indicative of a dispersion of the plurality of values; accessing a predetermined relationship between one or more stored certainty metric and one or more stored tumor fraction; and determining, from the certainty metric and the predetermined relationship, the tumor fraction of the sample. In one aspect, the present disclosure refers to a method of determining a tumor fraction of a sample from a subject, comprising: acquiring a plurality of values, each value indicative of an allele fraction at a corresponding locus within a subgenomic interval in the sample; determining a certainty metric indicative of a dispersion of the plurality of values, wherein the certainty metric comprises a quantitative measure of a deviation or an entropy of the plurality of values relative to an expected value for a non-tumorous sample; accessing a predetermined relationship between one or more stored certainty metric and one or more stored tumor fraction; and determining, from the certainty metric and the predetermined relationship, the tumor fraction of the sample.

[0011] In some examples, each value within the plurality of values is an allele fraction. In some examples, each value within the plurality of values comprises a ratio of the difference in abundance between a maternal allele and a paternal allele relative to abundance of the maternal allele or the paternal allele at the corresponding locus. In some examples, the certainty metric is indicative of a deviation of each of the plurality of values from an expected value. In some examples, the expected value is a locus-specific expected value.

[0012] In some examples, the certainty metric is a root mean squared deviation from the expected value. In some examples, the expected value is an expected allele frequency for a non-tumorous. In some examples, each value within the plurality of values is and allele fraction, and the expected value is 0.5.

[0013] In some examples, each value within the plurality of values is a ratio of the difference in abundance between a maternal allele and a paternal allele, relative to abundance of the maternal allele or the paternal allele at the corresponding locus, and the expected value comprises the expected ratio of the difference in abundance between a maternal allele and a paternal allele relative, to abundance of the maternal allele or the paternal allele, wherein the expected value is the expected ratio for a non-tumorous sample. In some examples, the expected value is 0.

[0014] In some examples, the plurality of values comprises a plurality of allele coverages.

[0015] In some examples, the method further comprising determining a probability distribution function for the plurality of values; wherein the certainty metric is determined using the probability distribution function. In some examples, the certainty metric is an entropy of the probability distribution function.

[0016] In some examples, the corresponding loci comprise one or more loci having a different maternal allele and paternal allele. In some examples, the corresponding loci consist of loci having a different maternal allele and paternal allele. In some examples, the corresponding loci comprise one or more loci having the same maternal allele and paternal allele.

[0017] In some examples, a method of determining a tumor fraction of a sample from a subject, comprises: acquiring a plurality of values, each value indicative of a difference between an allele coverage of a locus in a tumor sample and an allele coverage of the same locus in a non-tumor sample at a plurality of loci within a subgenomic interval; determining a certainty metric indicative of a dispersion of the plurality of values; accessing a predetermined relationship between one or more stored certainty metric and one or more stored tumor fraction; and determining, from the certainty metric and the predetermined relationship, the tumor fraction of the sample. In one aspect, the present disclosure refers to a method of determining a tumor fraction of a sample from a subject, comprising: acquiring a plurality of values, each value indicative of a difference between an allele coverage of a locus in a tumor sample and an allele coverage of the same locus in a non-tumor sample at a plurality of loci within a subgenomic interval; determining a certainty metric indicative of a dispersion of the plurality of values, wherein the certainty metric comprises a quantitative measure of a deviation or an entropy of the plurality of values relative to an expected value for a non-tumorous sample; accessing a predetermined relationship between one or more stored certainty metric and one or more stored tumor fraction; and determining, from the certainty metric and the predetermined relationship, the tumor fraction of the sample.

[0018] In some examples, each value within the plurality of values comprises a ratio of an allele coverage of a locus in the tumor sample compared to an allele coverage of the same locus in the non-tumor sample.

[0019] In some examples, each value within the plurality of values comprises a log ratio of an allele coverage of a locus in the tumor sample compared to an allele coverage of the same locus in the non-tumor sample. In some examples, the log ratio is a log2 ratio.

[0020] In some examples, each value within the plurality of values comprises a ratio of the difference in an allele coverage the locus in the tumor sample and same locus in the non-tumor sample, relative to an allele coverage of the same locus in the non-tumor sample.

