

SYSTEMS AND METHODS FOR DETERMINING MICROSATELLITE INSTABILITY

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The present application claims priority to and the benefit of U.S. Provisional Application No. 62/587,350, entitled “MICROSATELLITE INSTABILITY ASSESSMENT TECHNIQUES” FILED ON November 16, 2017, the disclosure of which is incorporated by reference in its entirety herein for all purposes. The present application also claims priority to and the benefit of U.S. Provisional Application No. 62/652,151, entitled “MICROSATELLITE INSTABILITY ASSESSMENT TECHNIQUES WITH REDUCED BIAS” filed on April 3, 2018, the disclosure of which is incorporated by reference in its entirety herein for all purposes.

BACKGROUND

[0002] The present disclosure relates generally to the field of data acquired from biological samples, such as sequence data. More particularly, the disclosure relates to techniques for assessing microsatellite instability via analysis of sequence data of biological samples that are independent of the presence of matched normal samples.

[0003] Genetic sequencing has become an increasingly important area of genetic research, promising future uses in diagnostic and other applications. Genetic sequencing data may be used to, among other applications, identify genetic mutations, modifications, variants, or polymorphisms that are associated with certain clinical outcomes. For example, certain genetic variants may be associated with a positive or negative disease outcome. Further, a subject’s genetic changes over time or relative to a matched normal sample may provide clinically useful information. However, matched normal samples may not be available for every subject.

BRIEF DESCRIPTION

[0004] The present disclosure provides improved techniques for detecting and characterizing microsatellite instability using sequence data from samples of interest. As provided herein, microsatellite instability may refer to the presence of nucleic acid replication errors in microsatellite repeat regions, which are short tandem repeat sequences (e.g., one to six base pairs in length) that are present throughout the genome. While microsatellite repeats may occur in untranslated regions of the genome, microsatellites may also be present in coding regions. During DNA replication, cells with microsatellite instability fail to repair DNA replication errors, which in turn may result in frame-shift mutations in the replicated daughter strand.

[0005] The presence of microsatellite instability may be associated with certain clinical conditions. For example, microsatellite instability is a hallmark of hereditary cancer syndrome, called Lynch Syndrome, based on germline mutations of mismatch repair genes such as MLH1, PMS2, MSH2 and MSH6. Microsatellite instability status is typically assessed in clinical labs as an independent prognostic factor for favorable survival in cancer types such as colorectal and endometrial tumors. Further, certain treatment protocols or treatment options may be initiated to administer nivolumab or pembrolizumab for patients with solid tumors that have microsatellite instability high (MSI-H) designations or that are mismatch repair deficient (dMMR). Further, the treatment option may be to not administer pembrolizumab for patients with solid tumors that have microsatellite stable designations per a microsatellite instability score as determined herein. In another embodiment, the MSI typing (high, low, stable) may be used to determine whether a patient may benefit from adjuvant 5-fluorouracil (5-FU) chemotherapy. For colorectal cancer patients, adjuvant 5-fluorouracil (5-FU) chemotherapy may provide limited benefits in MSI-H patients. Therefore, an MSI-H designation may lead to cessation of or contraindication of adjuvant 5-fluorouracil (5-FU) chemotherapy. Such patients may instead be offered folinic acid, 5-FU and oxaliplatin.

In another example, the MSI type of the patient may be used to determine if immunotherapy or chemotherapy is provided.

[0006] Accordingly, as provided herein, sequence data of samples of interest may be analyzed to determine a presence, absence, and/or degree of microsatellite instability in the sample of interest. Samples of interest with assessed microsatellite instability may be designated as MSI-H, microsatellite instability low (MSI-L), or microsatellite stable (MSS). The samples of interest may be tumor samples, and the microsatellite instability or stability designations may provide additional clinical information. As such, the present techniques may be used as part of or in conjunction with diagnosis, prognosis, and/or treatment protocols for cancer patients.

[0007] In certain embodiments, the present techniques permit assessment of samples of interest that do not have matched normal tissue samples. As provided herein, a reference sample dataset may be generated that is representative of a hypothetical matched normal sample for the sample of interest. The reference sample dataset may function as a universal matched normal sample. The reference sample dataset is generated from sequence data of the normal tissue of a plurality of individuals. When a tumor sample is tested, the appropriate reference sample dataset may be selected based on the tissue type, the sample origin, and other factors.

[0008] In certain embodiments, to generate a universal matched normal sample that may be applied to samples of interest independent of the ethnic background of the individual providing the sample, a reference sample dataset formed from samples of a multi-ethnic plurality of individuals (i.e., including individuals of a plurality of different ethnic backgrounds) may be assessed for microsatellite sites having relatively higher levels of variability between ethnic groups. Such sites may be eliminated or masked in the reference sample dataset, thus eliminating these highly variable sites from the analysis used to generate the overall microsatellite instability score representative of the sample of interest. In this manner, sites that are variable in normal samples due to variability between ethnic groups and not as a result of microsatellite instability do not introduce

potentially erroneous results into the microsatellite instability score. Accordingly, the present techniques provide a more accurate microsatellite instability assessment for samples without a matched normal and independent of the ethnic background of the samples. In one example, the present techniques may be used to assess microsatellite instability for samples for which no ethnic background identification information is present. In another example, the reference sample dataset used as the hypothetical matched normal and that is generated with ethnically variable microsatellite regions filtered out of the dataset may be generally application to a wide variety of samples, thus eliminating additional processing steps or selection of an appropriate reference sample based on the ethnic background of the individual providing the sample of interest.

