heterozygosity is not observed, for each chromosome in a plurality of samples, wherein each chromosome comprises a plurality of genomic regions, determining expected read coverage data for each chromosome in the plurality of samples, determining a deviation between the read coverage data and the expected read coverage data for at least one chromosome in the plurality of samples, determining a deviation in the allele balance distribution from an expected ratio of 1:1 for a plurality of bi-allelic SNPs for at least one chromosome in a plurality of samples, determining whether the deviation occurs over the entire chromosome or only a portion of the identified chromosome, using the read coverage and allele balance data in complement to further refine and validate identified deviations for at least one chromosome in a plurality of samples, and identifying the at least one chromosome as an abnormal karyotype.

[0004A] According to a first aspect, a method may include: determining, for each chromosome in a plurality of samples, if a chromosome is an autosome or a sex chromosome; wherein if a chromosome is an autosome: detecting, based on determining a deviation between read coverage data and expected read coverage data for the chromosome, a read coverage anomaly; determining a chromosome-wide heterozygous allele balance (ChromHetAB); determining Runs-of-Homozygosity (ROH); and determining local median heterozygous allele balance (LocalHetAB); wherein if a chromosome is a sex chromosome, wherein the sample is associated with a male: detecting, based on determining a deviation between read coverage data and expected read coverage data for the chromosome, a read coverage anomaly; and determining ChromHetAB; and identifying the autosome or the sex chromosome as an abnormal karyotype based on one or more of the read coverage anomaly, a ChromHetAB anomaly, a ROH anomaly, or a LocalHetAB anomaly.

[0005] Additional advantages will be set forth in part in the description which follows or may be learned by practice. The advantages will be realized and attained by means of the elements and combinations particularly pointed out in the appended claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0006] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate embodiments and together with the description, serve to explain the principles of the methods and systems:

Figure 1 is flowchart illustrating an example abnormal karyotype detection method;

Figure 2 is a graph illustrating an example linear regression model;

Figure 3 is a graph illustrating abnormal karyotypes exhibiting large residuals;

Figure 4 is another flowchart illustrating an example abnormal karyotype detection method;

Figure 5 shows a graph illustrating the relationship of GC content and coverage;

Figure 6 is a graph illustrating identified abnormal karyotypes and outliers;

Figures 7A, 7B, 7C, 7D, 7E, and 7F are allele balance plots showing anomalies on chromosomes 9, 13 and 20 for a sample. The subplot number is the chromosome number. The shaded bar (**701**) denotes the normal range of variability expected for a heterozygous SNP allele balance of 0.5. The solid line (**702**) denotes the whole chromosome median allele balance. The dashed line (**703**) denotes the local median allele balance in an approximately 20 SNP rolling window. The line (**704**)

CLAIMS

What is claimed is:

1. A method comprising:

determining, for each chromosome in a plurality of samples, if a chromosome is an autosome or a sex chromosome;

wherein if a chromosome is an autosome:

detecting, based on determining a deviation between read coverage data and expected read coverage data for the chromosome, a read coverage anomaly;

determining a chromosome-wide heterozygous allele balance (ChromHetAB);

determining Runs-of-Homozygosity (ROH); and

determining local median heterozygous allele balance (LocalHetAB);

wherein if a chromosome is a sex chromosome, wherein the sample is associated with a male:

detecting, based on determining a deviation between read coverage data and expected read coverage data for the chromosome, a read coverage anomaly; and

determining ChromHetAB; and

identifying the autosome or the sex chromosome as an abnormal karyotype based on one or more of the read coverage anomaly, a ChromHetAB anomaly, a ROH anomaly, or a LocalHetAB anomaly.

2. The method of claim 1, further comprising filtering the read coverage data, optionally wherein filtering the read coverage data comprises filtering the read coverage data based on a level of guanine-cytosine (GC) content in one or more genomic regions of the plurality of genomic regions, and/or wherein filtering the read coverage data comprises filtering one or more genomic regions of the plurality of genomic regions based on a mappability score of the one or more genomic regions of the plurality of genomic regions.

3. The method of claim 1, wherein determining that the sample is associated with a male comprises generating a plot of chromosome X versus chromosome Y coverage ratios for all samples and identifying occurrences above a threshold as male and occurrences below the threshold as female, and/or

wherein determining the ChromHetAB for a sample on a chromosome comprises determining a summary statistic representing the chromosome-wide heterozygous SNP allele balance among all unfiltered variants within a chromosome, optionally wherein the summary statistic is a median.

4. The method of claim 3, further comprising detecting a ChromHetAB anomaly, where a sample's ChromHetAB value is significantly smaller than a ChromHetAB value expected for a karyotypically normal sample, optionally wherein detecting the ChromHetAB anomaly comprises:

fitting a linear regression on the ChromHetAB for all samples on a given chromosome relative to a quality control metric;

calculating a Z-score for a residual of every sample's ChromHetAB;

transforming the Z-score into a p-value; and

identifying any occurrence of p-value less than a threshold as a ChromHetAB anomaly.

