CLAIMS

1. A method for detecting the presence or absence of contamination of a first sample with a second sample among a plurality of samples, comprising:

(a) processing the first sample and the second sample, wherein the processing comprises:

i. tagging a set of cell-free nucleic acid molecules in each sample with a set of molecular barcodes to generate tagged polynucleotides, wherein the set of molecular barcodes comprise 5-200 different molecular barcode sequences;

ii. amplifying a portion of the tagged polynucleotides to generate progeny polynucleotides;

(b) for each of the first sample and the second sample, sequencing a portion of the progeny polynucleotides to generate sequencing reads;

(c) aligning a plurality of sequencing reads from the first sample and the second sample to a reference sequence whereby a genomic start position and a genomic stop position of the cell-free nucleic acid molecule is determined from the alignment;

(d) for each of the first sample and the second sample, grouping the plurality of sequencing reads into a plurality of families based on grouping features, which comprise at least one of (i) one or more molecular barcodes attached to a cell-free nucleic acid molecule in the sample, (ii) start genomic position and (iii) stop genomic position of the cell-free nucleic acid molecule, wherein each family in the sample comprises sequencing reads of progeny polynucleotides amplified from a unique cell-free nucleic acid molecule among the set of cell-free nucleic acid molecules in the sample;

(e) generating family identifiers for the plurality of families;

(f) screening for a set of shared family identifiers, wherein a given shared family identifier is a family identifier of the first sample that is identical or substantially identical to a family identifier of the second sample;

(g) determining a quantitative measure of the set of shared family identifiers; and

(h) classifying the first sample as being contaminated with the second sample if the quantitative measure of the set of shared family identifiers is above a predetermined threshold, or as not being contaminated with the second sample if the quantitative measure of the set of shared family identifiers is at or below the predetermined threshold, thereby detecting the presence or absence of contamination.

2. The method of claim 1, wherein the quantitative measure of the set of shared family identifiers is a number of shared family identifiers in the first sample.

3. The method of claim 1, wherein the quantitative measure of the set of shared family identifiers excludes shared family identifiers in the first sample for which the number of sequencing reads in the family of the first sample is greater than the number of sequencing reads in the corresponding family of the second family.

4. The method of claim 1, wherein the quantitative measure of the set of shared family identifiers in the first sample excludes shared family identifiers at over-represented pairs of genomic start positions and genomic stop positions.

5. The method of claim 4, wherein the over-represented pairs of genomic start positions and genomic stop positions are determined by:

(a) providing sets of sequencing reads from the plurality of samples, wherein the sets of sequencing reads comprise a distribution of genomic start positions and genomic stop positions that are identical or substantially identical to the first sample;

(b) determining family identifiers in the sets of sequencing reads;

(c) quantifying number of family identifiers in the sets of sequencing reads sharing a pair of genomic start position and genomic stop position; and

(d) categorizing the pair of genomic start position and genomic stop position as over-represented if the number of family identifiers exceeds a set threshold.

6. The method of claim 1, wherein one or more sample indexes are attached to one or both ends of the progeny polynucleotides prior to the sequencing, wherein the one or more sample indexes distinguishes the first sample and the second sample.

7. The method of claim 5, wherein the set threshold is at least 5, at least 10, at least 15 or at least 20.

8. The method of claim 1, wherein the one or more molecular barcodes are attached to both ends of the cell-free nucleic acid molecule.

9. The method of claim 6, wherein the first sample and the second sample are sequenced in a same flow cell.

10. The method of claim 1, wherein the processing, further comprises, enriching a portion of the progeny polynucleotides for specific regions of interest to generate enriched molecules.

11. The method of claim 6, wherein the plurality of samples comprises samples processed in a same flow cell as the first sample.

12. The method of claim 1, wherein the predetermined threshold is at least 0.5% or at least 1% of total number of families in the first sample.

13. The method of claim 1, wherein the first sample is obtained from a bodily fluid of a subject and the second sample is obtained from the bodily fluid of another subject.

14. The method of claim 13, wherein the bodily fluid is plasma.

15. A computer-implemented method for detecting the presence or absence of contamination of a first sample with a second sample among a plurality of samples, comprising:

(a) obtaining sequence information comprising a plurality of sequencing reads derived from a set of cell-free nucleic acid molecules from the first sample and another set of cell-free nucleic acid molecules from the second sample;

(b) aligning the plurality of sequencing reads to a reference sequence whereby a genomic start position and a genomic stop position of the cell-free nucleic acid molecule is determined from the alignment;

(c) for each of the first sample and the second sample, grouping the plurality of sequencing reads into a plurality of families based on grouping features, which comprise at least one of (i) one or more molecular barcodes attached to a cell-free nucleic acid molecule in the sample, (ii) start genomic position and (iii) stop genomic position of the cell-free nucleic acid molecule, wherein each family in the sample comprises sequencing reads of progeny polynucleotides amplified from a unique cell-free nucleic acid molecule among the set of cell-free nucleic acid molecules in the sample;

(d) generating family identifiers for the plurality of families;

(e) screening for a set of shared family identifiers, wherein a given shared family identifier is a family identifier of the first sample that is identical or substantially identical to a family identifier of the second sample;

(f) determining a quantitative measure of the set of shared family identifiers; and

(g) classifying the first sample as being contaminated with the second sample if the quantitative measure of the set of shared family identifiers is above a predetermined threshold, or as not being contaminated with the second sample if the quantitative measure of the set of shared family identifiers is at or below the predetermined threshold, thereby detecting the presence or absence of contamination.

16. The method of claim 15, wherein the quantitative measure of the set of shared family identifiers is a number of shared family identifiers in the first sample.

17. The method of claim 15, wherein the quantitative measure of the set of shared family identifiers excludes shared family identifiers in the first sample for which the number of sequencing reads in the family of the first sample is greater than the number of sequencing reads in the corresponding family of the second family.

18. The method of claim 15, wherein the quantitative measure of the set of shared family identifiers in the first sample excludes shared family identifiers at over-represented pairs of genomic start positions and genomic stop positions.

19. The method of claim 18, wherein the over-represented pairs of genomic start positions and genomic stop positions are determined by:

(a) providing sets of sequencing reads from the plurality of samples, wherein the sets of sequencing reads comprise a distribution of genomic start positions and genomic stop positions that are identical or substantially identical to the first sample;

(b) determining family identifiers in the sets of sequencing reads;

(c) quantifying number of family identifiers in the sets of sequencing reads sharing a pair of genomic start position and genomic stop position; and

(d) categorizing the pair of genomic start position and genomic stop position as over-represented if the number of family identifiers exceeds a set threshold.

20. The method of claim 19, wherein the plurality of samples comprises samples comprising sample indexes processed in a same flow cell as the first sample.