CLAIMS

1. A method of determining a microsatellite instability (MSI) status of a sample, wherein the sample comprises cell-free nucleic acids, the method comprising:

(a) quantifying a number of different repeat lengths present at each of a plurality of microsatellite loci from sequence information to generate a site score for each of the plurality of the microsatellite loci,

wherein the sequence information is from a population of microsatellite loci in the sample,

wherein the site scores of the plurality of the microsatellite loci comprise likelihood scores, and wherein the likelihood scores comprise probabilistic log likelihood-based scores that discriminate biological signal derived from a number of cfDNA fragments of somatic origin in the sample from noise arising from post-sample collection artifacts in the sample,

wherein the method comprises determining the probabilistic log likelihood-based score for an individual microsatellite locus in the sequence information from the sample using at least two parameters, wherein at least a first parameter comprises allele frequencies and at least a second parameter comprises at least one error mode;

(b) comparing the site score of a given microsatellite locus to a site specific trained threshold for the given microsatellite locus for each of the plurality of the microsatellite loci;

(c) calling the given microsatellite locus as being unstable when the site score of the given microsatellite locus exceeds the site specific trained threshold for the given microsatellite locus to generate a microsatellite instability score comprising a number of unstable microsatellite loci from the plurality of the microsatellite loci; and

(d) classifying the MSI status of the sample as being unstable when the microsatellite instability score exceeds a population trained threshold for the population of microsatellite loci in the sample, thereby determining the MSI status of the sample;

2. The method of claim 1, the method further comprising:

(e) comparing the MSI status of the sample to one or more comparator results that are indexed with one or more therapies; and,

(f) identifying one or more customized therapies for treating a disease in a subject when there is a substantial match between the microsatellite instability status of the sample and the comparator results.

3. The method of claim 1 or claim 2,

wherein the allele frequencies comprise frequencies of nucleic acids comprising different repeat lengths in the sequence information from the sample; and/or

wherein the at least one error mode comprises a random error mode and a strand specific error mode.

4. The method of any one of the preceding claims, wherein the site scores of the plurality of the microsatellite loci

(a) comprise a difference between or ratio of:

(i) a score measuring a support of observed sequences for a null hypothesis that the given microsatellite locus is stable, and

(ii) a score measuring a support of observed sequences for an alternate hypothesis that the given microsatellite locus is unstable; and/or

(b) are generated using one or more of: a likelihood criterion, a log-likelihood criterion, a posterior probability criterion, an Akaike information criterion (AIC), and a Bayesian information criterion.

5. The method of any one of the preceding claims, wherein the site scores of the plurality of the microsatellite loci comprise Akaike Information Criterion (AIC)-based site scores that test for a presence of somatic indels at the plurality of the microsatellite loci, wherein a given AIC-based site score is calculated using the formula of:

$$AIC = k - \log\text{-likelihood},$$

where k is the number of parameters used in a model, the method comprising calculating:

(a) a null hypothesis score of the model using the formula of:

$$AIC_0 = k - \log(Pr(obs|\boldsymbol{\beta}, \boldsymbol{\gamma})),$$

where AIC_0 is the null hypothesis, k is the number of parameters used in the model, Pr is probability, obs comprises repeat lengths of observed sequencing reads covering the given microsatellite locus, $\boldsymbol{\beta}$ is at least one strand specific error parameter, and $\boldsymbol{\gamma}$ is at least one random error parameter; and/or

(b) an alternate hypothesis score of the model using the formula of:

$$AIC_{min} = min_{\alpha}(k - \log(Pr(obs|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha})),$$

where AIC_{min} is the alternate hypothesis, min_{α} an effect of minimizing over all values of $\boldsymbol{\alpha}$, k is the number of parameters used in the model, Pr is probability, obs comprises repeat lengths of observed sequencing reads covering the given microsatellite locus, $\boldsymbol{\beta}$ is at least

one strand specific error parameter, γ is at least one random error parameter, and α is at least one allele frequency, wherein α is a vector of allele frequencies such that the sum of one or more α_i is equal to one.

