wherein the computer further comprises machine-executable code which, when executed by the one or more computer processors, copies nucleic acid (e.g., DNA) sequence data stored on the database, writes the copied data to memory in the bioinformatics computer and performs steps including: generating from the nucleic acid (e.g., DNA) sequence data a first data set comprising for each genetic locus in a plurality of genetic loci a quantitative measure related to sequencing read coverage ("read coverage"); correcting the first data set by performing saturation equilibrium correction and probe efficiency correction; determining a baseline read coverage for the first data set, wherein the baseline read coverage relates to saturation equilibrium and probe efficiency; and determining a copy number state for each genetic locus in the plurality of genetic loci relative to the baseline read coverage. In some embodiments, the database is connected to a DNA sequencer.

[0015a] In another aspect, the present disclosure provides a method, comprising: (a) obtaining sequencing reads of deoxyribonucleic acid (DNA) molecules of a cell-free DNA sample of a subject, wherein the DNA molecules of the cell-free DNA sample is enriched for a set of genetic loci using one or more oligonucleotide probes that are complementary to at least a portion of the genetic locus from the set of genetic loci; (b) generating from the sequence reads a first data set comprising, for one or more genetic loci in a plurality of genetic loci, a quantitative measure related to sequencing read coverage ("read coverage"), wherein the first data set comprises, for one or more genetic loci in a plurality of genetic loci, a quantitative measure related to guanine-cytosine content ("GC content"); (c) removing from the first data set genetic loci that are high-variance genetic loci, wherein removing comprises: (i) fitting a model relating the quantitative measures related to guanine-cytosine content and the quantitative measures of sequencing read coverage of the genetic loci; and (ii) removing from the first data set between 10% and 30% of the genetic loci that most differ from the model, thereby providing the first data set of baselining genetic loci; (d) correcting the first data set of baselining genetic loci by performing saturation equilibrium correction and probe efficiency correction, wherein performing saturation equilibrium correction comprises transforming the first data set of baselining genetic loci into a saturation corrected data set by (i) determining for one or more genetic loci from the first data set of baselining genetic loci a quantitative measure related to the probability that a strand of DNA molecule from the sample derived from the genetic locus is represented within the sequencing reads; (ii) determining a first transformation for the read coverage by relating the read coverage in the first data set of baselining genetic loci to both the GC content of the first data set of baselining genetic loci and the quantitative measure related to the probability that a strand of DNA derived from each locus in the first data set of baselining genetic loci is represented within the sequencing reads; and (iii) applying the first transformation to the read coverage

of one or more genetic loci from the first data set of baselining genetic loci to provide the saturation corrected data set, wherein the saturation corrected data set comprises a first set of transformed read coverages of the first data set of baselining genetic loci; wherein performing probe efficiency correction comprises transforming the saturation corrected data set into a probe efficiency corrected data set by: (i) removing from the saturation corrected data set genetic loci that are high-variance genetic loci with respect to the first set of transformed read coverages, thereby providing a second data set of baselining genetic loci; (ii) determining a second transformation for the first set of transformed read coverages related to probe efficiency of the second data set of baselining genetic loci; and (iii) transforming the first set of transformed read coverages of the second data set of baselining genetic loci with the second transformation, thereby providing the probe efficiency corrected data set, wherein the probe efficiency corrected data set comprises a second set of transformed read coverages of the second data set of baselining genetic loci; (e) determining a baseline read coverage for the first data set, wherein the baseline read coverage relates to saturation equilibrium and probe efficiency; and (f) determining a copy number state for one or more genetic loci in the plurality of genetic loci using the probe efficiency corrected data set relative to the baseline read coverage.