[0021] In some examples, the certainty metric is indicative of a deviation of each value within the plurality of values from an expected value across the corresponding loci, wherein the expected value is the value that would be expected if the tumor sample were a non-tumor sample.

[0022] In some examples, each value comprises a ratio of an allele coverage of a locus in the tumor sample compared to an allele coverage of the same locus in the non-tumor sample, and the expected value is 1; each value comprises a log ratio of an allele coverage of a locus in the tumor sample compared to an allele coverage of the same locus in the non-tumor sample, and the expected value is 0; or each value comprises a ratio of the difference in an allele coverage the locus in the tumor sample and same locus in the non-tumor sample, relative to an allele coverage of the same locus in the non-tumor sample, and the expected value is 0.

[0023] In some examples, the certainty metric is a root mean squared deviation from the expected value.

stored predetermined relationship; and (d) determine, from the certainty metric and the predetermined relationship, the tumor fraction of the sample. In one aspect, the present disclosure refers to a computer system comprising: a processor; and a memory communicatively coupled to the processor, configured to store: a predetermined relationship between a one or more stored certainty metric and one or more associated stored tumor fraction; and instructions that, when executed by the processor cause the processor to: (a)(i) acquire a plurality of values, each value indicative of an allele fraction at a corresponding locus within a subgenomic interval in the sample, or (ii) acquire plurality of values, each value indicative of a difference between an allele coverage of a locus in a tumor sample and an allele coverage of the same locus in a non-tumor sample at a plurality of loci within a subgenomic interval; (b) determine a certainty metric indicative of a dispersion for the plurality of values, wherein the certainty metric comprises a quantitative measure of a deviation or an entropy of the plurality of values relative to an expected value for a non-tumorous sample; (c) access the stored predetermined relationship; and (d) determine, from the certainty metric and the predetermined relationship, the tumor fraction of the sample.

[0045] In some examples, of the computer system, the memory further comprises instructions that, when executed by the processor, cause the processor to: access a training dataset comprising a plurality of relationships between a plurality of training certainty metrics and associated training tumor fractions; and apply a machine learning process to the training dataset to determine the predetermined relationship between the training certainty metrics and the training tumor fractions.

[0046] In some examples of the computer system, the instructions, when executed by the processor, cause the processor to perform any one of the methods described above.

BRIEF DESCRIPTION OF THE DRAWINGS

[0047] Various aspects of at least one example are discussed below with reference to the accompanying figures, which are not intended to be drawn to scale. The figures are included to provide an illustration and a further understanding of the various aspects and examples, and are incorporated in and constitute a part of this specification, but are not intended as a definition of the limits of a particular example. The drawings, together with the remainder of the specification, serve to explain principles and operations of the described and claimed aspects and examples. In the figures, each identical or nearly identical component that is illustrated in various figures is represented by a like numeral. For purposes of clarity, not every component may be labeled in every figure.

[0048] FIG. 1 depicts a process according to one embodiment. The disclosed process may be used to estimate a tumor fraction from a sample.

[0049] FIG. 2 shows an example computer system with which various aspects of the present disclosure may be practiced.

[0050] FIG. 3 shows an example storage system capable of implementing various aspects of the present disclosure.

CLAIMS

1. A method of determining a tumor fraction of a sample from a subject, comprising:

acquiring a plurality of values, each value indicative of an allele fraction at a corresponding locus within a subgenomic interval in the sample;

determining a certainty metric indicative of a dispersion of the plurality of values, wherein the certainty metric comprises a quantitative measure of a deviation or an entropy of the plurality of values relative to an expected value for a non-tumorous sample;

accessing a predetermined relationship between one or more stored certainty metric and one or more stored tumor fraction; and

determining, from the certainty metric and the predetermined relationship, the tumor fraction of the sample.

2. The method of claim 1, wherein:

each value within the plurality of values is an allele fraction; or

each value within the plurality of values comprises a ratio of the difference in abundance between a maternal allele and a paternal allele relative to abundance of the maternal allele or the paternal allele at the corresponding locus.

