

[0030] In some embodiments, the different portion of the test nucleic acid represented by the bin corresponding to the first set of sequence reads is from a first chromosome of the biological sample, and the different portion of the test nucleic acid represented by the bin corresponding to the second set of sequence reads is from a second chromosome of the biological sample, where the second chromosome is other than the first chromosome. In some such embodiments, the first chromosome is a paternal chromosome and the second chromosome is a maternal chromosome. In some such embodiments, the biological sample is human and the first chromosome is chromosome 21, 18 or 13.

[0031] In some embodiments, the structural variation is deemed to have occurred and the method further comprises treating a subject that originated the biological sample with a treatment regimen responsive to the structural variation. In some embodiments, the treatment regimen comprises a diet modification. In some embodiments, the treatment regimen comprises application of a pharmaceutical composition that inhibits or augments a biological pathway associated with the structural variation.

[0032] Also disclosed is a computing system comprising one or more processors and memory storing one or more programs to be executed by the one or more processors. The one or more programs comprise instructions for obtaining a plurality of sequence reads. Each respective sequence read in the plurality of sequence reads comprises a first portion that corresponds to a subset of the test nucleic acid and a second portion that encodes a barcode for the respective sequence read. The barcode is independent of the sequencing data of the test nucleic acid. Each respective bin in the plurality of bins represents a different portion of the test nucleic acid. The bin information identifies, for each respective bin in the plurality of bins, a set of sequence reads in a plurality of sets of sequence reads. Each sequence read in each set of sequence reads in the plurality of sets of sequence reads is in the plurality of sequence reads. Moreover, each respective sequence read, in each respective set of sequence reads in the plurality of sets of sequence reads, has a respective first portion that corresponds to a subset of the test nucleic acid that at least partially overlaps the different portion of the test nucleic acid that is represented by the bin corresponding to the respective set of sequence reads. A determination is made of the number of unique barcodes in a first set of sequence reads in the plurality of sequence reads, where the first set of sequence reads is of a first bin in the plurality of bins, that are also found in a second set of sequence reads in the plurality of sequence reads, where the second set of sequence reads is a second bin the plurality of bins. A probability or likelihood that this number is attributable to chance is made by

comparison of a metric based upon the number to a threshold criterion. When the metric satisfies the threshold criterion, a structural variation is deemed to have occurred in (i) the different portion of the test nucleic acid that is represented by the first set of sequence reads and/or (ii) the different portion of the test nucleic acid that is represented by the second set of sequence reads.

[0033] Also disclosed is a non-transitory computer readable storage medium storing one or more programs configured for execution by a computer. The one or more programs comprise instructions for obtaining a plurality of sequence reads. Each respective sequence read in the plurality of sequence reads comprises a first portion that corresponds to a subset of the test nucleic acid and a second portion that encodes an barcode for the respective sequence read. The barcode is independent of the sequencing data of the test nucleic acid. Bin information for a plurality of bins is obtained. Each respective bin in the plurality of bins represents a different portion of the test nucleic acid. The bin information identifies, for each respective bin in the plurality of bins, a set of sequence reads in a plurality of sets of sequence reads. Each sequence read in each set of sequence reads in the plurality of sets of sequence reads is in the plurality of sequence reads. Moreover, each respective sequence read, in each respective set of sequence reads in the plurality of sets of sequence reads, has a respective first portion that corresponds to a subset of the test nucleic acid that at least partially overlaps the different portion of the test nucleic acid that is represented by the bin corresponding to the respective set of sequence reads. A determination is made of the number of unique barcodes in a first set of sequence reads in the plurality of sequence reads, where the first set of sequence reads is of a first bin in the plurality of bins, that are also found in a second set of sequence reads in the plurality of sequence reads, where the second set of sequence reads is of a second bin in the plurality of bins. A probability or likelihood that this number is attributable to chance is made by comparison of a metric based upon the number to a threshold criterion. When the metric satisfies the threshold criterion, a structural variation is deemed to have occurred in (i) the different portion of the test nucleic acid that is represented by the first set of sequence reads and/or (ii) the different portion of the test nucleic acid that is represented by the second set of sequence reads.

[0034] Part B, Additional Embodiment for Structural Variation. Disclosed is a method of determining a likelihood of a structural variation occurring in a test nucleic acid obtained from a single biological sample. The method comprises, at a computer system having one or more processors, and memory storing one or

respective first subset of sequence reads, the different first calculated fragment of the respective fragment pair originates with a first sequence read having the barcode corresponding to the respective fragment pair in the first bin, and each sequence read in the respective first subset of sequence reads is from the first bin. The different second calculated fragment consists of a respective second subset of sequence reads in the plurality of sequence reads having the barcode corresponding to the respective fragment pair, where each sequence read in the respective second subset of sequence reads is within a predefined genetic distance of another sequence read in the respective second subset of sequence reads, the different second calculated fragment of the respective fragment pair originates with a second sequence read having the barcode corresponding to the respective fragment pair in the second bin, and each sequence read in the respective second subset of sequence reads is from the second bin.

[0039] The method further comprises computing a respective likelihood based upon a probability of occurrence of a first model and a probability of occurrence of a second model regarding the one or more fragment pairs to thereby provide a likelihood of a structural variation in the test nucleic acid. Here, the first model specifies that the respective first calculated fragments and the respective second calculated fragments of the one or more fragment pairs are observed given no structural variation in the target nucleic acid sequence and are part of a common molecule. Further, the second model specifies that the respective first calculated fragments and the respective second calculated fragments of the one or more fragment pairs are observed given structural variation in the target nucleic acid sequence.