[0015b] In another aspect, the present disclosure provides a system comprising: (a) a network; (b) a database comprising computer memory configured to store nucleic acid sequence data which are connected to the network; (c) a bioinformatics computer comprising a computer memory and one or more computer processors, which computer is connected to the network; wherein the computer further comprises machine-executable code which, when executed by the one or more computer processors, copies said nucleic acid sequence data stored in the database, writes the copied data to memory in the bioinformatics computer and performs steps including: (i) generating from the nucleic acid sequence data a first data set comprising, for one or more genetic loci in a plurality of genetic loci a quantitative measure related to sequencing read coverage ("read coverage"), wherein the first data set comprises, for one or more genetic loci in a plurality of genetic loci, a quantitative measure related to guanine-cytosine content ("GC content"); (ii) removing from the first data set genetic loci that are high-variance genetic loci, wherein removing comprises: (1) fitting a model relating the quantitative measures related to guanine-cytosine content and the quantitative measures of sequencing read coverage of the genetic loci; and (2) removing from the first data set between 10% and 30% of the genetic loci that most differ from the model, thereby providing the first data set of baselining genetic loci; (iii) correcting the first data set of baselining genetic loci by performing saturation equilibrium correction and probe efficiency correction, wherein performing saturation equilibrium correction comprises transforming the first data set of baselining genetic loci

into a saturation corrected data set by (1) determining for one or more genetic loci from the first data set of baselining genetic loci a quantitative measure related to the probability that a strand of DNA molecule from the sample derived from the genetic locus is represented within the sequencing reads; (2) determining a first transformation for the read coverage by relating the read coverage in the first data set of baselining genetic loci to both the GC content of the first data set of baselining genetic loci and the quantitative measure related to the probability that a strand of DNA derived from each locus in the first data set of baselining genetic loci is represented within the sequencing reads; and (1) applying the first transformation to the read coverage of one or more genetic loci from the first data set of baselining genetic loci to provide the saturation corrected data set, wherein the saturation corrected data set comprises a first set of transformed read coverages of the first data set of baselining genetic loci; wherein performing probe efficiency correction comprises transforming the saturation corrected data set into a probe efficiency corrected data set by: (1) removing from the saturation corrected data set genetic loci that are high-variance genetic loci with respect to the first set of transformed read coverages, thereby providing a second data set of baselining genetic loci; (2) determining a second transformation for the first set of transformed read coverages related to probe efficiency of the second data set of baselining genetic loci; and (3) transforming the first set of transformed read coverages of the second data set of baselining genetic loci with the second transformation, thereby providing the probe efficiency corrected data set, wherein the probe efficiency corrected data set comprises a second set of transformed read coverages of the second data set of baselining genetic loci; (iv) determining a baseline read coverage for the first data set, wherein the baseline read coverage relates to saturation equilibrium and probe efficiency; and (v) determining a copy number state for one or more genetic loci in the plurality of genetic loci using the probe efficiency corrected data set relative to the baseline read coverage.

[0015c] In another aspect, the present disclosure provides a method comprising: (a) receiving into computer memory sequencing reads of deoxyribonucleic acid (DNA) molecules of a cell-free DNA sample of a subject; (b) executing code with a computer processor to perform the following steps: (i) generating from the sequence reads a first data set comprising, for one or more genetic loci in a plurality of genetic loci a quantitative measure related to sequencing read coverage ("read coverage"), wherein the first data set comprises, for one or more genetic loci in a plurality of genetic loci, a quantitative measure related to guanine-cytosine content ("GC content"); (ii) removing from the first data set genetic loci that are high-variance genetic loci, wherein removing comprises: (1) fitting a model relating the quantitative measures related to guanine-cytosine content and the quantitative measures of sequencing read coverage of the genetic loci; and (2) removing from the first data set between 10% and 30% of the genetic loci that most differ from the model, thereby providing the first data set of baselining genetic loci; (iii) correcting the first data set of baselining genetic loci by performing saturation equilibrium correction and probe efficiency correction,

