SYSTEMS AND METHODS FOR DETERMINING STRUCTURAL VARIATION AND PHASING USING VARIANT CALL DATA

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to United States Provisional Patent Application 62/238,077, entitled "Systems and Methods for Determining Structural Variation Using Probabilistic Models," filed October 6, 2015, which is hereby incorporated by reference in its entirety.

[0002] This application also claims priority to United States Provisional Patent Application 62/113,693, entitled "Systems and Methods for Determining Structural Variation," filed February 9, 2015, which is hereby incorporated by reference in its entirety.

[0003] This application also claims priority to United States Provisional Patent Application 62/120,247, entitled "Systems and Methods for Implementing Linked Read Algorithms for Haplotype Phasing and Structural Variant Detection," filed February 24, 2015, which is hereby incorporated by reference in its entirety.

[0004] This application also claims priority to United States Provisional Patent Application 62/120,330, entitled "Detecting Structural Variants and Phasing Haplotypes from Cancer Exome Sequencing Using 1ng Dna Input," filed February 24, 2015, which is hereby incorporated by reference in its entirety.

TECHNICAL FIELD

[0005] This specification describes technologies relating to haplotype phasing and structural variant detection using nucleic acid sequencing data.

BACKGROUND

[0006] Haplotype assembly from experimental data obtained from human genomes sequenced using massively parallelized sequencing methodologies has emerged as a prominent source of genetic data. Such data serves as a cost-effective way of implementing genetics based diagnostics as well as human disease study, detection, and personalized treatment.

[0007] The long-range information provided by platforms such as those disclosed in United States Patent Application No. 62/072,214, filed October 29, 2014, entitled "Analysis of Nucleic Acid Sequences" greatly facilitates the detection of large-scale structural variations of the genome, such as translocations, large deletions, or gene fusions. Other examples include, but are not limited to the sequencing-by-synthesis platform (ILLUMINA), Bentley*et al.*, 2008, "Accurate whole human genome sequencing using reversible terminator chemistry, Nature 456:53-59; sequencing-by-litigation platforms (POLONATOR; ABI SOLiD), Shendure *et al.*, 2005, "Accurate Multiplex Polony Sequencing of an Evolved bacterial Genome" Science 309:1728-1732; pyrosequencing platforms (ROCHE 454), Margulies *et al.*, 2005, "Genome sequencing in microfabricated high-density picoliter reactors," Nature 437:376-380; and single-molecule sequencing platforms (HELICOS HELISCAPE); Pushkarev *et al.*, 2009, "Single-molecule sequencing of an individual human genome," Nature Biotech 17:847-850, (PACIFIC BIOSCIENCES) Eid *et al.*, "Real-time sequencing form single polymerase molecules," Science 323:133-138, each of which is hereby incorporated by reference in its entirety.

[0008] Several algorithms have been developed for detecting such events from whole genome sequencing (WGS) data. See, for example, Chen *et al.*, 2009, "BreakDancer: an algorithm for high-resolution mapping of genomic structural variation," Nature Methods 6(9), pp, 677-681 and Layer *et al.*, 2014, "LUMPY: A probabilistic framework for structural variant discovery," Genome Biology 15(6):R84. The goal of these algorithms is to detect the endpoints of structural variants (e.g., the endpoints of a deletion or a gene fusion). These endpoints are also referred to as "breakpoints" and the terms endpoints and breakpoints are used interchangeably. In order to detect breakpoints, existing algorithms rely on the detection of read pairs that are mapped to the genome at unexpected orientations with respect to each other or at unexpected distances (too far from each other or too close to each other relative to the insert size). This implies that, in order for the breakpoint to be detected by conventional algorithms, it must be spanned by read pairs. This limitation makes existing algorithms not applicable to targeted sequencing data, such as whole exome sequencing (WES) data. This is because the breakpoints would be spanned by read pairs only if they were very close to the target regions. This is usually not the case. For example many gene fusions in cancer happen on gene introns rather than exons, so they would not be detectable with WES.