[0039a] In one aspect, there is provided a method of determining a likelihood of a structural variation occurring in a test nucleic acid obtained from a single biological sample, the method comprising:

at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors:

(A) obtaining a plurality of sequence reads from a plurality of sequencing reactions in which the test nucleic acid is fragmented, wherein

each respective sequence read in the plurality of sequence reads comprises a first portion that corresponds to a subset of the test nucleic acid and a second portion that encodes a respective barcode for the respective sequence read in a plurality of barcodes, and

each respective barcode is independent of the sequencing data of the test nucleic acid, and

the plurality of sequence reads collectively include the plurality of barcodes;

(B) obtaining bin information for a plurality of bins, wherein each respective bin in the plurality of bins represents a different portion of the test nucleic acid,

the bin information identifies, for each respective bin in the plurality of bins, a set of sequence reads in a plurality of sets of sequence reads that are in the plurality of sequence reads, and

the respective first portion of each respective sequence read in each respective set of sequence reads in the plurality of sets of sequence reads corresponds to a subset of the test nucleic acid that at least partially overlaps the different portion of the test nucleic acid that is represented by the bin corresponding to the respective set of sequence reads;

(C) identifying, from among the plurality of bins, a first bin and a second bin that correspond to portions of the test nucleic acid that are nonoverlapping, wherein the first bin is represented by a first set of sequence reads in the plurality of sequence reads and the second bin is represented by a second set of sequence reads in the plurality of sequence reads;

(D) determining a first value that represents a numeric probability or likelihood that the number of barcodes common to the first set and the second set is attributable to chance;

(E) responsive to a determination that the first value satisfies a predetermined cut-off value, for each barcode that is common to the first bin and the second bin, obtaining a fragment pair thereby obtaining one or more fragment pairs, each fragment pair in the one or more fragment pairs (i) corresponding to a different barcode that is common to the first bin and the second bin and (ii) consisting of a different first calculated fragment and a different second calculated fragment, wherein, for each respective fragment pair in the one or more fragment pairs:

the different first calculated fragment consists of a respective first subset of sequence reads in the plurality of sequence reads having the barcode corresponding to the respective fragment pair, wherein

each sequence read in the respective first subset of sequence reads is within a predefined genetic distance of another sequence read in the respective first subset of sequence reads,

the different first calculated fragment of the respective fragment pair originates with a first sequence read having the barcode corresponding to the respective fragment pair in the first bin, and

each sequence read in the respective first subset of sequence reads is from the first bin, and

the different second calculated fragment consists of a respective second subset of sequence reads in the plurality of sequence reads having the barcode corresponding to the respective fragment pair, wherein

each sequence read in the respective second subset of sequence reads is within a predefined genetic distance of another sequence read in the respective second subset of sequence reads,

the different second calculated fragment of the respective fragment pair originates with a second sequence read having the barcode corresponding to the respective fragment pair in the second bin, and

each sequence read in the respective second subset of sequence reads is from the second bin; and

(F) computing a respective likelihood based upon a probability of occurrence of a first model and a probability of occurrence of a second model regarding the one or more fragment pairs to thereby provide a likelihood of a structural variation in the test nucleic acid, wherein

(i) the first model specifies that the respective first calculated fragments and the respective second calculated fragments of the one or more fragment pairs are observed given no structural variation in the target nucleic acid sequence and are part of a common molecule, and

(ii) the second model specifies that the respective first calculated fragments and the respective second calculated fragments of the one or more fragment pairs are observed given structural variation in the target nucleic acid sequence.

[0040] In some embodiments, the computed likelihood term is a ratio score between the probability of occurrence of the first model and the probability of occurrence of the second model.

[0041] In some embodiments, the first bin and the second bin are at least a predetermined number of kilobases apart on the test nucleic acid.

[0042] In some embodiments, the first bin and the second bin are at least 50 kilobases apart on the test nucleic acid.

[0043] In some embodiments, a binomial test to compute the first value. In some embodiments, this binomial test has the form:

$$p = 1 - P_{Binom}(n; n_1 n_2 / B)$$

mass function of a Poisson distribution with parameter b for r_2 , $P_p(0; a_b(m - l_1 - l_2))$ is a probability mass function of a Poisson distribution with parameter b , and $P_L(m)$ is a pre-estimated probability that the true common molecule length is m . In some such embodiments, $2d'$ is estimated by computing the maximum extent d' such that $P_p(0; a_b d') \geq 0.75$.

[0065] In some embodiments, the plurality of sequence reads represents whole genome sequencing data. In some embodiments, the plurality of sequence reads represents targeted sequencing of a subset of a genome, a first subset of the plurality of sequence reads sequence reads are from inside the subset of the genome and have a first read rate of a_b , a second subset of the plurality of sequence reads sequence reads are from outside the subset of the genome and have a first read rate of $\tilde{a}_b$, wherein $\tilde{a}_b$ is different than a_b , and wherein the likelihood in the computing (F) corrects for the different read rates of the respective first and second subset of the plurality of sequence reads.

[0066] Disclosed is a computing system, comprising one or more processors, memory storing one or more programs to be executed by the one or more processors. The one or more programs comprising instructions for obtaining a plurality of sequence reads from a plurality of sequencing reactions in which the test nucleic acid is fragmented. Each respective sequence read in the plurality of sequence reads comprises a first portion that corresponds to a subset of the test nucleic acid and a second portion that encodes a barcode for the respective sequence read. The barcode is independent of the sequencing data of the test nucleic acid.

[0067] The one or more programs further comprise instructions for obtaining bin information for a plurality of bins. Each respective bin in the plurality of bins represents a different portion of the test nucleic acid. The bin information identifies, for each respective bin in the plurality of bins, a set of sequence reads in a plurality of sets of sequence reads that are in the plurality of sequence reads. The respective first portion of each respective sequence read in each respective set of sequence reads in the plurality of sets of sequence reads corresponds to a subset of the test nucleic acid that at least partially overlaps the different portion of the test nucleic acid that is represented by the bin corresponding to the respective set of sequence reads. The one or more programs further comprise instructions for identifying, from among the plurality of bins, a first bin and a second bin that correspond to portions of the test nucleic acid that are nonoverlapping, where the first bin is represented by a first set of sequence reads in the

first calculated fragments and the respective second calculated fragments of the one or more fragment pairs are observed given structural variation in the target nucleic acid sequence.