wherein performing saturation equilibrium correction comprises transforming the first data set of baselining genetic loci into a saturation corrected data set by (1) determining for one or more genetic loci from the first data set of baselining genetic loci a quantitative measure related to the probability that a strand of DNA molecule from the sample derived from the genetic locus is represented within the sequencing reads; (2) determining a first transformation for the read coverage by relating the read coverage in the first data set of baselining genetic loci to both the GC content of the first data set of baselining genetic loci and the quantitative measure related to the probability that a strand of DNA derived from each locus in the first data set of baselining genetic loci is represented within the sequencing reads; and (3) applying the first transformation to the read coverage of one or more genetic loci from the first data set of baselining genetic loci to provide the saturation corrected data set, wherein the saturation corrected data set comprises a first set of transformed read coverages of the first data set of baselining genetic loci; wherein performing probe efficiency correction comprises transforming the saturation corrected data set into a probe efficiency corrected data set by: (1) removing from the saturation corrected data set genetic loci that are high-variance genetic loci with respect to the first set of transformed read coverages, thereby providing a second data set of baselining genetic loci; (2) determining a second transformation for the first set of transformed read coverages related to probe efficiency of the second data set of baselining genetic loci; and (3) transforming the first set of transformed read coverages of the second data set of baselining genetic loci with the second transformation, thereby providing the probe efficiency corrected data set, wherein the probe efficiency corrected data set comprises a second set of transformed read coverages of the second data set of baselining genetic loci; (iv) determining a baseline read coverage for the first data set, wherein the baseline read coverage relates to saturation equilibrium and probe efficiency; and (v) determining a copy number state for one or more genetic loci in the plurality of genetic loci using the probe efficiency corrected data set relative to the baseline read coverage.

INCORPORATION BY REFERENCE

[0016] All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017] The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings of which:

CLAIMS

WHAT IS CLAIMED IS:

1. A method, comprising:

(a) obtaining sequencing reads of deoxyribonucleic acid (DNA) molecules of a cell-free DNA sample of a subject, wherein the DNA molecules of the cell-free DNA sample is enriched for a set of genetic loci using one or more oligonucleotide probes that are complementary to at least a portion of the genetic locus from the set of genetic loci;

(b) generating from the sequence reads a first data set comprising, for one or more genetic loci in a plurality of genetic loci, a quantitative measure related to sequencing read coverage ("read coverage"), wherein the first data set comprises, for one or more genetic loci in a plurality of genetic loci, a quantitative measure related to guanine-cytosine content ("GC content");

(c) removing from the first data set genetic loci that are high-variance genetic loci, wherein removing comprises:

(i) fitting a model relating the quantitative measures related to guanine-cytosine content and the quantitative measures of sequencing read coverage of the genetic loci; and

(ii) removing from the first data set between 10% and 30% of the genetic loci that most differ from the model, thereby providing the first data set of baselining genetic loci;

(d) correcting the first data set of baselining genetic loci by performing saturation equilibrium correction and probe efficiency correction,

wherein performing saturation equilibrium correction comprises transforming the first data set of baselining genetic loci into a saturation corrected data set by

(i) determining for one or more genetic loci from the first data set of baselining genetic loci a quantitative measure related to the probability

that a strand of DNA molecule from the sample derived from the genetic locus is represented within the sequencing reads;

(ii) determining a first transformation for the read coverage by relating the read coverage in the first data set of baselining genetic loci to both the GC content of the first data set of baselining genetic loci and the quantitative measure related to the probability that a strand of DNA derived from each locus in the first data set of baselining genetic loci is represented within the sequencing reads; and

(iii) applying the first transformation to the read coverage of one or more genetic loci from the first data set of baselining genetic loci to provide the saturation corrected data set, wherein the saturation corrected data set comprises a first set of transformed read coverages of the first data set of baselining genetic loci;

wherein performing probe efficiency correction comprises transforming the saturation corrected data set into a probe efficiency corrected data set by:

(i) removing from the saturation corrected data set genetic loci that are high-variance genetic loci with respect to the first set of transformed read coverages, thereby providing a second data set of baselining genetic loci;

(ii) determining a second transformation for the first set of transformed read coverages related to probe efficiency of the second data set of baselining genetic loci; and

(iii) transforming the first set of transformed read coverages of the second data set of baselining genetic loci with the second transformation, thereby providing the probe efficiency corrected data set, wherein the probe efficiency corrected data set comprises a second set of transformed read coverages of the second data set of baselining genetic loci;

(e) determining a baseline read coverage for the first data set, wherein the baseline read coverage relates to saturation equilibrium and probe efficiency; and

(f) determining a copy number state for one or more genetic loci in the plurality of genetic loci using the probe efficiency corrected data set relative to the baseline read coverage.