[0009] In an embodiment, a computer-implemented method of processing microsatellite instability is provided that includes the steps of acquiring reference sequence data from a plurality of reference biological samples corresponding to respective individuals, each reference biological sample being associated with one of a plurality of ethnic groups, the reference sequence data comprising nucleotide identity information for a plurality of microsatellite regions; analyzing, using a microprocessor, the reference sequence data to generate a distribution at each of the plurality of microsatellite regions for the plurality of reference biological samples; determining, using a microprocessor, 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 being based on assessing reference sequence data associated with each ethnic group relative to other ethnic groups of the plurality of ethnic groups; determining ethnically unbiased microsatellite regions of the plurality of microsatellite regions having distributions with ethnic group variability below a threshold; generating, using a microprocessor, a reference sample dataset from the distribution at each of the determined ethnically unbiased microsatellite regions of the plurality of microsatellite regions; determining, using a microprocessor, 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 without using a matched normal sample

from the individual to the sample of interest; and outputting information on treatment options based on the determined microsatellite instability.

[0010] In another embodiment, a computer-implemented method is provided that includes the steps of 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.

[0011] In an embodiment, a method is provided that includes the steps of acquiring reference sequence data from a plurality of reference biological samples corresponding to respective individuals, and the reference sequence data comprising nucleotide identity information for a plurality of microsatellite regions; analyzing the reference sequence data to generate a distribution at each of the plurality of microsatellite regions; generating a reference sample dataset from the distribution at each of the plurality of microsatellite regions; and 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.

[0012] In another embodiment, a system is provided that includes 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, 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 for the sample of interest based on the classification.

[0013] In another embodiment, a system is provided that includes a processor; and a memory storing instructions that, when executed by the processor, cause the processor to access 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 matched sample information for the sample of interest; and analyze the sequence data according to a first microsatellite analysis technique when the matched sample information is indicative of an absence of a matched normal tissue sample to the sample of interest to generate a first output indicative of microsatellite instability of the sample of interest; and analyze the sequence data according to a second microsatellite analysis technique when the matched sample information is indicative of a presence of a matched normal tissue sample to the sample of interest to generate a second output indicative of microsatellite instability of the sample of interest.

[0014] In another embodiment, a sequencing device is provided that is configured to acquire tumor sequence data of a tumor sample. The device includes a memory device including executable application instructions stored therein; and a processor configured to execute the application instructions stored in the memory device. The application instructions include 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

DRAWINGS

[0015] FIG. 1 is a schematic illustration of microsatellite instability in accordance with the present techniques;

[0016] FIG. 2 is a block diagram of a sequencing device configured to acquire sequencing data in accordance with the present techniques;

[0017] FIG. 3 is a flow diagram of methods of assessing microsatellite instability of a sample in accordance with the present techniques;

[0018] FIG. 4 is a flow diagram of a workflow for assessing microsatellite instability in matched or unmatched samples of interest in accordance with the present techniques;

[0019] FIG. 5A is a schematic diagram of an example of a sequence read mapped to the microsatellite regions extracted from sequence data of colorectal tumor tissue and matched normal samples;

[0020] FIG. 5B upper panel shows for mapped reads of a microsatellite instability high (MSI-H) sample and the lower panel shows repeat unit length distributions for both tumor and normal samples in a MSI-H sample;

[0021] FIG. 5C upper panel shows mapped reads of a microsatellite stable (MSS) sample and the lower panel shows repeat unit length distributions for both tumor and normal samples in the MSS sample;

[0022] FIG. 6 shows prediction accuracy for single microsatellite sites for tumor only samples (y-axis) and tumor/normal pairs (x-axis);

[0023] FIG. 7A is boxplot of a microsatellite instability score based on tumor/normal pairs;

[0024] FIG. 7B is an ROC curve for tumor/normal pairs;

[0025] FIG. 7C is boxplot of a microsatellite instability score based on tumor only samples;

[0026] FIG. 7D is an ROC curve for tumor only samples;

[0027] FIG. 7E is a boxplot showing MSI-H samples with higher nonsynonymous tumor mutational burden (TMB) compared to MSS samples;

[0028] FIG. 8A is boxplot of a microsatellite instability score for 232 tumor/normal samples from a variety of tissue types and using a 58 sample normal colorectal cancer reference sample dataset matched to some of the tumor samples, with the red circled portion indicating false positives for MSI-H status per previous characterization as MSS based on MSI-PCR;

[0029] FIG. 8B is boxplot of a microsatellite instability score for 116 colorectal cancer matched tumor/normal samples using the 58 sample matched normal colorectal cancer reference sample dataset;

[0030] FIG. 9 is boxplot of a microsatellite instability score for 140 normal samples including samples from individuals associated with one of four different ethnic groups (African, South American, East Asian, and European) using the 58 sample normal colorectal cancer reference sample dataset;

[0031] FIG. 10 is a flow diagram of a method of removing bias based on ethnic variability from a reference sample dataset;

[0032] FIG. 11A shows the distribution of ethnicities in 140 samples used to assess ethnic variability in a reference sample dataset;

[0033] FIG. 11B shows results from an example technique of identifying microsatellite regions with relatively high ethnic variability in a reference sample dataset using calculated delta Jensen Shannon distances;

[0034] FIG. 12 is boxplot of a microsatellite instability score for 140 normal samples including samples from individuals associated with one of four different ethnic groups (African, South American, East Asian, and European) using the 58 sample normal colorectal cancer reference sample dataset with the identified microsatellite regions with relatively high ethnic variability filtered out of the reference sample dataset prior to the analysis;

[0035] FIG. 13 is boxplot of a microsatellite instability score of 232 tumor/normal samples from a variety of tissue types using the 58 sample normal colorectal cancer reference sample dataset post filtering with the red circle denoting potential false positives;

[0036] FIG. 14 is boxplot of a microsatellite instability score for normal samples associated with one of four different ethnic groups (African, South American, East Asian, and European) using 58 unmatched cell lines samples as the reference sample dataset pre and post filtering of the identified microsatellite regions with relatively high ethnic variability;

