

- 1 -

Mitochondrial DNA Quality Control

Field

The present disclosure is directed, in part, to methods of identifying unreliable
5 biological samples that may be mislabeled or contaminated.

Background

For over 10 years, next-generation sequencing (NGS) has become an important
component in biological and biomedical researches because it makes sequencing of large
10 batches of DNA or RNA samples feasible. NGS has broad applications, such as whole genome
and whole exome sequencing for large cohort genetic investigation, bulk RNA-seq for disease
gene expression signatures identification in clinical evaluation, tissue biopsy sequencing in
tumor study/diagnostic and recently emerged single-cell sequencing research, providing
answers and solutions to many different problems and questions. However, in the studies
15 involving large-scale samples, sample identity complication is a common and almost inevitable
problem. The estimate sample identity error rate can range from 0.2% to 6% in practice (Pfeifer
et al., Amer. J. Clin. Pathol., 2013, 139, 93-100; Costello et al., BMC Genomics, 2018, 19, 332;
Lerner et al., Cancer Res., 2015, 75, Abstract P5-02-08; and Sehn et al., Amer. J. Clin. Pathol.,
2015, 144, 667-674). The errors can happen at different degrees: 1) a complete swapping
20 between samples, and/or 2) contamination of one sample with one or more other samples.
Various steps during the sample processing can introduce the errors, such as sample
mislabeling during sample collection, material spillage during pipetting, index swapping in
pooled libraries when performing sequencing and many other unexpected situations. Sample
swapping/contamination will subsequently reduce the quality and accuracy of the downstream
25 analysis. For example, sample swapping in a whole transcriptome analysis may lead to false
discovery or lose power to detect differentially expressed genes. In cancer studies, somatic
mutation identification is routinely used, given that many of those mutations were present at
very low frequency (<5%), thus even low levels (1% to 5%) of contamination may result in false
positive mutation callings. For those reasons, accurate detection of sample swapping and
30 contamination is an important quality control step in large scale NGS studies.

Mitochondria are essential organelles in most eukaryotic cells. Human mitochondrial
DNA (mtDNA) are 16.5 kb circular DNA molecules located in mitochondria, and encode gene

- 2 -

products that are essential for mitochondrial functions. There are hundreds to thousands of mtDNA copies within a single cell. mtDNA is maternally inherited with negligible recombination. Because mtDNA is uniparentally inherited and undergoes negligible recombination at a population level, mutations acquired over time have subdivided the human population into several discrete mtDNA haplogroups. On average, two random individuals will have 30 to 40 nucleotide differences in their mitochondrial genome (Gunnarsdóttir et al., Nature Commun., 2011, 2, 228; Slatkin et al., Genetics, 1991, 129, 555-562; and Ye et al., Proc. Nat'l Acad. Sci. USA, 2014, 111, E4548-E4550). Because of its multi-copy nature, mtDNA mutations often only present in a small proportion of the cell's mtDNA, a state termed as heteroplasmy. The percentage of mtDNA carrying the mutation is referred as heteroplasmy frequency. In contrast, if a mutation is found in all mtDNA molecules, this mutation will be termed as homoplasmy. Previous studies demonstrated that in general healthy populations, most individuals will harbor less than 5 heteroplasmies (with frequency > 1 to 2%) in their mitochondrial genome (Zhang et al., BMC Genomics, 2017, 18, 890; and Ye et al., Proc. Nat'l Acad. Sci. USA, 2014, 111, 10654-10659). For a batch of samples, samples collected from the same individual should all belong to the same haplogroup.

Summary

The present disclosure provides methods of identifying an unreliable biological sample, the method comprising: a) performing a nucleic acid sequencing assay on each biological sample of a plurality of biological samples obtained from a single individual to obtain mitochondrial DNA (mtDNA) sequencing reads for each biological sample; b) identifying a heteroplasmy and a homoplasmy in the mtDNA sequencing reads from the previous step for each of the biological samples; and c) assigning a primary mtDNA haplogroup to each biological sample, wherein any biological sample having an assigned primary mtDNA haplogroup that is different than the primary mtDNA haplogroup assigned to a majority of the biological samples from the same individual is an unreliable biological sample that is a mislabeled biological sample.

The present disclosure also provides methods of identifying an unreliable biological sample, the method comprising: a) performing a nucleic acid sequencing assay on each biological sample of a plurality of biological samples obtained from a single individual to obtain mitochondrial DNA (mtDNA) sequencing reads for each biological sample; b) identifying a

- 3 -

heteroplasmy and a homoplasmy in the mtDNA sequencing reads from the previous step for each of the biological samples; c) assigning a primary mtDNA haplogroup to each biological sample; and d) determining the total heteroplasmy number for each biological sample, wherein when a biological sample has a high heteroplasmy number, the biological sample is assigned a
5 secondary mtDNA haplogroup based on the minor alleles in the heteroplasmy sites, wherein a biological sample having an assigned secondary mtDNA haplogroup that is different than the assigned primary mtDNA haplogroup is an unreliable sample that is contaminated.

The present disclosure also provides methods of identifying an unreliable biological sample, the method comprising: a) performing a nucleic acid sequencing assay on each
10 biological sample of a plurality of biological samples obtained from a single individual to obtain mitochondrial DNA (mtDNA) raw sequencing reads for each biological sample; b) processing the mtDNA raw sequencing reads for quality control and adaptor sequence removal to produce quality controlled mtDNA sequencing reads; c) mapping the quality controlled mtDNA sequencing reads to a mitochondrial reference genome to produce candidate mtDNA
15 sequencing reads; d) re-mapping the candidate mtDNA sequencing reads to a human reference genome and retaining the candidate mtDNA sequencing reads when: i) the candidate mtDNA sequencing read is uniquely mapped to the mitochondrial reference genome or has fewer mismatches to the mitochondrial reference genome than to the human reference genome; and ii) the alignment mismatch count of the candidate mtDNA sequencing read is less than 5; e)
20 performing post-mapping processing of the retained candidate mtDNA sequencing reads for sorting and duplicate removal; f) identifying heteroplasmy and homoplasmy in the retained candidate mtDNA sequencing reads for each of the biological samples; and g) assigning a primary mtDNA haplogroup to each biological sample, wherein any biological sample having an assigned primary mtDNA haplogroup that is different than the primary mtDNA haplogroup
25 assigned to a majority of the biological samples from the same individual is an unreliable biological sample that is a mislabeled biological sample.