2. The method of claim 1, wherein (c)(ii) comprises removing 10% of the genetic loci.

3. The method of claim 1, wherein determining the first transformation comprises (i) determining a measure related to central tendency of the read coverage of the first data set of baselining genetic loci; (ii) determining a function that fits the measure related to central tendency of the read coverage of the first data set of baselining genetic loci based on the GC content of the genetic locus and the quantitative measure related to the probability that a strand of DNA derived from the genetic locus is represented within the sequencing reads; and (iii) for each genetic locus of the first data set of baselining genetic loci, determining a difference between the read coverage predicted by the function and the read coverage, wherein the difference is the transformed read coverage.

4. The method of claim 3, wherein the function is a surface approximation.

5. The method of claim 4, wherein the surface approximation is a two-dimensional second degree polynomial.

6. The method of claim 1, wherein removing from the saturation corrected data set genetic loci that are high-variance genetic loci comprises:

(i) fitting a model relating the GC content and the first set of transformed read coverages of the saturation corrected data set; and

(ii) removing from saturation corrected data set between 10% and 30% of the genetic loci, wherein the removing the genetic loci comprises removing genetic loci that most differ from the model, thereby providing the second data set of baselining genetic loci.

7. The method of claim 6, comprising removing 10% of the genetic loci.

8. The method of claim 1, wherein the probe efficiency is determined by performing the saturation equilibrium correction on one or more reference samples, wherein the probe efficiency is the transformed read coverage obtained by performing the saturation equilibrium correction.

9. The method of claim 8, wherein the one or more reference samples are cell-free DNA samples from a subject without cancer.

10. The method of claim 8, wherein the one or more reference samples are cell-free DNA samples from a subject with cancer, wherein the corresponding genetic locus has not undergone copy number alteration.

11. The method of claim 8, wherein determining the second transformation comprises (i) fitting the probe efficiency determined for the genetic loci from the one or more reference samples to the first set of read coverages from the second data set of baselining genetic loci; (ii) dividing the transformed read coverages of each genetic locus of the second data set of baselining genetic loci by a predicted probe efficiency based on the fitting of (i).

12. The method of claim 1, further comprising:

(g) determining a third transformation for the second set of transformed read coverages by relating the transformed read coverages of the second data set of baselining genetic loci to both the GC content of the second data set of baselining genetic loci and the quantitative measure related to the probability that a strand of DNA derived from the each locus in the second data set of baselining genetic loci is represented within the sequencing reads;

(h) applying the third transformation to the second set of transformed read coverages to provide a third data set, wherein the third data set comprises a third set of transformed quantitative read coverages.

13. The method of claim 1, wherein the GC content of each genetic locus from the set of genetic loci is a measure related to central tendency of guanine-cytosine content of the one or more oligonucleotide probes that are complementary to at least a portion of the genetic locus from the set of genetic loci.

14. The method of claim 1, wherein the read coverage of the genetic locus is a measure related to central tendency of the read coverage of regions of the genetic locus corresponding to the one or more oligonucleotide probes.

15. The method of claim 1, wherein the performing saturation equilibrium correction and the performing probe efficiency correction comprises fitting a Langmuir model, wherein the Langmuir model comprises probe efficiency (K) and saturation equilibrium constant (I_{sat}).

16. The method of claim 15, wherein K and I_{sat} are determined empirically for each oligonucleotide probe in the one or more oligonucleotide probes.

17. The method of claim 16, wherein performing saturation equilibrium correction and performing probe correction comprises fitting the read coverages of the genetic loci to the Langmuir model assuming that the genetic loci are present in identical copy number states, thereby providing a baseline read coverage.

18. The method of claim 17, wherein the identical copy number states are diploid.

19. The method of claim 17, wherein the baseline read coverage is a function dependent on the probe efficiency and the saturation equilibrium.

20. The method of claim 17, wherein determining a copy number state comprises comparing the read coverage of the genetic loci to the baseline read coverage.