[0037] FIG. 15 is a comparison of the ethnic diversity of the unmatched cell lines samples reference dataset with the normal colorectal cancer reference sample dataset;

[0038] FIG. 16A is boxplot of a microsatellite instability score of 232 tumor/normal samples from a variety of tissue types using the unmatched cell lines samples as the reference sample dataset post filtering of the identified microsatellite regions with relatively high ethnic variability;

[0039] FIG. 16B shows the sensitivity and specificity of the results of FIG. 16A;

[0040] FIG. 17 is a comparison of an original and repeat run of 78 colorectal cancer samples;

[0041] FIG. 18 shows MSI score results correlation for reference sample datasets with varying numbers of samples;

[0042] FIG. 19 shows MSI score results correlation for reference sample datasets with varying numbers of samples;

[0043] FIG. 20 shows MSI scores for different titration levels of cell lines;

[0044] FIG. 21 is boxplot of a microsatellite instability score of 46 cell line samples including four MSI-H cell lines;

[0045] FIG. 22 shows limits of detection for titrated levels of Lovo cells using a microsatellite analysis technique according to embodiments of the disclosure;

[0046] FIG. 23 shows limits of detection for titrated levels of SW48 cells using a microsatellite analysis technique according to embodiments of the disclosure;

[0047] FIG. 24 shows limits of detection for titrated levels of Lovo cells using an improved and more stringent microsatellite analysis technique according to embodiments of the disclosure; and

[0048] FIG. 25 shows limits of detection for titrated levels of SW48 cells using an improved and more stringent microsatellite analysis technique according to embodiments of the disclosure.

DETAILED DESCRIPTION

[0049] Assessing microsatellite instability of a tumor sample may provide information about potential prognosis or treatment options for a patient. However, in the clinical

setting, matching normal tissues are not always available for samples of interest. For example, matched normal samples are often unavailable in retrospective studies when performing analysis with human material from clinical trials, pathology archives, and legacy bio-banks. In these cases, there is a need to identify and/or assess microsatellite instability from tumor tissues that do not have matched normal samples. Further, using a matched normal sample taken from the same individual as the biological sample presents certain challenges. For example, variation in sample collection (sample quality, selected tissue sites) may mean that reference sample is not truly representative of normal tissue. In addition, not all test samples have available matched tissue or matched tissue of sufficiently high quality for sequencing. Still further, samples of interest for a given assessment may be provided by individuals having a variety of ethnic backgrounds. Such variety is often desirable in studies to show effects of treatment protocols across the global population.

[0050] Microsatellite instability is typically detected by PCR (MSI-PCR) of certain microsatellites (e.g., using $n=5$ or 10 markers) followed by fragment length analysis through PCR and capillary electrophoresis to separate PCR amplicons. With MSI-PCR, each individual marker is evaluated by comparing how tumor markers shift from matched normal markers. That is, the instability is detected by a change in the characteristics of amplified alleles between the normal and the tumor sample. If more than 30% of microsatellites are shifted in the tumor sample compared to its matched normal sample, the tumor sample is categorized as MSI-high. If 10-20% of microsatellites are shifted, the tumor sample is categorized as MSI-low. If no microsatellite is shifted relative to the matched normal, the tumor sample is categorized as microsatellite stable.

[0051] In another example, immunohistochemical analysis (IHC) may be used to identify samples with microsatellite instability through identification of mismatch repair deficiencies. However mismatch repair IHC and microsatellite instability do not always correlate because other loss of function genes result in samples that exhibit the microsatellite instability phenotype (POLE). Samples that exhibit microsatellite

instability due to other loss of function genes would not be identified when screening for mismatch repair genes using IHC. Further, mutations in the mismatch repair gene MSH6 tend to result in weaker or no microsatellite instability in the tumors. Such MSH6 cases may be missed by microsatellite instability testing but can be detectable by MSH6 mutation screening. In general, IHC is reliable in screening for mutations that result in truncation or degradation of the protein. IHC, however, cannot distinguish between mutant proteins commonly resulting from missense mutation and wild-type polypeptides. MSI-PCR and other microsatellite instability assessment techniques require comparison of tumor DNA with a matched normal sample. Further, the small number of assessed markers may impact test sensitivity.

[0052] Provided herein are techniques for determining microsatellite instability that use sequence data from a sample of interest. The techniques may include analyzing the sample relative to a reference sample dataset that functions as a hypothetical matched normal sample, even if no matched normal sample is available for the sample of interest. The reference sample dataset may be generated from sequence data from an unmatched normal cohort (i.e., sequence data from different individuals than the individual from whom the sample of interest was generated). The unmatched normal cohort may act as a universal matched normal for samples of interest. The sequence data may be assessed for any suitable number of microsatellite markers. The disclosed techniques provide a reference sample dataset that may be used without relying on the presence of a matched normal sample from the individual from whom the test sample is obtained. The disclosed techniques also provide a reference sample dataset that is screened for microsatellite regions having high variability between the unmatched cohort normal samples as a result of variability in ethnic background in the cohort. In this manner, the reference sample dataset serves as a hypothetical matched normal to any sample of interest, regardless of the ethnic background of the individual providing the sample of interest. In this manner, assessment of samples via identification of microsatellite instability in the samples may be expanded to a wider number of samples, e.g., samples without a matched normal, relative to other techniques. Further, by using a universal matched normal, the potential

for user error via mismatching of tumor/normal samples in the analysis is reduced. That is, because the universal normal is the same sample for many different tumor samples, there is reduced possibility of misassignment of a tumor sample to its matched normal.

[0053] Accordingly, the disclosed techniques facilitate more accurate microsatellite assessment without using a matched sample. A universal or representative unmatched normal sample is generated using a set or cohort of unmatched reference biological samples. The representative unmatched normal sample information represents a virtual reference that may serve as a normal sample against which an individual tumor sample may be compared. The representative unmatched normal sample information represents a set of microsatellite regions having relatively low variability (e.g., lower than a pre-defined threshold) as a result of ethnic background variability in the cohort from which the unmatched normal sample information is generated.

