

CLAIMS

WHAT IS CLAIMED IS:

1. A method for enriching multiple genomic regions, comprising:
 - (a) bringing a predetermined amount of nucleic acid from a sample in contact with a bait mixture comprising:
 - (i) a first bait set that selectively hybridizes to a first set of genomic regions of the nucleic acid from the sample, which first bait set is provided at a first concentration that is less than a saturation point of the first bait set, and
 - (ii) a second bait set that selectively hybridizes to a second set of genomic regions of the nucleic acid from the sample, which second bait set is provided at a second concentration that is at or above a saturation point of the second bait set; and
 - (b) enriching the nucleic acid from the sample for the first set of genomic regions and the second set of genomic regions, thereby producing an enriched nucleic acid.
2. The method of claim 1, wherein the second bait set has a saturation point that is larger than substantially all of the saturation points associated with baits in the second bait set when a bait of the second bait set is subjected to a titration curve generated by
 - (i) measuring capture efficiency of a bait of the second bait set as a function of the concentration of the bait, and
 - (ii) identifying an inflection point within the titration curve, thereby identifying a saturation point associated with the bait.
3. The method of claim 1, wherein the saturation point of the first bait set is selected such that an observed capture efficiency increases by less than 10% at a concentration of the bait twice that of the first concentration.
4. The method of claim 1, wherein the first bait set or the second bait set selectively enrich for one or more nucleosome-associated regions of a genome, the nucleosome-associated regions comprising genomic regions having one or more genomic base positions with differential nucleosomal occupancy, wherein the differential nucleosomal occupancy is characteristic of a cell or tissue type of origin or disease state.
5. The method of claim 1, further comprising:
 - (c) sequencing the enriched nucleic acid to produce a plurality of sequence reads.

6. The method of claim 5, further comprising:
 - (d) producing an output comprising nucleic acid sequences representative of the nucleic acid from the sample.
7. A method for generating a bait set comprising:
 - (a) identifying one or more predetermined backbone genomic regions, wherein the identifying the one or more backbone genomic regions comprises maximizing a ranking function of sequencing load and utility associated with each of the backbone genomic regions;
 - (b) identifying one or more predetermined hotspot genomic regions, wherein the one or more hotspots are selected using one or more of the following:
 - (i) maximizing a ranking function of sequencing load and utility associated with each of the hotspot genomic regions,
 - (ii) nucleosome profiling across the one or more predetermined genomic regions,
 - (iii) predetermined cancer driver mutations or prevalence across a relevant patient cohort, and
 - (iv) empirically identified cancer driver mutations;
 - (c) creating a first bait set that selectively captures the predetermined backbone genomic regions; and
 - (d) creating a second bait set that selectively captures the predetermined hotspot genomic regions, wherein the second bait set has a higher capture efficiency than the first bait set.
8. The method of claim 7, wherein the identifying the one or more predetermined hotspots comprises using a programmed computer processor to rank a set of hotspot genomic regions based on a ranking function of sequencing load and utility associated with each of the hotspot genomic regions.
9. The method of claim 7, wherein the identifying the one or more predetermined backbone genomic regions comprises:
 - (i) ranking a set of backbone genomic regions based on a ranking function of sequencing load and utility associated with each of the predetermined backbone genomic regions;
 - (ii) utilizing a set of empirically determined minor allele frequency (MAF) values or clonality of a variant measured by its MAF in relationship to the highest presumed driver or clonal mutation in a sample; or
 - (iii) a combination of (i) and (ii).

10. The method of claim 7, wherein the sequencing load of a genomic region is calculated by multiplying together one or more of:

- (i) size of the genomic region in base pairs,
- (ii) relative fraction of reads spent on sequencing fragments mapping to the genomic region,
- (iii) relative coverage as a result of sequence bias of the genomic region,
- (iv) relative coverage as a result of amplification bias of the genomic region, and
- (v) relative coverage as a result of capture bias of the genomic region.

11. The method of claim 7, wherein the utility of a genomic region is calculated by multiplying together one or more of:

- (i) frequency of one or more actionable mutations in the genomic region,
- (ii) frequency of one or more mutations associated with above-average minor allele frequencies (MAFs) in the genomic region,
- (iii) fraction of patients in a cohort harboring a somatic mutation within the genomic region,
- (iv) sum of MAF for variants in patients in a cohort, said patients harboring a somatic mutation within the genomic region, and
- (v) ratio of (1) MAF for variants in patients in a cohort, said patients harboring a somatic mutation within the genomic region, to (2) maximum MAF for a given patient in the cohort

12. The method of claim 11, wherein the actionable mutations comprise one or more of:

- (i) druggable mutations,
- (ii) mutations for therapeutic monitoring,
- (iii) disease specific mutations,
- (iv) tissue specific mutations,
- (v) cell type specific mutations,
- (vi) resistance mutations, and
- (vii) diagnostic mutations.

13. The method of claim 11, wherein the mutations associated with higher minor allele frequencies comprise one or more driver mutations or are known from external data or annotation sources.

14. A method comprising:

(a) providing a plurality of bait mixtures, wherein each of the plurality of bait mixtures comprises a first bait set that selectively hybridizes to a first set of genomic regions and a second bait set that selectively hybridizes to a second set of genomic regions, wherein the first bait set is at different concentrations across the plurality of bait mixtures and the second bait set is at the same concentration across the plurality of bait mixtures;

(b) contacting each of the plurality of bait mixtures with a nucleic acid sample to capture nucleic acids from the nucleic acid sample with the first bait set and the second bait set, wherein the second bait set in each bait mixture is provided at a first concentration that is at or above a saturation point of the second bait set, wherein nucleic acids from the nucleic acid sample are captured by the first bait set and the second bait set;

(c) sequencing a portion of the nucleic acids captured with each bait mixture to produce sets of sequence reads within an allocated number of sequence reads;

(d) determining the read depth of sequence reads for the first bait set and the second bait set for each bait mixture; and

(e) identifying at least one bait mixture that provides read depths for the second set of genomic regions;

wherein the read depths for the second set of genomic regions provides a sensitivity of detecting of a genetic variant of at least 0.0001% minor allele frequency (MAF).