6. The method of any one of the preceding claims, comprising estimating the parameters of the model using a maximum likelihood estimation (MLE), optionally comprising determining the MLE using a Nelder-Mead algorithm.

7. The method of any one of the preceding claims, comprising detecting the site score using the formula of:

$$\Delta AIC = AIC_0 - AIC_{min}.$$

8. The method of claim 5,

wherein γ comprises:

(a) a rate of read-level errors where a microsatellite length observed within a sequencing read is one repeat unit longer than an expected microsatellite length for a strand of an originating nucleic acid molecule; and/or

(b) a rate of read-level errors where a microsatellite length observed within a sequencing read is one repeat unit shorter than an expected microsatellite length for a strand of an originating nucleic acid molecule;

and/or wherein β comprises:

(c) a rate of strand-level errors where an expected microsatellite length of a sense strand is one repeat unit longer than an expected microsatellite length of an nucleic acid originating molecule;

(d) a rate of strand-level errors where an expected microsatellite length of an antisense strand is one repeat unit longer than an expected microsatellite length of an nucleic acid originating molecule;

(e) a rate of strand-level errors where an expected microsatellite length of a sense strand is one repeat unit shorter than an expected microsatellite length of an nucleic acid originating molecule; and/or,

(f) a rate of strand-level errors where an expected microsatellite length of an antisense strand is one repeat unit shorter than an expected microsatellite length of an nucleic acid originating molecule.

9. The method of any one of the preceding claims, comprising determining the site specific trained threshold and/or the population trained threshold from sequence information from a population of

microsatellite loci in one or more training DNA samples, wherein the training DNA samples comprise

(a) non-tumor cfDNA samples; or

(b) DNA from one or more tumor types.

10. The method of any one of the preceding claims, wherein

(a) the method comprises a sensitivity of at least 94% at a limit of detection (LOD) of a 0.2% tumor fraction of nucleic acids in the sample;

(b) the method comprises a specificity of at least 99% for non-tumor DNA in the sample; and/or

(c) the determined MSI status of the sample comprises at least 95%, 96%, 97%, 98%, or 99% concordance with a corresponding MSI status of the sample determined using a PCR-based MSI assessment technique across a tumor fraction range of 1.4% to 15%, wherein the concordance is 100%.

11. The method of any one of the preceding claims, comprising classifying the MSI status of the sample as MSI-high (MSI-H) when

(a) the microsatellite instability score is greater than 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30, 40, 50, or more unstable microsatellite loci from the plurality of the microsatellite loci; and/or

(b) the number of unstable microsatellite loci comprises 0.1%, 1%, 2%, 3%, 4%, 5%, 10%, 15%, 20%, or 25% of the plurality of the microsatellite loci.

12. The method of claim 2,

(a) wherein the disease comprises a cancer comprising at least one tumor type selected from the group consisting of: biliary tract cancer, bladder cancer, transitional cell carcinoma, urothelial carcinoma, brain cancer, gliomas, astrocytomas, breast carcinoma, metaplastic carcinoma, cervical cancer, cervical squamous cell carcinoma, rectal cancer, colorectal carcinoma, colon cancer, hereditary nonpolyposis colorectal cancer, colorectal adenocarcinomas, gastrointestinal stromal tumors (GISTs), endometrial carcinoma, endometrial stromal sarcomas, esophageal cancer, esophageal squamous cell carcinoma, esophageal adenocarcinoma, ocular melanoma, uveal melanoma, gallbladder carcinomas, gallbladder adenocarcinoma, renal cell carcinoma, clear cell renal cell carcinoma, transitional cell carcinoma, urothelial carcinomas, Wilms tumor, leukemia, acute lymphocytic leukemia (ALL), acute myeloid leukemia (AML), chronic lymphocytic (CLL), chronic myeloid (CML), chronic myelomonocytic (CMML), liver cancer, liver carcinoma, hepatoma, hepatocellular carcinoma, cholangiocarcinoma, hepatoblastoma, Lung cancer, non-small cell lung cancer (NSCLC), mesothelioma, B-cell lymphomas, non-Hodgkin lymphoma, diffuse large B-cell lymphoma, Mantle cell lymphoma, T cell lymphomas,