[0054] To that end, FIG. 1 is a schematic illustration of microsatellite instability, represented as regions having different alleles caused by unrepaired replication errors. For example, as a result of polymerase slippage during replication, a parent strand (shown as strand 12a by way of example) may have a microsatellite region 14 having the sequence of $N(n)$, where n is the number of repeat motifs 16, while daughter strand may have a sequence of $N(n+1)$, e.g., as in the strand 12b, or $N(n-1)$, e.g., as in the strand 12c, depending on the nature of the error, which results in alleles of different lengths at the microsatellite region. As provided herein, the assessment of microsatellite instability may determine whether there is divergence between allele distribution for samples of interest and allele distribution for matched normal, if available, or a representative unmatched normal sample. As shown, the distribution of strands 12 varies based on the variability of the microsatellite regions 14.

[0055] FIG. 2 is a schematic diagram of a sequencing device 60 that may be used in conjunction with FIG. 1 for acquiring sequencing data (e.g., sample of interest sequencing data, unmatched cohort sequencing data) this is used for assessing microsatellite instability. The sequence device 60 may be implemented according to any

sequencing technique, such as those incorporating sequencing-by-synthesis methods described in U.S. Patent Publication Nos. 2007/0166705; 2006/0188901; 2006/0240439; 2006/0281109; 2005/0100900; U.S. Pat. No. 7,057,026; WO 05/065814; WO 06/064199; WO 07/010,251, the disclosures of which are incorporated herein by reference in their entireties. Alternatively, sequencing by ligation techniques may be used in the sequencing device 60. Such techniques use DNA ligase to incorporate oligonucleotides and identify the incorporation of such oligonucleotides and are described in U.S. Pat. No. 6,969,488; U.S. Pat. No. 6,172,218; and U.S. Pat. No. 6,306,597; the disclosures of which are incorporated herein by reference in their entireties. Some embodiments can utilize nanopore sequencing, whereby target nucleic acid strands, or nucleotides exonucleolytically removed from target nucleic acids, pass through a nanopore. As the target nucleic acids or nucleotides pass through the nanopore, each type of base can be identified by measuring fluctuations in the electrical conductance of the pore (U.S. Patent No. 7,001,792; Soni & Meller, *Clin. Chem.* 53, 1996–2001 (2007); Healy, *Nanomed.* 2, 459–481 (2007); and Cockroft, et al. *J. Am. Chem. Soc.* 130, 818–820 (2008), the disclosures of which are incorporated herein by reference in their entireties). Yet other embodiments include detection of a proton released upon incorporation of a nucleotide into an extension product. For example, sequencing based on detection of released protons can use an electrical detector and associated techniques that are commercially available from Ion Torrent (Guilford, CT, a Life Technologies subsidiary of/ThermoFisher) or sequencing methods and systems described in US 2009/0026082 A1; US 2009/0127589 A1; US 2010/0137143 A1; or US 2010/0282617 A1, each of which is incorporated herein by reference in its entirety. Particular embodiments can utilize methods involving the real-time monitoring of DNA polymerase activity. Nucleotide incorporations can be detected through fluorescence resonance energy transfer (FRET) interactions between a fluorophore-bearing polymerase and γ -phosphate-labeled nucleotides, or with zeromode waveguides as described, for example, in Levene et al. *Science* 299, 682–686 (2003); Lundquist et al. *Opt. Lett.* 33, 1026–1028 (2008); Korlach et al. *Proc. Natl. Acad. Sci. USA* 105, 1176–1181 (2008), the disclosures of

which are incorporated herein by reference in their entireties. Other suitable alternative techniques include, for example, fluorescent in situ sequencing (FISSEQ), and Massively Parallel Signature Sequencing (MPSS). In particular embodiments, the sequencing device 16 may be a HiSeq, MiSeq, or HiScanSQ from Illumina (La Jolla, CA).

[0056] In the depicted embodiment, the sequencing device 60 includes a separate sample processing device 62 and an associated analysis device 64. However, as noted, these may be implemented as a single device. Further, the analysis device 64 may be local to or networked with the sample processing device 62. In the depicted embodiment, the biological sample may be loaded into the sample processing device 62 as a sample slide 70 that is imaged to generate sequence data. For example, reagents that interact with the biological sample fluoresce at particular wavelengths in response to an excitation beam generated by an imaging module 72 and thereby return radiation for imaging. For instance, the fluorescent components may be generated by fluorescently tagged nucleic acids that hybridize to complementary molecules of the components or to fluorescently tagged nucleotides that are incorporated into an oligonucleotide using a polymerase. As will be appreciated by those skilled in the art, the wavelength at which the dyes of the sample are excited and the wavelength at which they fluoresce will depend upon the absorption and emission spectra of the specific dyes. Such returned radiation may propagate back through the directing optics. This retrobeam may generally be directed toward detection optics of the imaging module 72.

[0057] The imaging module detection optics may be based upon any suitable technology, and may be, for example, a charged coupled device (CCD) sensor that generates pixilated image data based upon photons impacting locations in the device. However, it will be understood that any of a variety of other detectors may also be used including, but not limited to, a detector array configured for time delay integration (TDI) operation, a complementary metal oxide semiconductor (CMOS) detector, an avalanche photodiode (APD) detector, a Geiger-mode photon counter, or any other suitable detector. TDI mode detection can be coupled with line scanning as described in U.S. Patent No. 7,329,860,

which is incorporated herein by reference. Other useful detectors are described, for example, in the references provided previously herein in the context of various nucleic acid sequencing methodologies.