21. The method of any one of the preceding claims, wherein the cell-free DNA sample is obtained from a bodily fluid, wherein the bodily fluid is selected from the group consisting of serum, plasma, urine, and cerebrospinal fluid.

22. The method of any one of the preceding claims, wherein the read coverage is determined by mapping the sequencing reads to a reference genome.

23. The method of any one of the preceding claims, wherein obtaining the sequencing reads comprises ligating adaptors to the DNA molecules of the cell-free DNA sample from the subject.

24. The method of claim 23, wherein the DNA molecules are duplex DNA molecules and the adaptors are ligated to the duplex DNA molecules such that each adaptor differently tags complementary strands of the DNA molecule to provide tagged strands.

25. The method of claim 24, wherein determining the quantitative measure related to the probability that a strand of DNA derived from the genetic locus is represented within the sequencing reads comprises sorting sequencing reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in said set, and (ii) each unpaired read represents a first tagged strand having no second differently tagged complementary strand derived from a double-stranded polynucleotide molecule represented among said sequence reads in said set of sequence reads.

26. The method of claim 25, further comprising determining quantitative measures of (i) said paired reads and (ii) said unpaired reads that map to each of one or more genetic loci to determine a quantitative measure related to total double-stranded DNA molecules in said sample that map to each of said one or more genetic loci based on said quantitative measure related to paired reads and unpaired reads mapping to each locus.

27. The method of claim 23, wherein the adaptors comprise barcode sequences.

28. The method of claim 27, wherein determining the read coverage comprises collapsing the sequencing reads based on position of the mapping of the sequencing reads to the reference genome and the barcode sequences.

29. The method of any one of the preceding claims, wherein the genetic loci comprise one or more oncogenes.

30. The method of any one of the preceding claims, further comprising determining that at least a subset of the baselining genetic loci have undergone copy number alteration in tumor cells of the subject by determining relative quantities of variants within the baselining genetic loci for which the germline genome of the subject is heterozygous.

31. The method of claim 30, wherein the relative quantities of the variants are not approximately equal.

32. The method of claim 31, wherein the baselining genetic loci for which the relative quantities of the variants are not approximately equal are removed from the baselining genetic loci, thereby providing allelic-frequency corrected baselining genetic loci.

33. The method of claim 32, wherein the allelic-frequency corrected baselining genetic loci are used as the baselining loci in the methods of any one of the preceding claims.

34. A method comprising:

(a) receiving into computer memory sequencing reads of deoxyribonucleic acid (DNA) molecules of a cell-free DNA sample of a subject;

(b) executing code with a computer processor to perform the following steps:

(i)generating from the sequence reads a first data set comprising, for one or more genetic loci in a plurality of genetic loci a quantitative measure related to sequencing read coverage ("read coverage"), wherein the first data set comprises, for one or more genetic loci in a plurality of genetic loci, a quantitative measure related to guanine-cytosine content ("GC content");

(ii) removing from the first data set genetic loci that are high-variance genetic loci, wherein removing comprises:

(1) fitting a model relating the quantitative measures related to guanine-cytosine content and the quantitative measures of sequencing read coverage of the genetic loci; and

(2) removing from the first data set between 10% and 30% of the genetic loci that most differ from the model, thereby providing the first data set of baselining genetic loci;

(iii) correcting the first data set of baselining genetic loci by performing saturation equilibrium correction and probe efficiency correction, wherein performing saturation equilibrium correction comprises transforming the

first data set of baselining genetic loci into a saturation corrected data set by

(1) determining for one or more genetic loci from the first data set of baselining genetic loci a quantitative measure related to the probability that a strand of DNA molecule from the sample derived from the genetic locus is represented within the sequencing reads;

(2) determining a first transformation for the read coverage by relating the read coverage in the first data set of baselining genetic loci to both the GC content of the first data set of baselining genetic loci and the quantitative measure related to the probability that a strand of DNA derived from each locus in the first data set of baselining genetic loci is represented within the sequencing reads; and

