CLAIMS

WHAT IS CLAIMED IS:

1. A method for identifying sets of validated phased variants from a tumor sample of a subject, the method comprising:

(a) obtaining, by a computer system, sequencing data from a tumor sample of a subject;

(b) obtaining, by the computer system, sequencing data from a matched non-tumor sample of the subject;

(c) analyzing, using the computer system, the sequencing data from the tumor sample and the sequencing data from the matched non-tumor sample to identify a plurality of regions of a genome of the subject that include a first putative phased variant and a second putative phased variant, wherein each of the plurality of regions is no more than 170 base pairs in length;

(d) after (c), performing targeted sequencing on nucleic acids from the tumor sample of the subject that are from the plurality of regions identified in (c) to a depth of at least 250×;

(e) identifying, using the computer system, sets of validated phased variants from the tumor sample of the subject based on sequencing data from the targeted sequencing in (d).

2. The method of claim 1, wherein the first putative phased variant and the second putative phased variant are separated by at least one nucleotide.

3. The method of claim 1, wherein (c) further comprises:

determining a plurality of single nucleotide variants in the genome of the subject that are present in the tumor sample relative to the matched non-tumor sample; and

from the identified plurality of single nucleotide variants, identifying the plurality of regions of the genome that include the first phased variant and the second phased variant based at least in part on one or more of (1) a proximity of two or more of the identified plurality of single nucleotide variants within a genome, and (2) a presence of two or more of the identified plurality of the single nucleotide variants on a single read from the sequence data from the tumor sample of the subject.

4. The method of claim 3, wherein the sequencing data from the tumor sample of the subject in (a) and (b) each comprises data for at least 50% of the genome of the subject.

5. The method of claim 4, wherein (c) further comprises aligning the sequencing data from the tumor sample and the sequencing data from the matched non-tumor sample to a genome.

6. The method of claim 5, wherein the sequencing data from the tumor sample and the sequencing data from the matched non-tumor sample are obtained at a depth of between 10× and 100×.

7. The method of claim 6, wherein the identified plurality of regions of the genome in (c) has at least 400 regions and no more than 15,000 regions.

8. The method of claim 1, wherein identifying the plurality of regions of the genome in (c) comprises determining one or more of (i) a presence of phased variants in individual reads from the tumor sample of the subject, (ii) a presence or an absence of read support in the matched non-tumor sample, (iii) a presence of other non-reference bases on supporting reads, (iv) base quality, (v) mapping quality, and (vi) uniqueness of genomic positions.

9. The method of claim 1, wherein performing the targeted sequencing of (d) comprises (1) performing hybridization capture of nucleic acids from the tumor sample of the subject that correspond to the identified plurality of regions of the genome from (c) to obtain captured nucleic acids, and (2) sequencing the captured nucleic acids from the tumor sample of the subject.

10. The method of claim 1, wherein the number of sets of validated phased variants is less than 1000 and greater than 10.

11. The method of claim 1, wherein the plurality of regions identified in (c) are non-overlapping regions.

12. The method of claim 1, wherein the nucleic acids of (d) are DNA molecules.

13. The method of claim 1, wherein the first phased variant and the second phased variant are each single nucleotide variants.

14. The method of claim 1, wherein the sets of validated phased variants are found to have an allele fraction of greater than 5% and no read support in the sequencing data from the matched non-tumor sample.

15. The method of claim 1, wherein (d) further comprises: after (c), performing targeted sequencing on nucleic acids from the matched non-tumor sample of the subject that are from the plurality of regions identified in (c) to a depth of at least 500×.

16. The method of claim 1, wherein (d) further comprises: after (c), performing targeted sequencing on nucleic acids from the matched non-tumor sample of the subject that are from the plurality of regions identified in (c) to a depth of at least 250×.

17. The method of claim 1, wherein the sets of validated phased variants have a background signal of less than one in a million.

18. The method of claim 1, wherein the tumor sample is a solid tumor sample.

19. The method of claim 1, wherein:

the sequencing data from the tumor sample is obtained at a depth of between 10× and 100× and comprises at least 50% of the genome of the subject;

the sequencing data from the matched non-tumor sample is obtained at a depth of between 10× and 100× and comprises at least 50% of the genome of the subject;

(c) further comprises:

aligning the sequencing data from the tumor sample and the sequencing data from the matched non-tumor sample to a genome;

determining a plurality of single nucleotide variants in the genome of the subject that are present in the tumor sample relative to the matched non-tumor sample; and

from the identified plurality of single nucleotide variants, identifying the plurality of regions of the genome that include the first phased variant and the second phased variant based at least in part on one or more of (1) a proximity of two or more of the identified plurality of single nucleotide variants within a genome, and (2) a presence of two or more of the identified plurality of the single nucleotide variants on a single read from the sequence data from the tumor sample of the subject;

wherein the plurality of regions identified in (c) are non-overlapping regions;

performing the targeted sequencing of (d) comprises (1) performing hybridization capture of nucleic acids from the tumor sample of the subject that correspond to the identified plurality of regions of the genome from (c) to obtain captured nucleic acids, and (2) sequencing the captured nucleic acids from the tumor sample of the subject; and

the first putative phased variant and the second putative phased variant are single nucleotide variants separated by at least one nucleotide.

20. The method of claim 1, further comprising combining a set of nucleic acid probes with a plurality of cell-free nucleic acid molecules that are obtained or derived from a subject to form a mixture, wherein each nucleic acid probe of the set of nucleic acid probes is configured to hybridize to a cell-free nucleic acid molecule that comprises one or more of the identified sets of validated phased variants.

21. The method of claim 20, wherein each nucleic acid probe of the set of nucleic acid probes is configured to hybridize to a target nucleic acid molecule comprising a plurality of phased variants such that the nucleic acid probe is complementary to at least a region of the target nucleic acid molecule that

extends from a first phased variant of the set of phased variants to a second phased variant of the set of phased variants.

22. The method of claim 20, wherein each nucleic acid probe of the set of nucleic acid probes comprises a biotin pull-down tag.

23. The method of claim 20, further comprising:

separating target nucleic acid molecules that hybridize to at least one nucleic acid probe of the set of nucleic acid probes from nucleic acid molecules that do not hybridize to at least one nucleic acid probe of the set of nucleic acid probes, thereby capturing target nucleic acid molecules; and

sequencing the target nucleic acid molecules that hybridize to at least one of the set of nucleic acid probes.

24. The method of claim 23, further comprising identifying one or more cell-free nucleic acid molecules as being a cancer-derived molecule with a limit of detection of less than about 1 out of 50,000 observations from sequencing data obtained from the sequencing of the cell-free nucleic acid molecules.

25. The method of claim 24, wherein the limit of detection of is less than 1 out of 1,000,000 observations from the sequencing data.

26. The method of claim 23, wherein the sequencing does not comprise use of molecular barcodes.