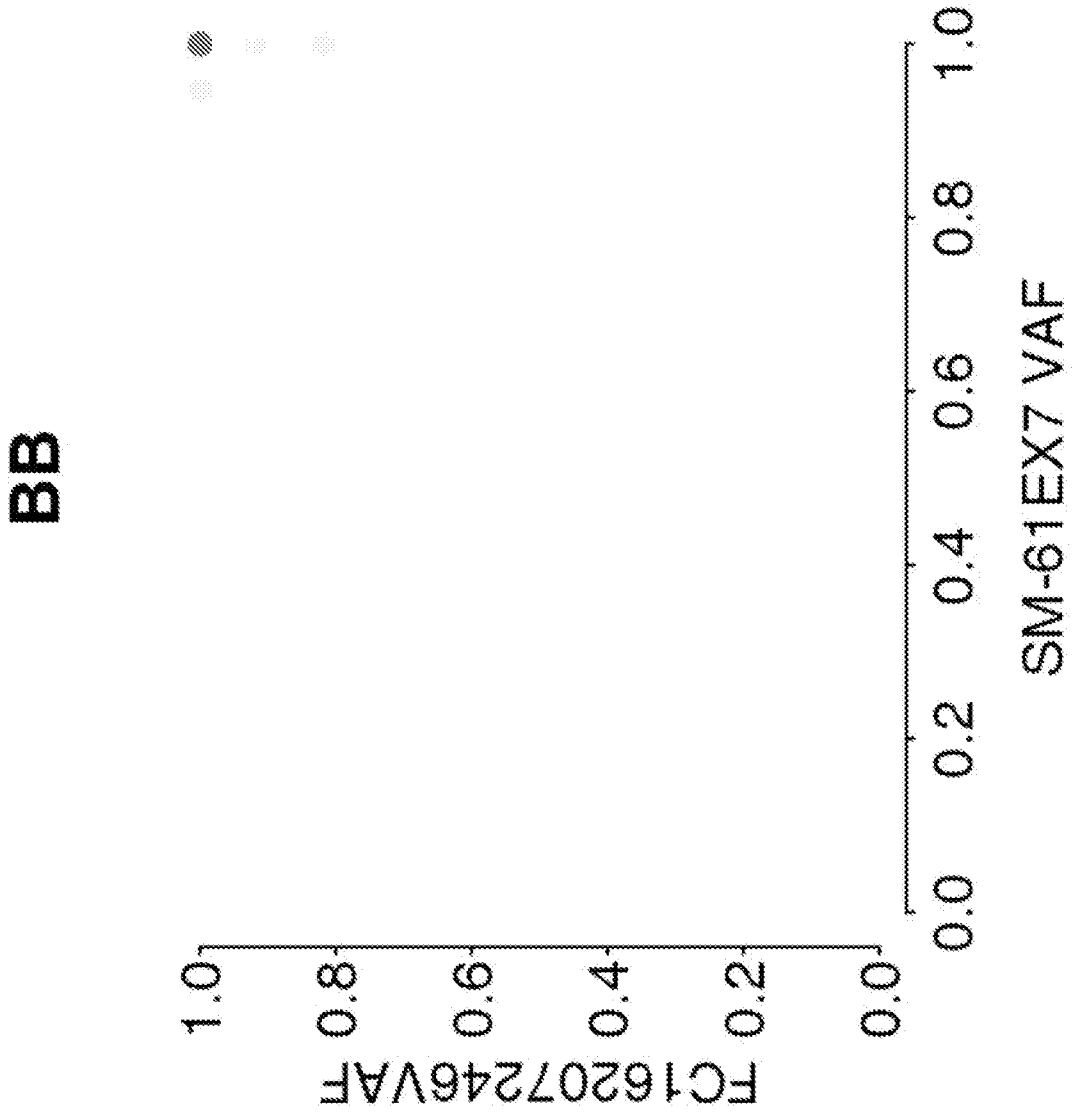

Fig. 12H

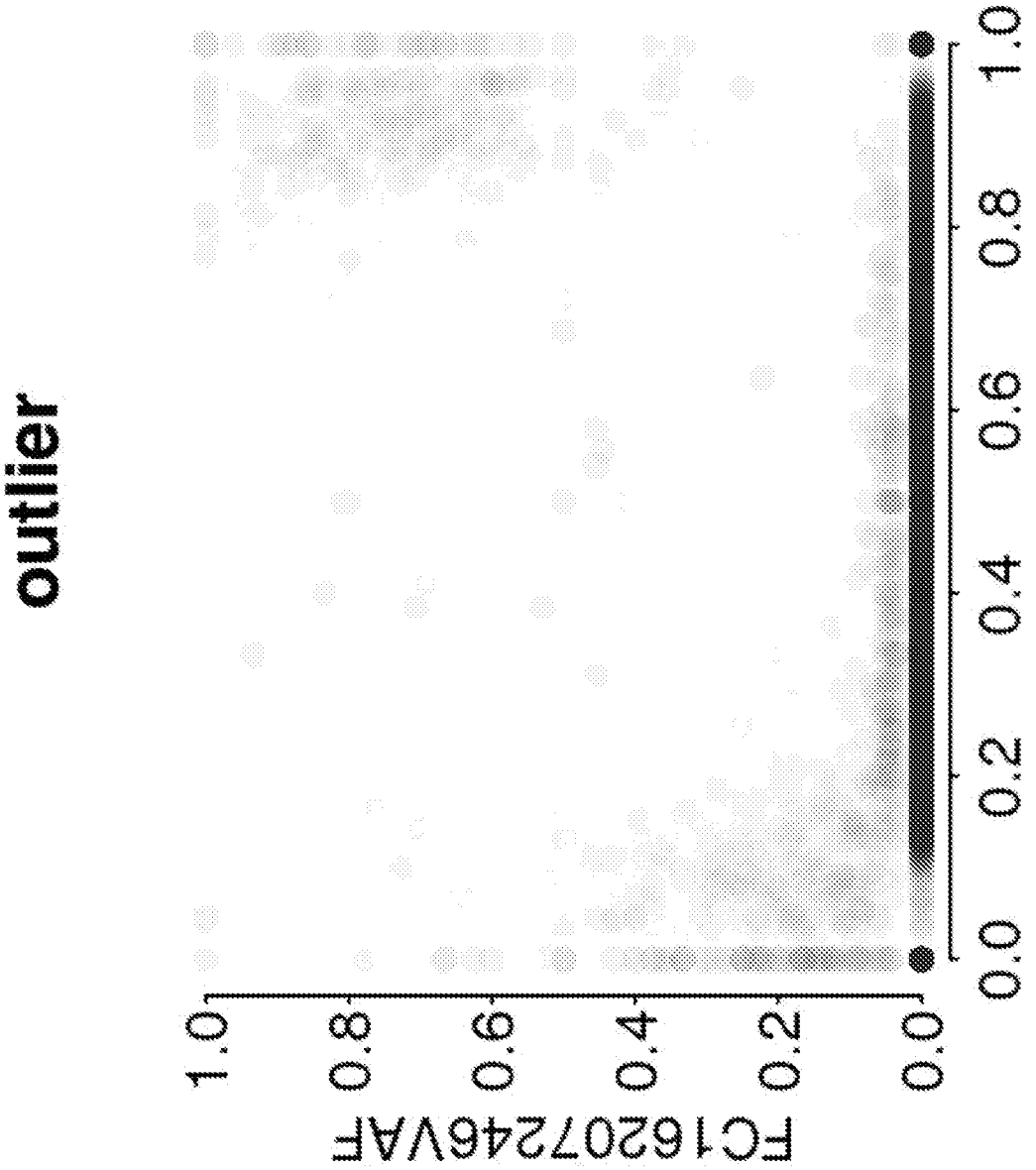
outlier
FC16207246VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
SM-61EX7 VAF

Fig. 12I

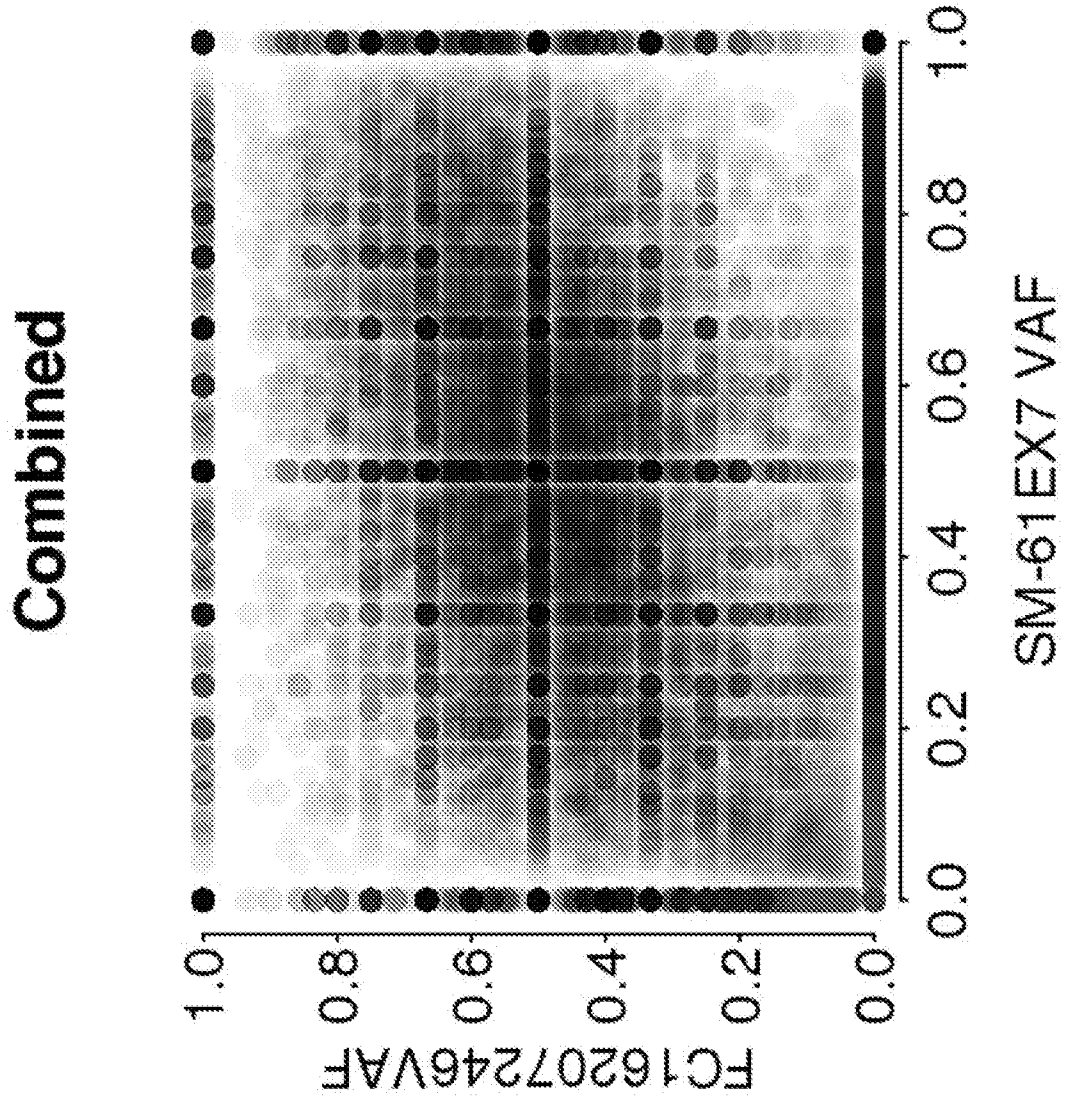

Fig. 12J

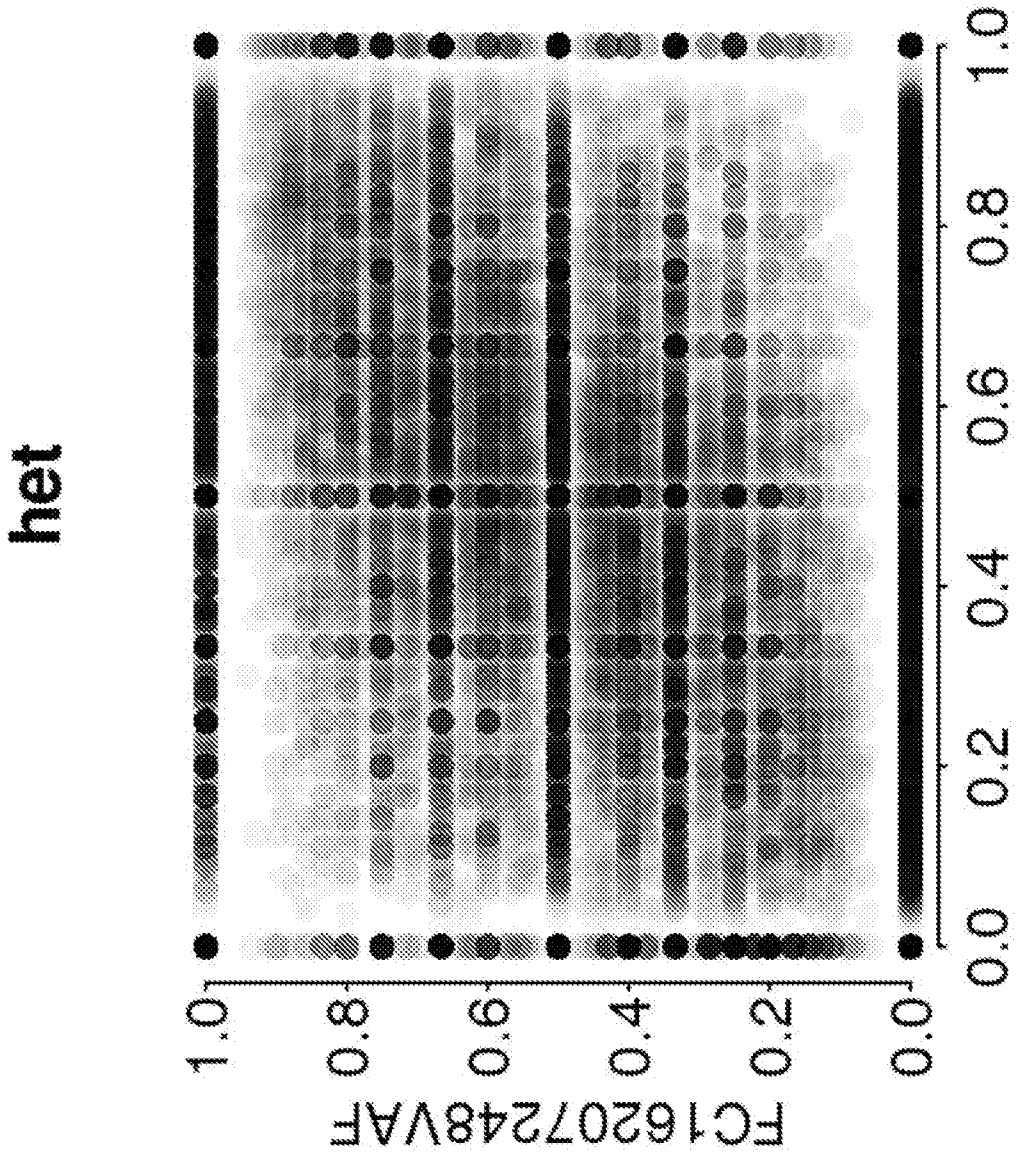
het
FC16207248VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
SM-61EX7 VAF