[0058] The imaging module 72 may be under processor control, e.g., via a processor 74 (e.g., a microprocessor), and the sample receiving device 18 may also include I/O controls 76, an internal bus 78, non-volatile memory 80, RAM 82 and any other memory structure such that the memory is capable of storing executable instructions, and other suitable hardware components that may be similar to those described with regard to FIG. 2. Further, the associated computer 20 may also include a processor 84, I/O controls 86, a communications module 84, and a memory architecture including RAM 88 and non-volatile memory 90, such that the memory architecture is capable of storing executable instructions 92. The hardware components may be linked by an internal bus 94, which may also link to the display 96. In embodiments in which the sequencing device is implemented as an all-in-one device, certain redundant hardware elements may be eliminated.

[0059] FIG. 3 is a flow diagram of a method 100 of assessing microsatellite instability. The steps of the method 100 may be performed by a user and/or a provider as shown. For example, a user may be an end user of the sequencing device, such as an owner of the sequencing device, a contractor of the sequencing device, a user of the sequencing device. The user may be a user interested in identifying microsatellite instability in one or more samples. The provider may be a provider of the universal matched normal reference sequence as provided herein. Further, in certain embodiments, the user and the provider may be the same entity. That is, the microsatellite assessment may be performed by the provider of the universal matched normal reference sequence.

[0060] At step 102, a sample of interest is acquired and sample preparation for sequencing occurs at step 104. Sample preparation may be based on sample type (e.g., liquid sample, solid sample, FFPE sample, plasma sample). Sequence data may be acquired at step 106 using a sequencing device 60 as provided herein. In other

embodiments, previously acquired sequence data may be accessed. It should be understood that the biological sample sequencing data (i.e., the sample of interest, the representative unmatched normal samples, matched normal samples) as provided herein may be in the form of raw data, base call data providing nucleotide identities, or data that has gone through primary or secondary analysis (sequence alignment maps, binary alignment maps).

[0061] The analysis device may, via the display 96, provide a graphical user interface that facilitates user input of information related to sequencing reactions using the microsatellite instability assessment techniques as provided herein. For example, the user may provide input relating to a name or identification of each sample in the sequencing run, the sample origin (i.e., nucleic acids prepared from FFPE samples, fresh frozen samples, cell lines), a sequencing panel (i.e., a set of sequencing probes) used to acquire the sequence data, and the tissue type of the sample of interest. The user may also provide input related to whether a matched normal sample is available.

[0062] The present techniques facilitate detecting or assessment of microsatellite instability in biological samples (e.g., tumor samples) at step 108 without sequencing data from a matched normal sample. Accordingly, the method 100 acquires sequence data from a normal cohort at step 110. In certain embodiments, after generation and storage, the universal or representative normal sample sequence data, generated from a cohort of multiple samples, is used in the analysis of a plurality of samples of interest at different and/or subsequent time points. The user may access the stored files based on the cohort that most closely aligns with the sample of interest characteristics. To that end, multiple different normal cohort sequence data sets 112 may be acquired. Different normal cohort sequence data sets 112 may represent different sample types (nucleic acids prepared from normal FFPE samples, fresh frozen samples, cell lines), sequencing panels, tissue types, etc. The normal cohort sequence data 112 may be acquired from a suitable size cohort (at least 10 individuals, at least 20 individuals, at least 50 individuals) to provide a sufficient number of usable sequences for each microsatellite region examined. The individuals (or

samples representative of different individuals) in each cohort may provide samples from normal cells or tissues that may be used to acquire the normal cohort sequence data (i.e., representative normal sample sequence data). The cohorts represent individuals that are not matched to the samples of interest, i.e., are different individuals.

[0063] In one embodiment, the representative unmatched normal sample sequence data, once generated, is fixed for a particular sample preparation technique. That is, the representative unmatched normal sample sequence data is associated with the type of samples from which the data was generated. Different representative unmatched normal sample sequence data sets may be generate for FFPE samples, cell lines, fresh frozen, etc. Further, the representative unmatched normal sample sequence data sets may be stored by the provider and sent to the user as part of an analysis package at step 116. The analysis package may also be capable of receiving updates from a remote server if the microsatellite instability analysis is refined by the provider.

[0064] In one embodiment, the normal cohort sequence data may include sequence data from a plurality of individuals. Each individual sequence may be assessed according to certain quality metrics (e.g. depth of sequencing) at each individual microsatellite region. For example, the sequence data of each individual sequence may only be used when there are at least a predetermined number (e.g., 20) sequencing reads at the individual microsatellite region. Accordingly, each individual sequence may pass at a subset of the microsatellite regions and fail for others, depending on the available sequencing depth. The passing regions are used for further analysis, while the failing regions are masked or not used. After quality assessment, the individual sequences of the cohort may be pooled to generate a distribution at each microsatellite region. The distribution of the pooled normal cohort serves as a reference sample dataset that represents a hypothetical matched normal sample.

[0065] The analysis maybe used to generate a microsatellite instability score at step 120. The microsatellite instability score may be based on a comparison of a distance (i.e., a statistical distance, a Jensen-Shannon distance) between a distribution at each

microsatellite region between the sample of interest and the reference sample dataset. In one embodiment, a microsatellite instability score is based on a number of microsatellite regions having a distance above a threshold, where a larger distance is indicative of greater divergence from the reference sample dataset and being associated with a positive score, relative to a total number of microsatellite regions. Samples having a percentage greater than a predetermined number (e.g., 5%) having a positive score may be classified as having microsatellite instability while samples having a percentage lower than the predetermined number may be classified as microsatellite stable. Further, microsatellite instability may be designated as high or low based on the percentage.