3. The method of claim 1 or 2, wherein the certainty metric is indicative of a deviation of each of the plurality of values from an expected value:

optionally wherein:

the expected value is a locus-specific expected value; or

the certainty metric is a root mean squared deviation from the expected value; or

the expected value is an expected allele frequency for a non-tumorous sample; or

each value within the plurality of values is an allele fraction, and the expected value is 0.5; or

each value within the plurality of values is a ratio of the difference in abundance between a maternal allele and a paternal allele, relative to abundance of the

maternal allele or the paternal allele at the corresponding locus, and the expected value comprises the expected ratio of the difference in abundance between a maternal allele and a paternal allele relative, to abundance of the maternal allele or the paternal allele, wherein the expected value is the expected ratio for a non-tumorous sample; or

the plurality of values comprises a plurality of allele coverages.

4. The method of any one of claims 1-3, the method further comprising determining a probability distribution function for the plurality of values; wherein the certainty metric is determined using the probability distribution function;

optionally wherein the certainty metric is an entropy of the probability distribution function.

5. The method of any one of claims 1-4, wherein:

the corresponding loci comprise one or more loci having a different maternal allele and paternal allele; or

the corresponding loci consist of loci having a different maternal allele and paternal allele; or

the corresponding loci comprise one or more loci having the same maternal allele and paternal allele.

6. A method of determining a tumor fraction of a sample from a subject, comprising:

acquiring a plurality of values, each value indicative of a difference between an allele coverage of a locus in a tumor sample and an allele coverage of the same locus in a non-tumor sample at a plurality of loci within a subgenomic interval;

determining a certainty metric indicative of a dispersion of the plurality of values, wherein the certainty metric comprises a quantitative measure of a deviation or an entropy of the plurality of values relative to an expected value for a non-tumorous sample;

accessing a predetermined relationship between one or more stored certainty metric and one or more stored tumor fraction; and

determining, from the certainty metric and the predetermined relationship, the tumor fraction of the sample.

7. The method of claim 6, wherein:

each value within the plurality of values comprises a ratio of an allele coverage of a locus in the tumor sample compared to an allele coverage of the same locus in the non-tumor sample; or

each value within the plurality of values comprises a log ratio of an allele coverage of a locus in the tumor sample compared to an allele coverage of the same locus in the non-tumor sample; or

each value within the plurality of values comprises a ratio of the difference in an allele coverage the locus in the tumor sample and same locus in the non-tumor sample, relative to an allele coverage of the same locus in the non-tumor sample; or

the certainty metric is indicative of a deviation of each value within the plurality of values from an expected value across the corresponding loci, wherein the expected value is the value that would be expected if the tumor sample were a non-tumor sample, optionally wherein:

(i) each value comprises a ratio of an allele coverage of a locus in the tumor sample compared to an allele coverage of the same locus in the non-tumor sample, and the expected value is 1; or

(ii) each value comprises a log ratio of an allele coverage of a locus in the tumor sample compared to an allele coverage of the same locus in the non-tumor sample, and the expected value is 0; or

(iii) each value comprises a ratio of the difference in an allele coverage of the locus in the tumor sample and same locus in the non-tumor sample, relative to an allele coverage of the same locus in the non-tumor sample, and the expected value is 0.

8. The method of claim 7, the method further comprising determining a probability distribution function for the plurality of values; wherein the certainty metric is determined using the probability distribution function;

optionally wherein the certainty metric is an entropy of the probability distribution function.

9. The method of claim 7 or 8, wherein:

the allele coverage comprises an allele coverage of a maternal allele and a paternal allele; or

the allele coverage consists of an allele coverage of a maternal allele and a paternal allele; or

the allele coverage consists of an allele coverage of a maternal allele and a paternal allele.

10. The method of any one of claims 1-9, wherein:

the plurality of loci comprises at least one nucleotide associated with a single nucleotide polymorphism (SNP); optionally wherein (i) the plurality of loci comprises two or more nucleotides each associated with a single nucleotide polymorphism (SNP), and/or (ii) the SNP is associated with a cancer; or

at least a portion of the plurality of loci is associated with a copy number variation (CNV), optionally wherein the CNV is associated with a cancer.