Fig. 12K

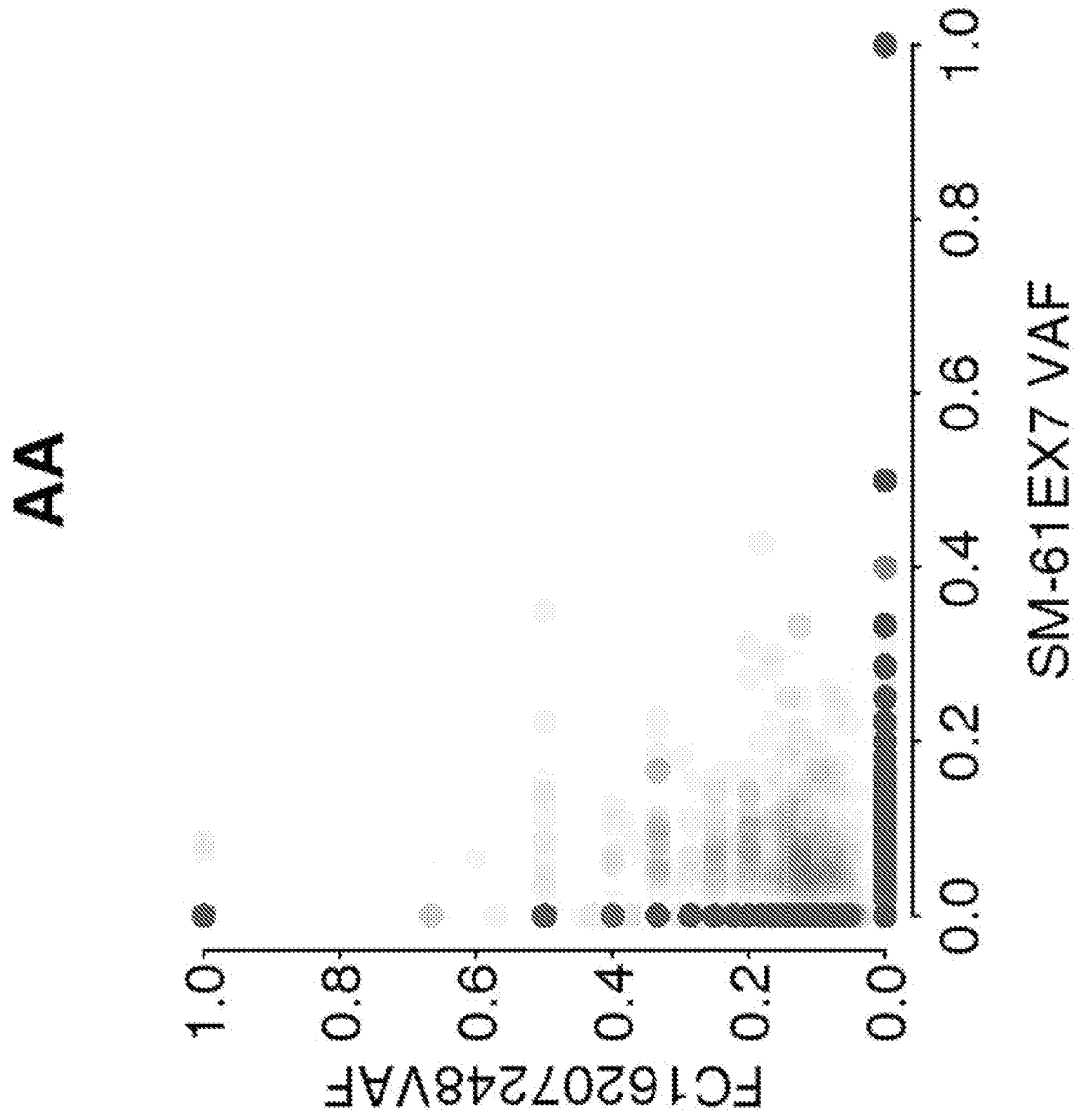

Fig. 12L

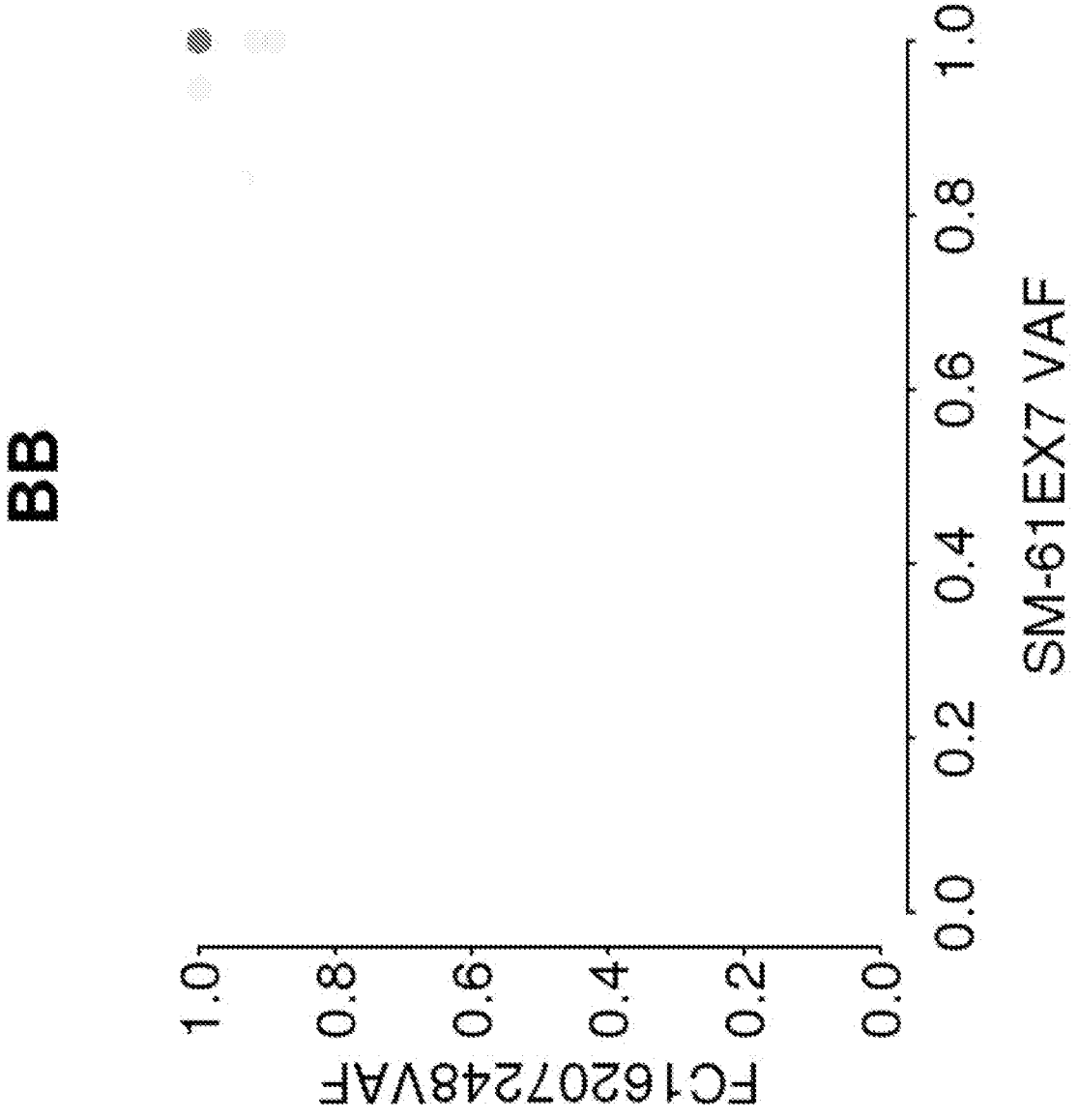

Fig. 12M

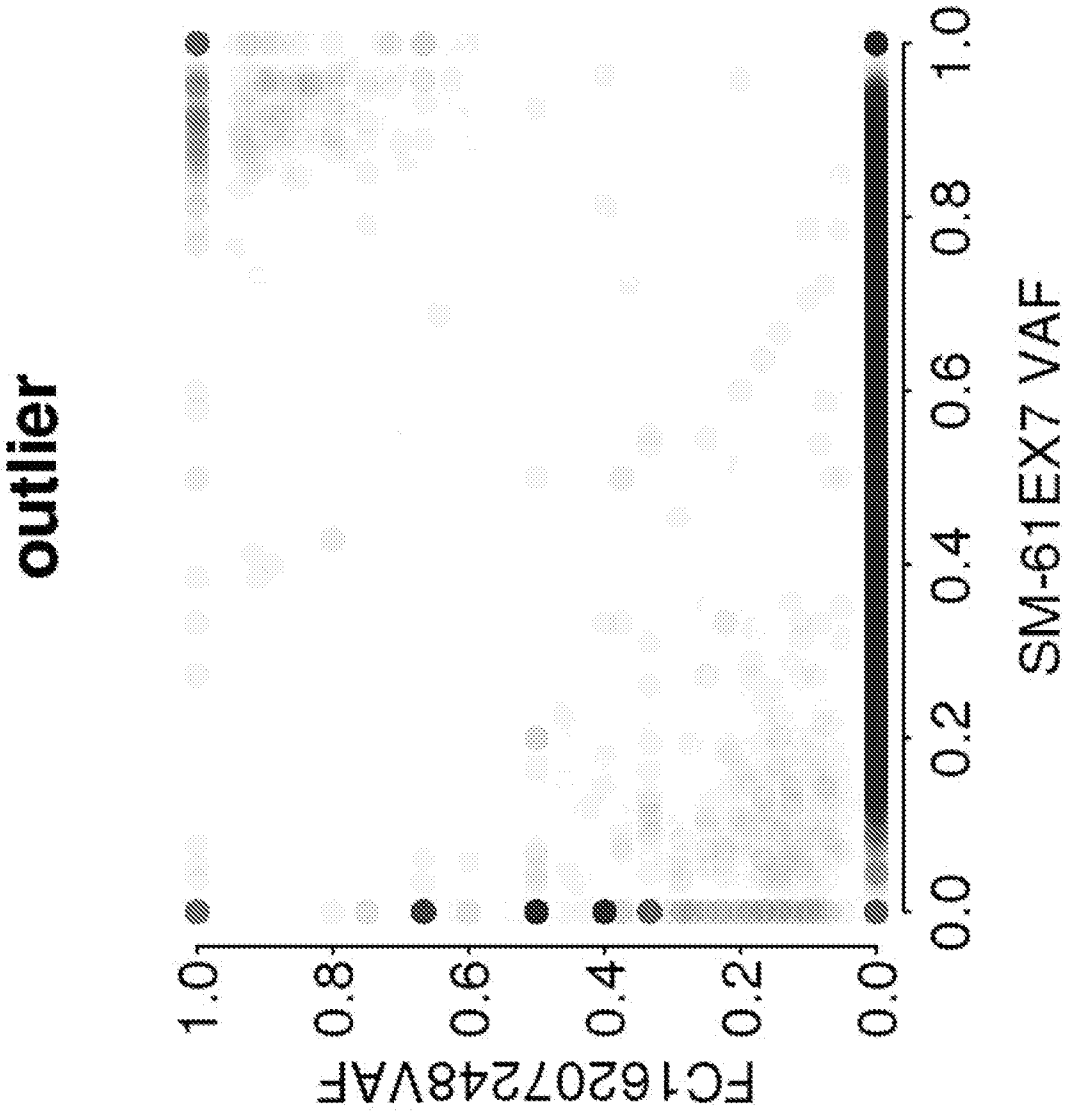
outlier
FC16207248VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
SM-61EX7 VAF

Fig. 12N

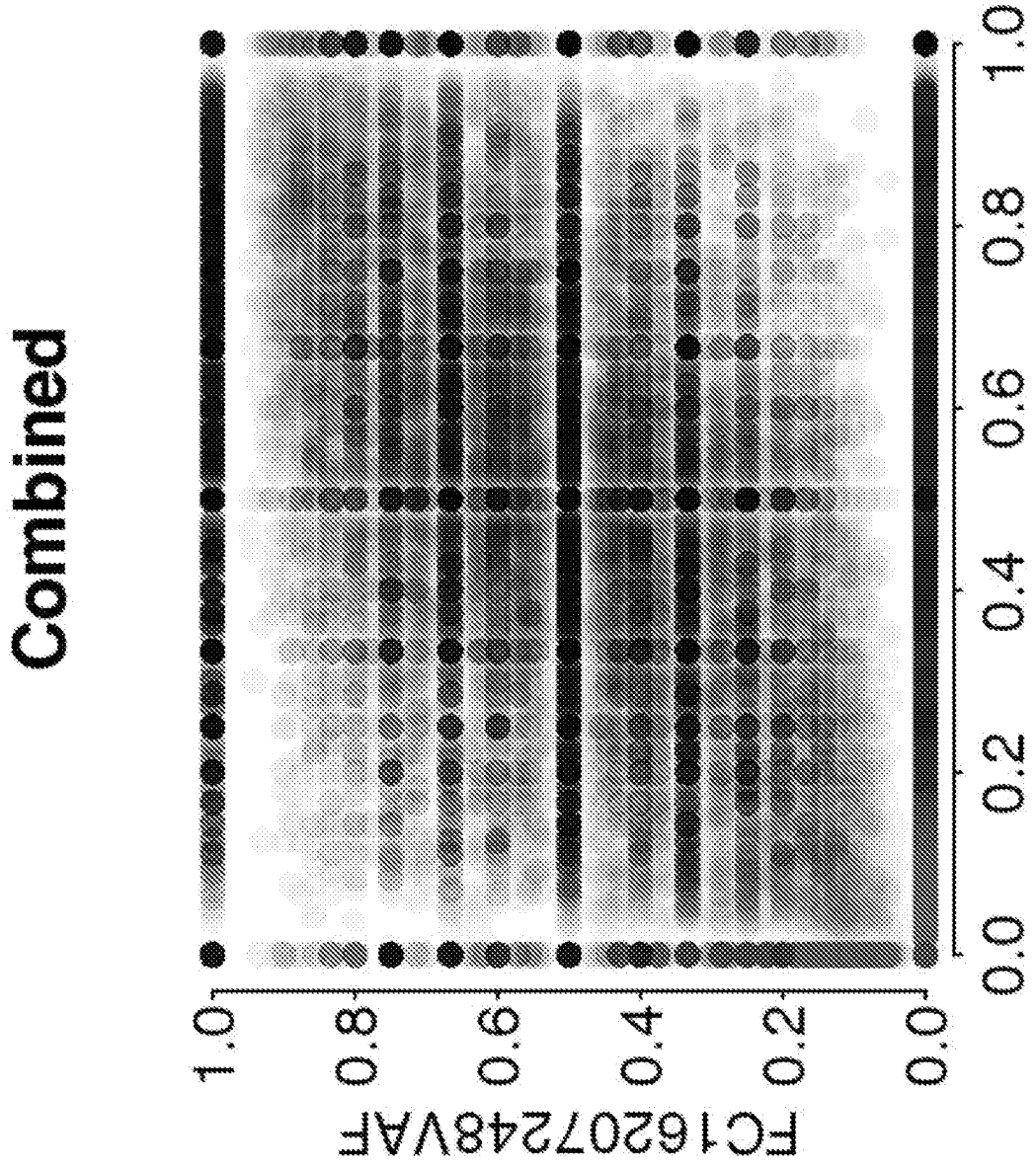
Combined
FC16207248VAF
1.0
0.8
0.6
0.4
0.2
0.0
SM-61EX7 VAF
0.0
0.2
0.4
0.6
0.8
1.0

Fig. 12O

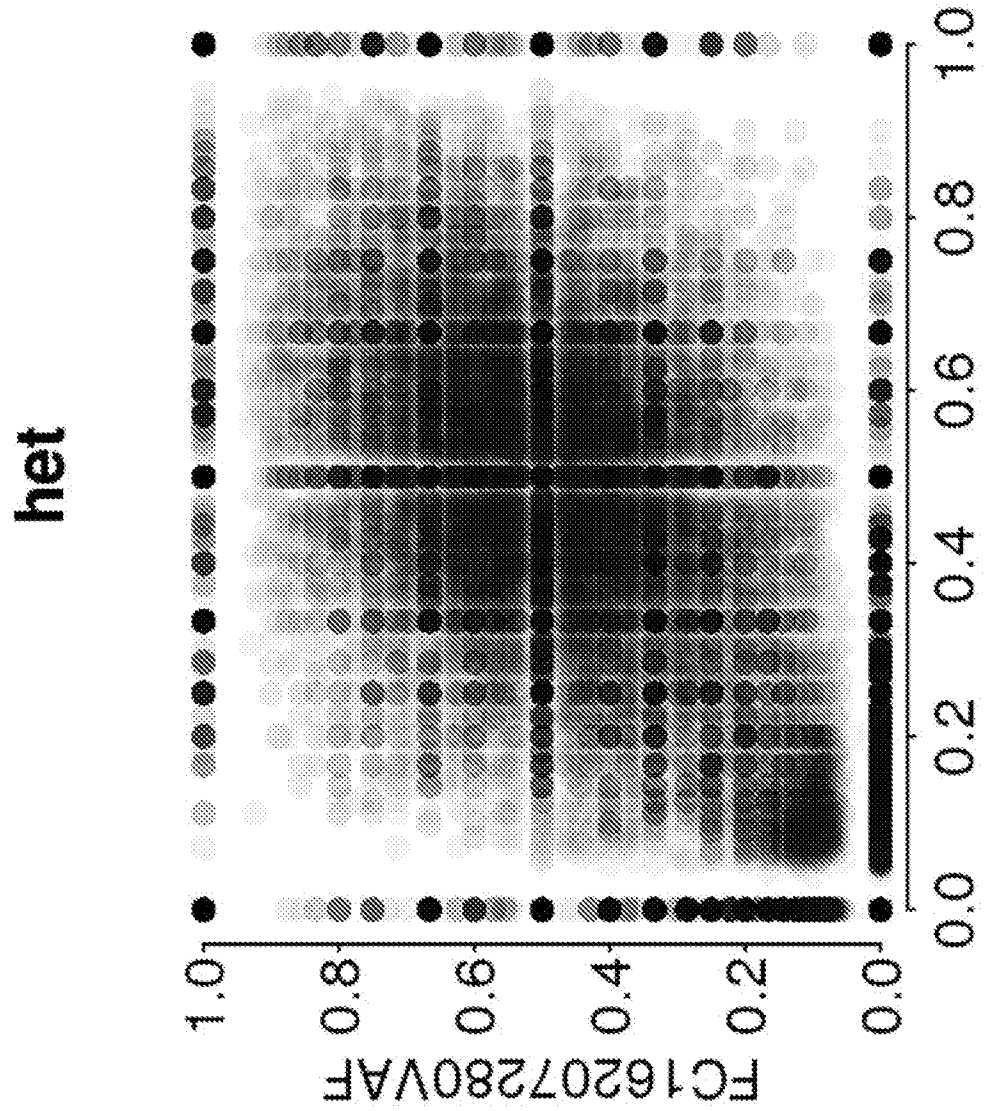
het
FC16207280VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FC16207281 VAF

Fig. 13A

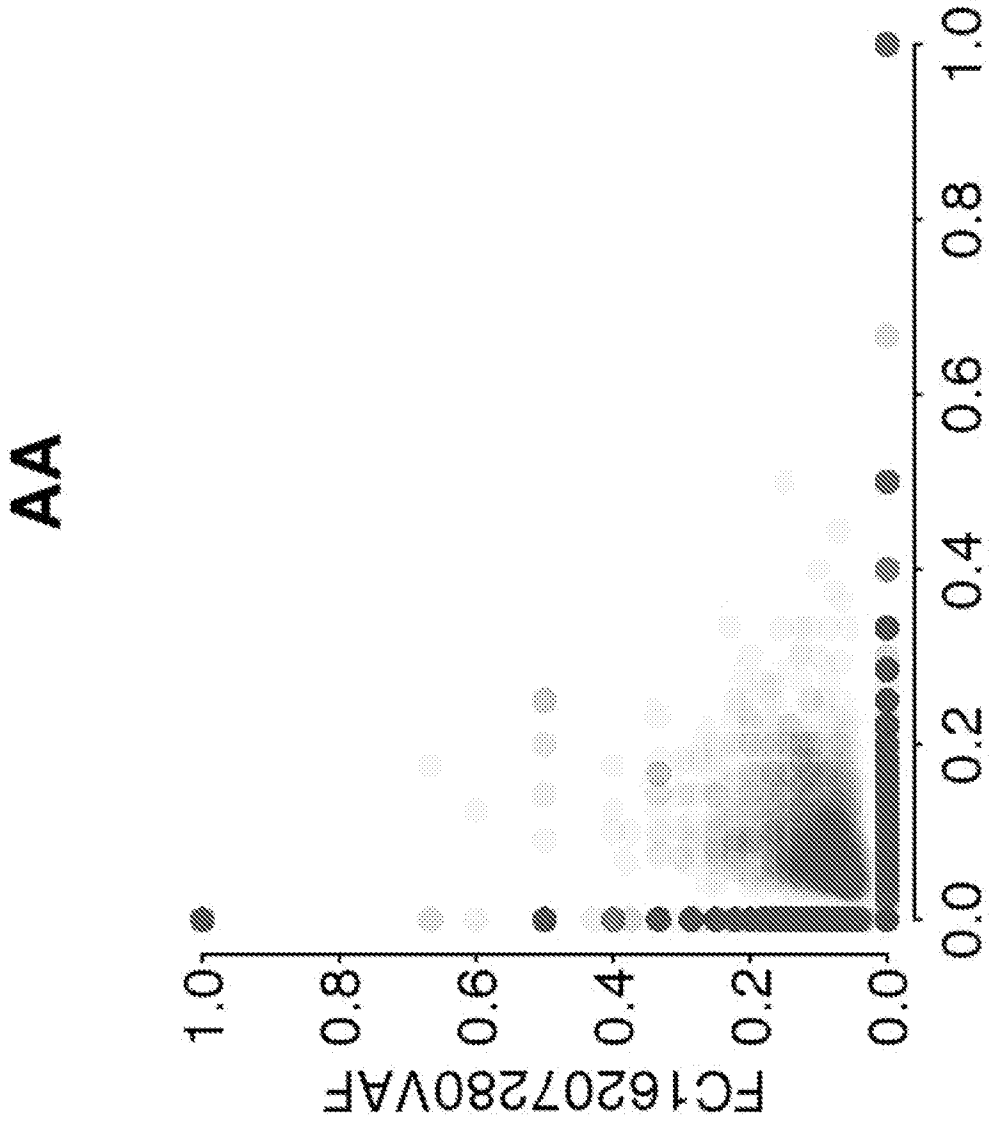
AA
FC16207280VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FC16207281 VAF

