

We claim:

1. A method for detecting residual disease in a subject in need thereof, comprising,

(A) receiving a first subject- specific genome wide compendium of reads associated with genetic markers from a first biological sample of a subject, the first biological sample comprising a baseline sample and a normal cell sample, wherein the first compendium of reads each comprise reads of a single base pair length and wherein the baseline sample comprises a tumor sample or a plasma sample;

(B) filtering artefactual sites from the first compendium of reads, wherein the filtering comprises removing, from the first compendium of genetic markers, recurring sites generated over a cohort of reference healthy samples, and/or identifying germ line mutations in peripheral blood mononuclear cells of the normal cell sample and removing said germ line mutations from the from the first compendium of genetic markers;

(C) detecting reads from a second subject- specific genome wide compendium of genetic markers in a second biological sample of the subject to generate a tumor-associated genome-wide representation of genetic markers in the second sample;

(D) filtering noise from the first and second genome-wide compendium of reads using at least one error suppression protocol to produce a first filtered read set for the first genome-wide compendium of reads and a second filtered read set for the second genome-wide compendium of reads, wherein the at least one error suppression protocol comprises (a) calculating the probability that any single nucleotide variation in the first and second compendium is an artefactual mutation, and removing said mutation, wherein the probability is calculated as a function of features selected from the group consisting of mapping-quality (MQ), variant base-quality (MBQ), position-in-read (PIR), mean read base quality (MRBQ), and combinations thereof; and/or (b) removing artefactual mutations using discordance testing between independent replicates of the same DNA fragment generated from polymerase chain reaction or sequencing processing, and/or duplication consensus wherein artefactual mutations are identified and removed when lacking concordance across a majority of a given duplication family;

(E) computing an estimated tumor fraction (eTF) of the first and second biological sample using the first and second filtered read sets by applying a background noise model to one or more integrative mathematical models;

and

(F) detecting a residual disease in the subject if the estimated tumor fraction in the second biological sample exceeds an empirical threshold.

2. A method for detecting residual disease in a subject in need thereof, comprising,

(A) receiving a first subject- specific genome wide compendium of reads associated with genetic markers from a first biological sample of a subject, the first biological sample comprising a baseline sample, wherein the first compendium of reads each comprise a copy number variation (CNV) or structural variations (SVs) and wherein the baseline sample comprises a tumor sample or a plasma sample;

(B) receiving a second subject- specific genome wide compendium of reads associated with genetic markers from a second biological sample of a subject, the second biological sample comprising a peripheral blood mononuclear cell sample (PBMC), wherein the second compendium of genetic markers each comprise CNVs or SVs;

(C) filtering artefactual sites from the first and second compendium of reads, wherein the filtering comprises removing, from the first and second compendium of reads, recurring sites generated over a cohort of reference healthy samples; identifying shared CNVs/SVs between the first and second compendium as germ line mutations and removing said mutations from the first and second compendium of reads;

(D) detecting reads from a third subject- specific genome wide compendium of genetic markers in a third biological sample of the subject to generate a tumor-associated genome-wide representation of genetic markers in the third sample;

(E) normalizing each of the first, second and third compendium of reads to produce a first filtered read set for the first genome-wide compendium of reads, a second filtered read set for the second genome-wide compendium of reads, and a third filtered read set for the third genome-wide compendium of reads;

(F) computing an estimated tumor fraction (eTF) of the third biological samples, using the third filtered read set, by applying a background noise model to one or more integrative mathematical models, the one or more models producing a first eTF using the first filtered read set, and/or the one or more models producing a second eTF using the second filtered read set;

and

(G) detecting a residual disease in the subject if the estimated tumor fraction in the third biological sample exceeds an empirical threshold.

3. A system for detecting residual disease in a subject in need thereof, comprising,

an analyzing unit, the analyzing unit comprising

a pre-filter engine configured and arranged to

receive a first subject- specific genome wide compendium of reads associated with genetic markers from a first biological sample of a subject, the first biological sample comprising a baseline sample and a normal sample, wherein the first compendium of reads each comprise reads of a single base pair length and wherein the baseline sample comprises a tumor sample or a plasma sample; and

filter artefactual sites from the first compendium of reads, wherein the filtering comprises removing, from the first compendium of genetic markers, recurring sites generated over a cohort of reference healthy samples, and/or identifying germ line mutations in peripheral blood mononuclear cells of the normal cell sample and removing said germ line mutations from the from the first compendium of genetic markers;