non-Hodgkin lymphoma, precursor T-lymphoblastic lymphoma/leukemia, peripheral T cell lymphomas, multiple myeloma, nasopharyngeal carcinoma (NPC), neuroblastoma, oropharyngeal cancer, oral cavity squamous cell carcinomas, osteosarcoma, ovarian carcinoma, pancreatic cancer, pancreatic ductal adenocarcinoma, pseudopapillary neoplasms, acinar cell carcinomas. prostate cancer, prostate adenocarcinoma, skin cancer, melanoma, malignant melanoma, cutaneous melanoma, small intestine carcinomas, stomach cancer, gastric carcinoma, gastrointestinal stromal tumor (GIST), uterine cancer, and uterine sarcoma;

(b) wherein the therapies comprise at least one immunotherapy, wherein the immunotherapy comprises:

at least one checkpoint inhibitor antibody;

an antibody against PD-1, PD-2, PD-L1, PD-L2, CTLA-4, OX40, B7.1, B7He, LAG3, CD137, KIR, CCR5, CD27, CD40, or CD47;

administration of a pro-inflammatory cytokine against at least one tumor type; or

administration of T cells against at least one tumor type.

13. The method of any one of the preceding claims, further comprising

(a) obtaining the sample from a subject, optionally wherein the sample is selected from the group consisting of: tissue, blood, plasma, serum, sputum, urine, semen, vaginal fluid, feces, synovial fluid, spinal fluid, and saliva:

wherein the subject is a mammalian subject, the mammalian subject comprising a human subject; and/or

wherein the sample comprises circulating tumor nucleic acids; and/or

(b) receiving the sequence information generated from the sample, wherein the sequence information comprises cfDNA sequencing reads from the population of microsatellite loci in the sample.

14. The method of claim 13,

(a) comprising amplifying one or more segments of nucleic acids in the sample to generate at least one amplified nucleic acid;

(b) further comprising sequencing nucleic acids from the sample to generate the sequence information, optionally wherein the sequencing is selected from the group consisting of: targeted sequencing, intron sequencing, exome sequencing, and whole genome sequencing, the method comprising sequencing at least 50, 100, 150, 200, 250, 500, 750, 1,000, 1,500, 2,000, or more targeted genomic regions in the nucleic acids of the sample to generate the sequence information;

(c) wherein the sequence information is obtained from targeted segments of nucleic acids in the sample, wherein the targeted segments are obtained by selectively enriching one or more regions from the nucleic acids in the sample prior to sequencing, the method optionally further comprising amplifying the obtained targeted segments prior to sequencing; and/or

(d) further comprising attaching one or more adapters comprising barcodes to the nucleic acids prior to sequencing.

15. The method of any one of the preceding claims,

(a) comprising calling the given microsatellite locus as being unstable when the site score of the given microsatellite locus statistically exceeds the site specific trained threshold for the given microsatellite locus;

(b) wherein the sample comprises a mutant allele fraction and/or wherein the sample comprises a tumor fraction, optionally wherein the tumor fraction comprises a maximum mutant allele fraction (MAF) of all somatic mutations identified in the nucleic acids in the sample, wherein the tumor fraction is below 0.05%, 0.1%, 0.2%, 0.5%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 11%, 12%, 13%, 14%, or 15% of all nucleic acids in the sample;

(c) wherein the plurality of microsatellite loci comprises (i) all of the population of microsatellite loci or (ii) a subset of the population of microsatellite loci;

(d) wherein the number of different repeat lengths comprises a frequency of each different repeat length present at each of the plurality of microsatellite loci;

(e) wherein at least a portion of the method is computer implemented;

(f) further comprising generating a report which optionally includes information on the instability status of the sample, information on the instability score of the sample and/or information on the one or more customized therapies for treating the disease in the subject, the method optionally further comprising communicating the report to a third party, the third party comprising the subject or a health care provider; and/or

(g) further comprising classifying the microsatellite instability status of the sample as being stable if the microsatellite instability score is below or at the population trained threshold for the population of microsatellite loci in the sample.