Fig. 13B

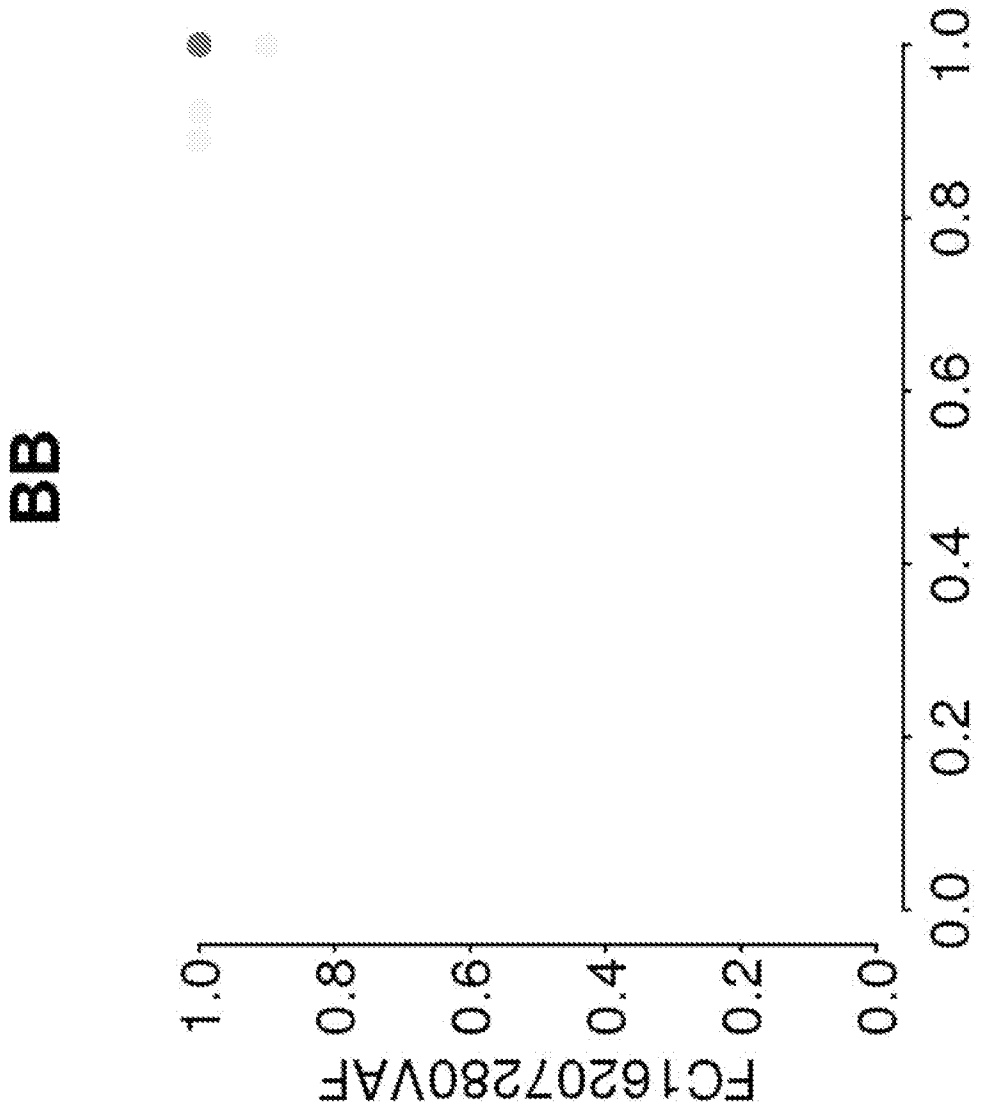
BB
FC16207280VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FC16207281 VAF

Fig. 13C

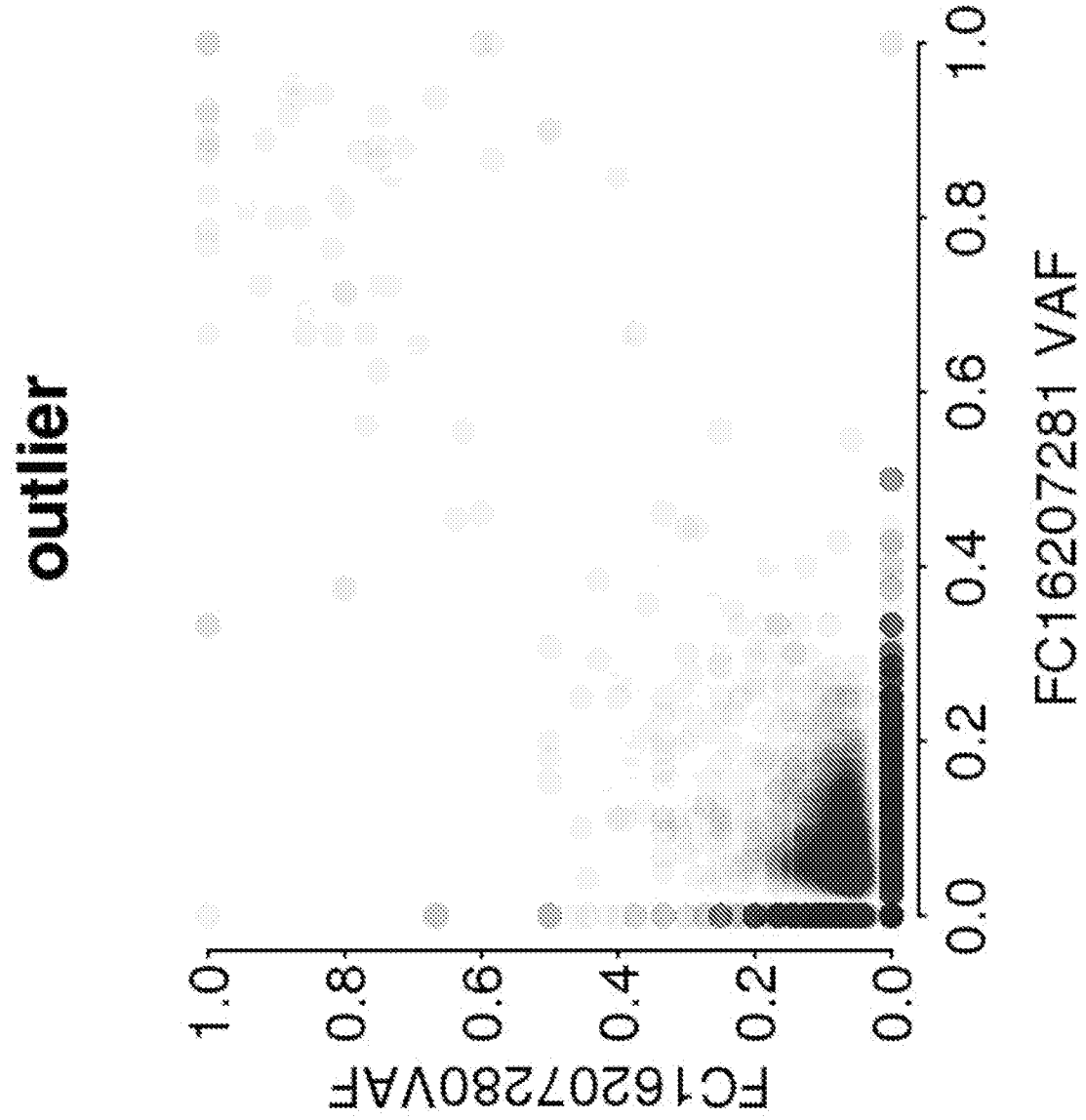
outlier
FC16207280VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FC16207281 VAF

Fig. 13D

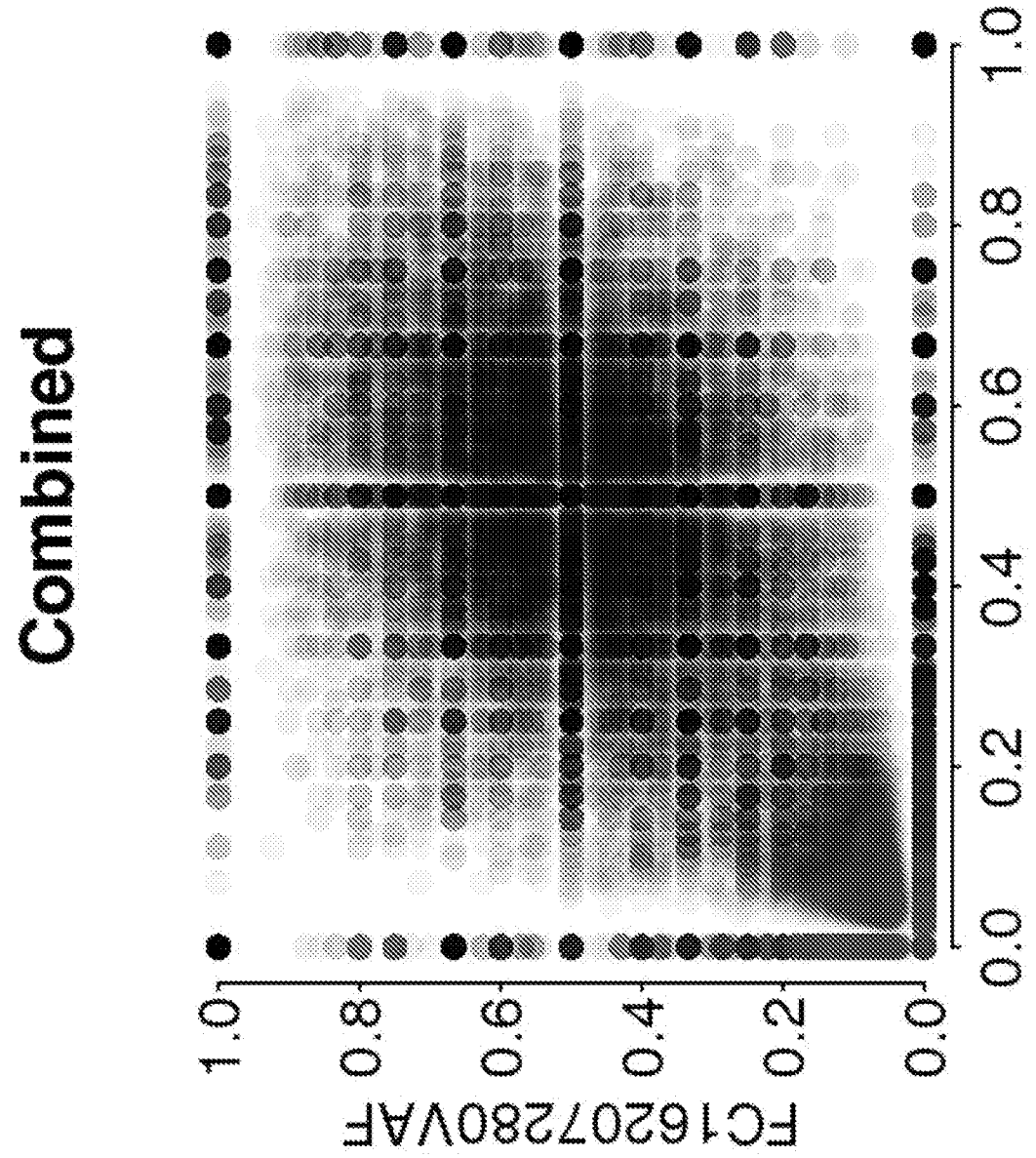
Combined
FC16207280VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FC16207281 VAF

Fig. 13E

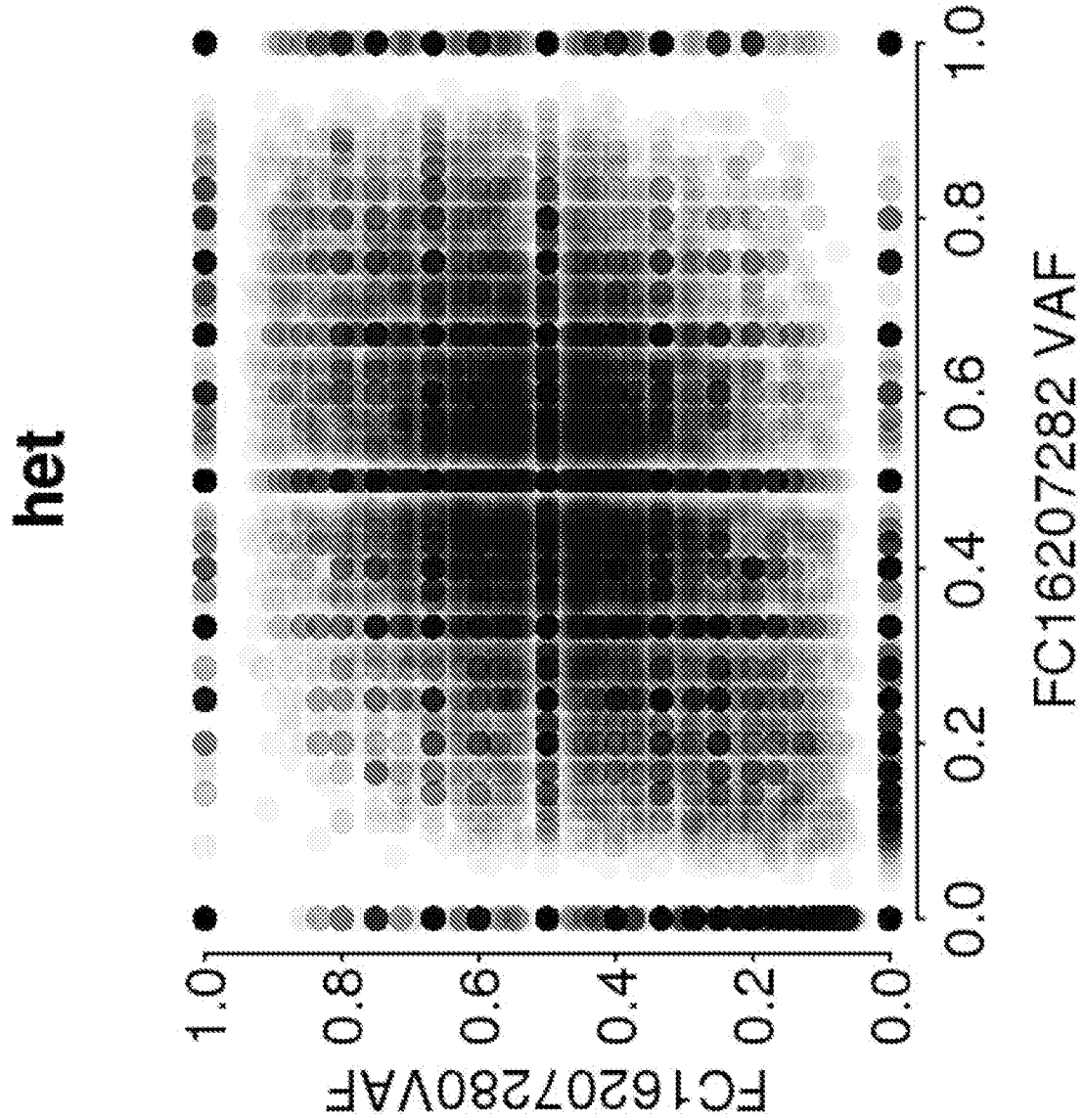
het
FC16207280VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FC16207282 VAF

Fig. 13F

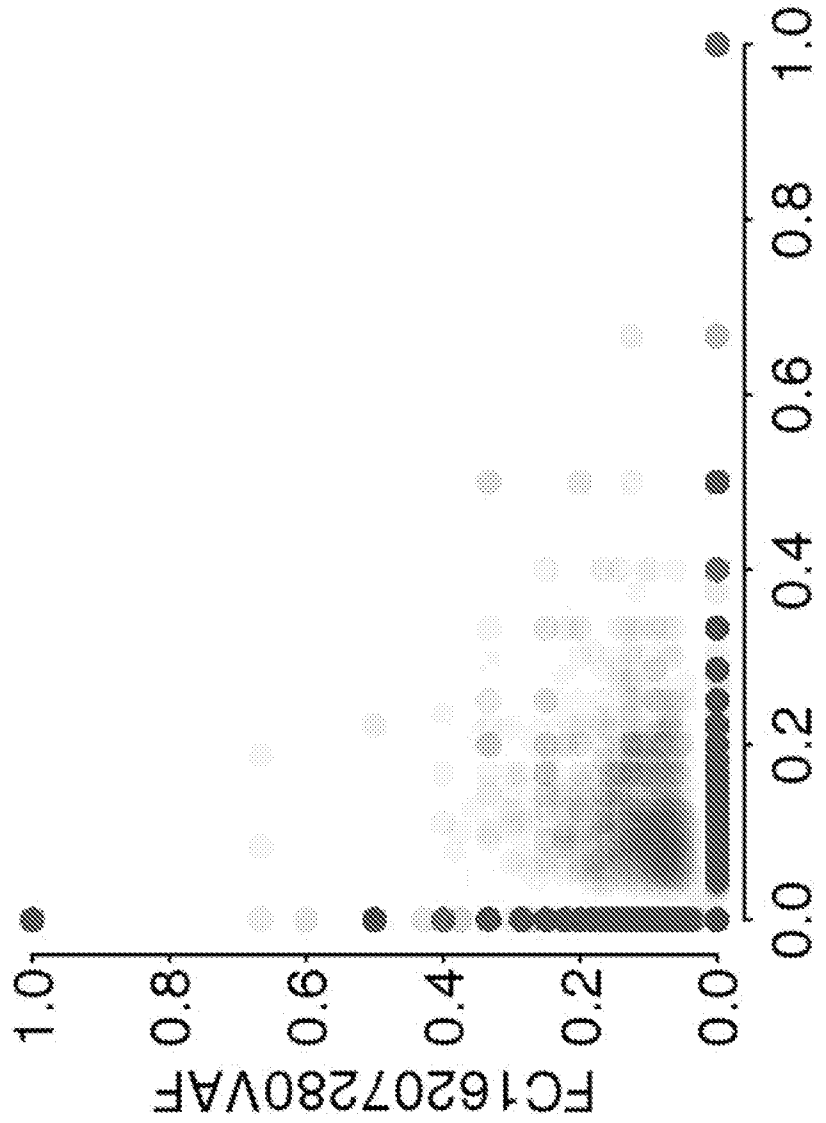
AA
FC16207280VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FC16207282 VAF