The present disclosure also provides methods of identifying an unreliable biological sample, the method comprising: a) performing a nucleic acid sequencing assay on each
biological sample of a plurality of biological samples obtained from a single individual to obtain
30 mitochondrial DNA (mtDNA) raw sequencing reads for each biological sample; b) processing the mtDNA raw sequencing reads for quality control and adaptor sequence removal to produce quality controlled mtDNA sequencing reads; c) mapping the quality controlled mtDNA

- 4 -

sequencing reads to a mitochondrial reference genome to produce candidate mtDNA sequencing reads; d) re-mapping the candidate mtDNA sequencing reads to a human reference genome and retaining the candidate mtDNA sequencing reads when: i) the candidate mtDNA sequencing read is uniquely mapped to the mitochondrial reference genome or has fewer
5 mismatches to the mitochondrial reference genome than to the human reference genome; and ii) the alignment mismatch count of the candidate mtDNA sequencing read is less than 5; e) performing post-mapping processing of the retained candidate mtDNA sequencing reads for sorting and duplicate removal; f) identifying heteroplasmy and homoplasmy in the retained candidate mtDNA sequencing reads for each of the biological samples; g) assigning a primary
10 mtDNA haplogroup to each biological sample; and h) determining the total heteroplasmy number for each biological sample, wherein when a biological sample has a high heteroplasmy number, the biological sample is assigned a secondary mtDNA haplogroup, wherein a biological sample having an assigned secondary mtDNA haplogroup that is different than the assigned primary mtDNA haplogroup is an unreliable sample that is contaminated.

15

Brief Description Of The Drawings

The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

20 Figure 1 shows a representative schematic showing suitable steps for carrying out the quality control analysis described herein. In Box 1, mtDNA homoplasmies and heteroplasmies are identified from fastq files. An optional down sampling step can be applied for sample with high mtDNA coverage. After reads QC, mtDNA reads are selected by a two-step mapping strategy. mtDNA variants were identified from the mtDNA mapping results and primary and
25 secondary mtDNA haplogroup are assigned to each sample based on the variants information. In Box 2, sample swapping/mislabeled can be detected by comparing haplogroup assignments of samples from the given individual. In Box 3, sample contamination can be detected by unusual high mtDNA heteroplasmies number and unmatched primary and secondary haplogroup groups.

30 Figure 2 shows the performance of the methods described herein on virtual contamination samples. Virtual contamination samples were created by mixing two samples from 1000 Genomes Project at different ratios. The X-axis indicates the theoretical

- 5 -

contamination level and the Y-axis indicates the heteroplasmy frequencies identified from each virtual contamination sample. Each colored dot represents one heteroplasmy, the black dots represent the means of the heteroplasmy frequencies in the samples and error bars represent the frequency standard errors. The mean of the frequencies is significantly correlated with
5 theoretical contamination level (Pearson correlation = 0.996781, P value = 6.212e-09).

Figure 3 shows results of contamination detection in virtual contamination samples.

Figure 4 shows results from sample swapping and contamination detection in RNA-seq data 1.

Figure 5 shows results from sample swapping and contamination detection in RNA-seq
10 data 2.

Description Of Embodiments

Methods are presented herein for leveraging mtDNA sequences information to detect potential sample mislabeling and contaminations in NGS data. mtDNA polymorphisms and
15 mutations can be used to infer the identity of a particular biological sample, and act as an indicator of sample mislabeling. In addition, when a biological sample is contaminated by DNA/RNA from another biological sample, unusual mtDNA mutation patterns will be revealed, which can help identify and further quantify the contaminant. Compared to nuclear DNA mutation-based approaches, the methods described herein allows for higher sensitivity even in
20 low coverage sequencing data.

The methods described herein can take any NGS data containing sufficient mtDNA reads as input, identify mtDNA variants (heteroplasmasy and homoplasmasy) from the data, and use the variants information to assign haplogroups to each sample to detect potential sample swapping or mislabeling. By evaluating the samples' heteroplasmy information, the methods
25 described herein can further detect cross-individual contamination.

The terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting.

A human will have wild type (or reference) mtDNA molecules and may have mutant mtDNA molecules. If a human has no mutant mtDNA molecules, such a human is considered to
30 be homoplasmic wild type (or homoplasmic reference). If a human has no wild type mtDNA molecules (i.e., only has mutant mtDNA), such a human is considered to be homoplasmic mutant. Homoplasmy is, thus, a measure of possessing all or no copies of mutant mtDNA.

- 6 -

If a human possesses a mixture of wild type and mutant mtDNA molecules, the human is considered to possess a heteroplasmy. The fraction of mutated copies is referred to herein as the "heteroplasmy frequency." For example, assuming a human has 8 copies of mtDNA molecules, and possesses a single copy of the eight mtDNA molecules that has a particular mutation in gene A, such a human is considered to have a heteroplasmy frequency of 12.5% (i.e., 1/8). Heteroplasmy can be determined for each mutation within the mtDNA genome for a particular individual. Thus, an individual having 2 mtDNA mutations (relative to wild type mtDNA) can have two heteroplasmies. Each heteroplasmy is associated with its own heteroplasmy frequency.

10 The present disclosure provides methods of identifying an unreliable biological sample. The methods comprise performing a nucleic acid sequencing assay on each biological sample of a plurality of biological samples obtained from a single individual to obtain mitochondrial DNA (mtDNA) sequencing reads for each biological sample. The methods also comprise identifying the presence of one or more heteroplasmies and homoplasmies in the mtDNA sequencing
15 reads for each of the biological samples. The methods also comprise assigning a primary mtDNA haplogroup to each biological sample. A biological sample that has an assigned primary mtDNA haplogroup that is different than the primary mtDNA haplogroup assigned to a majority of the biological samples from the same individual is an unreliable biological sample. Such an unreliable biological sample may have been, for example, mislabeled or swapped with another
20 biological sample.

 The nucleic acid sequencing assay is any nucleic acid sequencing protocol. In some embodiments, the sequencing assay comprises next generation sequencing (NGS). In some embodiments, the NGS comprises whole genome sequencing. In some embodiments, the NGS comprises whole exome sequencing. In some embodiments, the NGS comprises RNA
25 sequencing. In some embodiments, the NGS comprises bisulfite sequencing.