11. The method of any of claims 1-10, further comprising:

sequencing the sample, to determine an allele abundance or coverage at each locus; or

performing array hybridization on the sample to determine an allele abundance or coverage at each locus.

12. The method of any of claims 1-11, further comprising:

accessing a training dataset comprising a plurality of relationships between a plurality of training certainty metrics and associated training tumor fractions; and

applying a machine learning process to the training dataset to determine the predetermined relationship between the training certainty metrics and the training tumor fractions.

13. The method of any one of claims 1-12, comprising generating a report comprising information identifying the subject and the determined tumor fraction; and

optionally comprising providing the report to the subject or a healthcare provider, and/or formatting the report for an electronic health record.

14. A method of monitoring tumor progression or recurrence in a subject, comprising:

(a) determining a first tumor fraction of a first sample obtained from the subject at a first time point according to the method of any one of claims 1-13;

(b) determining a second tumor fraction of a second sample obtained from the subject at a second time point; and

(c) comparing the first tumor fraction to the second tumor fraction, thereby monitoring the tumor progression.

15. The method of claim 14, wherein:

(a) determining the second tumor fraction comprises:

acquiring a second plurality of values, each value indicative of an allele fraction at a corresponding locus within a subgenomic interval in the second tumor sample, wherein the subgenomic interval in the second sample is the same or different than the subgenomic interval in the first sample;

determining a second certainty metric indicative of a dispersion of the second plurality of values, wherein the second certainty metric comprises a quantitative measure of a deviation or an entropy of the second plurality of values relative to an expected value for a non-tumorous sample;

accessing the predetermined relationship between one or more stored certainty metrics and one or more stored tumor fractions; and

determining, from the second certainty metric and the predetermined relationship, the second tumor fraction of the second sample; or

(b) determining the second tumor fraction comprises:

acquiring a second plurality of values, each value indicative of a difference between an allele coverage of a locus in the second tumor sample and an allele coverage of the same locus in a non-tumor sample at a plurality of loci within a subgenomic interval in the sample, wherein the subgenomic interval used to determine the second tumor fraction is the same or different than the subgenomic interval used to determine the first tumor fraction;

determining a second certainty metric indicative of a dispersion of the second plurality of values, wherein the second certainty metric comprises a quantitative measure

of a deviation or an entropy of the second plurality of values relative to an expected value for a non-tumorous sample;

accessing the predetermined relationship between one or more stored certainty metrics and one or more stored tumor fractions; and

determining, from the second certainty metric and the predetermined relationship, the second tumor fraction of the second tumor sample.

16. The method of claim 14 or 15, wherein the first time point is before the subject has been administered a tumor therapy, and wherein the second time point is after the subject has been administered the tumor therapy.

17. The method of any one of claims 1-16, wherein:

the subject has a cancer, is at risk of having a cancer, or is suspected of having a cancer; or

the sample is a liquid sample; or

the sample is a solid sample; or

the sample comprises cell-free DNA (cfDNA) or circulating tumor DNA (ctDNA); or the one or more stored certainty metrics comprises a plurality of stored certainty metrics, and the one or more stored tumor fractions comprises plurality of stored tumor fractions.

18. The method of any one of claims 1-17, wherein the dispersion is:

(a) determined as a root mean squared deviation (RMSD) defined by:

$$RMSD = \left[\frac{1}{N}\sum_{i=0}^{N}\left(af_i - af_{i,expected}\right)^2\right]^{(1/2)}$$

wherein af_i is an allele frequency or value indicative of the allele frequency, at locus *i*, $af_{expected}$is an expected allele frequency at locus *i*, and *N* is the number of loci in the plurality of corresponding loci; or

(b) determined as a root mean squared deviation (RMSD) defined by:

$$RMSD = \left[\frac{1}{N}\sum_{i=0}^{N}\left(\frac{Cvg_{i,a} - Cvg_{i,b}}{Cvg_{i,b}}\right)^2\right]^{(1/2)}$$

wherein $Cvg_{i,a}$ is the coverage of allele *a* at locus *i*, and $Cvg_{i,b}$ is the coverage of allele *b* at locus *i*; or