Fig. 13G

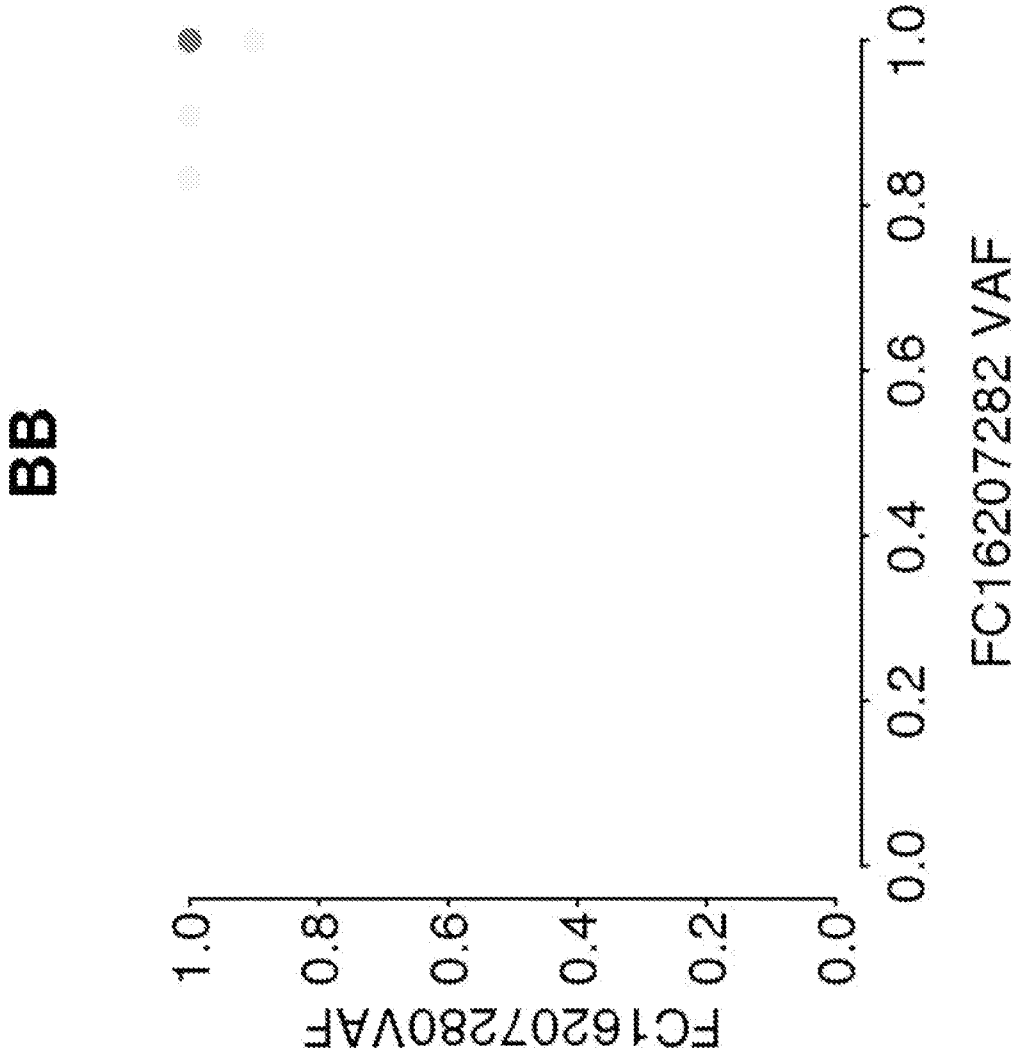
BB
FC16207280VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FC16207282 VAF

Fig. 13H

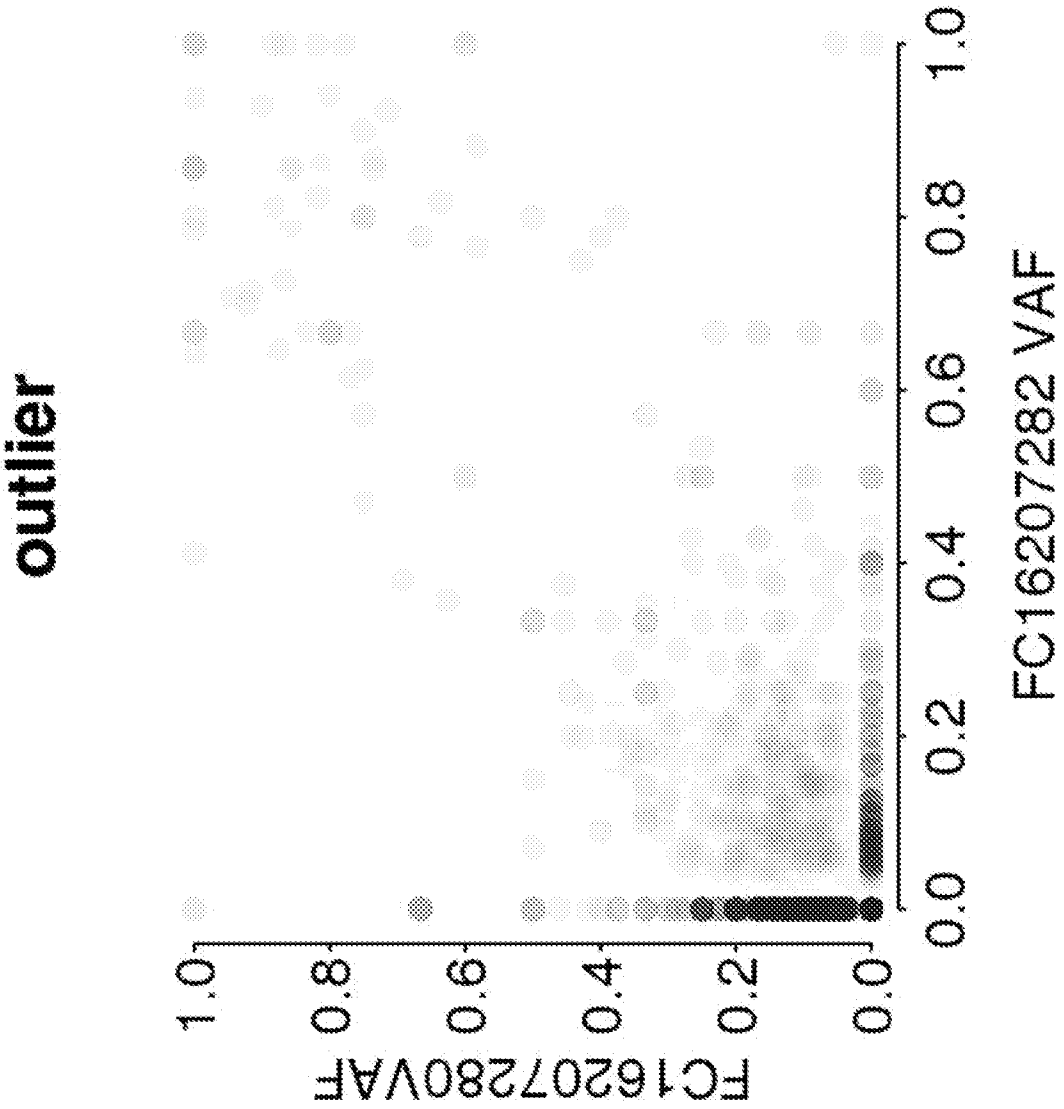
outlier
FC16207280VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FC16207282 VAF

Fig. 13I

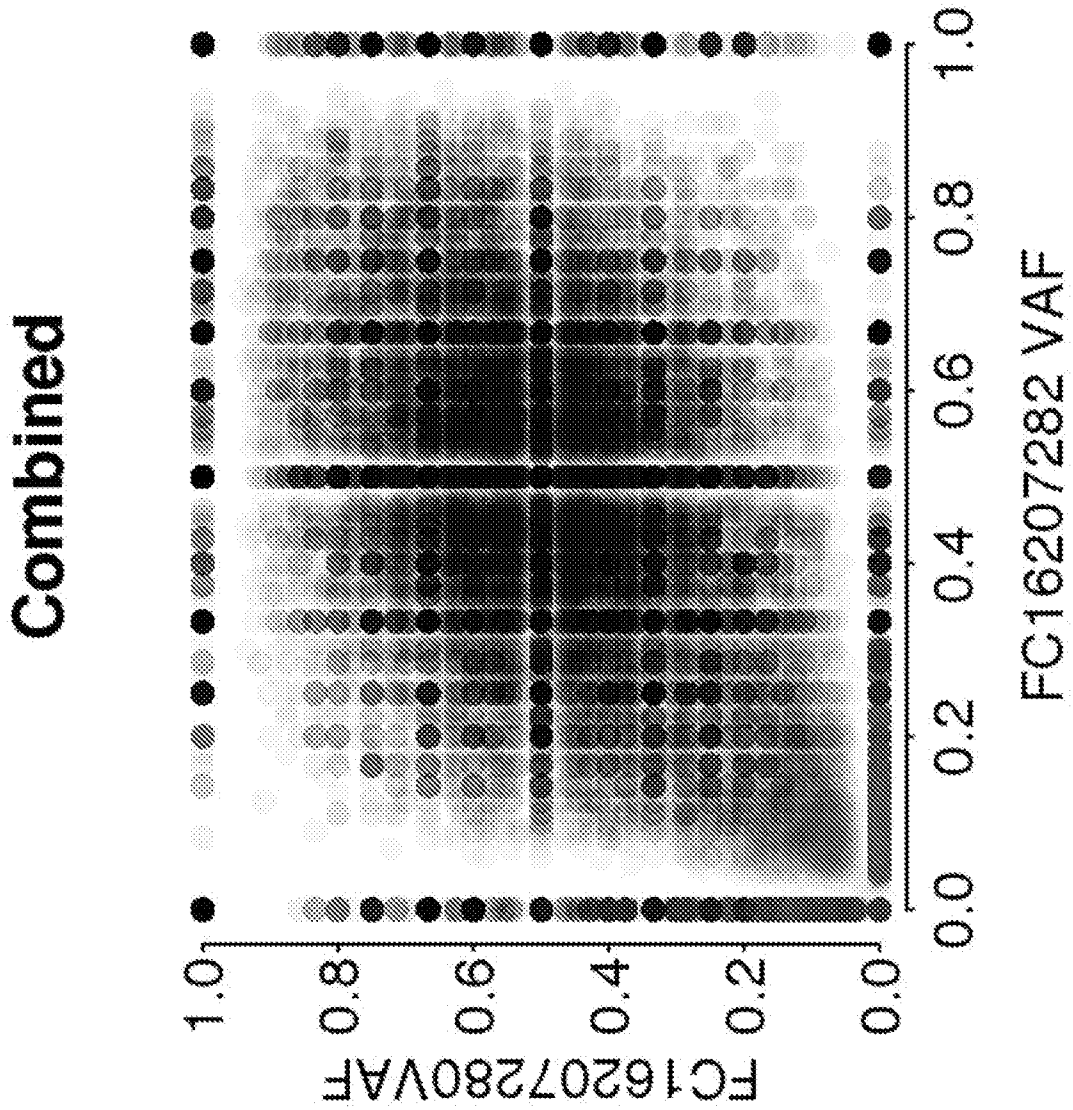
Combined
FC16207280VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FC16207282 VAF

Fig. 13J

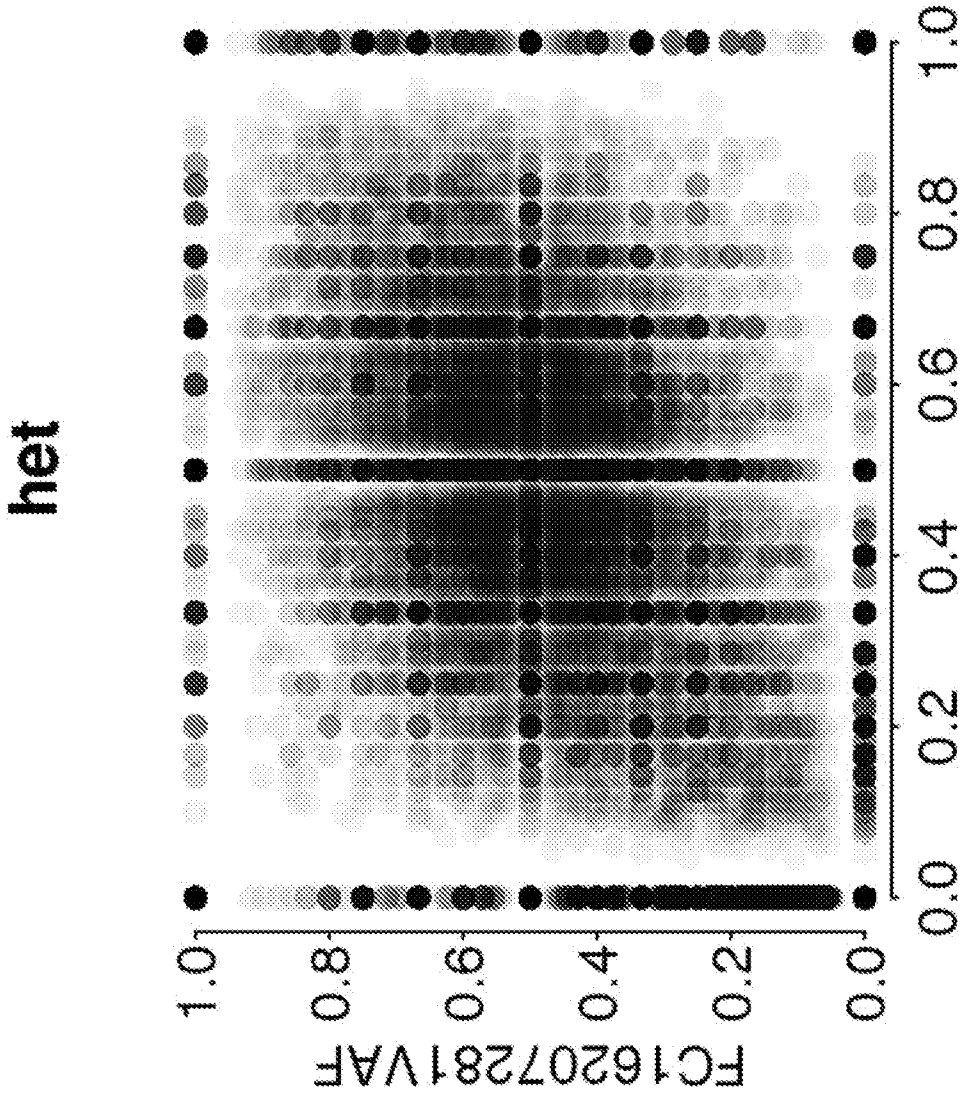
het
FC16207281VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FC16207282 VAF

Fig. 13K

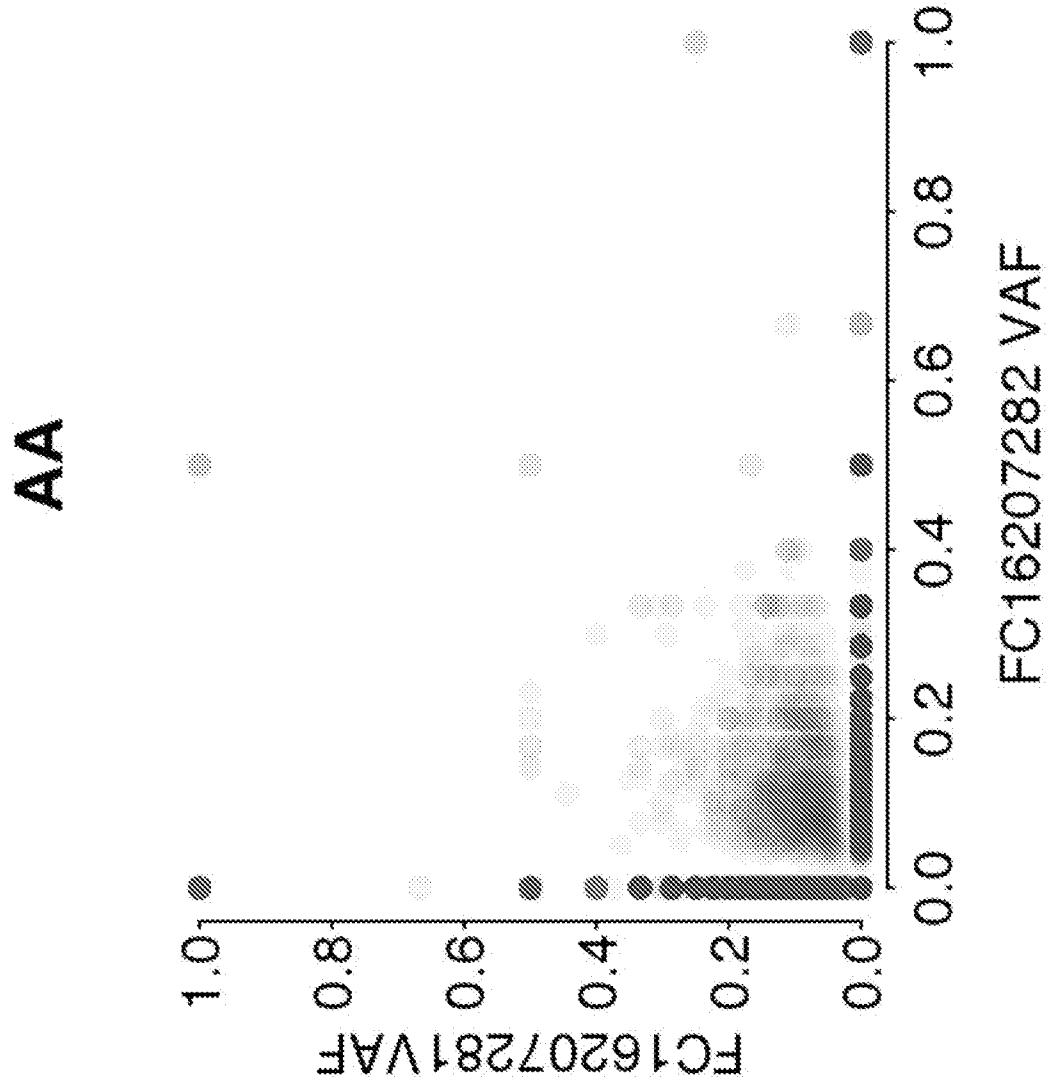
AA
FC16207281VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FC16207282 VAF