 The nucleic acid sequencing assay is performed on each biological sample of a plurality of biological samples obtained from a single individual. In some embodiments, the plurality of biological samples can number from as low as 2 to thousands of samples. In some
30 embodiments, the plurality of biological samples can number from as low as 2 to hundreds of samples. In some embodiments, the plurality of biological samples is obtained from one or more clinical studies. In some embodiments, the single individual's plurality of biological

- 7 -

samples may be intermixed or batched with a plurality of biological samples from another individual. mtDNA sequencing reads for each biological sample are obtained.

The presence of one or more heteroplasmies and homoplasmies in the mtDNA sequencing reads for each of the biological samples is determined. Thus, for each mutation identified in the mtDNA sequencing reads, the heteroplasmy and homoplasmy analysis is performed. The sum of all the heteroplasmies is represented by the total heteroplasmy number for a particular biological sample. The mtDNA sequencing read information for each mtDNA mutational site is compiled to provide a summary of the sequencing information of mapped reads at each single site. In some embodiments, the compiling can be carried out using, for example, the samtools mpileup function (Li et al., Bioinformatics, 2009, 25, 2078-2079). The mtDNA sequencing read information for each mtDNA mutational site is filtered by sequence quality to, for example, remove sequencing bases with low sequencing quality to reduce sequencing errors. In some embodiments, a sequence quality score (Q) is determined, which is a property that is logarithmically related to the sequencing error probability ($Q = -10 \cdot \log_{10}(P)$, where P is the probability of a sequencing error). In some embodiments, the sequence quality Q is ≥ 20 . When Q is 20, there is a 1% probability of a sequencing error.

In some embodiments, the heteroplasmy is identified by determining the sequencing coverage, the presence of a minor allele, and minor allele frequency. The sequencing coverage represents the number of reads that align to a known mtDNA reference base. In some embodiments, the sequencing coverage is ≥ 50 . The sequencing coverage is generated by mpileup minus the bases with $Q < 20$. In some embodiments, the minor allele frequency is $\geq 1\%$ for DNA sequencing data and $\geq 5\%$ for RNA sequencing data. In some embodiments, the minor allele is observed at least twice from each DNA strand, or the minor allele is observed at least three times for RNA. For example, the following mtDNA sequencing reads may be obtained (the first sequence is the reference sequence):

```

N1-N2-N3-N4-N5-N6-N7-N8-G9-N10-N11-N12-N13-N14-N15-N16-N17-N18-N19-N20
5'-N1-N2-N3-N4-N5-N6-N7-N8-A9-N10-N11-N12-N13-N14-N15-3'
3'-N2-N3-N4-N5-N6-N7-N8-A9-N10-N11-N12-N13-N14-N15-N16-5'
5'-N2-N3-N4-N5-N6-N7-N8-G9-N10-N11-N12-N13-N14-N15-N16-3'
30 5'-N3-N4-N5-N6-N7-N8-G9-N10-N11-N12-N13-N14-N15-N16-N17-3'
5'-N3-N4-N5-N6-N7-N8-G9-N10-N11-N12-N13-N14-N15-N16-N17-3'
3'-N3-N4-N5-N6-N7-N8-G9-N10-N11-N12-N13-N14-N15-N16-N17-5'

```

- 8 -

3'-N4-N5-N6-N7-N8-**G9**-N10-N11-N12-N13-N14-N15-N16-N17-N18-5'5'-N7-N8-**G9**-N10-N11-N12-N13-N14-N15-N16-N17-N18-N19-N20-3'

The heteroplasmy frequency for this candidate mtDNA heteroplasmy site is 25% (2/8). In this particular analysis, the sequencing quality is > 20 and the sequencing coverage is > 50. The
 5 minor allele is observed in both strands of the DNA. Accordingly, this particular mutational site (i.e., a candidate mtDNA heteroplasmy site) is a mtDNA heteroplasmy.

In some embodiments, the homoplasmy is identified by determining the sequencing coverage and the presence of one or more alleles. A homoplasmy is present when: i) the sequencing coverage is ≥ 10 ; and ii) only one allele is observed at a particular nucleic acid
 10 mutation site and it is different than the corresponding reference allele, or multiple alleles are observed at a particular nucleic acid mutation site and the major allele is different than the corresponding reference allele, and the particular nucleic acid site is not a heteroplasmy and does not meet the heteroplasmy identification criteria.

In some embodiments, assigning the primary mtDNA haplogroup to each biological
 15 sample comprises constructing a mtDNA sequence for each biological sample. In some embodiments, the mtDNA sequence for each biological sample is constructed using the homoplasmy and major alleles of the heteroplasmy. In some embodiments, the primary mtDNA haplogroups are assigned based on the constructed mtDNA sequences using HaploGrep2 (Weissensteiner et al., Nuc. Acids Res., 2016, 44, W58-W63). HaploGrep2 is an algorithm
 20 whereby the haplogroup is classified based on precalculated phylogenetic weights that correspond to the mutation occurrence per position in Phylotree. Similar tools for assigning primary mtDNA haplogroups include mthap (world wide web at "dna.jameslick.com/mthap/") and haplofind (world wide web at "haplofind.unibo.it/").

A biological sample that has an assigned primary mtDNA haplogroup that is different
 25 than the primary mtDNA haplogroup assigned to a majority of the biological samples from the same individual is an unreliable biological sample. In some embodiments, the unreliable biological sample has been mislabeled. In some embodiments, the unreliable biological sample has been swapped with another biological sample. In some embodiments, the one or more mislabeled samples are re-labeled correctly. In some embodiments, the one or more
 30 mislabeled samples are discarded.

In some embodiments, the methods further comprise determining a heteroplasmy number for each biological sample. In some embodiments, the heteroplasmy frequency is

- 9 -

determined for each mutation identified in the mtDNA sequences for each biological sample. When a biological sample has a high heteroplasmy number, the biological sample is assigned a secondary mtDNA haplogroup. In some embodiments, a threshold for having a high heteroplasmy number is ≥ 10 heteroplasmies.

5 In some embodiments, assigning the secondary mtDNA haplogroup comprises constructing a secondary mtDNA sequence using the homoplasmy and minor alleles of the heteroplasmy. A biological sample having an assigned secondary mtDNA haplogroup that is different than the assigned primary mtDNA haplogroup is an unreliable sample that is contaminated. The choices for the primary haplogroup are identical to the choices for the
10 secondary haplogroup.

