

CLAIMS:

1. A system for determining microsatellite instability, comprising:
a processor; and
a memory storing instructions that, when executed by the processor, cause the processor to:
 - access genomic sequence data of a sample of interest, the sample of interest being derived from a tumor sample for which a matched normal sample is unavailable, wherein the sequence data comprises nucleotide identity information for a plurality of microsatellite regions;
 - receive sample information related to the sample of interest;
 - select an associated reference sample dataset from a plurality of reference sample datasets based on the sample information, wherein each of the reference sample datasets are generated from nucleotide identity information for the plurality of microsatellite regions and from a plurality of individuals;
 - classify microsatellite instability for the sample of interest based on a comparison of the sequence data from the sample of interest to the associated reference sample dataset; and
 - provide an indication representative of microsatellite instability of the sample of interest based on the classification.
2. The system of claim 1, wherein the sample information comprises sample of interest origin information, wherein the plurality of reference sample datasets differ from one another based on origin, and wherein the associated reference sample dataset is selected based on a match between the sample of interest origin information and the origin of the associated reference sample dataset.

3. The system of claim 1 or claim 2, wherein the associated reference sample dataset is generated from FFPE samples from a plurality of individuals and the sample of interest is an FFPE sample.

4. The system of claim 1 or claim 2, wherein the associated reference sample dataset is generated from fresh frozen samples from a plurality of individuals and the sample of interest is a fresh frozen sample.

5. The system of claim 1 or claim 2, wherein the associated reference sample dataset is generated from cell lines from a plurality of individuals and the sample of interest is a cell line.

6. The system of any one of claims 1 to 5, wherein the sample information comprises tissue type information, wherein the plurality of reference sample datasets differ from one another based on tissue type, and wherein the associated reference sample dataset is further selected based on a match between the tissue type information and the tissue type of the associated reference sample dataset.

7. The system of any one of claims 1 to 6, wherein the sample information comprises sequencing panel information used to generate the sequence data, wherein the plurality of reference sample datasets differ from one another based on a sequencing panel used to generate the reference sample datasets, and wherein the associated reference sample dataset is further selected based on a match between the sequencing panel information and the sequencing panel used to generate the associated reference sample dataset.

8. The system of any one of claims 1 to 7, wherein the associated reference sample dataset is a pooled dataset from the plurality of individuals.

9. The system of any one of claims 1 to 8, wherein the plurality of reference sample datasets are generated from normal tissue of the plurality of individuals.
10. The system of any one of claims 1 to 9, wherein the sample of interest is not matched to samples used to generate the plurality of reference sample datasets.
11. A computer-implemented method, comprising:
- acquiring, using a microprocessor, genomic reference sequence data from a plurality of reference biological samples corresponding to respective individuals;
 - analyzing the reference sequence data to generate a distribution of sequences at each of a plurality of microsatellite regions;
 - determining ethnic group variability of the distribution at each of the plurality of microsatellite regions for the plurality of reference biological samples, the ethnic group variability including genomic sequence differences;
 - identifying ethnically biased microsatellite regions of the plurality of microsatellite regions based on the ethnic group variability at each of the plurality of microsatellite regions; and
 - generating a reference sample dataset by removing or filtering the ethnically biased microsatellite regions from the reference sequence data of the plurality of reference biological samples.
12. The method of claim 11, further comprising:
- acquiring second reference sequence data from a second plurality of reference biological samples corresponding to respective individuals; and
 - removing the ethnically biased microsatellite regions from the second to generate a second reference sample dataset.

13. The method of claim 11 or claim 12, further comprising providing instructions to assess microsatellite instability based on a comparison of sequence data from a sample of interest to the reference sample dataset,

wherein the sample of interest is derived from a tumor sample of an individual and wherein a matched normal sample from the individual to the sample of interest is not available.

14. The method of any one of claims 11 to 13, wherein the plurality of reference biological samples are derived from normal tissue that is not from the individual.

15. A sequencing device configured to acquire tumor sequence data of a tumor sample, comprising:

a memory device including executable application instructions stored therein; and

a processor configured to execute the application instructions stored in the memory device, wherein the application instructions comprise instructions that cause the processor to:

receive the tumor sequence data from sequencing device;

identify a distribution of a plurality of microsatellite regions in the tumor sequence data;

determine that the tumor sample is not associated with a matched normal sample;

access reference sequence data;

determine a microsatellite instability type of the tumor sample based on a comparison of the distribution of the tumor sample to a reference distribution of the reference sample dataset; and

provide an indication of a treatment option based on a determination that the tumor sample is a microsatellite instability high type.

16. The sequencing device of claim 15, wherein the reference dataset comprises distribution data of a plurality of microsatellite regions and wherein determining the microsatellite instability type of the tumor sample based on the comparison of the distribution comprises comparing only a subset of the plurality of microsatellite regions.

17. The sequencing device of claim 16, wherein the subset is selected based on a sample type of the tumor sample.

18. The sequencing device of claim 17, wherein the subset is a first subset is selected based on a frozen solid tumor sample type and a second subset is selected based on a plasma tumor sample type, wherein the first subset is different than the second subset.

19. The sequencing device of claim 17 or claim 18, wherein the subset is selected based on a cancer type of the tumor sample.

20. The sequencing device of claim 16, wherein the subset is selected based on a ranking of a distance of distribution of individual microsatellite regions of the plurality of microsatellite regions.

21. The sequencing device of claim 20, wherein the subset is selected based on the individual microsatellite regions of the plurality of microsatellite regions having lowest distances of distribution.

22. The sequencing device of claim 21, wherein the subset represents less than 20% of the plurality of microsatellite regions.

23. The sequencing device of any one of claims 16 to 22, wherein the application instructions comprise instructions that cause the processor to provide an indication of a different treatment option based on a determination that the tumor sample is a microsatellite instability stable type.

24. A computer-implemented method for detecting microsatellite instability in a sample of interest, comprising:

providing sequence data of the sample of interest, the sample of interest being derived from a tumor sample for which a matched normal sample is unavailable, wherein the sequence data comprises nucleotide identity information for a plurality of microsatellite regions;

providing sample information related to the sample of interest;

selecting an associated reference sample dataset from a plurality of reference sample datasets based on the sample information, wherein each of the reference sample datasets are generated from nucleotide identity information for the plurality of microsatellite regions and from a plurality of individuals; and

assessing microsatellite instability of the sample of interest based on a computer-implemented comparison of the sequence data from the sample of interest to the associated reference sample dataset.

25. The method of claim 24, wherein the method uses a system according to any one of claims 1 to 10 or a sequencing device according to any one of claims 15 to 23.

26. Use of a system according to any one of claims 1 to 10 or a sequencing device according to any one of claims 15 to 23 to detect microsatellite instability in a sample of interest and/or to determine a microsatellite instability type in the sample of interest, wherein the sample has been derived from a tumor sample for which a matched normal sample is unavailable.