and

a correction engine configured and arranged to

receive reads from a second subject- specific genome wide compendium of genetic markers in a second biological sample of the subject to generate a tumor-associated genome-wide representation of genetic markers in the second sample; and

filter noise from the first and second genome-wide compendium of reads using at least one error suppression protocol to produce a first filtered read set for the first genome-wide compendium of reads and a second filtered read set for the second genome-wide compendium of reads, wherein the at least one error suppression protocol comprises (a) calculating the probability that any single nucleotide variation in the first and second compendium is an artefactual mutation, and removing said mutation, wherein the probability is calculated as a function of features selected from the group consisting of mapping-quality (MQ), variant base-quality (MBQ), position-in-read (PIR), mean read base quality (MRBQ), and combinations thereof; and/or (b) removing artefactual mutations using discordance testing between independent replicates of the same DNA fragment generated from polymerase chain reaction or sequencing processing, and/or duplication consensus wherein artefactual mutations are identified and removed when lacking concordance across a majority of a given duplication family;

and

a computing unit configured and arranged to

- compute an estimated tumor fraction (eTF) of the first and second biological sample using the first and second filtered read sets by applying a background noise model to one or more integrative mathematical models; and
- detect a residual disease in the subject if the estimated tumor fraction in the second biological sample exceeds an empirical threshold.

4. A system for detecting residual disease in a subject in need thereof, comprising,
 - a pre-filter engine configured and arranged to
 - receive a first subject- specific genome wide compendium of reads associated with genetic markers from a first biological sample of a subject, the first biological sample comprising a baseline sample, wherein the first compendium of reads each comprise reads of a single base pair length and wherein the baseline sample comprises a tumor sample or a plasma sample;
 - receive a second subject- specific genome wide compendium of reads associated with genetic markers from a second biological sample of a subject, the second biological sample comprising a peripheral blood mononuclear cell sample (PBMC), wherein the second compendium of genetic markers each comprise a copy number variation (CNV); and
 - filter artefactual sites from the first and second compendium of reads, wherein the filtering comprises removing, from the first and second compendium of reads, recurring sites generated over a cohort of reference healthy samples; identifying shared CNVs between the first and second compendium as germ line mutations and removing said mutations from the first and second compendium of reads;
 - and
 - a correction engine configured and arranged to
 - receive reads from a third subject- specific genome wide compendium of genetic markers in a second biological sample of the subject to generate a tumor-associated genome-wide representation of genetic markers in the third sample;
 - and
 - normalize each of the first, second and third compendium of reads to produce a first filtered read set for the first genome-wide compendium of reads, a

second filtered read set for the second genome-wide compendium of reads, and a third filtered read set for the third genome-wide compendium of reads;

and

a computing unit configured and arranged to

compute an estimated tumor fraction (eTF) of the third biological samples, using the third filtered read set, by applying a background noise model to one or more integrative mathematical models, the one or more models producing a first eTF using the first filtered read set, and/or the one or more models producing a second eTF using the second filtered read set; and

detect a residual disease in the subject if the estimated tumor fraction in the third biological sample exceeds an empirical threshold.

5. The method of claim 1, wherein the markers comprise single nucleotide variations (SNVs) or insertion/deletions (indels); preferably SNV.
6. The method of claim 1, wherein filtering recurring sites generated over a cohort of reference healthy samples comprises generating a panel of normal (PON) blacklist or mask.
7. The method of claim 1, wherein the normal sample comprises peripheral blood mononuclear cells (PBMC) and germ line mutations in PBMC are removed in the artefactual site filtration step (B).
8. The method of claim 1, wherein in step (A), the first biological sample comprises plasma sample that is obtained from the subject pre-surgery or pre-therapy.
9. The method of claim 1, wherein in step (C), the second biological sample comprises plasma sample which is obtained from the same subject post-therapy or post-surgery.
10. The method of claim 1, wherein step (D) comprises employing a machine learning (ML) algorithm, *e.g.*, deep convolutional neural network (CNN), recurrent neural network (RNN), random forest (RF), support vector machine (SVM), discriminant analysis, nearest neighbor analysis (KNN), ensemble classifier, or a combination thereof; preferably, support vector machine (SVM), to filter artefactual noise.
11. The method of claim 1, wherein in step (D), the second error suppression step includes correction of artefactual mutations generated by PCR or sequencing using the comparison of independent replicates of the same original nucleic acid fragment.