[0009] The availability of haplotype data spanning large portions of the human genome, the need has arisen for ways in which to efficiently work with this data in order to

advance the above stated objectives of diagnosis, discovery, and treatment, particularly as the cost of whole genome sequencing for a personal genome drops below $1000. To computationally assemble haplotypes from such data, it is necessary to disentangle the reads from the two haplotypes present in the sample and infer a consensus sequence for both haplotypes. Such a problem has been shown to be NP-hard. See Lippert *et al.*, 2002, "Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem," Brief. Bionform 3:23-31, which is hereby incorporated by reference.

[0010] Given the above background, what is needed in the art are improved systems and methods for haplotype phasing and structural variant detection using sequencing data from parallelized sequencing methodologies.

SUMMARY

[0011] Technical solutions (e.g., computing systems, methods, and non-transitory computer readable storage mediums) for identifying structural variations and for haplotype phasing are provided. With platforms such as those disclosed in United States Provisional Patent Application No. 62/072,214, filed October 29, 2014, entitled "Analysis of Nucleic Acid Sequences," or United States Provisional Patent Application 62/113,693, entitled "Systems and Methods for Determining Structural Variation," filed February 9, 2015, each of which is hereby incorporated by reference, the genome is fragmented and partitioned and barcoded prior to the target identification. Therefore the integrity of the barcode information is maintained across the genome. The barcode information is used to identify potential structural variation breakpoints by detecting regions of the genome that show significant barcode overlap. They are also used to obtain phasing information.

[0012] The following presents a summary of the invention in order to provide a basic understanding of some of the aspects of the invention. This summary is not an extensive overview of the invention. It is not intended to identify key/critical elements of the invention or to delineate the scope of the invention. Its sole purpose is to present some of the concepts of the invention in a simplified form as a prelude to the more detailed description that is presented later.

[0013] Various embodiments of systems, methods and devices within the scope of the appended claims each have several aspects, no single one of which is solely responsible for the desirable attributes described herein. Without limiting the scope of the appended claims, some

prominent features are described herein. After considering this discussion, and particularly after reading the section entitled "Detailed Description" one will understand how the features of various embodiments are used.

[0014] **Part A, Structural Variation.** In some implementations, there is provided a method of detecting a structural variant in sequencing data of a test nucleic acid obtained from a biological sample. The method comprises performing certain operations 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 plurality of sequence reads is obtained. 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. As used herein, the terms "sequence read" and "sequencing read" are used interchangeably. The barcode is independent of the sequencing data of the test nucleic acid. In some embodiments, a first sequence read in the plurality of sequence reads is from a subset of the test nucleic acid that is greater than 10 kilobase pairs (kbp), 20 kbp, 30 kbp, 40 kbp, 50 kbp, 60 kbp, 70 kbp, 80 kbp, 90 kbp, or 100 kbp. In some embodiments, a first sequence read in the plurality of sequence reads is 2x36 bp, 2x50 bp, 2x76 bp, 2x100 bp, 2x150 bp or 2x250 bp, where the terminology 2 x N bp means that the sequence read has two reads of length N base pairs from a single piece of nucleic acid (e.g., from a text nucleic acid obtained from a biological sample) that are separated by an unspecified length. In some embodiments this unspecified length is between 200 to 1200 base pairs. In some embodiments, a first sequence read in the plurality of sequence reads represents at least 25 bp, at least 30 bp, at least 50 bp, at least 100 bp, at least 200 bp, at least 250 bp, at least 500 bp, less than 500 bp, less than 400 bp, or less than 300 bp of a single piece of nucleic acid (e.g., from a text nucleic acid obtained from a biological sample).

[0015] Bin information for a plurality of bins is also 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.