 In some embodiments, the methods further comprise determining the level of contamination of a biological sample. In some embodiments the level of contamination is indicated by determining the median of the heteroplasmy frequencies of all heteroplasmies in the contaminated sample. The greater the median heteroplasmy frequency, the greater the
15 level of contamination. There is strong correlation between the real contamination percent and heteroplasmy frequency median/mean. For example, if the median heteroplasmy frequency is 6%, the contamination level is also about 6%.

 In some embodiments, the methods further comprise processing the mtDNA sequencing reads obtained from the nucleic acid sequencing assay for quality control and
20 adaptor sequence removal prior to identifying a heteroplasmy and a homoplasmy. In such embodiments, the mtDNA sequencing reads obtained from the nucleic acid sequencing assay are mtDNA raw sequencing reads. In carrying out the processing of the mtDNA raw sequencing reads, quality controlled mtDNA sequencing reads are produced. In some embodiments, the processing of the mtDNA sequencing reads obtained from the nucleic acid sequencing assay for
25 quality control and adaptor sequence removal can be carried out by using "Trimmomatic" (Bolger et al., Bioinformatics, 2014, 30, 2114-2120). This processing step improves the accuracy of the subsequent mtDNA variant identification. Another tool that can be used for processing is cutadapt (world wide web at "cutadapt.readthedocs.io/en/stable/").

 In some embodiments, the method further comprises a two-step mapping process
30 prior to identifying a heteroplasmy and a homoplasmy. In some embodiments, the mtDNA sequencing reads obtained from the nucleic acid sequencing assay can be used in the two-step mapping process. In some embodiments, the quality controlled mtDNA sequencing reads

- 10 -

obtained from the quality control and adaptor sequence removal process can be used in the two-step mapping process. In these embodiments, the mtDNA sequencing reads (obtained from the nucleic acid sequencing assay) or the quality controlled mtDNA sequencing reads (obtained from the quality control and adaptor sequence removal process) are mapped to a
5 mitochondrial reference genome to produce candidate mtDNA sequencing reads. In some embodiments, the mitochondrial reference genome is the revised Cambridge Reference Sequence (rCRS) for the mitochondrial genome. In some embodiments, the mapping step can be carried out using “bowtie2” (Langmead et al., Nature Methods, 2012, 9, 357-359) or bwa. The candidate mtDNA sequencing reads obtained from the first mapping step are re-mapped to
10 an entire human reference genome. In some embodiments, the human reference genome is GRCh38 for the nuclear genome. In addition, GRCh37 can also be used. In some embodiments, the mapping step can be carried out using “bowtie2”.

Upon carrying out the two-step mapping process, the candidate mtDNA sequencing reads are retained under two circumstances: 1) the candidate mtDNA sequencing reads are
15 retained when the candidate mtDNA sequencing read is uniquely mapped to the mitochondrial reference genome, or has fewer mismatches to the mitochondrial reference genome than to the human reference genome; and 2) the candidate mtDNA sequencing reads are retained when the alignment mismatch count of the candidate mtDNA sequencing read is less than 5 mismatched bases.

20 In some embodiments, the methods further comprise processing the mtDNA sequencing reads (obtained from the nucleic acid sequencing assay) for sorting and duplicate removal. In some embodiments, the methods further comprise processing the quality controlled mtDNA sequencing reads (obtained from the quality control and adaptor sequence removal process) for sorting and duplicate removal. In some embodiments, the methods
25 further comprise performing post-mapping processing of the retained candidate mtDNA sequencing read for sorting and duplicate removal. In some embodiments, the processing for sorting and duplicate removal can be carried out by using the “samtools toolkit” (Li et al., Bioinformatics, 2009, 25, 2078-2079). These processing steps are standard Next Generation Sequencing (NGS) data processing steps. The GATK toolkit can also be used.

30 In some embodiments, the methods further comprise down-sampling the mtDNA sequencing reads obtained from the nucleic acid sequencing assay to a desired depth prior to identifying the heteroplasmy and the homoplasmy. In some embodiments, the methods further

- 11 -

comprise down-sampling the mtDNA sequencing reads obtained from the nucleic acid sequencing assay to a desired depth prior to processing the mtDNA raw sequencing reads for quality control and adaptor sequence removal. In some embodiments, the mtDNA raw sequencing reads from a whole transcriptome data set can be down-sampled to 10 million
5 reads. In some embodiments, the down-sampling can be carried out by using “seqtk” (world wide web at “github.com/lh3/seqtk”). RNA seq data usually has very high mtDNA content. Thus, not all of the sequences are required to perform the methodology described herein because the greater the mtDNA coverage, the longer the computational time will be. In some embodiments, the desired depth is about 1000, but can be as low as about 200. Additional
10 tools that can be used include, for example, FASTQ-SAMPLE (world wide web at “homes.cs.washington.edu/~dcjones/fastq-tools/fastq-sample.html”).

In some embodiments, the methods described herein further comprise obtaining the plurality of biological samples from the individual prior to performing the nucleic acid sequencing assay on the plurality of samples. In some embodiments, the biological samples are
15 blood, tissue, or tumor biopsy. In some embodiments, the methods described herein further comprise amplifying nucleic acid molecules in the biological samples prior to performing the nucleic acid sequencing assay on the plurality of samples.

The present disclosure also provides methods of identifying an unreliable biological sample, the methods comprising: a) performing a nucleic acid sequencing assay on each
20 biological sample of a plurality of biological samples obtained from a single individual to obtain DNA raw sequencing reads for each biological sample; b) processing the DNA raw sequencing reads for quality control and adaptor sequence removal to produce quality controlled DNA sequencing reads; c) mapping the quality controlled DNA sequencing reads to a mitochondrial reference genome to produce candidate mtDNA sequencing reads; d) re-mapping the
25 candidate mtDNA sequencing reads to a human reference genome and retaining the candidate mtDNA sequencing reads when: i) the candidate mtDNA sequencing read is uniquely mapped to the mitochondrial reference genome or has fewer mismatches to the mitochondrial reference genome than to the human reference genome; and ii) the alignment mismatch count of the candidate mtDNA sequencing read is less than 5; e) performing post-mapping processing
30 of the retained candidate mtDNA sequencing reads for sorting and duplicate removal; f) identifying heteroplasmy and homoplasmy in the retained candidate mtDNA sequencing reads for each of the biological samples; and g) assigning a primary mtDNA haplogroup to each

- 12 -

biological sample, wherein any biological sample having an assigned primary mtDNA haplogroup that is different than the primary mtDNA haplogroup assigned to a majority of the biological samples from the same individual is an unreliable biological sample that is a mislabeled biological sample. The steps of this method can be carried out by the processes
5 described herein.