[0066] The microsatellite instability assessment may be provided to a clinician as an input for determining a treatment protocol. In recent years, immune checkpoint inhibitors have shown great promise in treating various cancer types; however, only a fraction of patients respond to this type of immunotherapy. PD-L1 protein expression measured by quantitative immunohistochemistry (IHC) is an FDA approved companion diagnostic or complementary assay for some immune checkpoint inhibitors. Pembrolizumab (KEYTRUDA, Merck & Co.) may be provided to patients with solid tumors that have microsatellite instability high (MSI-H) or mismatch repair deficient (dMMR).

[0067] FIG. 4 shows an example workflow for a sample of interest for an example tumor sample 150. If a matched normal sample 154 is available, the workflow proceeds to sequence analysis 156 of the tumor sample sequence data 158 and sequence analysis 160 of the normal sample sequence data 162. The sequence data may be in the form of BAM files, base call data, image data, etc. The sequence analysis may be via a microsatellite instability analysis technique in which the matched normal sample data is used as a basis for comparison (block 164). If no matched normal sample is available, the workflow proceeds to analysis via the microsatellite instability analysis technique as provided herein using the reference sample dataset from the appropriate unmatched normal cohort (block 166). Both techniques yield a microsatellite instability score, either a tumor only microsatellite instability score 168 or a tumor/normal microsatellite instability score 170.

Further, for matched samples, the sample of interest may nonetheless be fed into the unmatched analysis and the results compared to the matched analysis for quality purposes.

[0068] In a specific embodiment, a sequencing panel was provided covering 170 genes associated with solid tumors. Designed to capture mutational changes, including single nucleotide variants, indels, amplifications, splice variants and fusions, the panel was designed to target both DNA and RNA variants from the same FFPE tumor sample in a single sequencing run. The performance of the panel to assess 103 microsatellite loci was evaluated using 53 colon cancer samples (28 MSI-H and 25 MSS as determined by MSI-PCR) and showed that the panel achieved 100% concordance for microsatellite instability status with matched tumor/normal pairs. Additionally the microsatellite instability analysis may be used for unmatched tumor sample only achieving 98% concordance with MSI-PCR. Furthermore, MSI-H samples had significantly higher tumor mutational burden compared to MSS samples in this cohort of colon samples. In summary, a microsatellite targeted panel may accurately determine microsatellite instability status from FFPE tumor samples. FIGS. 5, 6, and 7A-E show results from the experiments.

[0069] For each microsatellite locus, flanking sequences of microsatellite repeats were anchored to determine the number of repeat units supported by reads mapped to the region. Subsequently the distribution of the repeat unit lengths determines the microsatellite instability status of each site. The final microsatellite instability score was calculated with number of unstable microsatellite sites divided by number of total sites evaluated. FIG. 5A shows reads mapped to the microsatellite regions that are extracted from a binary alignment map file. FIG. 5B in the upper panel shows for mapped reads of a microsatellite instability high (MSI-H) sample and the lower panel shows repeat unit length distributions for both tumor and normal samples in a MSI-H sample. FIG. 5C in the upper panel shows mapped reads of a microsatellite stable (MSS) sample and the lower panel shows repeat unit length distributions for both tumor and normal samples in the MSS sample. FIG. 6 shows single microsatellite region predictive value for each of

the 103 sites of the sequencing panel. Single sites were less accurate relative to the full panel. As provided herein, the number of microsatellite sites or microsatellite regions used in the present techniques to generate a microsatellite instability score may be 1 or more, 5 or more, 50 or more, or 100 or more. In certain embodiments, the present techniques may analyze 1-20, 5-20, 5-50, 10-20, 10-50, or 50-100 microsatellite sites in the analysis to generate a microsatellite instability score.

[0070] As provided, the Jensen-Shannon Distance between the sample of interest to the reference sample dataset was determined. The reference Jensen-Shannon Distance, $d1$, was calculated for all pairwise combinations of reference samples (BL_n , $n = 1..N$) as follows:

$$BL_n1 = \Pr [X=x]$$

$$BL_n2 = \Pr [X=x]$$

$$JS1 = 0.5 * (\text{sum}(BL_n1 * \log(BL_n1/m1)) + \text{sum}(BL_n1 * \log(BL_n1/m1)))$$

$$m1 = 0.5 * (BL_n1 + BL_n2)$$

$$d1 = \text{sqrt}(JS1)$$

[0071] The test Jensen-Shannon Distance, $d2$, was calculated between the sample of interest (T) and each sample of the reference dataset as follows:

$$BL_n = \Pr [X=x]$$

$$T = \Pr [X=x]$$

$$JS2 = 0.5 * (\text{sum}(BL_n * \log(BL_n/m2)) + \text{sum}(T * \log(T/m2)))$$

$$m2 = 0.5 * (BL1 + T)$$

$$d2 = \text{sqrt}(JS2)$$

[0072] Comparison between two Jensen-Shannon Distance distributions is performed via a one-sided t-test to establish $d1 < d2$, with $FDR < 0.05$ and $d2 - d1 > 0.1$.

[0073] FIG. 7A is boxplot of a microsatellite instability score based on tumor/normal pairs. FIG. 7B is an ROC curve for tumor/normal pairs. FIG. 7C is boxplot of a microsatellite instability score based on tumor only samples. FIG. 7D is an ROC curve for tumor only samples. FIG. 7E is a boxplot showing MSI-H samples with higher nonsynonymous tumor mutational burden (TMB) compared to MSS samples.

[0074] FIG. 8A is boxplot of a microsatellite instability score for 232 tumor/normal samples from a variety of tissue types and using a 58 sample normal colorectal cancer reference sample dataset matched to some of the tumor samples. The 58 sample normal colorectal cancer reference sample dataset was generated from the matched normal samples for the MSI-H and MSS CRC samples below. The samples included matched tumor normal pairs: $n=140$, with 92 pairs with MSI-PCR results:

MSI-H tumor: $n=35$ (32 CRC and 3 UCEC)

MSS tumor: $n=57$ (26 CRC and 31 UCEC)

Total test sample ($92 \text{ tumor} + 140 \text{ normal} = 232 \text{ samples}$)

[0075] The samples were characterized as MSI-H ($n=35$), MSS ($n=54$) tumor, or normal ($n=140$) based on MSI-PCR. The microsatellite instability score was determined as provided herein. While the results generally aligned with the MSI-PCR results, the red circled portion indicates false positives for MSI-H based on the MSI cutoff score.