15. The method of claim 14, wherein the second bait set has a saturation point when subjected to titration, which titration comprises generating a titration curve comprising:

(i) measuring capture efficiency of the second bait set as a function of the concentration of the baits; and

(ii) identifying an inflection point within the titration curve, thereby identifying a saturation point associated with the second bait set.

16. The method of claim 15, wherein the saturation point is selected such that an observed capture efficiency increases by less than 10% at a concentration of the bait set twice that of the first concentration.

17. The method of claim 14, wherein the first bait set or the second bait set selectively enrich for one or more nucleosome-associated regions of a genome, the nucleosome-associated regions comprising genomic regions having one or more genomic base positions with differential

nucleosomal occupancy, wherein the differential nucleosomal occupancy is characteristic of a cell or tissue type of origin or disease state.

18. The method of claim 14, wherein the first set of genomic regions comprises one or more actionable mutations, wherein the one or more actionable mutations comprise one or more of:

- (i) druggable mutations,
- (ii) mutations for therapeutic monitoring,
- (iii) disease specific mutations,
- (iv) tissue specific mutations,
- (v) cell type specific mutations,
- (vi) resistance mutations, and
- (vii) diagnostic mutations.

19. The method of claim 14, wherein the first genomic regions comprise at least a portion of each of at least 5 genes selected from Table 3 or Table 4.

20. The method of claim 14, wherein the first genomic regions have a size between 25 kilobases to 1,000 kilobases and a read depth of between 1,000 counts/base and 50,000 counts/base.

21. The method of claim 1, wherein the saturation point of the second bait set is selected such that an observed capture efficiency increases by less than 10% at a concentration of the bait twice that of the second concentration.

22. The method of claim 14, wherein the second set of genomic regions comprises one or more actionable mutations, wherein the one or more actionable mutations comprise one or more of:

- (i) druggable mutations,
- (ii) mutations for therapeutic monitoring,
- (iii) disease specific mutations,
- (iv) tissue specific mutations,
- (v) cell type specific mutations,
- (vi) resistance mutations, and
- (vii) diagnostic mutations.

23. The method of claim 14, wherein the second genomic regions comprise at least a portion of each of at least 5 genes selected from Table 3 or Table 4.
24. The method of claim 14, wherein the second genomic regions have a size between about 25 kilobases to 1,000 kilobases and a read depth of between 1,000 counts/base and 50,000 counts/base.
25. A method for improving accuracy of detecting an insertion or deletion (indel) from a plurality of sequence reads derived from cell-free deoxyribonucleic acid (cfDNA) molecules in a bodily sample of a subject, which plurality of sequence reads are generated by nucleic acid sequencing, the method comprising:
- (a) for each of the plurality of sequence reads associated with the cell-free DNA molecules, providing:
 - a predetermined expectation of an indel being detected in one or more sequence reads of the plurality of sequence reads;
 - a predetermined expectation that a detected indel is a true indel present in a given cell-free DNA molecule of the cell-free DNA molecules, given that an indel has been detected in the one or more of the sequence reads; and
 - a predetermined expectation that a detected indel is introduced by non-biological error, given that an indel has been detected in the one or more of the sequence reads;
 - (b) providing quantitative measures of one or more model parameters characteristic of sequence reads generated by nucleic acid sequencing;
 - (c) detecting one or more candidate indels in the plurality of sequence reads associated with the cell-free DNA molecules; and
 - (d) for each candidate indel, performing a hypothesis test using one or more of the model parameters to classify said candidate indel as a true indel or an introduced indel, thereby improving accuracy of detecting an indel.
26. The method of claim 25, further comprising enriching one or more loci from the cell-free DNA in the bodily sample before step (a) (i), thereby producing enriched polynucleotides.
27. The method of claim 26, further comprising amplifying the enriched polynucleotides to produce families of amplicons, wherein each family comprises amplicons originating from a single strand of the cell-free DNA molecules.
28. The method of claim 25, wherein the non-biological error comprises error in sequencing at a plurality of genomic base locations.

29. The method of claim 25, wherein the non-biological error comprises error in amplification at a plurality of genomic base locations.
30. The method of claim 25, wherein the model parameters comprise one or more of:
- (i) for each of one or more variant alleles, a frequency of the variant allele (α) and a frequency of non-reference alleles other than the variant allele (α');
 - (ii) a frequency of an indel error in the entire forward strand of a family of strands (β_1), wherein a family comprises a collection of amplicons originating from a single strand of the cell-free DNA molecules;
 - (iii) a frequency of an indel error in the entire reverse strand of a family of strands (β_2);
- and
- (iv) a frequency of an indel error in a sequence read (γ).
31. The method of claim 25, wherein the step of performing a hypothesis test comprises performing a multi-parameter maximization algorithm.
32. The method of claim 31, wherein the multi-parameter maximization algorithm comprises a Nelder-Mead algorithm.
33. The method of claim 25, wherein the classifying of (d) comprises:
- (a) maximizing a multi-parameter likelihood function,
 - (b) classifying a candidate indel as a true indel if the maximum likelihood function value is greater than a predetermined threshold value, and
 - (c) classifying a candidate indel as an introduced indel if the maximum likelihood function value is less than or equal to a predetermined threshold value.