(3) applying the first transformation to the read coverage of one or more genetic loci from the first data set of baselining genetic loci to provide the saturation corrected data set, wherein the saturation corrected data set comprises a first set of transformed read coverages of the first data set of baselining genetic loci;

wherein performing probe efficiency correction comprises transforming the saturation corrected data set into a probe efficiency corrected data set by:

(1) removing from the saturation corrected data set genetic loci that are high-variance genetic loci with respect to the first set of transformed read coverages, thereby providing a second data set of baselining genetic loci;

(2) determining a second transformation for the first set of transformed read coverages related to probe efficiency of the second data set of baselining genetic loci; and

(3) transforming the first set of transformed read coverages of the second data set of baselining genetic loci with the second transformation, thereby providing the probe efficiency corrected data set, wherein the probe efficiency corrected data set comprises

a second set of transformed read coverages of the second data set of baselining genetic loci;

(iv) determining a baseline read coverage for the first data set, wherein the baseline read coverage relates to saturation equilibrium and probe efficiency; and

(v) determining a copy number state for one or more genetic loci in the plurality of genetic loci using the probe efficiency corrected data set relative to the baseline read coverage.

35. A system comprising:

(a) a network;

(b) a database comprising computer memory configured to store nucleic acid sequence data which are connected to the network;

(c) a bioinformatics computer comprising a computer memory and one or more computer processors, which computer is connected to the network;

wherein the computer further comprises machine-executable code which, when executed by the one or more computer processors, copies said nucleic acid sequence data stored in the database, writes the copied data to memory in the bioinformatics computer and performs steps including:

(i)generating from the nucleic acid sequence data a first data set comprising, for one or more genetic loci in a plurality of genetic loci a quantitative measure related to sequencing read coverage ("read coverage"), wherein the first data set comprises, for one or more genetic loci in a plurality of genetic loci, a quantitative measure related to guanine-cytosine content ("GC content");

(ii) removing from the first data set genetic loci that are high-variance genetic loci, wherein removing comprises:

(1) fitting a model relating the quantitative measures related to guanine-cytosine content and the quantitative measures of sequencing read coverage of the genetic loci; and

(2) removing from the first data set between 10% and 30% of the genetic loci that most differ from the model, thereby providing the first data set of baselining genetic loci;

(iii) correcting the first data set of baselining genetic loci by performing saturation equilibrium correction and probe efficiency correction, wherein performing saturation equilibrium correction comprises transforming the first data set of baselining genetic loci into a saturation corrected data set by

(1) determining for one or more genetic loci from the first data set of baselining genetic loci a quantitative measure related to the probability that a strand of DNA molecule from the sample derived from the genetic locus is represented within the sequencing reads;

(2) determining a first transformation for the read coverage by relating the read coverage in the first data set of baselining genetic loci to both the GC content of the first data set of baselining genetic loci and the quantitative measure related to the probability that a strand of DNA derived from each locus in the first data set of baselining genetic loci is represented within the sequencing reads; and

(3) applying the first transformation to the read coverage of one or more genetic loci from the first data set of baselining genetic loci to provide the saturation corrected data set, wherein the saturation corrected data set comprises a first set of transformed read coverages of the first data set of baselining genetic loci;

wherein performing probe efficiency correction comprises transforming the saturation corrected data set into a probe efficiency corrected data set by:

(1) removing from the saturation corrected data set genetic loci that are high-variance genetic loci with respect to the first set of

transformed read coverages, thereby providing a second data set of baselining genetic loci;

(2) determining a second transformation for the first set of transformed read coverages related to probe efficiency of the second data set of baselining genetic loci; and

(3) transforming the first set of transformed read coverages of the second data set of baselining genetic loci with the second transformation, thereby providing the probe efficiency corrected data set, wherein the probe efficiency corrected data set comprises a second set of transformed read coverages of the second data set of baselining genetic loci;

(iv) determining a baseline read coverage for the first data set, wherein the baseline read coverage relates to saturation equilibrium and probe efficiency; and

(v) determining a copy number state for one or more genetic loci in the plurality of genetic loci using the probe efficiency corrected data set relative to the baseline read coverage.

36. The system of claim 35, wherein said database is connected to a nucleic acid sequencer.