Fig. 13L

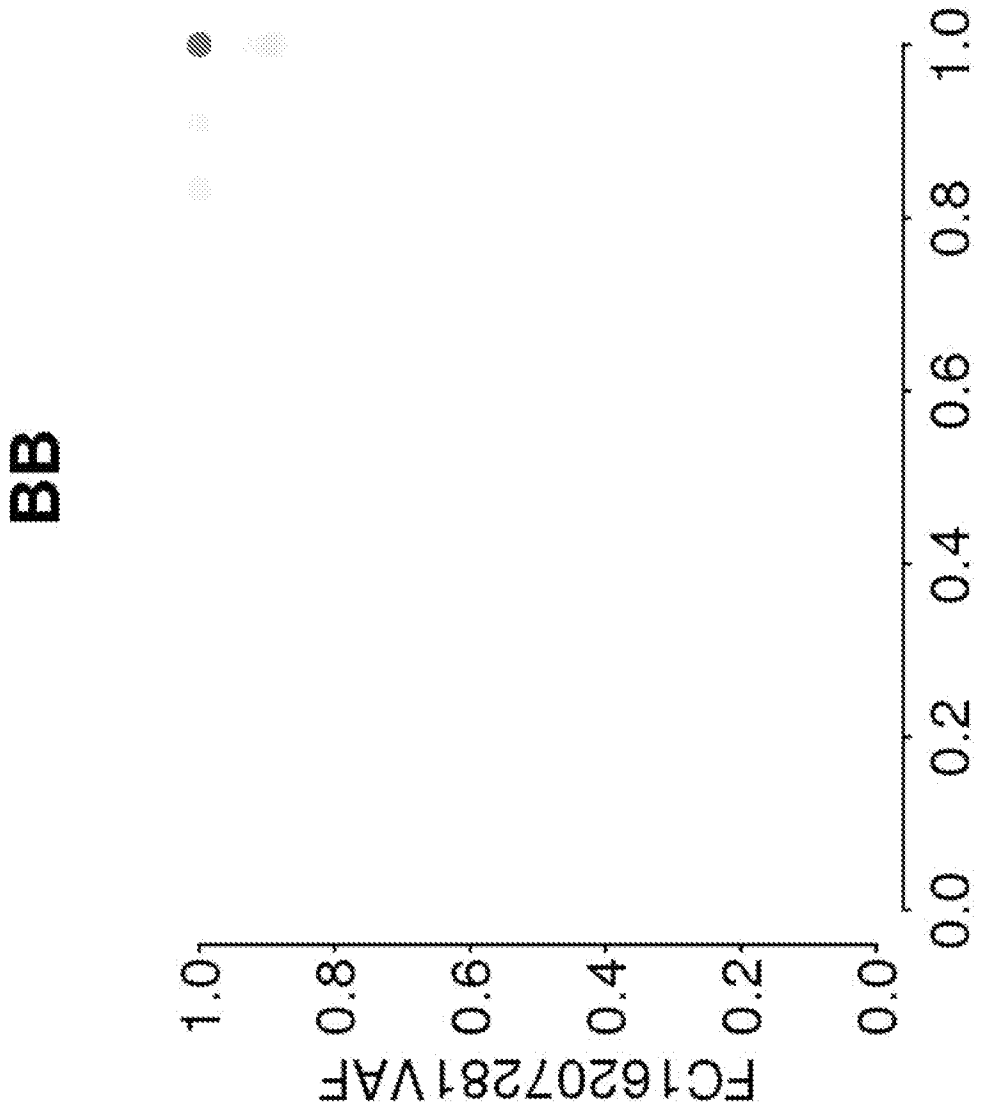
BB
FC16207281VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FC16207282 VAF

Fig. 13M

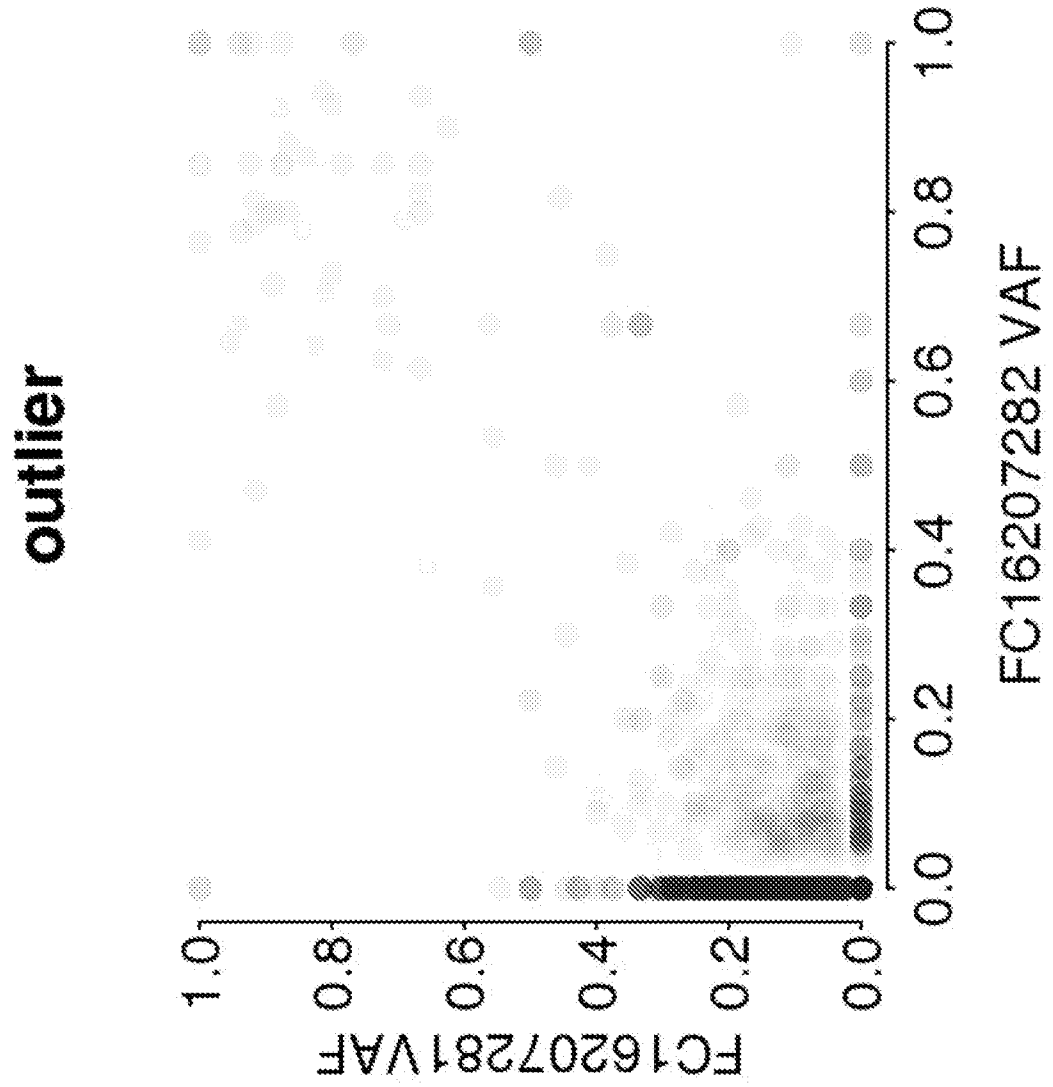
outlier
FC16207281VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FC16207282 VAF

Fig. 13N

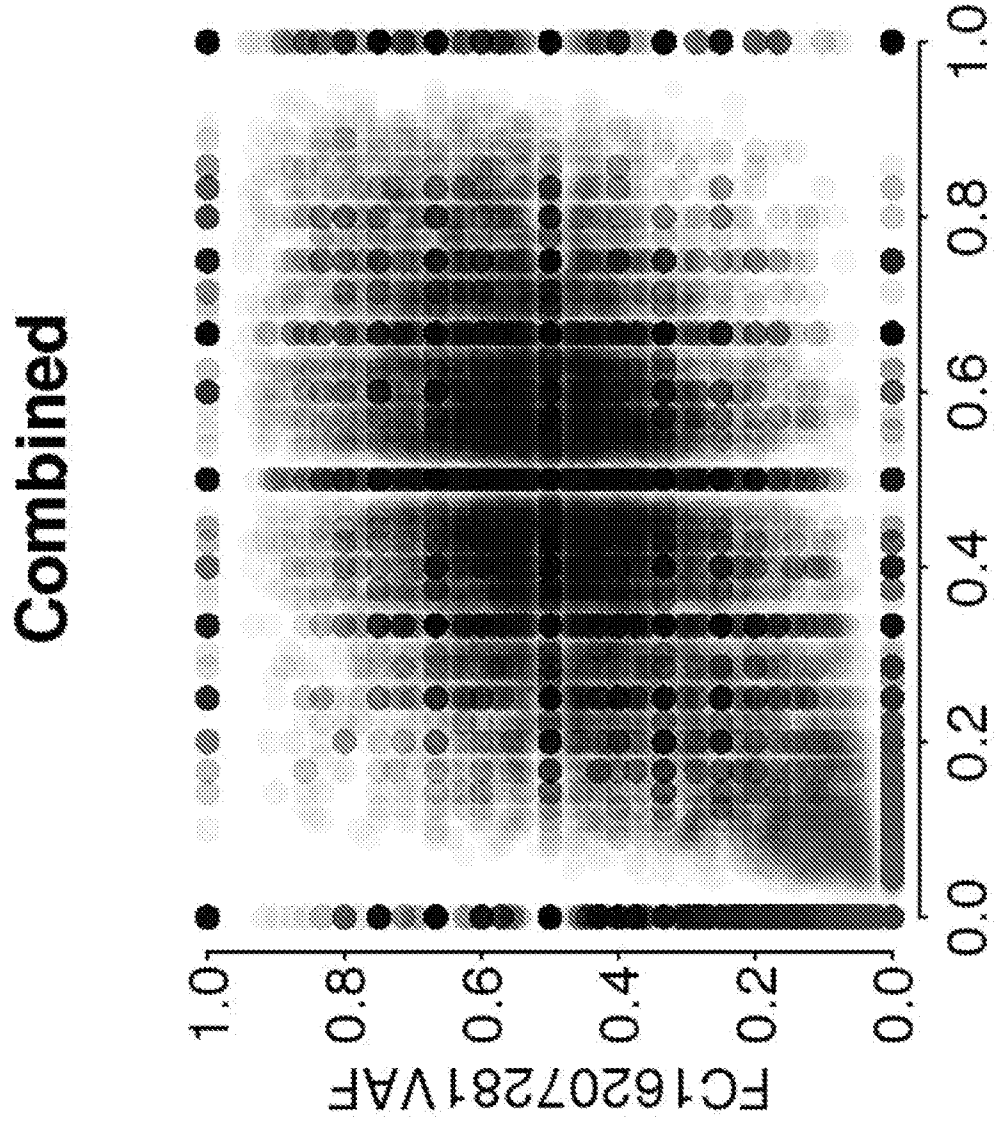
Combined
FC16207281VAF
1.0
0.8
0.6
0.4
0.2
0.0
FC16207282 VAF
0.0
0.2
0.4
0.6
0.8
1.0

Fig. 13O

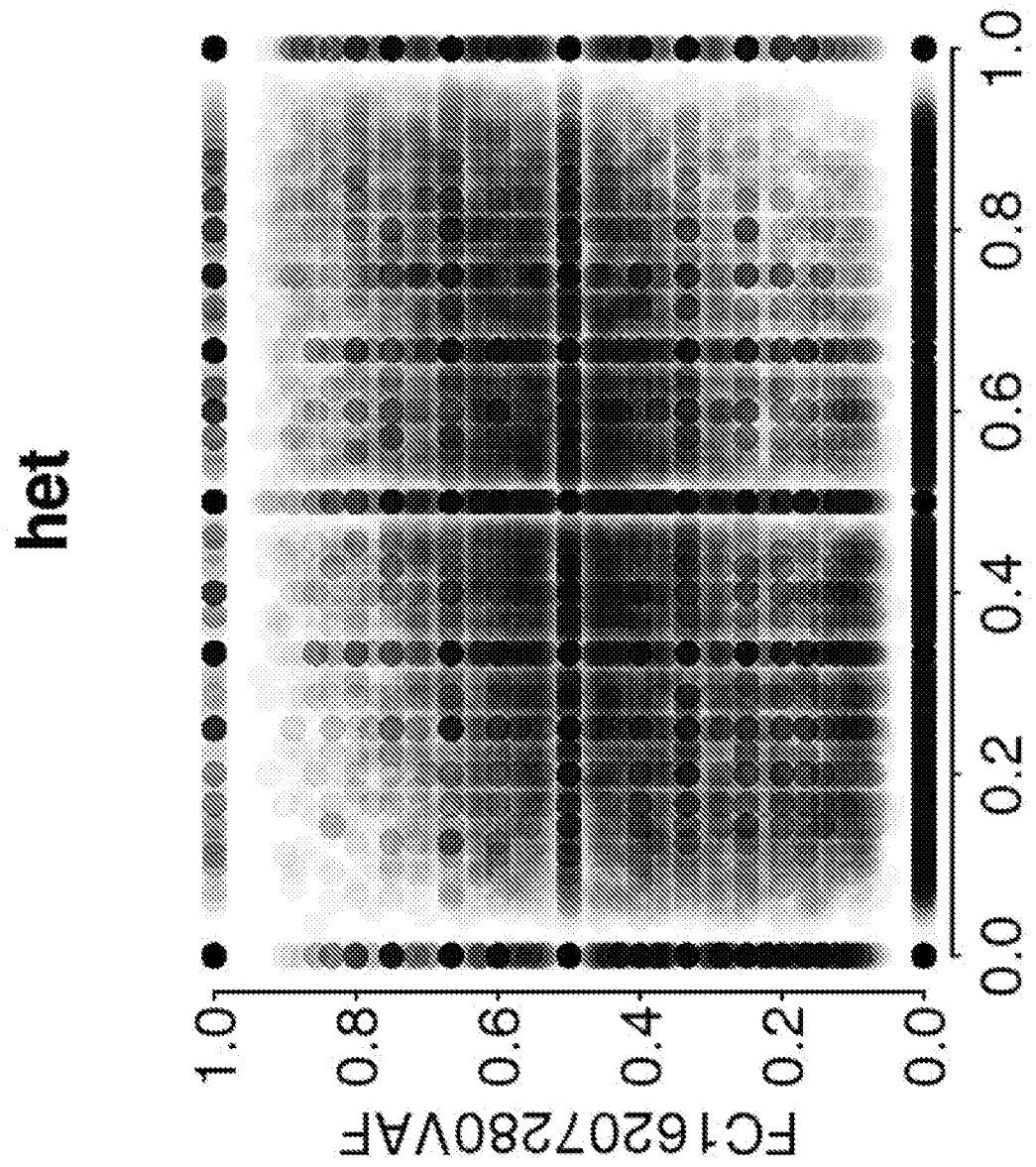
het
FC16207280VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
SM-9ZE13 VAF