5. The method of claim 4, wherein determining the LocalHetAB comprises determining a smoothed, sub-chromosome scale summary statistic of a sample's heterozygous SNP allele balance across all unfiltered variants on a chromosome, optionally wherein the smoothed, sub-chromosome scale summary statistic comprises a running median.

6. The method of claim 5, further comprising detecting a LocalHetAB anomaly, where the LocalHetAB value falls below the corresponding ChromHetAB value over a chromosomal region, indicating a possible partial chromosome anomaly, optionally wherein detecting the LocalHetAB anomaly comprises:

determining a LocalHetAB Event, wherein the LocalHetAB Event comprises a contiguous region of two or more SNPs, all having LocalHetAB < ChromHetAB;

determining a callable chromosome length, wherein the callable chromosome length comprises a number of base pairs between first and last unfiltered exons on the

chromosome – a number of overlapping centromeric bases;

determining a LocalHetAB Event Area by calculating sum(PairwiseArea for all N-1 neighboring SNP pairs in the LocalHetAB Event) / (callable chromosome length * ChromHetAB);

defining a linear function empirically fit to the variation data relating LocalHetAB Event Area and the number of SNPs included in the LocalHetAB Event; and

identifying any occurrence of the number of SNPs + (LocalHetAB Event Area * a first amount) >= a second amount as a LocalHetAB anomaly.

7. The method of claim 1, wherein each chromosome comprises a plurality of genomic regions comprises and wherein determining read coverage data for each chromosome in the plurality of samples comprises determining a sum of read depths over exomic regions with guanine-cytosine (GC) content within a range and a mappability score above a threshold.

8. The method of claim 2, wherein filtering the read coverage data based on a level of guanine-cytosine (GC) content in one or more genomic regions of the plurality of genomic regions comprises:

determining a level of GC content for each of the plurality of genomic regions; and

excluding one or more genomic regions of the plurality of genomic regions having a level of GC content outside a range.

9. The method of claim 2, wherein filtering one or more genomic regions of the plurality of genomic regions based on a mappability score of the one or more genomic regions of the plurality of genomic regions comprises:

determining a mappability score for each genomic region of the plurality of genomic regions; and

excluding one or more genomic regions of the plurality of genomic if the mappability score of the one or more genomic regions of the plurality of genomic regions is below a predetermined threshold.

10. The method of claim 1, further comprising normalizing the read coverage data, optionally wherein normalizing the read coverage data comprises determining an exome-wide ratio of read coverage for each chromosome relative to other autosomes.

11. The method of claim 10, wherein the exome-wide ratio (γ) is determined for each chromosome (i) by:

$$\gamma_i = \frac{r_i}{\sum_{\forall j \in (A-i)} r_j}$$

wherein A is the set of autosomes and r is read coverage.

12. The method of claim 1, further comprising determining expected read coverage data for each chromosome in the plurality of samples by applying a linear regression model to determine an expected exome-wide ratio for each chromosome, wherein a plurality of metrics are used as covariates, optionally wherein the plurality of metrics comprise sequencing quality control metrics (QC metrics) and the expected exome-wide ratio ($\hat{\gamma}$) is determined for each chromosome (i) by:

$$\hat{\gamma}_i = f(\text{QC metrics}, \gamma_j, \gamma_k)$$

wherein chromosomes j,k are defined as two autosomes with minimal D statistics relative to a GC content distribution of chromosome i.

13. The method of claim 12, wherein determining the deviation between the read coverage data and the expected read coverage data comprises:

determining, for each chromosome in the plurality of samples, a difference between the read coverage data and the expected read coverage data to generate a plurality of residuals; and

Z-score normalizing the plurality of residuals relative to a standard error of the mean estimate, $S(\hat{\gamma}_i)$, for an individual sample of the plurality of samples with covariates x:

$$S(\hat{\gamma}_i) = S_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{j=1..n} (x_j - \bar{x})^2}}$$

where S_e is the residual standard error, and:

$$Z_i = \frac{(\gamma_i - \hat{\gamma}_i)}{S(\hat{\gamma}_i)\sqrt{n}}.$$

14. The method of claim 13, further comprising determining a p-value based on the Z-score for each chromosome to identify significantly large residuals, representing abnormal karyotypes for chromosome *I*, optionally wherein significantly large residuals comprise residuals having a p-value less than 0.05.

15. The method of claim 14, further comprising:

detecting one or more outliers; and

removing the one or more outliers from consideration for identification as an abnormal karyotype, optionally wherein detecting one or more outliers comprises flagging one or more of the plurality of samples having leverage (h_i, where $1/n < h_i < 1$) above a threshold on the linear regression model for each chromosome, wherein leverage is determined as a function of *n* and *p*:

$$h_i(n, p) = \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{j=1..n} (x_j - \bar{x})^2} \right] \bullet \frac{n}{p+1}$$

where *p* is the number of covariates in the model, *n* is the number of samples modeled, x_i represents a vector of covariates for sample *i*, and $\bar{x}$ is the vector of covariate means across the sample population, preferably wherein the threshold is from about 3 to about 5.