[0069a] In one aspect, there is provided a computing system, comprising:

one or more processors;

memory storing one or more programs to be executed by the one or more processors;

the one or more programs comprising instructions for:

(A) obtaining a plurality of sequence reads from a plurality of sequencing reactions in which the test nucleic acid is fragmented, wherein

each respective sequence read in the plurality of sequence reads comprises a first portion that corresponds to a subset of the test nucleic acid and a second portion that encodes a respective barcode for the respective sequence read in a plurality of barcodes,

each respective barcode is independent of the sequencing data of the test nucleic acid, and

the plurality of sequence reads collectively include the plurality of barcodes;

(B) obtaining bin information for a plurality of bins, wherein

each respective bin in the plurality of bins represents a different portion of the test nucleic acid,

the bin information identifies, for each respective bin in the plurality of bins, a set of sequence reads in a plurality of sets of sequence reads that are in the plurality of sequence reads, and

the respective first portion of each respective sequence read in each respective set of sequence reads in the plurality of sets of sequence reads corresponds to a subset of the test nucleic acid that at least partially overlaps the different portion of the test nucleic acid that is represented by the bin corresponding to the respective set of sequence reads;

(C) identifying, from among the plurality of bins, a first bin and a second bin that correspond to portions of the test nucleic acid that are nonoverlapping, wherein the first bin is represented by a first set of sequence reads in the plurality of sequence reads and the second bin is represented by a second set of sequence reads in the plurality of sequence reads;

(D) determining a first value that represents a numeric probability or likelihood that the number of barcodes common to the first set and the second set is attributable to chance;

(E) responsive to a determination that the first value satisfies a predetermined cut-off value, for each barcode that is common to the first bin and the second bin, obtaining a fragment pair thereby obtaining one or more fragment pairs, each fragment pair in the one or more fragment pairs (i) corresponding to a different barcode that is common to the first bin and the second bin and (ii) consisting of a different first calculated fragment and a different second calculated fragment, wherein, for each respective fragment pair in the one or more fragment pairs:

the different first calculated fragment consists of a respective first subset of sequence reads in the plurality of sequence reads having the barcode corresponding to the respective fragment pair, wherein

each sequence read in the respective first subset of sequence reads is within a predefined genetic distance of another sequence read in the respective first subset of sequence reads,

the different first calculated fragment of the respective fragment pair originates with a first sequence read having the barcode corresponding to the respective fragment pair in the first bin, and

each sequence read in the respective first subset of sequence reads is from the first bin, and

the different second calculated fragment consists of a respective second subset of sequence reads in the plurality of sequence reads having the barcode corresponding to the respective fragment pair, wherein

each sequence read in the respective second subset of sequence reads is within a predefined genetic distance of another sequence read in the respective second subset of sequence reads,

the different second calculated fragment of the respective fragment pair originates with a second sequence read having the barcode corresponding to the respective fragment pair in the second bin, and

each sequence read in the respective second subset of sequence reads is from the second bin; and

(F) computing a respective likelihood based upon a probability of occurrence of a first model and a probability of occurrence of a second model regarding the one or more fragment pairs to thereby provide a likelihood of a structural variation in the test nucleic acid, wherein

(i) the first model specifies that the respective first calculated fragments and the respective second calculated fragments of the one or more fragment pairs are observed given no structural variation in the target nucleic acid sequence and are part of a common molecule, and

(ii) the second model specifies that the respective first calculated fragments and the respective second calculated fragments of the one or more fragment pairs are observed given structural variation in the target nucleic acid sequence.

[0070] Part C, Phasing Methods. Disclosed are methods for phasing sequencing data of a test nucleic acid sample. In some embodiments the test nucleic acid sample is obtained from a single biological sample from a single organism of a species. In some embodiments the test nucleic acid sample is obtained from a single biological sample but may represent more than a single species. Such a situation arises, for example, when a host has been infected by, for instance, a retrovirus.

[0071] The test nucleic acid sample comprises a first set of haplotypes (H_0) and a second set of haplotypes (H_1). In other words, the biological sample is diploid and inherits maternal and paternal haplotypes. For instance, some portions of the genome of the biological sample are paternally inherited while other portions of the genome are maternally inherited. If the portions that are maternally inherited were to arbitrarily be designated the haplotype H_0 , the portions that are paternally inherited are designated H_1 .

[0072] In typical embodiments, the method occurs at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors of the methods. In the disclosed methods, a reference consensus sequence for all or a portion of a genome of the species is obtained. In some embodiments, the reference consensus sequence is partial or incomplete. In some embodiments, the reference consensus sequence is the sequence of only a single organism of the species. In some embodiments, the reference consensus sequence is the consensus sequence of a plurality of organisms of the species.

[0073] In the method, a plurality of variant calls $A_{i,p}$ is obtained. Here, i is an index to a position in the reference consensus sequence. In some embodiments, $p \in \{0, 1\}$ in which label “0” assigns a respective variant call in $A_{i,p}$ to H_0 and label “1” assigns the respective variant call to H_1 , and n is the number of variants calls in $A_{i,p}$. For example, if n is 5, there are five positions i in $A_{i,p}$ and each such position is independently labeled as either 0 (indicating a first haplotype) or 1 (indicating a second haplotype).

[0074] In some alternative embodiments, $p \in \{0, 1, -1\}$ in which label “0” assigns a respective variant call in $A_{i,p}$ to H_0 , label “1” assigns the respective variant call to H_1 , and label “-1” provides the advantageous possibility of denoting an error condition in the assignment of the variant call to a haplotype. This alternative embodiment takes into consideration that

[0098] In some embodiments, a respective local block of variant calls consists of between 20 and 60 variants in $A_{i,p}$. In some embodiments, a respective local block of variant calls consists of between 30 and 80 variants in $A_{i,p}$. In some embodiments, an iteration of the beam search for the assignments of one of $X_k, X_{k+1}, \dots, X_{k+j}$ discards all but a predetermined number of solutions for $\hat{X}$. In some embodiments, the predetermined number of solutions for $\hat{X}$ is 1000 or less. In some embodiments, predetermined number of solutions for $\hat{X}$ is 5000 or less.