Fig. 13P

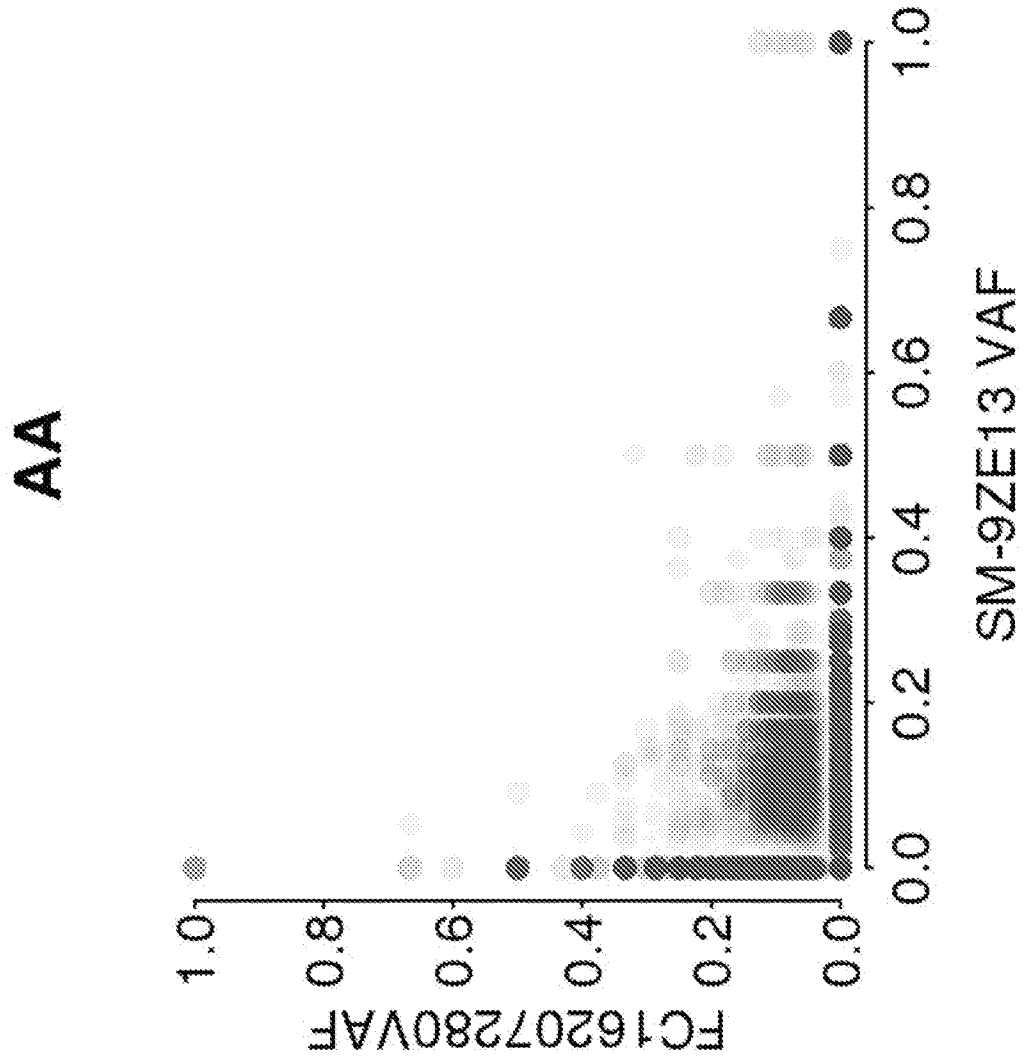

Fig. 13Q

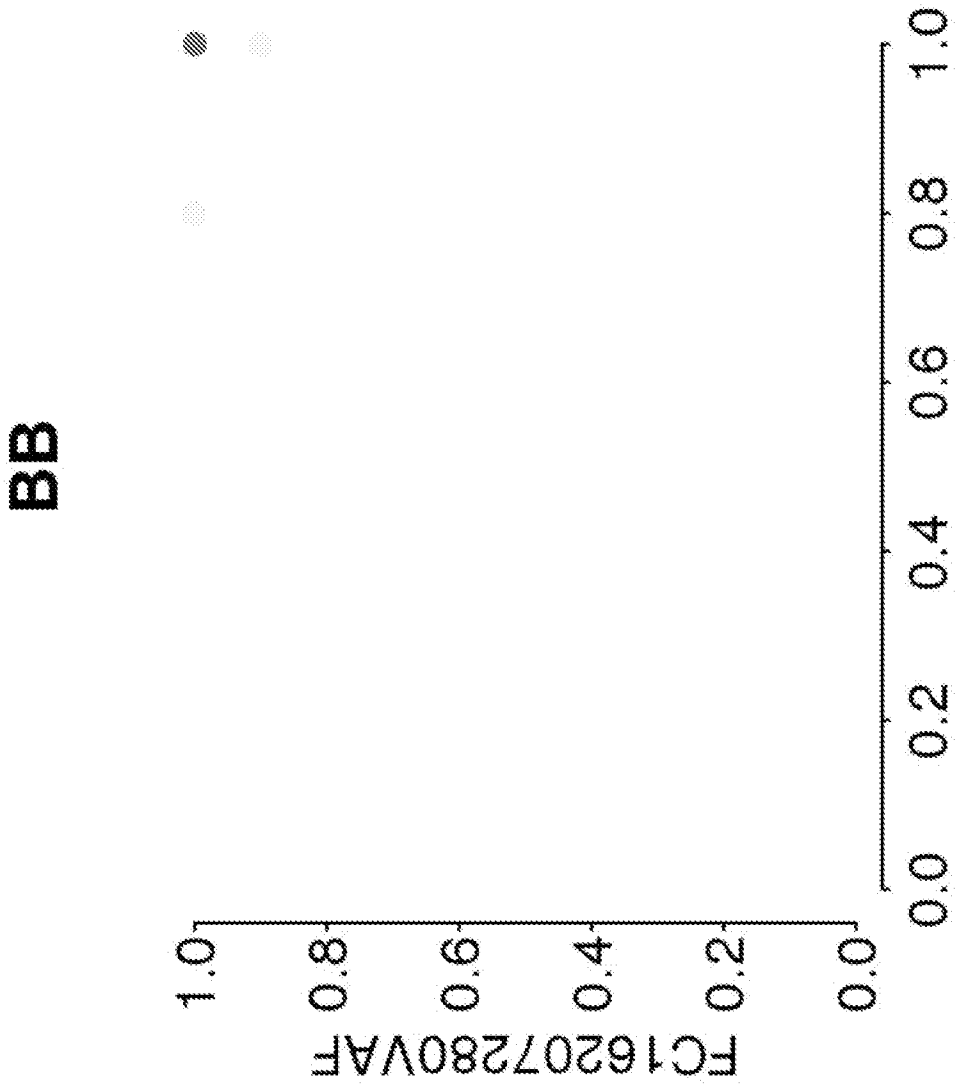

Fig. 13R

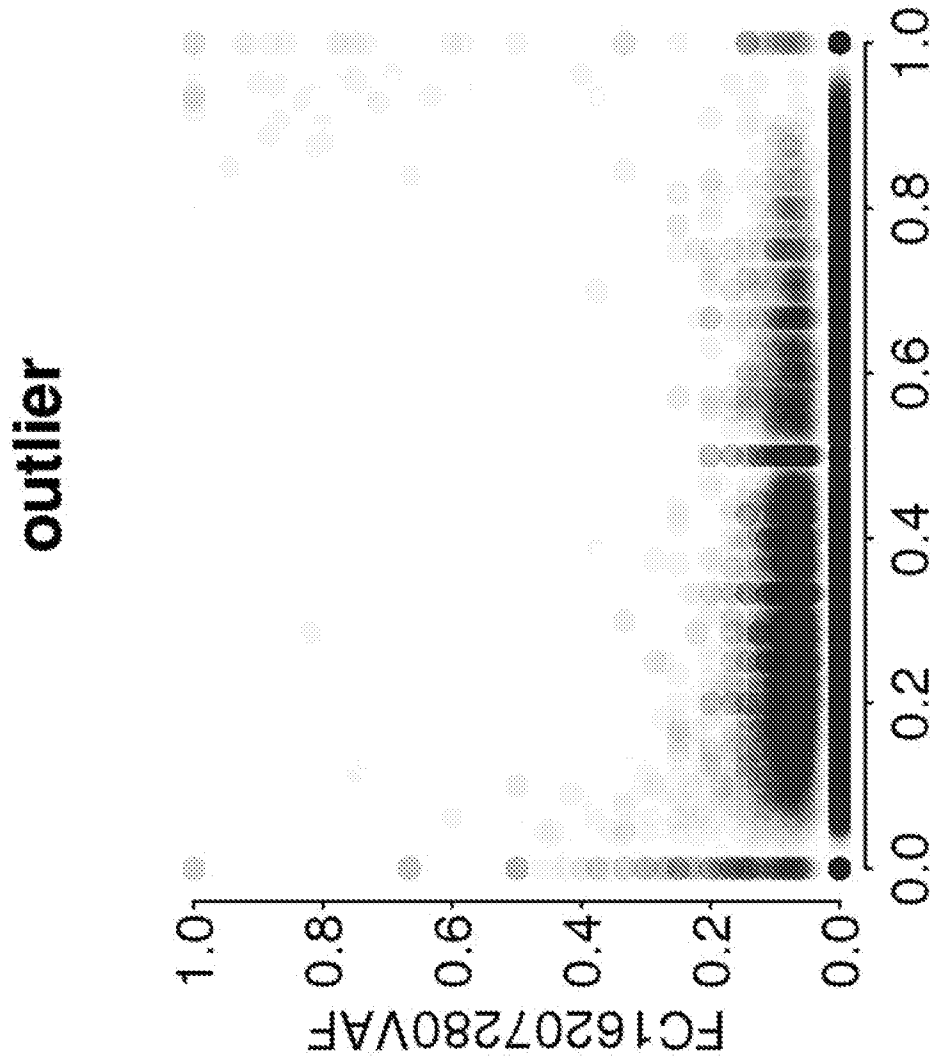
outlier
FC16207280VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
SM-9ZE13 VAF

Fig. 13S

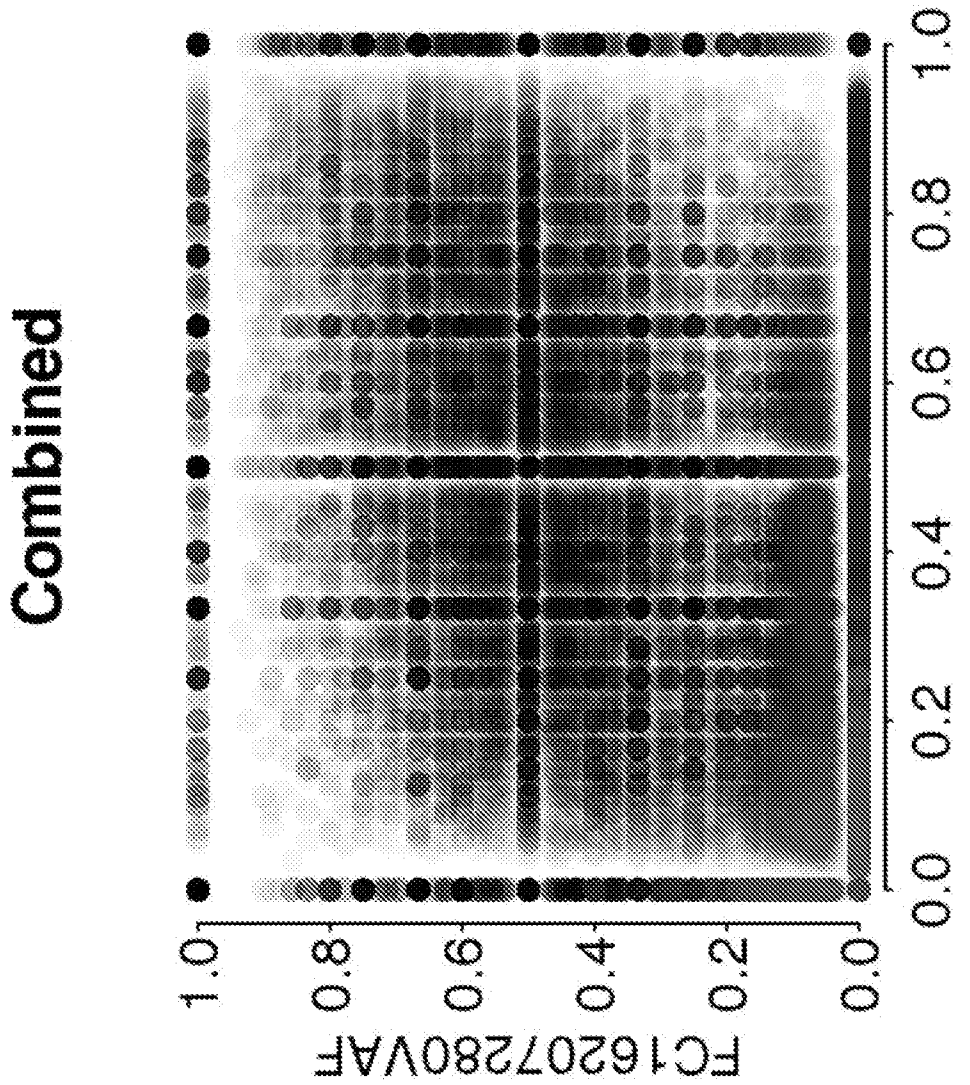
Combined
FC16207280VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
SM-9ZE13 VAF

Fig. 13T

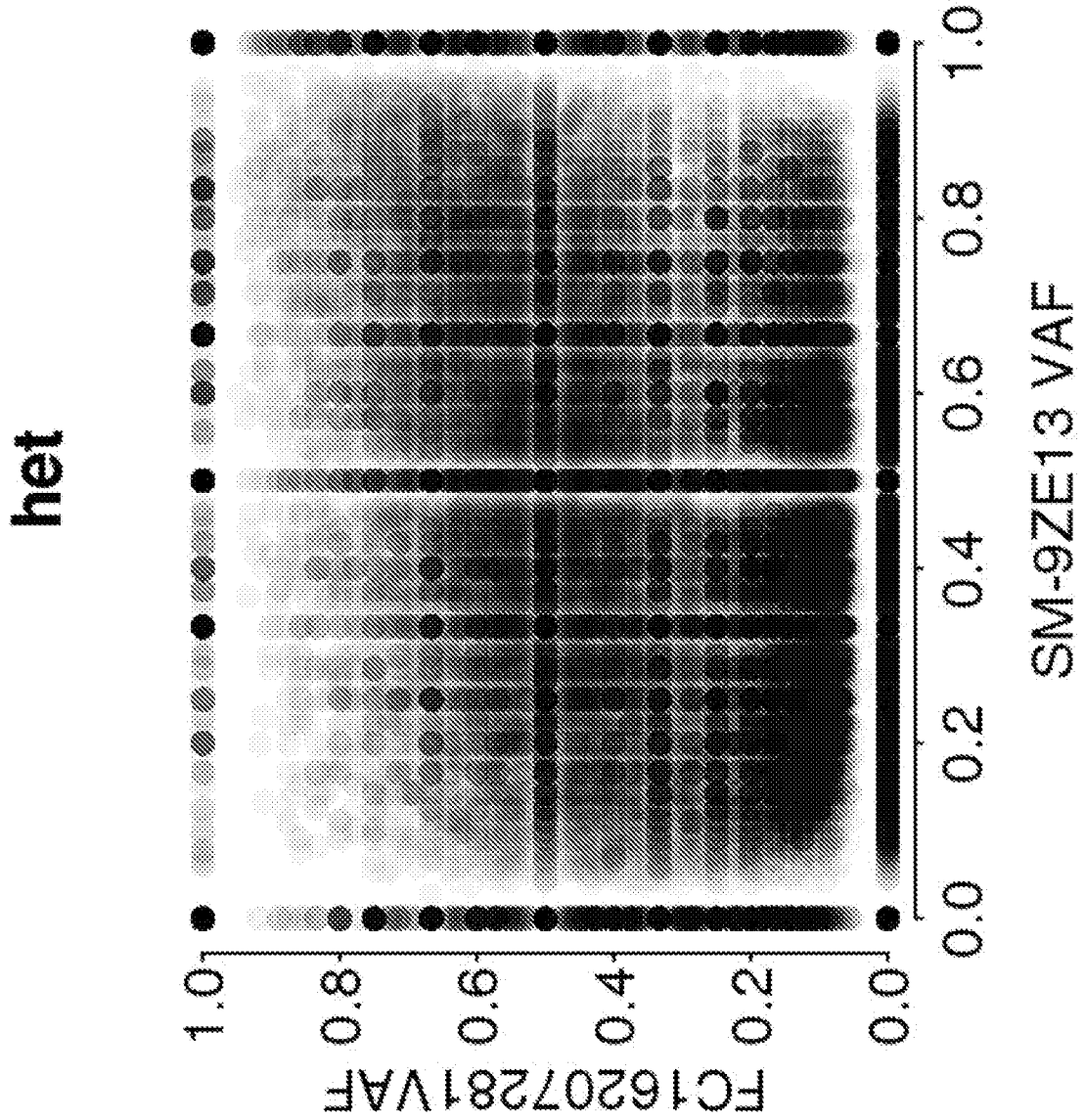
het
FC16207281VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
SM-9ZE13 VAF