[0076] FIG. 8B is boxplot of a microsatellite instability score for the 58 colorectal cancer matched tumor/normal samples using the 58 sample matched normal colorectal cancer reference sample dataset of FIG. 8A and showing a tighter grouping of the stable-identified MSI scores.

[0077] FIG. 9 is boxplot of a microsatellite instability score for 140 normal samples whereby samples from individuals are separated into their associated ethnic groups (African, South American, East Asian, and European) using the 58 sample normal colorectal cancer reference sample dataset of FIG. 8A. As shown, the microsatellite instability scores vary between ethnic groups, indicating the potential for ethnic bias to be present into the reference sample dataset.

[0078] FIG. 10 is a flow diagram of a method 200 of removing bias based on ethnic variability from a reference sample dataset. At step 202, reference sample sequence data may be acquired from reference samples of a cohort of a plurality of individuals, e.g., using a sequencing device 60 as provided herein. The plurality of individuals may be associated with a particular ethnic background (e.g., African, South American, East Asian, and European, in a non-limiting example). The association may be based on self-reporting, in one example. In other embodiments, previously acquired reference sample sequence data may be accessed. At step 204, the reference sequence data is analyzed to generate a distribution at each of a plurality of microsatellite regions of interest. In an initial quality control step, data for an individual microsatellite region from an individual reference sample having insufficient read coverage (e.g., fewer than 20 reads of a particular microsatellite region) may be filtered out of the reference sample sequence data (e.g., deleted, masked, or otherwise indicated as being not for consideration in further analysis steps).

[0079] At step 206, the ethnic group variability of the distribution (e.g., the allele distribution) at each of the plurality of microsatellite regions for the plurality of reference biological samples is determined. For example, for each of the individual reference sample sequence (e.g., 10 sequences, 20 sequences, or more) for which ethnic

background information is available, the sequence data is grouped into one of the ethnic groups represented in the cohort for analysis. It should be understood that the cohort may be selected to provide an advantageous mix of samples from a desired number of ethnic groups that may be generally evenly distributed or may unevenly distributed in the cohort. The ethnic variability may include a variability between a first distribution of the sequence data of the group of samples associated with a first ethnic group relative to a second distribution of the sequence data of the group of samples associated with a second ethnic group. The distribution may be a region-by-region distribution, such that the variability between two or more ethnic groups at each individual microsatellite region for which sequence data is available (e.g., after any quality assessments) is assessed. In one embodiment, after the first quality assessment of sufficient coverage, certain individual microsatellite regions may fail to qualify for an assessment of ethnic variability based on a low number of qualifying samples. That is, the method may include a cutoff of samples (e.g., 5 or more individual samples of sufficient quality or having sufficient read coverage) per ethnic group to qualify for variability assessment.

[0080] At step 208, ethnically biased and/or unbiased microsatellite regions of the plurality of microsatellite regions are identified based on the ethnic group variability at each of the plurality of microsatellite regions for a group of reference samples separated into each individual ethnic group. In one embodiment, a measure of the variability of a microsatellite region sequence data distribution for a particular microsatellite region within a particular ethnic group is determined. This variability is then compared with variability of another ethnic group. This variability is then compared with variability of another ethnic group. Microsatellite regions with relatively large differences in variability between ethnic groups may be indicative of inherent ethnic variability for the region. The variability may be assessed by any suitable method, e.g., range, mean, variance and/or standard deviation or by Jensen-Shannon distance as provided herein. After this analysis is performed for each microsatellite region, the regions having a variability metric above a threshold may be identified as having high ethnic bias or variability while the regions having the variability metric below the threshold may be

identified as having low or acceptable ethnic bias or variability. At step 210, a reference sample dataset is generated from the reference sequence data by removing/filtering the ethnically biased microsatellite regions from the reference sequence data of the plurality of reference biological samples. As a result of the method 200, the reference sample dataset is generated in which, for example, a portion of the sequenced microsatellite regions are excluded from further analysis because of identification of ethnic bias. The remaining microsatellite regions are less susceptible to bias related to ethnic background. Accordingly, the reference sample dataset has been made more robust and independent of bias to more accurately serve as a hypothetical matched normal to a wide range of samples of interest.

[0081] FIG. 11A shows the distribution of ethnicities in 140 samples used to assess ethnic variability in the reference sample dataset. FIG. 11B shows results from an example technique of identifying microsatellite regions with relatively high ethnic variability in a reference sample dataset using calculated delta Jensen Shannon distances. The delta Jensen Shannon distances were determined as follows:

$$\Delta\text{JSD} = \text{avg}(\text{JSD}_{\text{between}}) - \text{avg}(\text{JSD}_{\text{within}})$$

The delta Jensen Shannon distance for each microsatellite region is a measure of the average Jensen Shannon distance between two groups and the average Jensen Shannon distance within a group. A pair-wise comparison between three or more groups may be performed. The technique assessed 175 sites (microsatellite regions) with at least 20 supporting reads for minimum 5 samples of each ethnicity group. Based on the analysis, 44 sites with ≥ 0.1 ΔJSD based on pair-wised comparison between three ethnicity groups were identified as having high ethnic variability. These sites were filtered out of (e.g., eliminated or masked from) the sequence data used to generate the reference sample dataset.