[0099] In some embodiments, the species is human and the test nucleic acid sample comprises the genome of the biological sample. In some embodiments, the species is a multi-chromosomal species and the test nucleic acid sample comprises a plurality of nucleic acids collectively representing a plurality of chromosomes in the multi-chromosomal species.

[00100] In some embodiments, the barcode in the second portion of each respective sequence read in the plurality of sequence reads O encodes a unique predetermined value selected from the set $\{1, \dots, 1024\}$, selected from the set $\{1, \dots, 4096\}$, selected from the set $\{1, \dots, 16384\}$, selected from the set $\{1, \dots, 65536\}$, selected from the set $\{1, \dots, 262144\}$, selected from the set $\{1, \dots, 1048576\}$, selected from the set $\{1, \dots, 4194304\}$, selected from the set $\{1, \dots, 16777216\}$, selected from the set $\{1, \dots, 67108864\}$, or selected from the set $\{1, \dots, 1 \times 10^{12}\}$. In some embodiments, the barcode in the second portion of a respective sequence read in the plurality of sequence reads is localized to a contiguous set of oligonucleotides within the respective sequence read. In some embodiments, the contiguous set of oligonucleotides is an N-mer, where N is an integer selected from the set $\{4, \dots, 20\}$. In some embodiments, the barcode in the second portion of a sequence read in the plurality of sequence reads is localized to a predetermined noncontiguous set of nucleotides within the sequence read. In some embodiments, the predetermined noncontiguous set of nucleotides collectively consists of N nucleotides, wherein N is an integer in the set $\{4, \dots, 20\}$.

[00101] In some embodiments, the subset of sequence reads in the plurality of sequence reads having the same barcode corresponds to a portion of the reference consensus sequence that is greater than 10 kilobases. In some embodiments, the subset of sequence reads having the same barcode corresponds to a portion of the reference consensus sequence that is greater than 20 kilobases.

[00102] Disclosed is a computing system, comprising one or more processors and memory storing one or more programs to be executed

by the one or more processors. The one or more programs comprise instructions for phasing sequencing data of a test nucleic acid sample obtained from a biological sample from a single organism of a species. The test nucleic acid sample comprises a first set of haplotypes ($H = 0$) and a second set of haplotypes ($H = 1$). The one or more programs execute the phasing methods disclosed in the present disclosure.

[00103] Disclosed is a method of phasing sequencing data of a test nucleic acid sample obtained from a biological sample from a single organism of a species. The test nucleic acid sample comprises a first set of haplotypes ($H = 0$) and a second set of haplotypes ($H = 1$). The method comprises, at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors, obtaining a plurality of variant calls $A_{i,p}$, where i is an index to a position in a reference consensus sequence for all or a portion of a genome of the species, and $p \in \{0, 1\}$ in which label 0 assigns a respective variant call in $A_{i,p}$ to $H = 0$ and label 1 assigns the respective variant call to $H = 1$. For each respective local block of variant calls in $A_{i,p}$ that are localized to a corresponding subset of the reference consensus sequence, a beam search or equivalent search technique is used over the haplotype assignments of local phasing vectors $X_k, X_{k+1}, \dots, X_{k+j}$ in the respective local block of variant calls, where k is the first variant in the respective local block of variant calls, j is a number of variant calls in the respective local block of variant calls, assignments of $X_k, X_{k+1}, \dots, X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k, X_{k+1}, \dots, X_{k+j}$, and the objective function is calculated by matching observed sequence reads of the test nucleic acid sample against the respective local block of variant calls in $A_{i,p}$, thereby finding a phasing solution for each respective local block of variant calls in $A_{i,p}$. Upon completion of the beam search for each respective local block of variant calls in $A_{i,p}$, neighboring local blocks of variant calls in $A_{i,p}$ are greedily joined using the phasing solution for each respective local block of variant calls thereby obtaining a phasing configuration $\hat{\mathbf{X}}$ for the single organism of the species. In some embodiments, the method further comprises iteratively swapping the phase result of individual x_i in $\hat{\mathbf{X}}$ and recomputing the objective function, thereby obtaining $\hat{\mathbf{X}}$. In some embodiments, a respective local block of variant calls consists of between 20 and 60 variants in $A_{i,p}$. In some embodiments, a respective local block of variant calls consists of between 30 and 80 variants in $A_{i,p}$. In some embodiments, an iteration of the beam search for the assignments of one of $X_k, X_{k+1}, \dots, X_{k+j}$ discards all but a predetermined number of solutions for $\hat{\mathbf{X}}$ (e.g., 1000 or less, 5000 or less, etc.).

[00104] Disclosed is a method of addressing error in the zygoty of variant calls in phasing sequencing data of a test nucleic acid sample obtained from a biological sample from a single organism of a species. The test nucleic acid sample comprises a first set of haplotypes (H_0) and a second set of haplotypes (H_1). The method comprises, at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors, obtaining a reference consensus sequence for all or a portion of a genome of the species and obtaining a plurality of variant calls $A_{i,p}$ for the biological sample. Here, i is an index to a position in the reference consensus sequence and $p \in \{0, 1, -1\}$ in which label 0 assigns a respective variant call in $A_{i,p}$ to H_0 , label 1 assigns the respective variant call to H_1 , and label -1 assigns the respective variant call to the zygoty error condition H_{-1} . In the method, a plurality of sequence reads $\vec{O}$ for the biological sample is obtained. Each respective sequence read $\vec{O}_i$ in the plurality of sequence reads comprises a first portion that corresponds to a subset of the reference sequence and a second portion that encodes a respective barcode, independent of the reference sequence, for the respective sequence read, in a plurality of barcodes. Each respective sequence read $\vec{O}_i$ in the plurality of sequence reads is $\in \{0, 1, -1, -\}^n$. Here, (i) n is the number of variants calls in $A_{i,p}$, (ii) each respective label 0 for the respective sequence read $\vec{O}_i$ assigns a corresponding variant call in $A_{i,p}$ to H_0 , (iii) each respective label 1 for the respective sequence read $\vec{O}_i$ assigns a corresponding variant call in $A_{i,p}$ to H_1 , (iv) each respective label -1 for the respective sequence read $\vec{O}_i$ assigns a corresponding variant call in $A_{i,p}$ to H_{-1} , and (v) each respective label - for the respective sequence read $\vec{O}_i$ indicates that the corresponding variant call in $A_{i,p}$ is not covered. In the method, a phasing vector result $\vec{X}$ is refined by optimization of haplotype assignments at individual positions i in $A_{i,p}$ between H_0 , H_1 and H_{-1} for the plurality of sequence reads using an overall objective function:

$$[00105] \quad \hat{\vec{X}} = \underset{\vec{X}}{\operatorname{argmax}} P(\vec{X} | \vec{O}) = \frac{P(\vec{X})P(\vec{O} | \vec{X})}{c}$$

[00106] where

$$[00107] \quad P(\vec{X}) = \prod_i \frac{(1-\varepsilon_i)}{2} (X_i = H_1) + (X_i = H_0) + \varepsilon_i (X_i = H_{-1}),$$

[00108] ε_i is an estimate of incurring H_{-1} at position i , and

$$[00109] \quad P(\vec{O} | \vec{X}) = \prod_f P(O_{1,f}, \dots, O_{N,f}),$$

[00110] $\hat{\vec{X}}$ is the refined phasing vector result,

What is claimed is:

1. A method of determining a likelihood of a structural variation occurring in a test nucleic acid obtained from a single biological sample, the method comprising:

at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors:

(A) obtaining a plurality of sequence reads from a plurality of sequencing reactions in which the test nucleic acid is fragmented, wherein

each respective sequence read in the plurality of sequence reads comprises a first portion that corresponds to a subset of the test nucleic acid and a second portion that encodes a respective barcode for the respective sequence read in a plurality of barcodes, and

each respective barcode is independent of the sequencing data of the test nucleic acid, and

the plurality of sequence reads collectively include the plurality of barcodes;

(B) obtaining bin information for a plurality of bins, wherein

each respective bin in the plurality of bins represents a different portion of the test nucleic acid,

the bin information identifies, for each respective bin in the plurality of bins, a set of sequence reads in a plurality of sets of sequence reads that are in the plurality of sequence reads, and

the respective first portion of each respective sequence read in each respective set of sequence reads in the plurality of sets of sequence reads corresponds to a subset of the test nucleic acid that at least partially overlaps the different portion of the test nucleic acid that is represented by the bin corresponding to the respective set of sequence reads;

(C) identifying, from among the plurality of bins, a first bin and a second bin that correspond to portions of the test nucleic acid that are nonoverlapping, wherein the first bin is represented by a first set of sequence reads in the plurality of sequence reads and the second bin is represented by a second set of sequence reads in the plurality of sequence reads;

(D) determining a first value that represents a numeric probability or likelihood that the number of barcodes common to the first set and the second set is attributable to chance;

(E) responsive to a determination that the first value satisfies a predetermined cut-off value, for each barcode that is common to the first bin and the second bin, obtaining a

fragment pair thereby obtaining one or more fragment pairs, each fragment pair in the one or more fragment pairs (i) corresponding to a different barcode that is common to the first bin and the second bin and (ii) consisting of a different first calculated fragment and a different second calculated fragment, wherein, for each respective fragment pair in the one or more fragment pairs:

the different first calculated fragment consists of a respective first subset of sequence reads in the plurality of sequence reads having the barcode corresponding to the respective fragment pair, wherein

each sequence read in the respective first subset of sequence reads is within a predefined genetic distance of another sequence read in the respective first subset of sequence reads,

the different first calculated fragment of the respective fragment pair originates with a first sequence read having the barcode corresponding to the respective fragment pair in the first bin, and

each sequence read in the respective first subset of sequence reads is from the first bin, and

the different second calculated fragment consists of a respective second subset of sequence reads in the plurality of sequence reads having the barcode corresponding to the respective fragment pair, wherein

each sequence read in the respective second subset of sequence reads is within a predefined genetic distance of another sequence read in the respective second subset of sequence reads,

the different second calculated fragment of the respective fragment pair originates with a second sequence read having the barcode corresponding to the respective fragment pair in the second bin, and

each sequence read in the respective second subset of sequence reads is from the second bin; and

(F) computing a respective likelihood based upon a probability of occurrence of a first model and a probability of occurrence of a second model regarding the one or more fragment pairs to thereby provide a likelihood of a structural variation in the test nucleic acid, wherein

(i) the first model specifies that the respective first calculated fragments and the respective second calculated fragments of the one or more fragment pairs are observed

given no structural variation in the target nucleic acid sequence and are part of a common molecule, and

(ii) the second model specifies that the respective first calculated fragments and the respective second calculated fragments of the one or more fragment pairs are observed given structural variation in the target nucleic acid sequence.

2. The method of claim 1, wherein the computed likelihood term in the computing (F) is a ratio score between the probability of occurrence of the first model and the probability of occurrence of the second model.

3. The method of claim 1 or 2, wherein the first bin and the second bin are at least a predetermined number of kilobases apart on the test nucleic acid.

4. The method of claim 3, wherein the first bin and the second bin are at least 50 kilobases apart on the test nucleic acid.

5. The method of any one of claims 1-4, wherein the determining (D) uses a binomial test to compute the first value.

6. The method of any one of claims 1-4, wherein the determining (D) uses a binomial test to compute the first value of the form:

$$p = 1 - P_{Binom}(n, n_1 n_2 / B)$$

wherein,

p is the first value, expressed as a p-value,

n is the number of unique barcodes that is found in both in the first and second set of sequence reads,

n_1 is the number of unique barcodes in the first set of sequence reads,

n_2 is the number of unique barcodes in the second set of sequence reads, and

B is the total number of unique barcodes across the plurality of bins.