Fig. 13U

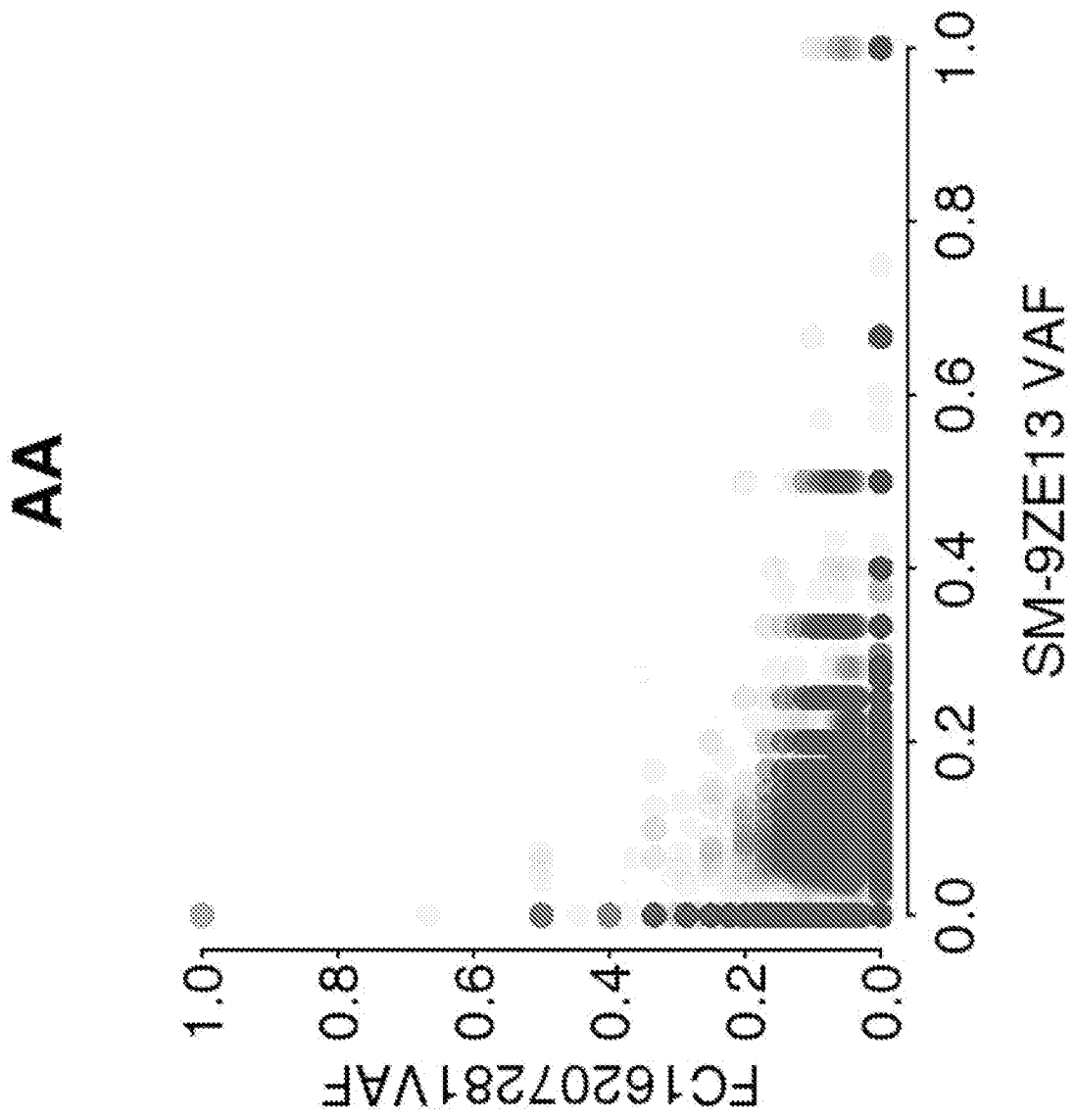

Fig. 13V

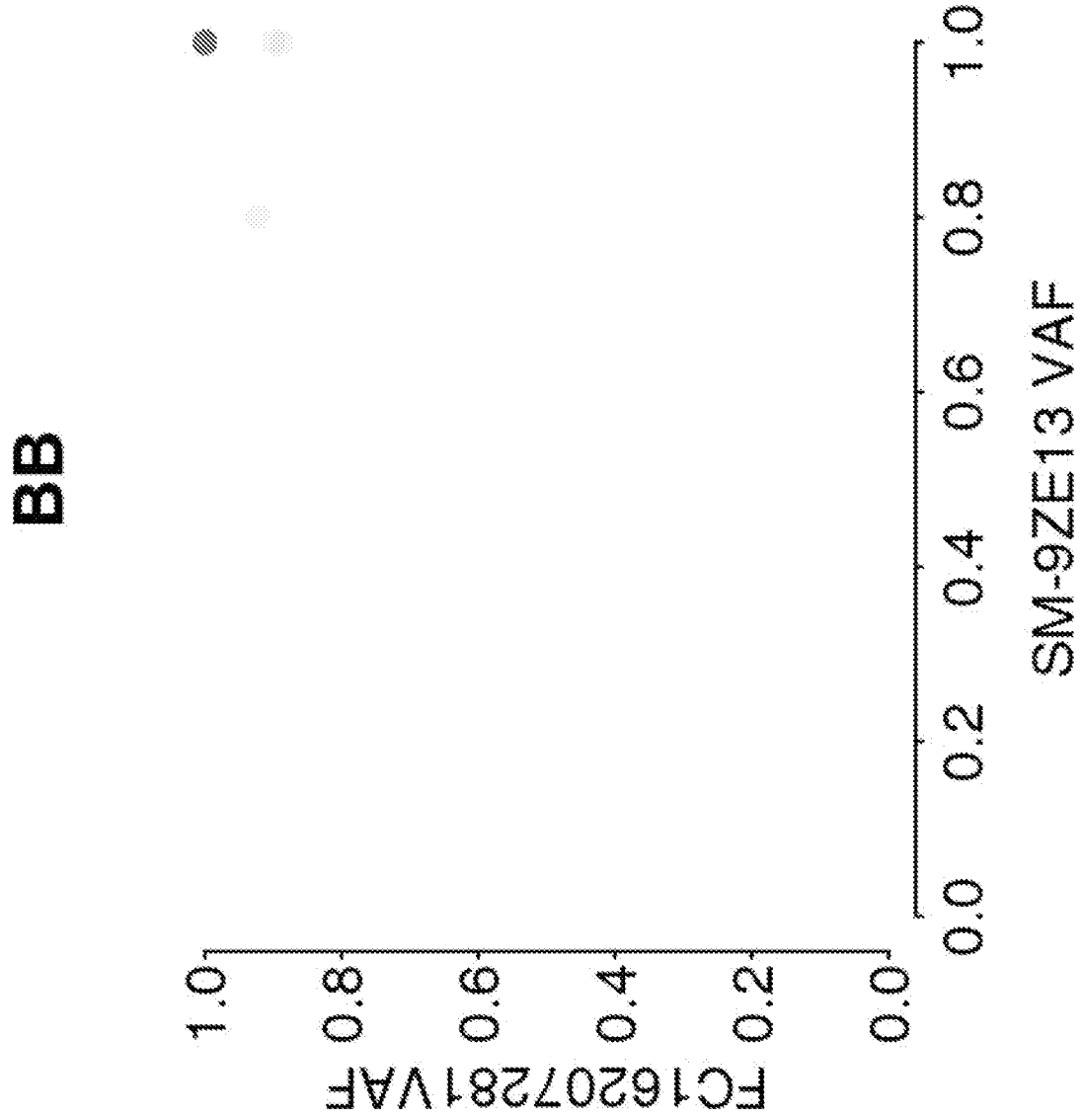

Fig. 13W

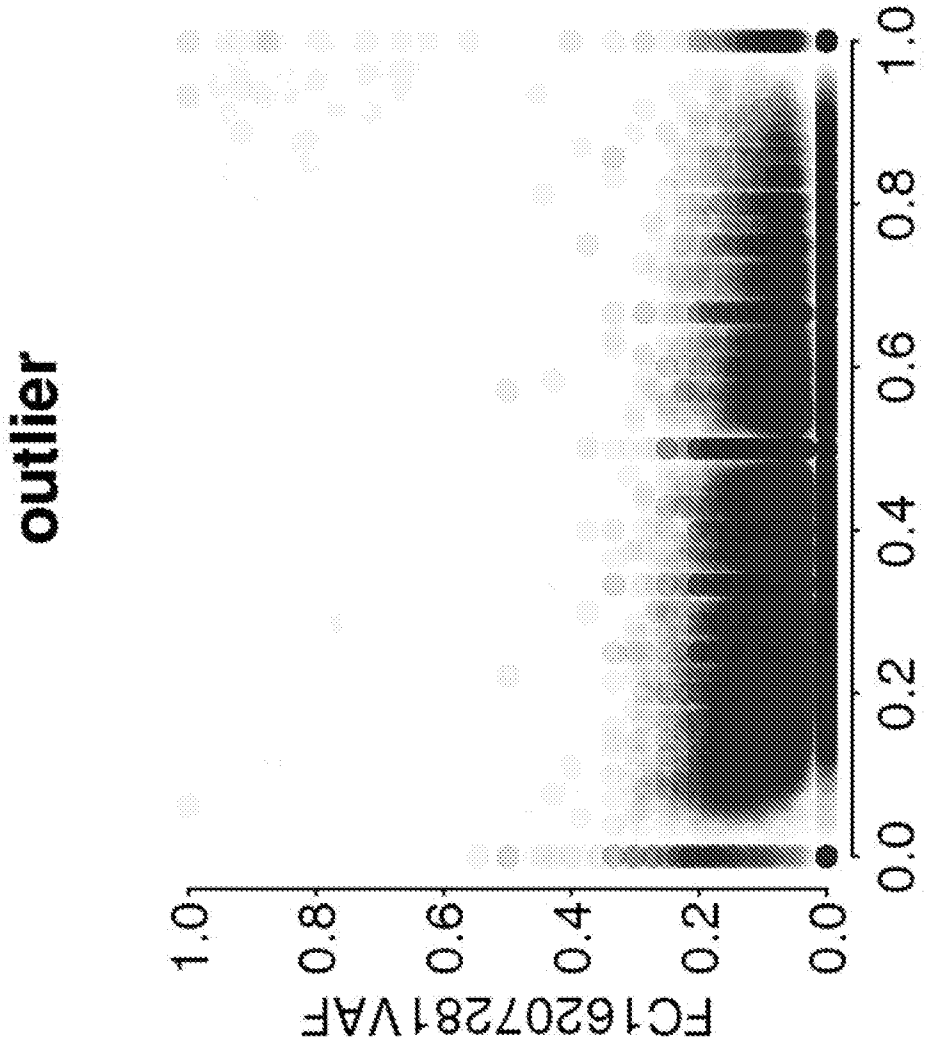
outlier
FC16207281VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
SM-9ZE13 VAF

Fig. 13X

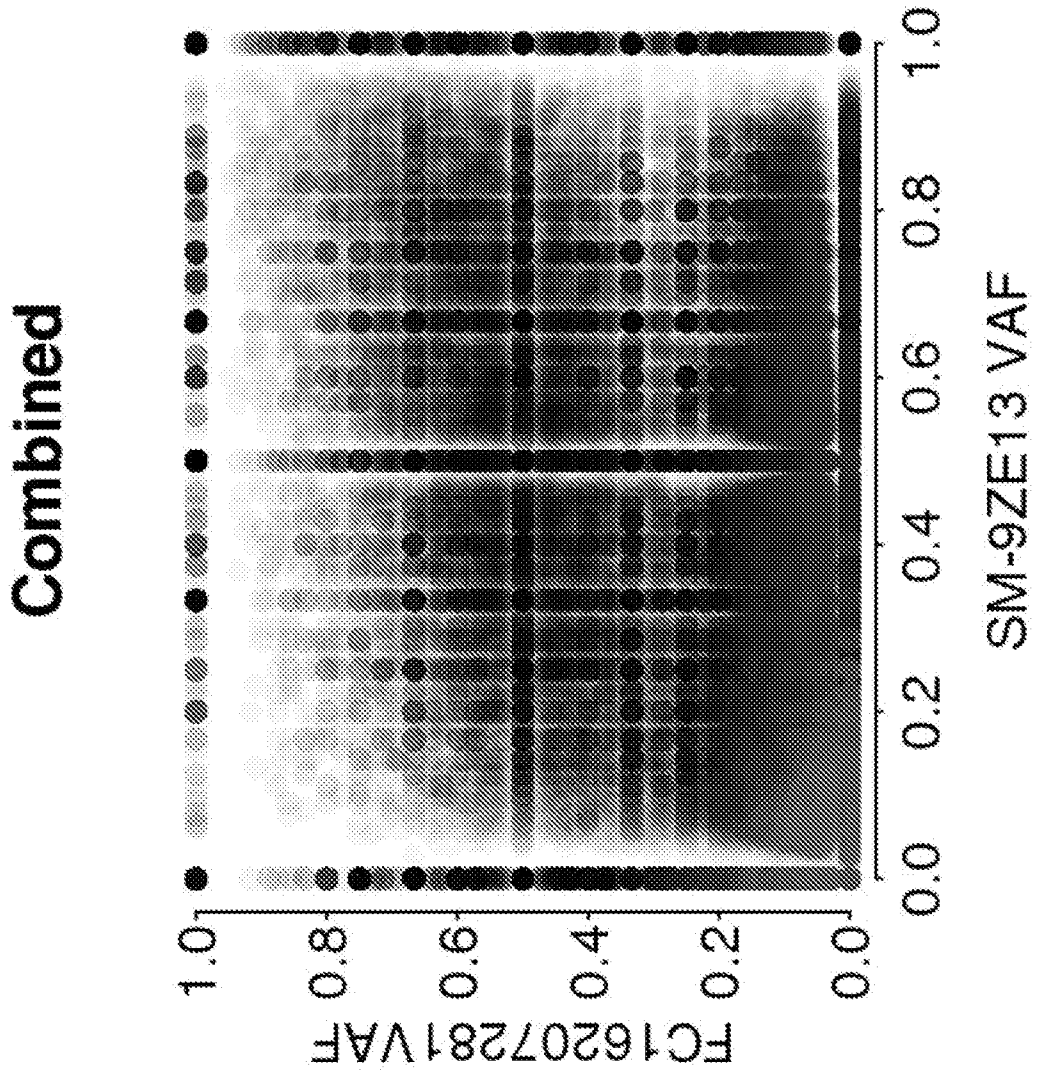
Combined
FC16207281VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
SM-9ZE13 VAF

Fig. 13Y

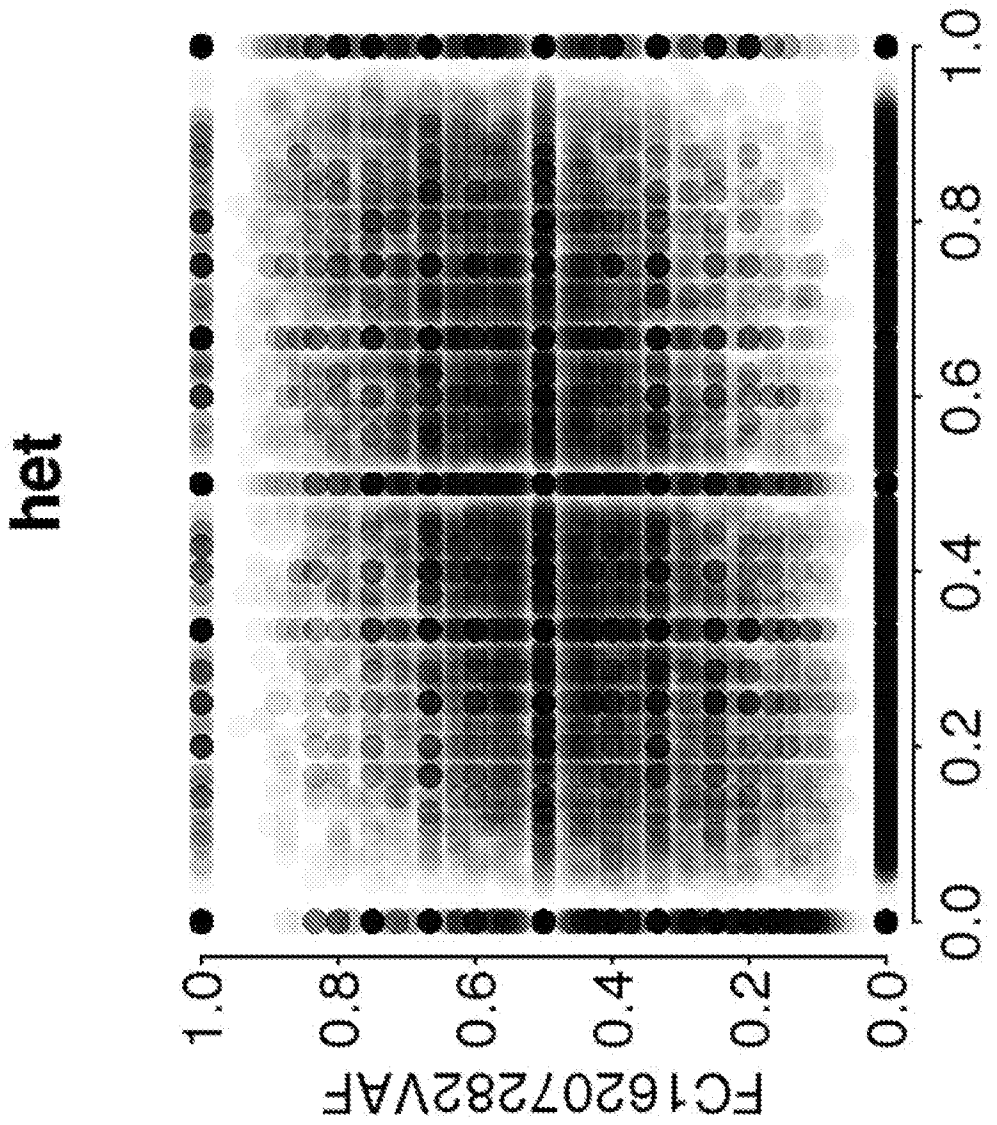
het
FC1620728 2VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
SM-9ZE13 VAF