The present disclosure also provides methods of identifying an unreliable biological sample, the methods comprising: a) performing a nucleic acid sequencing assay on each biological sample of a plurality of biological samples obtained from a single individual to obtain DNA raw sequencing reads for each biological sample; b) processing the DNA raw sequencing
10 reads for quality control and adaptor sequence removal to produce quality controlled DNA sequencing reads; c) mapping the quality controlled DNA sequencing reads to a mitochondrial reference genome to produce candidate mtDNA sequencing reads; d) re-mapping the candidate mtDNA sequencing reads to a human reference genome and retaining the candidate mtDNA sequencing reads when: i) the candidate mtDNA sequencing read is uniquely mapped
15 to the mitochondrial reference genome or has fewer mismatches to the mitochondrial reference genome than to the human reference genome; and ii) the alignment mismatch count of the candidate mtDNA sequencing read is less than 5; e) performing post-mapping processing of the retained candidate mtDNA sequencing reads for sorting and duplicate removal; f) identifying heteroplasmy and homoplasmy in the retained candidate mtDNA sequencing reads
20 for each of the biological samples; g) assigning a primary mtDNA haplogroup to each biological sample; and h) determining the total heteroplasmy number for each biological sample, wherein when a biological sample has a high heteroplasmy number, the biological sample is assigned a secondary mtDNA haplogroup, wherein a biological sample having an assigned secondary mtDNA haplogroup that is different than the assigned primary mtDNA haplogroup is an
25 unreliable sample that is contaminated. The steps of this method can be carried out by the processes described herein.

In some embodiments, the methods described herein can be carried out as a workflow. Numerous workflow management tools such as, for example, Pyflow (see, world wide web at "github.com/Illumina/pyflow"), can be used to streamline the steps together.

30 The methods described herein have several advantages. First, the methods do not require any nuclear DNA (nDNA) variant information for each sample or at population allele frequency level -- this kind of information is often not available for many studies, especially for

- 13 -

RNA-seq studies. Second, the methods do not require intensively preprocessed sequencing data as input, such as whole genome mapped bam files, whole genome variants VCF files. The methods can directly take fastq files as input. Third, the methods can be applied to low-coverage sequencing data. nDNA variants-based methods usually need high coverage (>50X) to
5 detect low-level contaminations. Due to multiple-copy nature of mtDNA, even for the low coverage data, for example, 2 to 4X for 1000 Genomes Project, mtDNA coverage still can be as high as 1000 to 2000X, which is sufficient to detect contamination level as low as 1%. Fourth, the methods do not require high computational power, a typical sample with 1000X mtDNA coverage can be processed in 10 to 20 minutes with a single processor and 4Gb memory.
10 Samples with high mtDNA contents can take longer to process but can be down-sampled to shorten the processing time. The methods can be easily incorporated into standard NGS data processing pipelines and serve as an important quality control step by identifying problematic samples and further improving the accuracy of downstream data analysis.

In order that the subject matter disclosed herein may be more efficiently understood,
15 examples are provided below. It should be understood that these examples are for illustrative purposes only and are not to be construed as limiting the claimed subject matter in any manner.

Examples

20 **Example 1: mtDNA Variation Identification and Haplogroup Assignment**

General Methodology

mtDNA variations (both homoplasmy and heteroplasmy) are identified from next generation sequencing data by, for example, performing an analysis represented in Figure 1 (see, Box 1). Upon performing a nucleic acid sequencing assay on a plurality of biological
25 samples, raw sequencing reads can be down-sampled to a desired depth to lower the computational burden (see, Figure 1, Box 1, "Step0") using "seqtk" that can be found at, for example, the world wide web at "github.com/lh3/seqtk". This step is optional and need not be performed.

The raw mtDNA sequencing reads obtained from the nucleic acid sequencing assay,
30 optionally from the previous down-sampling step, are processed for quality control and adaptor sequence removal by using "Trimmomatic" (Bolger et al., Bioinformatics, 2014, 30, 2114-2120) (see, Figure 1, Box 1, "Step1").

- 14 -

To retrieve candidate mtDNA sequencing reads, the quality controlled sequencing reads are mapped to the mitochondrial reference genome (revised Cambridge Reference Sequence, rCRS) with "bowtie2" (Langmead et al., *Nature Methods*, 2012, 9, 357-359) (see, Figure 1, Box 1, "Step2"). Nuclear mitochondrial DNA segments (NUMTs) in nuclear genome may be mismapped to the mitochondrial genome and counted as mtDNA reads. To minimize the effect of NUMTs, a second-round mapping can be performed whereby the mapped reads from first round are re-mapped to the entire human reference genome, GRCh38 for the nuclear genome and the revised Cambridge Reference Sequence (rCRS) for the mitochondrial genome. Reads (or read pairs) are retained if: a) reads (read pairs) are uniquely mapped to the mitochondrial genome or have less mismatches to the mitochondrial genome than to the nuclear genome; and b) the alignment mismatch count is less than 5.

The retained candidate mtDNA sequencing reads are further processed by "samtools toolkit" (Li et al., *Bioinformatics*, 2009, 25, 2078-2079), including sam to bam conversion, sorting and duplication removal (see, Figure 1, Box 1, "Step3").

The retained candidate mtDNA sequencing reads for each mtDNA site are compiled with "samtools mpileup function" (Li et al., *Bioinformatics*, 2009, 25, 2078-2079), and bases are further filtered by sequencing quality (≥ 20), and heteroplasmies and homoplasmies are identified (see, Figure 1, Box 1, "Step4"). Heteroplasmies are identified with the following criteria: a) sequencing coverage ≥ 50 ; b) minor allele frequency $\geq 1\%$; and c) for DNA data, a minor allele must be observed at least twice from each strand, and for RNA data, a minor allele must be observed at least three times. Homoplasmies are identified with following criteria: a) sequencing coverage > 10 ; and b1) only one allele is observed at the given site and it is different from the reference allele, or b2) multiple alleles are observed and the major allele is different from the reference, but the site fails heteroplasmy criteria.

mtDNA sequences for each sample are constructed with the homoplasmy information and the major alleles at heteroplasmy sites, and haplogroups are assigned based on the constructed sequences using "HaploGrep2" (Weissensteiner et al., *Nuc. Acids Res.*, 2016, 44, W58-W63) (see, Figure 1, Box 1, "Step5"). The assigned haplogroup at this step is referred to as the primary haplogroup for each sample.