7. The method of any one of claims 1-6, wherein the single biological sample is human, the test nucleic acid is the genome of the biological sample, and the first value satisfies the predetermined cut-off value when the first value is 10^{-14} or less.

8. The method of any one of claims 1-6, wherein the single biological sample is human, the test nucleic acid is the genome of the biological sample, and the first value satisfies the predetermined cut-off value when the first value is 10^{-15} or less.
9. The method of any one of claims 1-8, wherein the structural variation is an insertion or deletion of 50 consecutive bases or more into the different portion of the test nucleic acid that is represented by the first set of sequence reads.
10. The method of any one of claims 1-8, wherein the structural variation is an insertion or deletion of 500 consecutive bases or more into the different portion of the test nucleic acid that is represented by the first set of sequence reads.
11. The method of any one of claims 1-8, wherein the structural variation is an insertion or deletion of 5000 consecutive bases or more into the different portion of the test nucleic acid that is represented by the first set of sequence reads.
12. The method of any one of claims 1-11, wherein the structural variation is associated with a genetic disease.
13. The method of any one of claims 1-12, wherein each bin in the plurality of bins represents at least 20 kilobases of the test nucleic acid, at least 50 kilobases of the test nucleic acid, at least 100 kilobases of the test nucleic acid, at least 250 kilobases of the test nucleic acid, or at least 500 kilobases of the test nucleic acid.
14. The method of any one of claims 1-12, wherein each respective sequence read in each respective set of sequence reads in the plurality of sequence reads has a respective first portion that corresponds to a subset of the test nucleic acid that fully overlaps the different portion of the test nucleic acid that is represented by the bin corresponding to the respective set of sequence reads.
15. The method of any one of claims 1-14, wherein the plurality of bins comprises 10,000 or more bins.
16. The method of any one of claims 1-14, wherein the plurality of bins comprises 100,000 or more bins.

17. The method of any one of claims 1-14, wherein the plurality of bins comprises 1,000,000 or more bins.
18. The method of any one of claims 1-17, wherein the biological sample is from a multi-chromosomal species and the test nucleic acid comprises a plurality of nucleic acids collectively representing a plurality of chromosomes in the multi-chromosomal species.
19. The method of any one of claims 1-18, wherein the barcode in the second portion of each respective sequence read in the plurality of sequence reads encodes a unique predetermined value selected from the set {1, ..., 1024}, selected from the set {1, ..., 4096}, selected from the set {1, ..., 16384}, selected from the set {1, ..., 65536}, selected from the set {1, ..., 262144}, selected from the set {1, ..., 1048576}, selected from the set {1, ..., 4194304}, selected from the set {1, ..., 16777216}, selected from the set {1, ..., 67108864}, or selected from the set {1, ..., 1×10^{12} }.
20. The method of any one of claims 1-19, wherein the barcode in the second portion of a respective sequence read in the plurality of sequence reads is localized to a contiguous set of oligonucleotides within the respective sequence read.
21. The method of claim 20, wherein the contiguous set of oligonucleotides is an N-mer, wherein N is an integer selected from the set {4, ..., 20}.
22. The method of any one of claims 1-10, wherein the barcode in the second portion of a sequence read in the plurality of sequence reads is localized to a predetermined noncontiguous set of nucleotides within the sequence read.
23. The method of claim 22, wherein the predetermined noncontiguous set of nucleotides collectively consists of N nucleotides, wherein N is an integer in the set {4, ..., 20}.
24. The method of any one of claims 1-23, wherein the first sequence read corresponds to a first subset of the test nucleic acid that is greater than 10 kilobases.
25. The method of any one of claims 1-23, wherein the first sequence read corresponds to a first subset of the test nucleic acid that is greater than 20 kilobases.
26. The method of any one of claims 1-25, wherein an identity of the first and second bin is determined by the identifying (C) using sparse matrix multiplication.

27. The method of any one of claims 1-25, wherein an identity of the first and second bin is determined by the identifying (C) using sparse matrix multiplication of the form:

$$V = A_1^T A_2,$$

wherein,

A_1 is a first $B \times N_1$ matrix of barcodes that includes the first bin,

A_2 is a second $B \times N_2$ matrix of barcodes that includes the second bin,

B is the number of unique barcodes across the plurality of bins,

N_1 is the number of bins in A_1 ,

N_2 is the number of bins in A_2 , and

A_1^T is the transpose of matrix A_1 .

28. The method of claim 27, wherein

the first bin is associated with a first chromosome of the biological sample,

the second bin is associated with a second chromosome of the biological sample,

N_1 is the number of bins associated with the first chromosome, and

N_2 is the number of bins associated with the second chromosome.

29. The method of claim 27, wherein

the first and second bin are both associated with a first chromosome of the biological sample,

N_1 is the number of bins associated with the first chromosome, and

N_2 equals N_1 .

30. The method of any one of claims 1-29, the method further comprising:

maintaining a blackout list, the blackout list comprising a plurality of blackout regions of the test nucleic acid, the method further comprising prior to the identifying (C):

eliminating a sequence read from the plurality of sequence reads when the first portion of the sequence read overlaps a blackout region in the plurality of blackout regions.

31. The method of claim 1, wherein the computed likelihood in the computing (F) is computed as:

$$LR = \frac{P(\text{observed fragments} \mid SV)}{P(\text{observed fragments} \mid \text{no SV})}$$

wherein, LR is equal to a product of a plurality of terms, wherein each term in the plurality of terms (i) represents a respective fragment pair in the one or more fragment pairs and (ii) has the form:

$$= \frac{P(r_1, r_2, l_1, l_2, d \mid SV; a_b)}{P(r_1, r_2, l_1, l_2, d \mid no\ SV; a_b)}$$

r_1 is a number of sequence reads in the respective first subset of sequence reads in the first calculated fragment for the respective fragment pair,

l_1 is a length of the first calculated fragment as determined by the first subset of sequence reads of the respective fragment pair,

r_2 is a number of reads in the respective second subset of sequence reads in the second calculated fragment for the respective fragment pair,

l_2 is a length of the second calculated fragment as determined by the second subset of sequence reads of the respective fragment pair,

d is a distance between the first calculated fragment and the second calculated fragment of the respective fragment pair in the test nucleic acid,

a_b is a read rate of the first barcode across the plurality of sequence reads,

SV indicates the first calculated fragment and the second calculated fragment are observed in accordance with the first model, and

$no\ SV$ indicates the first calculated fragment and the second calculated fragment are observed in accordance with the second model.