Fig. 13Z

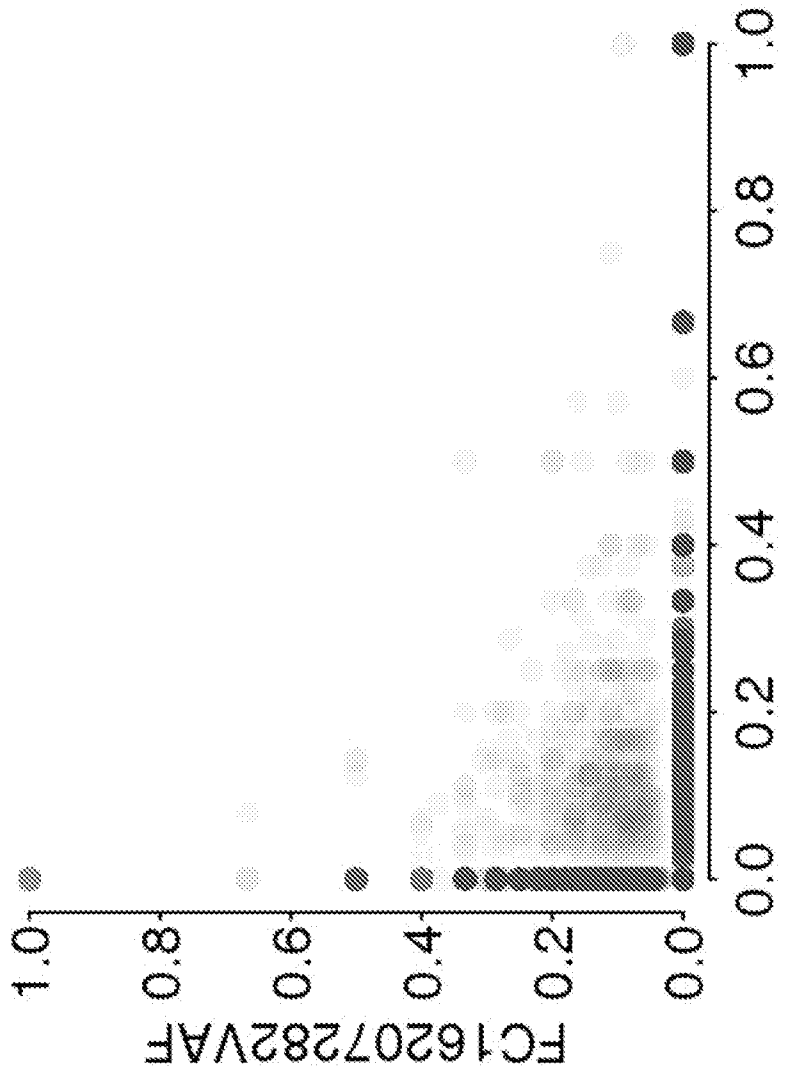

Fig. 13AA

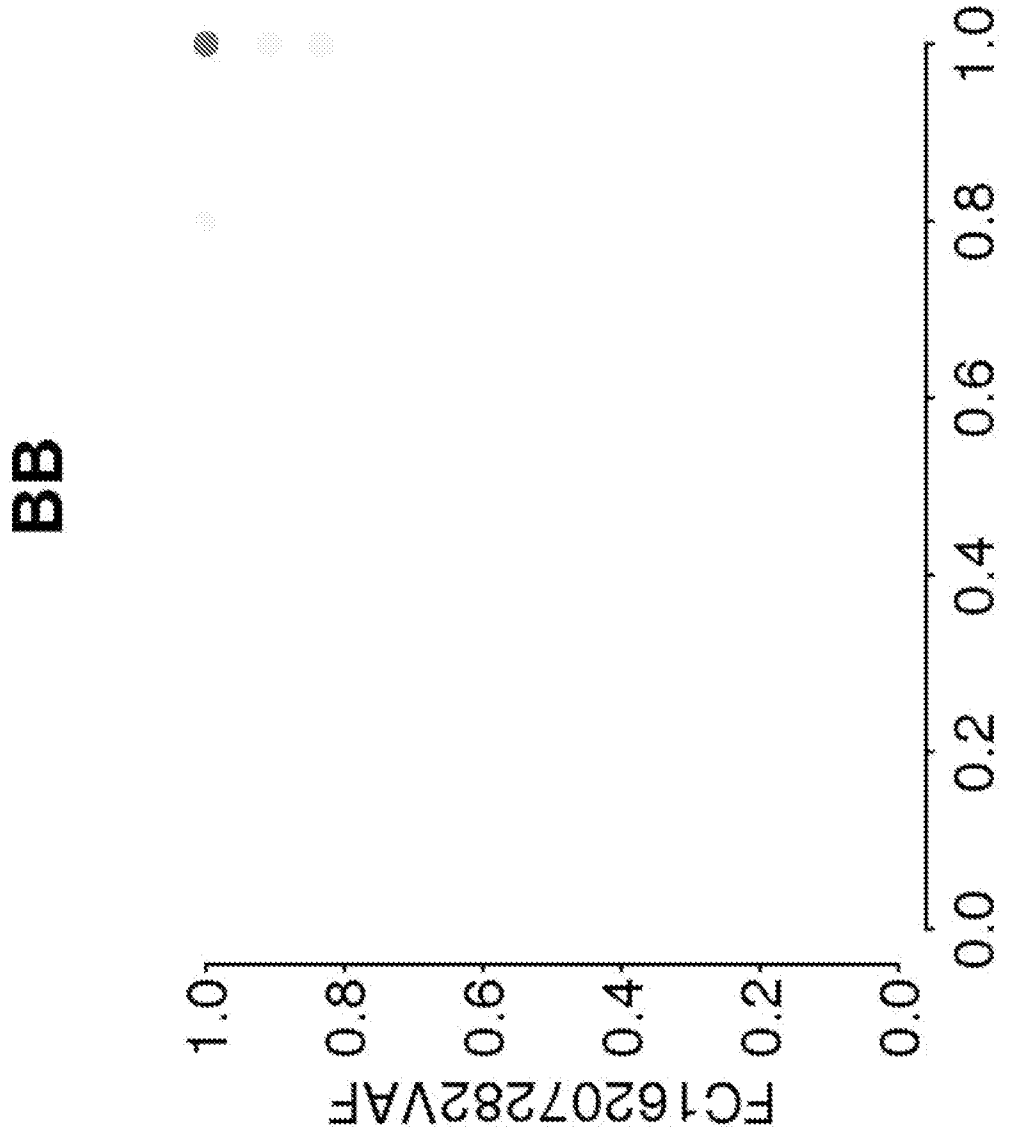

Fig. 13AB

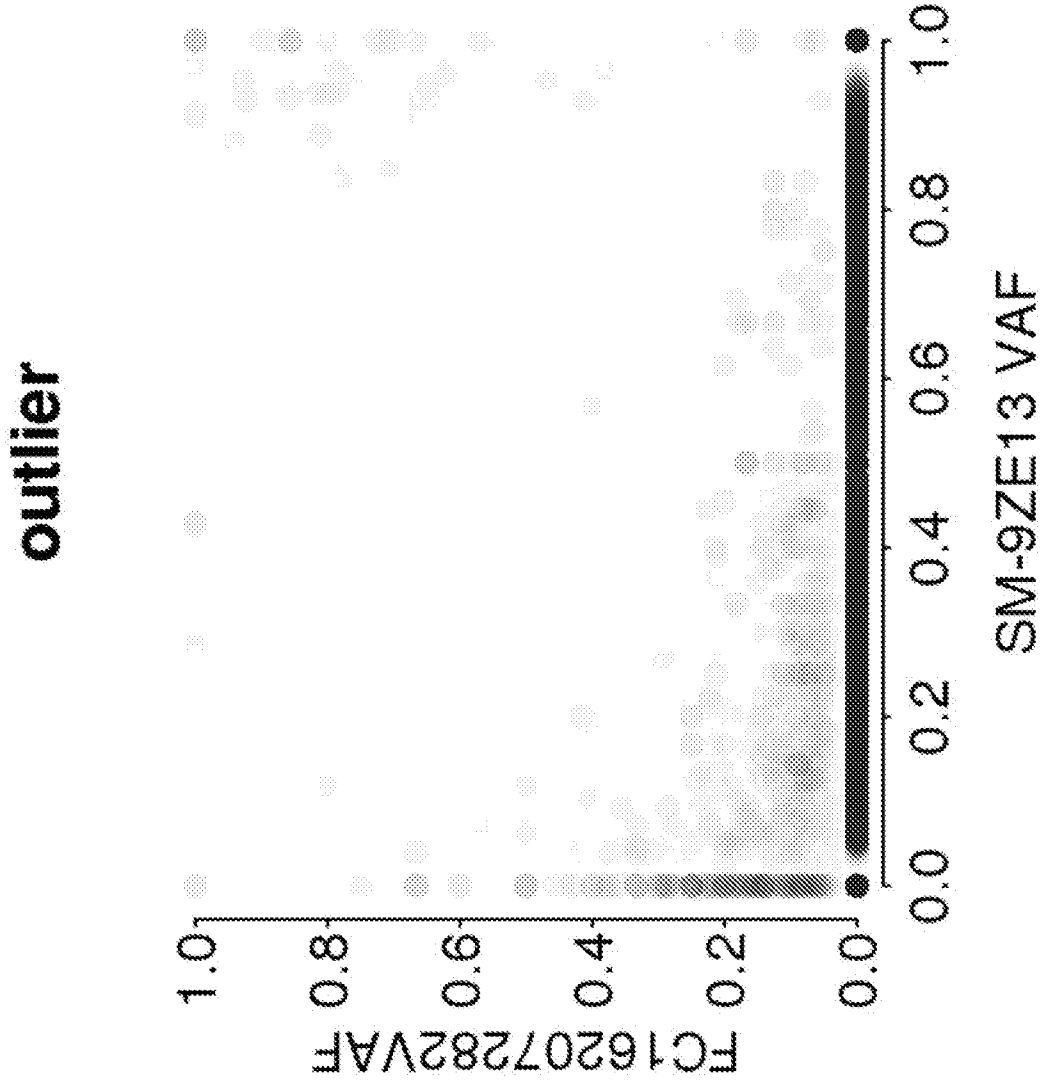
outlier
FC16207282VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
SM-9ZE13 VAF

Fig. 13AC

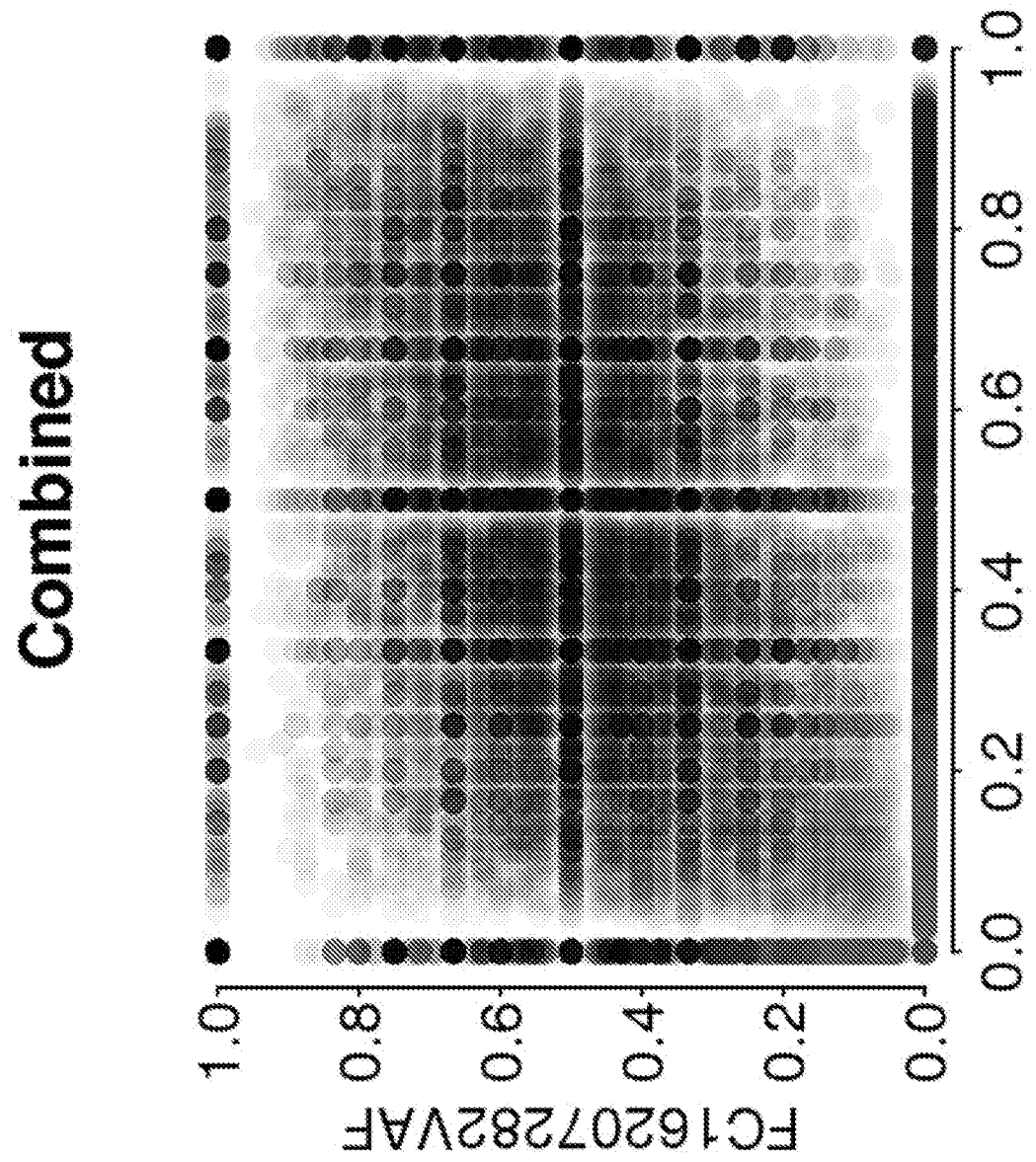
Combined
FC16207282VAF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
SM-9ZE13 VAF

Fig. 13AD

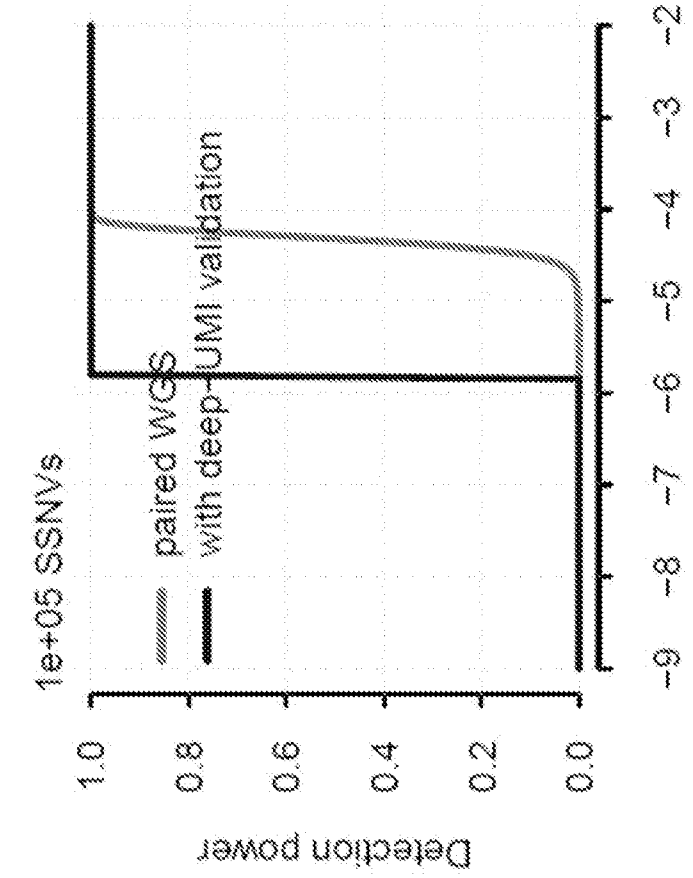

Fig. 14B

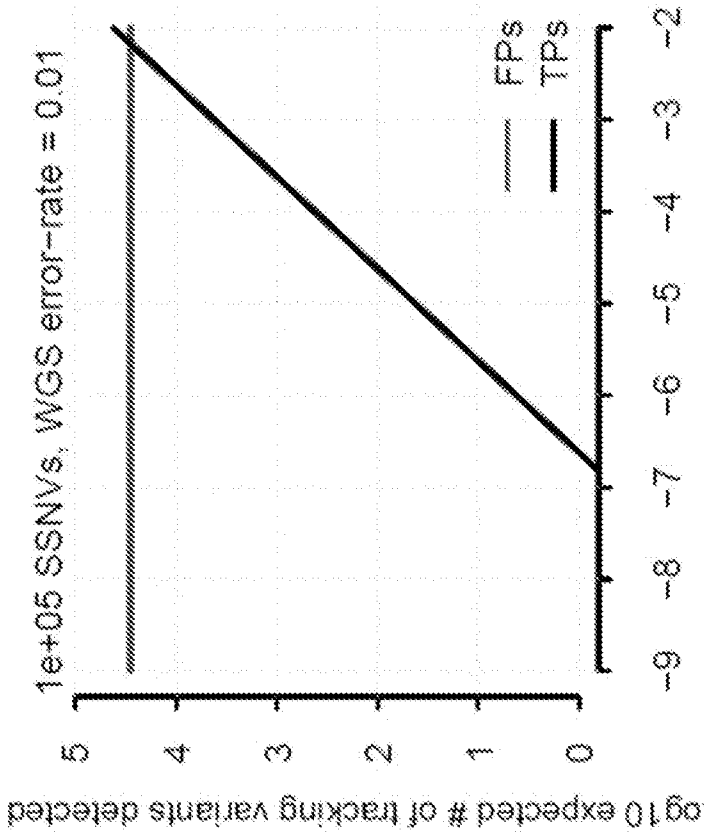

Fig. 14A

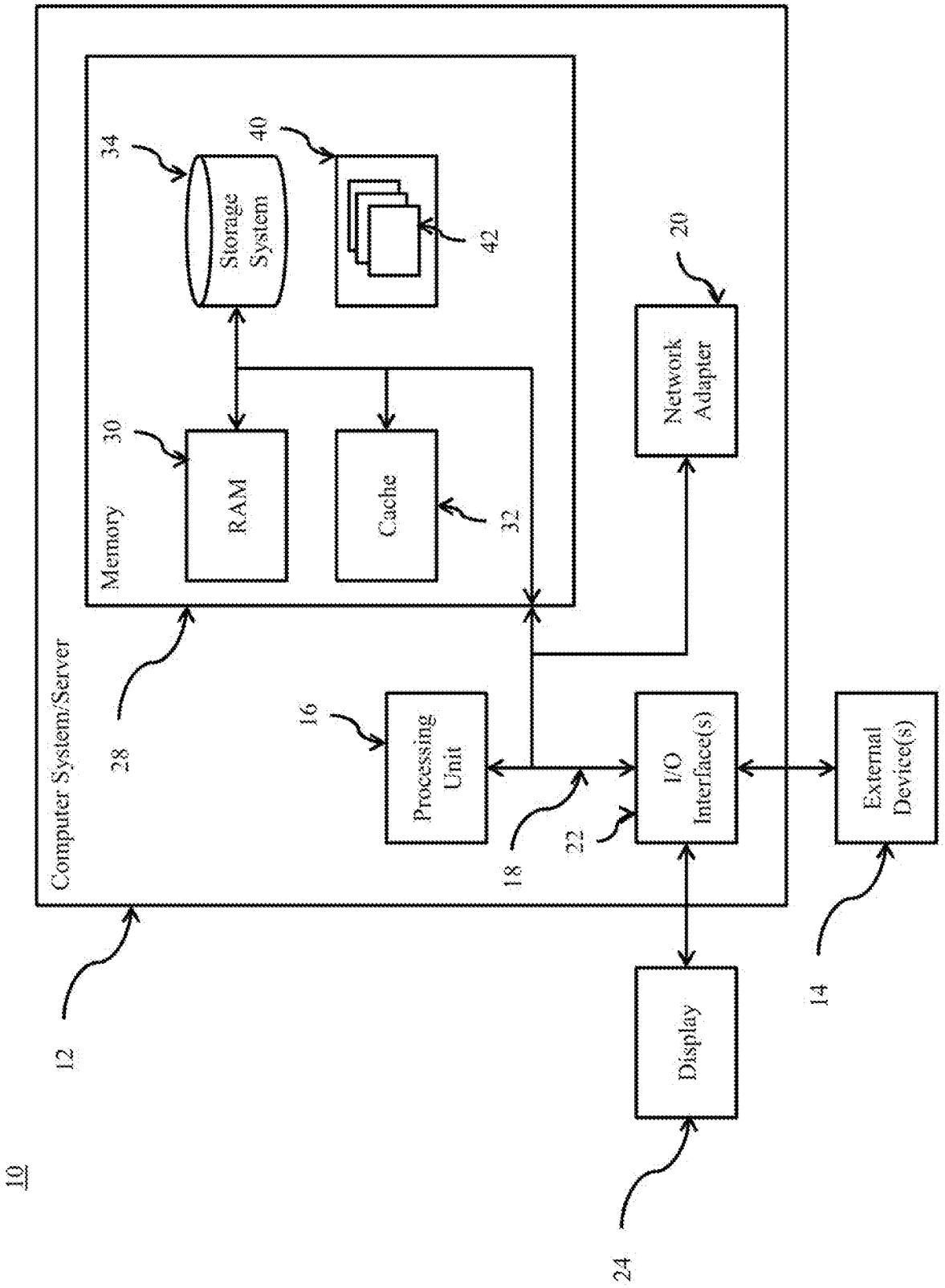
10
Computer System/Server
12
Memory
28
34
Storage System
40
42
30
RAM
Cache
32
16
Processing Unit
18
22
20
Network Adapter
I/O Interface(s)
External Device(s)
14
Display
24

Fig. 15

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2019/027525

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - C12Q 1/6827; C12Q 1/6886; G16B 30/00; G16B 40/00 (2019.01)
CPC - C12Q 1/6883; C12Q 1/6886; C12Q 2600/156; G16B 20/20 (2019.05)

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)
See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC - 435/6.14; 702/20 (keyword delimited)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2017/0061072 A1 (GUARDANT HEALTH, INC.) 02 March 2017 (02.03.2017) entire document	1, 2, 4, 5, 8-10, 13, 16, 18, 21-28, 30, 31, 34-36, 39, 42, 44, 47-54, 56, 57, 60-62, 65, 68, 70, 73-78
--		----------
Y		3, 6, 7, 11, 12, 14, 15, 17, 19, 20, 29, 32, 33, 37, 38, 40, 41, 43, 45, 46, 55, 58, 59, 63, 64, 66, 67, 69, 71, 72, 79-213
Y	US 2017/0073774 A1 (THE CHINESE UNIVERSITY OF HONG KONG) 16 March 2017 (16.03.2017) entire document	3, 20, 29, 46, 55, 72, 93, 120, 147
Y	US 2017/0199961 A1 (GRITSTONE ONCOLOGY, INC.) 13 July 2017 (13.07.2017) entire document	6, 11, 14, 15, 32, 37, 40, 41, 58, 63, 66, 67, 163, 166, 167, 181, 184, 185, 199, 202, 203
Y	US 2017/0107576 A1 (NATERA, INC.) 20 April 2017 (20.04.2017) entire document	7, 17, 19, 33, 43, 45, 59, 69, 71, 97, 99, 124, 126, 151, 153, 169, 171, 187, 189, 205, 207

☒ 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
01 July 2019	18 JUL 2019
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, VA 22313-1450 Facsimile No. 571-273-8300	Authorized officer Blaine R. Copenheaver PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2019/027525

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2017/0184597 A1 (NODALITY, INC.) 29 June 2017 (29.06.2017) entire document	12, 38, 64, 164, 182, 200
Y	US 2016/0371431 A1 (COUNSYL, INC.) 22 December 2016 (22.12.2016) entire document	79-159
Y	US 2009/0326832 A1 (HECKERMAN et al) 31 December 2009 (31.12.2009) entire document	81, 108, 135
Y	US 2017/0191998 A1 (NEUROINNOVATION OY) 06 July 2017 (06.07.2017) entire document	82, 109, 136
Y	US 2015/0254400 A1 (AGENCY FOR SCIENCE, TECHNOLOGY AND RESEARCH et al) 10 September 2015 (10.09.2015) entire document	83, 110, 137
Y	US 2017/0260585 A1 (CLINICAL GENOMICS PTY. LTD. et al) 14 September 2017 (14.09.2017) entire document	84, 111, 138
Y	US 2008/0172351 A1 (HECKERMAN et al) 17 July 2008 (17.07.2008) entire document	85, 112, 139
Y	EL-KEBIR et al. "Reconstruction of clonal trees and tumor composition from multi-sample sequencing data," Bioinformatics, 10 June 2015 (10.06.2015), Vol. 31, Iss. 12, Pgs. i62-i70. entire document	91, 118, 145
Y	US 2018/0075185 A1 (THE REGENTS OF THE UNIVERSITY OF CALIFORNIA) 15 March 2018 (15.03.2018) entire document	160-213