If a particular sample has an unusual high heteroplasmy number, a secondary mtDNA sequence is constructed with the homoplasmy information and the minor alleles at the

- 15 -

heteroplasmy sites, and a secondary haplogroup will be assigned based on this secondary mtDNA sequence. (see, Figure 1, Box 1, "Step6").

Sample mislabeling/swap detection (see, Figure 1, Box 2)

- 5 In a plurality of samples, each sample can be assigned a primary haplogroup as described herein. In the case where all samples are processed accurately, all samples from the same individual will be assigned to the same haplogroup. On the contrary, if two or more haplogroups were assigned among these samples, the sample(s) with a minority haplogroup assignment are considered haplogroup unmatched (i.e., mislabeled or swapped with another
- 10 sample). For example, in Table 1 below, sample 001 is considered to be swapped with sample 008.

Table 1

sample ID	subject ID	haplogroup	unmatched
001	A	J2b1a1	yes
002	A	H1ba	no
003	A	H1ba	no
004	A	H1ba	no
005	B	J2b1a1	no
006	B	J2b1a1	no
007	B	J2b1a1	no
008	B	H1ba	yes

Sample contamination detection and quantification (see, Figure 1, Box 3)

- 15 If an unusual high heteroplasmy number is observed in a particular sample, the sample is potentially contaminated. The primary and secondary haplogroups are assigned to the suspected sample based on the major alleles and minor alleles on the heteroplasmy sites, respectively (see, Figure 1, Box 1, "Step5" and "Step6"). If the primary and secondary haplogroups are different, the sample is considered to be a contaminated sample.
- 20 When a sample is determined to be contaminated, the median of the frequencies of all the heteroplasmies in the sample is used to represent the contamination level.

Example 2: Use of mtDNA Haplogroup to Detect a Mislabelled Sample

Samples are collected from several individuals and each individual has multiple samples. After obtaining RNA-seq data for a batch of clinical samples, mtDNA haplogroups are assigned to each sample (see, Table 2). Samples collected from the same individual should belong to the same mtDNA haplogroup. Unmatched mtDNA haplogroups suggests possible sample mislabeling. The sample with the haplogroup L3h1a1 should be considered a mislabeled sample.

Table 2

Sample ID	Subject ID	Haplogroup
001	A	T2+16189
02	A	T2+16189
003	A	T2+16189
004	A	L3h1a1
005	A	T2+16189

Example 3: Use of mtDNA Heteroplasmy to Detect Sample Contamination*Virtual Contamination Sample Preparation*

Whole genome sequencing fastq files of two individuals, HG00290 and NA19086, were downloaded from 1000 Genome Project (see, ftp site at "ftp.1000genomes.ebi.ac.uk/vol1/ftp/"). Sequencing reads were sampled from the two individuals, and NA19086 reads were mixed into HG00290 at different ratios (0.1%, 0.5%, 1%, 2%, 5%, 10%, 20%, 30%, and 40%) to create virtual contamination samples.

Real-World Datasets

DNA sequencing data were downloaded from 1000 Genome Project (see, ftp site at "ftp.1000genomes.ebi.ac.uk/vol1/ftp/"). For each individual, reads mapped to mitochondrial genome were extracted by samtools (Li et al., Bioinformatics, 2009, 25, 2078-2079) from the bam file and subsequently converted to pair end fastq files. The fastq files were used as the input for the methods described herein.

Fastq files of two RNA-seq studies were downloaded from GEO at GSE81266 and GSE127165. GSE81266 contained whole transcriptome data for 77 ileum and prepouch ileum

- 17 -

samples, including 61 pair end (2 x 75 bp) and 16 single end (50 bp) samples. GSE127165 contained whole transcriptome data from 57 laryngeal squamous cell carcinoma patients, each patient had a tumor sample and an adjacent normal sample. All samples were pair end with 150 bp read length.

5

Analytical Performance

Virtual contamination samples were analyzed. Whole genome sequencing data of two individuals, HG00290 and NA19086, were downloaded from 1000 Genomes Project (Auton et al., Nature, 2015, 526, 68-74). HG00290 belonged to haplogroup U5a2a1a and one
10 heteroplasmy was identified in this individual's mtDNA genome (2610T>C 1.4%), while NA19086 belonged to haplogroup D4b1a1 and two heteroplasmies were identified (1646T>C 2.1%, 12785T>T 21.3%). The two individuals had 45 nucleotide differences in their mitochondrial genome.

Virtual contamination samples were created by mixing the sequencing reads from the
15 two samples at a series ratio, ranging from 0.1% to 40%. HG00290 was treated as the original sample and NA19086 was treated as the contaminant. Each contamination sample contained 50 million read pairs with read length 100 bp. The virtual contaminated samples were processed by the methods described herein for contamination analysis and the results are summarized in Figure 3. When the contamination levels were above 2%, 45-46 heteroplasmies
20 were identified from the samples, much higher than the normal range (1 to 2 heteroplasmies in an individual). These heteroplasmies covered almost all the expected sites (the 45 segregating sites between the two individuals plus the original heteroplasmy 2610 T>C in HG00290). Only one expected site was missed in the 2% sample. The primary haplogroups were U5a2a1a for all 6 samples, which was same as the original sample HG00290 and the secondary haplogroup was
25 D4b1a1, same as the contaminant. When the contamination level was 1%, 29 heteroplasmies were detected. The 17 missing sites were manually checked and it was determined that these sites all showed some heteroplasmy signal, but because the heteroplasmy frequency identification cutoff was set at 1%, those sites did not make the cutoff. The secondary haplogroup of the 1% sample was correctly assigned to D4b1a1. Only 1 and 11 heteroplasmies
30 were detected when the contamination levels were 0.1% and 0.5%, and the secondary haplogroups for these two samples were still U5a2a1a, therefore, contamination cannot be confidently detected in these low contamination level samples. These results indicate that by

- 18 -

combining the heteroplasmy number and secondary haplogroup assignment, contaminations as low as 1% were able to be detected.