32. The method of claim 31, wherein

$$\begin{aligned} P(r_1, r_2, l_1, l_2, d \mid no\ SV; a_b) = \\ P(r_1, r_2, l_1, l_2, d \mid SM, no\ SV; a_b)P(SM \mid no\ SV) + \\ P(r_1, r_2, l_1, l_2, d \mid DM, no\ SV; a_b)P(DM \mid no\ SV), \end{aligned}$$

wherein,

SM is the hypothesis that the first calculated molecule and the second calculated molecule originated from the same fragment of the test nucleic acid in the plurality of sequencing reactions,

DM is the hypothesis that the first calculated molecule and the second calculated molecule originated from different fragments of the test nucleic acid in the plurality of sequencing reactions,

$$P(r_1, r_2, l_1, l_2, d \mid DM, no\ SV; a_b) = P_{frag}(r_1, l_1; a_b)P_{frag}(r_2, l_2; a_b), \text{ wherein}$$

$P_{frag}(r_1, l_1; a_b)$ is the probability of observing r_1 reads from a first molecule of unknown length such that the reads span an observed length of l_1 , and

$P_{frag}(r_2, l_2; a_b)$ is the probability of observing r_2 reads from a second molecule of unknown length such that the reads span an observed length of l_2 .

33. The method of claim 32, wherein $P_{frag}(r_1, l_1; a_b)$ and $P_{frag}(r_2, l_2; a_b)$ are each computed as

$$\begin{aligned} & \sum_{m:m \geq l} \left(r(r-1) \left(\frac{l}{m} \right)^{r-2} \frac{m-l}{m^2} \right) P_p(r; ma_b) P_L(m) \\ &= \sum_{m:m \geq l} (m-l) P_p(r-2; a_b l) P_p(0; a_b(m-l)) a_b^2 P_L(m) \end{aligned}$$

where,

$P_p(r; b)$ is the probability mass function of a Poisson distribution with parameter b , and

$P_L(m)$ is the (pre-estimated) probability that the true molecule length of the respective molecule is m .

34. The method of claim 32, wherein

$P(r_1, r_2, l_1, l_2, d \mid SM, \text{no } SV; a_b)$ is computed as

$$\begin{aligned} & \sum_{m:m \geq l_1+l_2+d} (m-l_1-l_2-d) P_p(r_1-2; a_b l_1) P_p(r_2 \\ & \quad -2; a_b l_2) P_p(0; a_b(m-l_1-l_2)) a_b^4 P_L(m) \end{aligned}$$

wherein,

m is the length of the true molecule length,

$P_p(r_1-2; a_b l_1)$ is a probability mass function of a Poisson distribution with parameter b for r_1 ,

$P_p(r_2-2; a_b l_2)$ is a probability mass function of a Poisson distribution with parameter b for r_2 ,

$P_p(0; a_b(m-l_1-l_2))$ is a probability mass function of a Poisson distribution with parameter b , and

$P_L(m)$ is a pre-estimated probability that the true common molecule length is m .

35. The method of claim 31, wherein

$$P(r_1, r_2, l_1, l_2, d \mid SV; a_b) = \\ P(r_1, r_2, l_1, l_2, 2d' \mid SM, \text{no } SV; a_b)P(SM \mid \text{no } SV) + \\ P(r_1, r_2, l_1, l_2, 2d' \mid DM, \text{no } SV; a_b)P(DM \mid \text{no } SV),$$

wherein,

SM is the hypothesis that the first calculated molecule and the second calculated molecule originated from the same fragment of the test nucleic acid in the plurality of sequencing reactions,

DM is the hypothesis that the first calculated molecule and the second calculated molecule originated from different fragments of the test nucleic acid in the plurality of sequencing reactions,

$$P(r_1, r_2, l_1, l_2, 2d' \mid DM, SV; a_b) = P_{frag}(r_1, l_1; a_b)P_{frag}(r_2, l_2; a_b), \text{ wherein}$$

$P_{frag}(r_1, l_1; a_b)$ is the probability of observing r_1 reads from a first molecule of unknown length such that the reads span an observed length of l_1 ,

$P_{frag}(r_2, l_2; a_b)$ is the probability of observing r_2 reads from a second molecule of unknown length such that the reads span an observed length of l_2 , and

$2d'$ is a distance between the first calculated fragment and the second calculated fragment of the respective fragment pair in the test nucleic acid taking into account an estimate of the breakpoints of a structural variation associated with the first calculated molecule and the second calculated molecule.

36. The method of claim 35, wherein $P_{frag}(r_1, l_1; a_b)$ and $P_{frag}(r_2, l_2; a_b)$ are each computed as

$$\sum_{m: m \geq l} \left(r(r-1) \left(\frac{l}{m} \right)^{r-2} \frac{m-l}{m^2} \right) P_p(r; ma_b) P_L(m) \\ = \sum_{m: m \geq l} (m-l) P_p(r-2; a_b l) P_p(0; a_b(m-l)) a_b^2 P_L(m)$$

where,

$P_p(r; b)$ is the probability mass function of a Poisson distribution with parameter b , and

$P_L(m)$ is the (pre-estimated) probability that the true molecule length of the respective molecule is m .