[0082] FIG. 12 is boxplot of a microsatellite instability score for 140 normal samples including samples from individuals associated with one of four different ethnic groups

(African, South American, East Asian, and European) using the 58 sample normal colorectal cancer reference sample dataset with the identified microsatellite regions with relatively high ethnic variability filtered out of the reference sample dataset prior to the analysis. Compared to FIG. 9, which shows the same analysis, but without the filtering out of the identified microsatellite regions, the MSI scores are more compressed for certain ethnic groups, indicating the effect of the filtering out of the relatively high ethnic variability regions.

[0083] FIG. 13 is boxplot of a microsatellite instability score of 232 tumor/normal samples from a variety of tissue types using the 58 sample normal colorectal cancer reference sample dataset post filtering with the red circle denoting potential false positives. The MSI-H samples have generally intact MSI scores post filtering while normal/MSS samples have lower MSI scores. FIG. 14 is boxplot of a microsatellite instability score for normal samples associated with one of four different ethnic groups (African, South American, East Asian, and European) using 58 unmatched cell lines samples as the reference sample dataset pre and post filtering of the identified microsatellite regions with relatively high ethnic variability. The 58 unmatched cell lines included 10 IHW and 48 coriell lines with the following ethnic group distributions:

P_AFR (n=22) (African)

P_AMR (n=25) (South American)

P_EUR (n=8) (European)

P_EAS (n=3) (East Asian)

[0084] FIG. 15 is a comparison of the ethnic diversity of the unmatched cell lines samples reference dataset with the normal colorectal cancer reference sample dataset. The cell line reference sample does not share a genotype with FFPE samples, which were matched to a portion of the CRC tumor samples. The cell line reference sample dataset represents a true unmatched sample. Further, the cell line reference sample dataset

represents different ethnicity composition from the original baseline or FFPE samples in general.

[0085] FIG. 16A is boxplot of a microsatellite instability score of 232 tumor/normal samples from a variety of tissue types using the unmatched cell lines samples as the reference sample dataset post filtering of the identified microsatellite regions with relatively high ethnic variability. FIG. 16B shows the sensitivity and specificity of the results of FIG. 16A.

[0086] FIG. 17 is a comparison of an original and repeat run of 78 colorectal cancer samples.

[0087] FIG. 18-19 show MSI score results correlation for reference sample datasets with varying numbers of samples. Performance was tested with random 10, 20, 30, 40, 50 baseline samples (randomly selected from 71 samples (58 cell lines + 13 FFPE normal) that is not overlapped with the current 232 testing set).

[0088] FIG. 20 shows MSI scores for different titration levels of cell lines, and FIG. 21 is boxplot of a microsatellite instability score of 46 cell line samples including four MSI-H cell lines.

[0089] The results showed that blocking 44 ethnicity specific/biased sites improved performance. Further, reference sample datasets do not need to be sample type/ethnicity constrained. The performance was robust with $n \geq 30$ samples in the reference sample dataset.

[0090] A subset of available microsatellite sites may be selected of the available sites to develop a more sensitive limit of detection. Stringency in MSI site selection may improve limits of detection. FIG. 22 shows a cell line titration and associated limit of detection of MSI scores for various titrated levels of Lovo cells titrated into a cell line with microsatellite stability. FIG. 23 shows a cell line titration and associated limit of detection of MSI scores for various titrated levels of SW48 cells titrated into a cell line

with microsatellite stability. The model of titration of cell lines in stable cells models the presence of tumor cells mixed with other cell types in a tumor sample. The depicted correlation between titration and loss of detection of MSI is based on an analysis of distribution of 100+ microsatellite sites and is limited by sequencing depth for the sample. While the depicted 2.5% or 5% titration at the limit of detection may be acceptable for solid samples, the detection of microsatellite instability in plasma (e.g., plasma DNA debris) may involve lower detection limits for accuracy.

[0091] However, in certain embodiments, a quality analysis of the available microsatellite sites may facilitate selection of a subset of sites with higher quality or lower variability to achieve improved limits of detection. With potential different DNA extraction methods, baseline, sequencing depth, microsatellite site quality may vary based on sample type (e.g., solid or FPE vs. liquid). FIG. 24 shows improved limits of detection for Lovo cells titrated into a cell line with microsatellite stability. FIG. 25 shows improved limits of detection for SW48 cells titrated into a cell line with microsatellite stability. FIGS. 24-25 represent analysis using only a subset of higher quality microsatellite sites, in the depicted example 16 out of 130 available sites. The sites of the reference dataset were selected as the lowest delta distribution at each individual site as assessed in the reference sequence data. Accordingly, in certain embodiments, the limit of detection is improved 5-fold by limiting the MSI sites to a subset of higher quality sites, e.g., using a ranking of the delta distribution and selecting those sites having a lowest distribution.

[0092] Further, the selection of the subset of MSI sites may be based on one or more user inputs and/or sample type. For samples, such as solid samples, a limit of detection achieved using all or most available MSI sites may be sufficient. Accordingly, indication of a solid sample may initiate analysis as provided herein using a larger subset of the available MSI sites than for a user input that the sample is a plasma or liquid sample. Further, the user may be able to input stringency or limit of detection settings in the sequencing or analysis device.

[0093] Technical effects of the disclosed embodiments include improved and more accurate microsatellite instability assessment for unmatched samples of interest. Additional technical effects include improved generation of reference sample datasets to serve as hypothetical matched normal samples for analysis of tumor only samples. The improved reference sample datasets used to determine the microsatellite instability score of an unmatched sample of interest are stripped of (e.g., via masking or eliminating) sites with high inter-ethnic group variability. In this manner, the reference sample dataset may be used as a hypothetical matched sample independent of the ethnic background of the individual associated with the unmatched sample of interest, providing a more robust technique for microsatellite instability assessment.

[0094] While only certain features of the disclosure have been illustrated and described herein, many modifications and changes will occur to those skilled in the art. It is, therefore, to be understood that the appended claims are intended to cover all such modifications and changes as fall within the true spirit of the disclosure.