The heteroplasmy frequencies in the artificial contamination samples were further evaluated. There were some fluctuations of the heteroplasmy frequencies in each sample, but the mean and median of the frequencies were significantly correlated with theoretical contamination level (see, Figure 2, and Figure 3; Pearson correlation = 0.996781, 0.9979935, P value = 6.212e-09, 1.189e-09 for mean and median, respectively). Therefore, when a given sample is detected as contaminated by the methods described herein, the contamination level can be relatively quantified by the mean/median of the heteroplasmy frequencies in the sample.

Real-World Data Application: RNA-seq Data

There are several factors that can make the low-frequency (<5%) heteroplasms identification more challenging in RNA-seq data than that in DNA-seq data: 1) errors introduced during reverse transcription step; 2) RNA editing/modification; and/or 3) uneven coverage across the mtDNA genome due to varied gene expression levels. Therefore, to reduce the false positive heteroplasms, only heteroplasms with frequency > 5% was considered as reliable heteroplasms in RNA data. In addition, three well defined mtDNA editing sites: 295, 2617 and 13710 (Bar-Yaacov et al., Genome Res., 2013, 23, 1789-1796; and Hodgkinson et al., Science, 2014, 344, 413-415) were excluded.

The methods described herein were applied to two bulk RNA-seq datasets to evaluate different disease or tissue type context. First, the methods described herein were applied to a dataset with 77 samples from 25 subjects (Huang et al., Inflamm. Bowel Dis., 2017, 23, 366-378). Most subjects in this study had samples from different tissues (ileum and prepouch ileum) and/or at different biopsy time points (4 months, 8 months, 12 months etc.). 16 samples in this dataset were single end samples with 50 bp read length and 61 were pair end samples with 75 read length. For each sample, 10 million reads (pairs) were randomly sampled to test. The primary haplogroup assignments to the samples were first evaluated. In this dataset, samples from the same subject were all assigned to the same mtDNA haplogroup (see, Figure 4), indicating that there was no sample swapping. Potential contaminations in those samples was evaluated next. At 5% heteroplasmy frequency cutoff, except sample SRR3493833, all other samples had at most 6 heteroplasms and the secondary haplogroup assignments were the

- 19 -

same as the primary haplogroup (see, Figure 4). In sample SRR3493833, 29 heteroplasmies were identified, much higher than the normal range, and the median frequency of the heteroplasmies was 14.8%. The secondary haplogroup of this sample was J1c8a, which was also different from the primary haplogroup U5b2a1a. These results indicated that sample
5 SRR3493833 was potentially contaminated by another sample from J1 haplogroup and the contamination level was about 14.8%.

The methods described herein were also applied to a dataset involving tumor samples (Wu et al., Molec. Cancer, 2020, 19, 99). This dataset contained samples from 57 laryngeal squamous cell carcinoma patients, each has a tumor sample and a paired adjacent normal
10 mucosa sample. In this dataset, the paired tumor sample and adjacent normal sample from the same patient were all assigned to the same haplogroup (see, Figure 5), no sample swap was detected. All samples have low heteroplasmy numbers and same primary and secondary haplogroup assignment -- therefore, there was also no detectable contamination. By this dataset, the methods described herein were demonstrated to be able to identify tumor sample
15 identities.

Early and accurate sample swapping and contamination detection is a critical quality control step for large scale NGS data, since it can filter out suspected samples and improve the quality for subsequent analysis. In these examples, an efficient method is presented to detect sample swapping and cross-individual contamination by using mtDNA variations identified from
20 the NGS data. The methodology can take demultiplexed fastq files as input without any data preprocessing. It will first detect any sample swapping for individual with multiple samples. It will further detect and quantify potential contamination then suggest the source sample of the contaminants. Although whole genome DNA sequencing data from 1000 Genomes Project and two bulk RNA-seq datasets were used as working examples for these examples, the methods
25 described herein can be generalized to any NGS dataset containing mtDNA reads, such as whole exome sequencing data with offsites mtDNA reads, single cell RNA-seq, ATAC-seq data, etc. The simulation results described herein show that the methods described herein effectively detected contamination as low as 1%.

- 20 -

Various modifications of the described subject matter, in addition to those described herein, will be apparent to those skilled in the art from the foregoing description. Such modifications are also intended to fall within the scope of the appended claims. Each reference (including, but not limited to, journal articles, U.S. and non-U.S. patents, patent application
5 publications, international patent application publications, gene bank accession numbers, and the like) cited in the present application is incorporated herein by reference in its entirety.

- 21 -

What Is Claimed Is:

1. A method of identifying an unreliable biological sample, the method comprising:
performing a nucleic acid sequencing assay on each biological sample of a plurality of
biological samples obtained from a single individual to obtain mitochondrial DNA (mtDNA)
5 sequencing reads for each biological sample;
identifying a heteroplasmy and a homoplasmy in the mtDNA sequencing reads from
the previous step for each of the biological samples; and
assigning a primary mtDNA haplogroup to each biological sample, wherein any
biological sample having an assigned primary mtDNA haplogroup that is different than the
10 primary mtDNA haplogroup assigned to a majority of the biological samples from the same
individual is an unreliable biological sample that is a mislabeled biological sample.
2. The method according to claim 1, wherein the heteroplasmy is identified by
determining the sequencing coverage, a minor allele frequency, and presence of the minor
allele, wherein a heteroplasmy is present when: i) the sequencing coverage is ≥ 50 ; ii) the minor
15 allele frequency is $\geq 1\%$; and iii) for DNA data, the minor allele is observed at least twice from
each strand, or for RNA data, the minor allele is observed at least three times.
3. The method according to claim 1 or claim 2, wherein the homoplasmy is identified by
determining the sequencing coverage and the presence of one or more alleles, wherein a
homoplasmy is present when: i) the sequencing coverage is ≥ 10 ; and ii) only one allele is
20 observed at a particular nucleic acid site and it is different than the corresponding reference
allele, or multiple alleles are observed at a particular nucleic acid site and the major allele is
different than the corresponding reference allele, and the particular nucleic acid site fails
heteroplasmy criteria.
4. The method according to any one of claims 1 to 3, wherein assigning the primary
25 mtDNA haplogroup to each biological sample comprises constructing a mtDNA sequence for
each biological sample using the homoplasmy and major alleles of the heteroplasmy.
5. The method according to any one of claims 1 to 4, the method further comprising
determining the total heteroplasmy number for each biological sample, wherein when a
biological sample has a high heteroplasmy number, the biological sample is assigned a
30 secondary mtDNA haplogroup.
6. The method according to claim 5, wherein assigning the secondary mtDNA haplogroup
comprises constructing a secondary mtDNA sequence using the homoplasmy and minor alleles

- 22 -

of the heteroplasmy, wherein a biological sample having an assigned secondary mtDNA haplogroup that is different than the assigned primary mtDNA haplogroup is an unreliable biological sample that is contaminated.