37. The method of claim 35, wherein

$P(r_1, r_2, l_1, l_2, 2d' \mid SM, SV; a_b)$ is computed as

$$\sum_{m:m \geq l_1 + l_2 + d} (m - l_1 - l_2 - 2d') P_p(r_1 - 2; a_b l_1) P_p(r_2 - 2; a_b l_2) P_p(0; a_b(m - l_1 - l_2)) a_b^4 P_L(m)$$

wherein,

m is the length of the true molecule length,

$P_p(r_1 - 2; a_b l_1)$ is a probability mass function of a Poisson distribution with parameter b for r_1 ,

$P_p(r_2 - 2; a_b l_2)$ is a probability mass function of a Poisson distribution with parameter b for r_2 ,

$P_p(0; a_b(m - l_1 - l_2))$ is a probability mass function of a Poisson distribution with parameter b , and

$P_L(m)$ is a pre-estimated probability that the true common molecule length is m .

38. The method of claim 35, wherein $2d'$ is estimated by computing the maximum extent d' such that $P_p(0; a_b d') \geq 0.75$.

39. The method of any one of claims 1-38, wherein the plurality of sequence reads represents whole genome sequencing data.

40. The method of any one of claims 1-38, wherein the plurality of sequence reads represents targeted sequencing of a subset of a genome,

a first subset of the plurality of sequence reads sequence reads are from inside the subset of the genome and have a first read rate of a_b ,

a second subset of the plurality of sequence reads sequence reads are from outside the subset of the genome and have a first read rate of $\tilde{a}_b$, wherein $\tilde{a}_b$ is different than a_b , and wherein the likelihood in the computing (F) corrects for the different read rates of the respective first and second subset of the plurality of sequence reads.

41. The method of any one of claims 1-40, wherein the (D) through (F) are computed for a plurality of first and second bins, thereby calling one or more structural variations in the test nucleic acid, the method further comprising refining a breakpoint location in the test nucleic acid using the plurality of sequence reads and the calling one or more structural variations.

42. The method of any one of claims 1-41, wherein the plurality of barcodes comprises 1000 or more barcodes.

43. The method of any one of claims 1-41, wherein the plurality of barcodes comprises 10,000 or more barcodes.

44. The method of any one of claims 1-41, wherein the plurality of barcodes comprises 100,000 or more barcodes.

45. The method of any one of claims 1-41, wherein the plurality of barcodes comprises 1×10^6 or more barcodes.

46. A computing system, comprising:

one or more processors;

memory storing one or more programs to be executed by the one or more processors;

the one or more programs comprising instructions for:

(A) obtaining a plurality of sequence reads from a plurality of sequencing reactions in which the test nucleic acid is fragmented, wherein

each respective sequence read in the plurality of sequence reads comprises a first portion that corresponds to a subset of the test nucleic acid and a second portion that encodes a respective barcode for the respective sequence read in a plurality of barcodes,

each respective barcode is independent of the sequencing data of the test nucleic acid, and

the plurality of sequence reads collectively include the plurality of barcodes;

(B) obtaining bin information for a plurality of bins, wherein

each respective bin in the plurality of bins represents a different portion of the test nucleic acid,

the bin information identifies, for each respective bin in the plurality of bins, a set of sequence reads in a plurality of sets of sequence reads that are in the plurality of sequence reads, and

the respective first portion of each respective sequence read in each respective set of sequence reads in the plurality of sets of sequence reads corresponds to a

subset of the test nucleic acid that at least partially overlaps the different portion of the test nucleic acid that is represented by the bin corresponding to the respective set of sequence reads;

(C) identifying, from among the plurality of bins, a first bin and a second bin that correspond to portions of the test nucleic acid that are nonoverlapping, wherein the first bin is represented by a first set of sequence reads in the plurality of sequence reads and the second bin is represented by a second set of sequence reads in the plurality of sequence reads;

(D) determining a first value that represents a numeric probability or likelihood that the number of barcodes common to the first set and the second set is attributable to chance;

(E) responsive to a determination that the first value satisfies a predetermined cut-off value, for each barcode that is common to the first bin and the second bin, obtaining a fragment pair thereby obtaining one or more fragment pairs, each fragment pair in the one or more fragment pairs (i) corresponding to a different barcode that is common to the first bin and the second bin and (ii) consisting of a different first calculated fragment and a different second calculated fragment, wherein, for each respective fragment pair in the one or more fragment pairs:

the different first calculated fragment consists of a respective first subset of sequence reads in the plurality of sequence reads having the barcode corresponding to the respective fragment pair, wherein

each sequence read in the respective first subset of sequence reads is within a predefined genetic distance of another sequence read in the respective first subset of sequence reads,

the different first calculated fragment of the respective fragment pair originates with a first sequence read having the barcode corresponding to the respective fragment pair in the first bin, and

each sequence read in the respective first subset of sequence reads is from the first bin, and

the different second calculated fragment consists of a respective second subset of sequence reads in the plurality of sequence reads having the barcode corresponding to the respective fragment pair, wherein

each sequence read in the respective second subset of sequence reads is within a predefined genetic distance of another sequence read in the respective second subset of sequence reads,

the different second calculated fragment of the respective fragment pair originates with a second sequence read having the barcode corresponding to the respective fragment pair in the second bin, and

each sequence read in the respective second subset of sequence reads is from the second bin; and

(F) computing a respective likelihood based upon a probability of occurrence of a first model and a probability of occurrence of a second model regarding the one or more fragment pairs to thereby provide a likelihood of a structural variation in the test nucleic acid, wherein

(i) the first model specifies that the respective first calculated fragments and the respective second calculated fragments of the one or more fragment pairs are observed given no structural variation in the target nucleic acid sequence and are part of a common molecule, and

(ii) the second model specifies that the respective first calculated fragments and the respective second calculated fragments of the one or more fragment pairs are observed given structural variation in the target nucleic acid sequence.

47. The computing system of claim 46, wherein the plurality of barcodes comprises 1000 or more barcodes.

48. The computing system of claim 46, wherein the plurality of barcodes comprises 10,000 or more barcodes.

49. The computing system of claim 46, wherein the plurality of barcodes comprises 100,000 or more barcodes.

50. The computing system of claim 46, wherein the plurality of barcodes comprises 1×10^6 or more barcodes.