(c) determined as an entropy of a probability distribution function defined by:

$$S[P(af)] = \sum_{af=0}^{0.5} P(af)log_n\big(P(af)\big)$$

wherein $P(af)$is the allele fraction probability distribution function, and *n* is the log base; or

(d) determined as a root mean squared deviation (RMSD) defined by:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i}^{N}\left[\left(\mathbf{log_2}\frac{\left(Cvg_{i,a}^{cancer} + Cvg_{i,b}^{cancer}\right)}{\left(Cvg_{i,a}^{normal} + Cvg_{i,b}^{normal}\right)}\right) - \mathbf{0}\right]^{\mathbf{2}}}$$

or

$$RMSD = \left[\frac{1}{N}\sum_{i=0}^{N}\left(\frac{\left(Cvg_{i,a}^{cancer} + Cvg_{i,b}^{cancer}\right) - \left(Cvg_{i,a}^{normal} + Cvg_{i,b}^{normal}\right)}{\left(Cvg_{i,a}^{normal} + Cvg_{i,b}^{normal}\right)}\right)^2\right]^{(1/2)}$$

wherein $Cvg_{i,a}^{Cancer}$ is the coverage of the maternal allele at the locus *i* within the tumor sample, $Cvg_{i,b}^{Cancer}$ is the coverage of the paternal allele at the locus *i* within the tumor sample, $Cvg_{i,a}^{Normal}$ is the coverage of the maternal allele at the locus *i* within the non-tumor sample, and $Cvg_{i,b}^{Normal}$ is the coverage of the paternal allele at the locus *i* within the non-tumor sample; or

(e) determined as an entropy of a probability distribution function defined by:

$$S[P(lnr)] = \sum_{lnr=0} P(lnr)log_n\big(P(lnr)\big)$$

wherein:

$$P(lnr) = P\left(\mathbf{log_n}\left(\frac{Cvg_{i,a}^{cancer} + Cvg_{i,b}^{cancer}}{Cvg_{i,a}^{normal} + Cvg_{i,b}^{normal}}\right)\right)$$

wherein log_n is a log having base *n*, $Cvg_{i,a}^{Cancer}$ is the coverage of the maternal allele at the locus *i* within the tumor sample, $Cvg_{i,b}^{Cancer}$ is the coverage of the paternal allele at the locus *i* within the tumor sample, $Cvg_{i,a}^{Normal}$ is the coverage of the maternal allele at the locus *i* within the non-tumor sample, and $Cvg_{i,b}^{Normal}$ is the coverage of the paternal allele at the locus *i* within the

non-tumor sample.

19. A computer system comprising:

a processor; and

a memory communicatively coupled to the processor, configured to store:

a predetermined relationship between a one or more stored certainty metric and one or more associated stored tumor fraction; and

instructions that, when executed by the processor cause the processor to:

(a)(i) acquire a plurality of values, each value indicative of an allele fraction at a corresponding locus within a subgenomic interval in the sample, or (ii) acquire plurality of values, each value indicative of a difference between an allele coverage of a locus in a tumor sample and an allele coverage of the same locus in a non-tumor sample at a plurality of loci within a subgenomic interval;

(b) determine a certainty metric indicative of a dispersion for the plurality of values, wherein the certainty metric comprises a quantitative measure of a deviation or an entropy of the plurality of values relative to an expected value for a non-tumorous sample;

(c) access the stored predetermined relationship; and

(d) determine, from the certainty metric and the predetermined relationship, the tumor fraction of the sample.

20. The computer system of claim 19, wherein the memory further comprises instructions that, when executed by the processor, cause the processor to:

access a training dataset comprising a plurality of relationships between a plurality of training certainty metrics and associated training tumor fractions; and

apply a machine learning process to the training dataset to determine the predetermined relationship between the training certainty metrics and the training tumor fractions.

21. The computer system of claim 19 or 20, wherein the instructions, when executed by the processor, cause the processor to perform the method of any one of claims 1-13.