7. The method according to claim 6, the method further comprising determining the level
5 of contamination of a biological sample by determining the median of the heteroplasmy frequencies of all heteroplasmies in the contaminated biological sample, wherein the greater the median of the heteroplasmy frequency, the greater the level of contamination.

8. The method according to any one of claims 1 to 7, the method further comprising,
prior to identifying a heteroplasmy and a homoplasmy, processing the mtDNA sequencing
10 reads obtained from the nucleic acid sequencing assay for quality control and adaptor sequence removal to produce quality controlled mtDNA sequencing reads.

9. The method according to claim 8, the method further comprising:
mapping the quality controlled mtDNA sequencing reads to a mitochondrial reference
genome to produce candidate mtDNA sequencing reads; and
15 re-mapping the candidate mtDNA sequencing reads to a human reference genome,
and retaining the candidate mtDNA sequencing reads when:

i) the candidate mtDNA sequencing read is uniquely mapped to the
mitochondrial reference genome or has fewer mismatches to the mitochondrial
reference genome than to the human reference genome; and
20 ii) the alignment mismatch count of the candidate mtDNA sequencing read is
less than 5.

10. The method according to claim 9, the method further comprising performing post-
mapping processing of the retained candidate mtDNA sequencing read for sorting and
duplicate removal.

25 11. The method according to any one of claims 1 to 10, the method further comprising
down-sampling the mtDNA sequencing reads obtained from the nucleic acid sequencing assay
to a desired depth prior to identifying the heteroplasmy and the homoplasmy and/or prior to
processing the mtDNA sequencing reads for quality control and adaptor sequence removal.

12. The method according to any one of claims 1 to 11, the method further comprising
30 obtaining the plurality of biological samples from the individual prior to performing the nucleic
acid sequencing assay on the plurality of biological samples.

- 23 -

13. The method according to any one of claims 1 to 12, wherein the biological samples are blood or tissue.
14. The method according to any one of claims 1 to 13, the method further comprising amplifying nucleic acid molecules in the biological samples prior to performing the nucleic acid
5 sequencing assay on the plurality of biological samples.
15. The method according to any one of claims 1 to 14, the method further comprising correctly labeling the one or more mislabeled biological samples or discarding the one or more mislabeled biological samples.
16. The method according to any one of claims 1 to 14, the method further comprising
10 discarding the one or more contaminated biological samples.
17. The method according to any one of claims 1 to 16, wherein the sequencing assay comprises next generation sequencing (NGS).
18. The method according to claim 17, wherein the NGS comprises whole genome sequencing.
- 15 19. The method according to claim 17, wherein the NGS comprises whole exome sequencing.
20. The method according to claim 17, wherein the NGS comprises RNA sequencing.
21. The method according to claim 17, wherein the NGS comprises bisulfite sequencing.
22. A method of identifying an unreliable biological sample, the method comprising:
20 performing a nucleic acid sequencing assay on each biological sample of a plurality of biological samples obtained from a single individual to obtain mitochondrial DNA (mtDNA) raw sequencing reads for each biological sample;
processing the mtDNA raw sequencing reads for quality control and adaptor sequence removal to produce quality controlled mtDNA sequencing reads;
25 mapping the quality controlled mtDNA sequencing reads to a mitochondrial reference genome to produce candidate mtDNA sequencing reads;
re-mapping the candidate mtDNA sequencing reads to a human reference genome and retaining the candidate mtDNA sequencing reads when:
30 i) the candidate mtDNA sequencing read is uniquely mapped to the mitochondrial reference genome or has fewer mismatches to the mitochondrial reference genome than to the human reference genome; and

- 24 -

- ii) the alignment mismatch count of the candidate mtDNA sequencing read is less than 5;
- performing post-mapping processing of the retained candidate mtDNA sequencing reads for sorting and duplicate removal;
- 5 identifying heteroplasmy and homoplasmy in the retained candidate mtDNA sequencing reads for each of the biological samples; and
- assigning a primary mtDNA haplogroup to each biological sample, wherein any biological sample having an assigned primary mtDNA haplogroup that is different than the primary mtDNA haplogroup assigned to a majority of the biological samples from the same
- 10 individual is an unreliable biological sample that is a mislabeled biological sample.
23. A method of identifying an unreliable biological sample, the method comprising:
- performing a nucleic acid sequencing assay on each biological sample of a plurality of biological samples obtained from a single individual to obtain mitochondrial DNA (mtDNA) raw sequencing reads for each biological sample;
- 15 processing the mtDNA raw sequencing reads for quality control and adaptor sequence removal to produce quality controlled mtDNA sequencing reads;
- mapping the quality controlled mtDNA sequencing reads to a mitochondrial reference genome to produce candidate mtDNA sequencing reads;
- re-mapping the candidate mtDNA sequencing reads to a human reference genome and
- 20 retaining the candidate mtDNA sequencing reads when:
- i) the candidate mtDNA sequencing read is uniquely mapped to the mitochondrial reference genome or has fewer mismatches to the mitochondrial reference genome than to the human reference genome; and
- ii) the alignment mismatch count of the candidate mtDNA sequencing read is
- 25 less than 5;
- performing post-mapping processing of the retained candidate mtDNA sequencing reads for sorting and duplicate removal;
- identifying heteroplasmy and homoplasmy in the retained candidate mtDNA sequencing reads for each of the biological samples;
- 30 assigning a primary mtDNA haplogroup to each biological sample; and
- determining the total heteroplasmy number for each biological sample, wherein when a biological sample has a high heteroplasmy number, the biological sample is assigned a

- 25 -

secondary mtDNA haplogroup, wherein a biological sample having an assigned secondary mtDNA haplogroup that is different than the assigned primary mtDNA haplogroup is an unreliable sample that is contaminated.