12. The method of claim 11, wherein in step (D), the second error suppression step includes correction of artefactual mutations generated by paired-end 150 bp sequencing, resulting in overlapping paired reads (R1 and R2), and discordance between R1 and R2 pairs are corrected back to the corresponding reference genome.

13. The method of claim 1, wherein in step (D), the second error suppression step includes correction of duplication families generated during sequencing and/or PCR amplification, wherein the duplication families are recognized by 5' and 3' similarity as well as alignment position and wherein each duplication family is used to check the consensus of a specific mutation across independent replicates, thereby correcting artefactual mutations that do not show concordance in a majority of the duplication family.

14. The method of claim 1, wherein in step (E), the mathematical model integrates a relationship between the coverage, mutation load, number of detected mutations and the tumor fraction (TF).

15. The method of claim 1, wherein in step (E), the background noise calculation includes using patient specific mutation signature to calculate (1) the expected noise distribution over a cohort of healthy plasma samples (panel-of-normal or PON) or (2) the expected noise distribution across other patients (cross-patient analysis).

16. The method of claim 15, wherein the background noise model provides an estimated mean and standard-deviation (μ , σ) of artefactual mutation detection rate.

17. The method of any one of claims 1-16, further comprising orthogonal integration of a secondary feature comprising fragment size shift.

18. The method of claim 17, wherein intra-patient fragment size shifts in the list of tumor-specific markers and random markers are analyzed using statistical methods, *e.g.*, tests for significance or Gaussian mixture model (GMM).

19. The method of claim 2, wherein the markers comprise copy number variations (CNVs).

20. The method of claim 2, wherein filtering recurring sites generated over a cohort of reference healthy samples comprises generating a panel of normal (PON) blacklist or mask.

21. The method of claim 2, wherein germ line events in PBMC are removed in the artefactual site filtration step (C).

22. The method of claim 2, wherein in step (A), the first biological sample comprises plasma sample that is obtained from the subject pre-surgery or pre-therapy and the second biological sample comprises PBMCs obtained from the same subject pre-surgery or pre-therapy.

23. The method of claim 2, wherein in step (C), the third biological sample comprises plasma sample which is obtained from the same subject post-therapy or post-surgery.

24. The method of claim 2, wherein in step (C) comprises binning (to ≥ 500 bp windows) a region-of-interest (ROI) containing all the genomic segments of the somatic tumor CNV (sT_CNV) and somatic PBMC CNV (sP_CNV); estimating the depth coverage (read count) in each window from a follow-up plasma sample; and calculating median depth coverage per window.

25. The method of claim 2, wherein the follow-up plasma sample is obtained after surgery, during treatment, or at follow-up.

26. The method of claim 2, wherein the normalization step includes normalizing depth coverage values to correct for GC-content and mappability biases by performing two LOESS regression curve-fitting on the bin-wise GC-fraction and mappability score.

27. The method of claim 2, wherein the normalization step includes batch-effect correction using a robust-zscore normalization, which is applied to each sample separately.

28. The method of claim 27, wherein the zscore normalization includes calculation of median and median-absolute-deviation (MAD) based on the neutral regions of each sample and normalizing all CNV bins are normalized by subtracting the median value and dividing the differential by MAD.

29. The method of claim 2, wherein step (E) includes calculating depth coverage skew and/or fragment size center-of-mass (COM) skew in the third sample in comparison to a panel of normal (PON) healthy plasma samples.

30. The method of claim 2, wherein step (E) includes calculation of tumor fraction by checking a linear dilution ratio between the cumulative signal detected at the follow-up plasma sample in comparison to the cumulative signal detected in the tumor sample.

31. The method of claim 2, wherein in step (F), the background noise model includes using patient specific CNV/SV signature to calculate (1) the expected noise distribution over a cohort

of healthy plasma samples (panel-of-normal or PON) or (2) the expected noise distribution across other patients (cross-patient analysis).

32. The method of claim 31, wherein the background noise model provides an estimated mean and standard-deviation (μ , σ) of artefactual SNV/SV detection rate.

33. The method of claims 2, further comprising orthogonal integration of a secondary feature comprising fragment size shift.

34. The method of claim 33, wherein correlation between depth coverage skew and fragment size skew in CNV segments are analyzed to infer tumor fraction, *e.g.*, using a generalized linear model (GLM).