[0016] A determination is made as to 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. From this, a probability or likelihood that this number is attributable to chance is determined by comparison of (a) a metric based upon the determined number to (b) 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. In some embodiments, this metric is computed as:

$$p = \prod_{i}^{n} f_{b_i}$$

in which $\{b_1, b_2, \ldots, b_n\}$ is the set of n unique barcodes that is found in both the first and second set of sequence reads, i is an integer index to n, and f_{b_i} is the fraction of the plurality of bins in which the first portion of sequence read b_i appears. In some such embodiments, the metric is deemed to satisfy the threshold criterion when p is 10^{-2} or less, 10^{-3} or less, 10^{-4} or less, 10^{-5} or less, 10^{-6} or less, or 10^{-7} or less.

[0017] In some embodiments, the structural variation is an insertion or deletion of 50 consecutive bases or more, 500 consecutive bases or more, 5000 consecutive bases or more, or 10000 consecutive bases or more into the different portion of the test nucleic acid that is represented by the first set of sequence reads. In some embodiments, the structural variation is a single nucleotide polymorphism.

[0018] In some embodiments, the metric is deemed to satisfy the threshold criterion, and the method further comprises aligning each respective sequence read in the number of respective sequence reads to (i) the subset of the test nucleic acid corresponding to the first set of sequence reads and (ii) the subset of the test nucleic acid corresponding to the second set of sequence reads. A first alignment quality is determined for each respective sequence read in the number of respective sequence reads against the subset of the test nucleic acid corresponding to the first set of sequence reads based on such aligning. Then, a second alignment quality is determined for each respective sequence read in the number of respective sequence reads against the subset of the test nucleic acid corresponding to the second set of sequence reads based on the aligning. Each sequence read having a first alignment quality and

a second alignment quality that are similar is eliminated from the number of respective sequence reads. With this new number of respective sequence reads, the operation of determining a probability or likelihood that the number is attributable to chance by is recomputed as a comparison of (A) the metric based upon the newly reduced number to (B) a threshold criterion. When the recomputed 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.

[0019] In some embodiments, a blackout list is maintained. This blackout list comprises a plurality of blackout regions of the test nucleic acid. In such embodiments, a sequence read is eliminated from the number of respective sequence reads used to evaluate against the threshold criterion when the first portion of the sequence read overlaps a blackout region in the plurality of blackout regions.

[0020] In some embodiments, each bin in the plurality of bins represents at least 20 kbp, at least 50 kbp, at least 100 kbp, at least 250 kbp, or at least 500 kbp.

[0021] In some embodiments, the different portion of the test nucleic acid represented by the first bin overlaps the different portion of the test nucleic acid represented by the second bin. In some such embodiments, at least 50 percent, at least 80 percent, or at least 95 percent of the different portion of the test nucleic acid represented by the first bin overlaps the different portion of the test nucleic acid represented by the second bin.

[0022] In some embodiments there is no overlap between each different portion of the test nucleic acid represented by each respective bin in the plurality of bins.

[0023] In some embodiments, 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.

[0024] In some embodiments, the plurality of bins comprises 10,000 or more bins, 100,000 or more bins, or 1,000,000 or more bins. In some embodiments 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 chromosome from the multi-chromosomal species.

[0025] In some embodiments, 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}, the set {1, ..., 4096}, the set {1, ..., 16384}, the set {1, ..., 65536}, the set {1, ..., 262144}, the set {1, ..., 1048576}, the set {1, ..., 4194304}, the set {1, ..., 16777216}, the set {1, ..., 67108864}, or the set {1, ..., 1×10^{12}}.

[0026] In some embodiments, the barcode in the second portion of a sequence read in the plurality of sequence reads is localized to a contiguous set of oligonucleotides. In some such embodiments, the contiguous set of oligonucleotides is an N-mer, where N is an integer selected from the set {4, ..., 20}.

[0027] 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. For instance, the noncontiguous set is localized, in various embodiments, to two noncontiguous portions, three noncontiguous portions, four noncontiguous portions, five noncontiguous portions or more of the sequence read. In some embodiments, the predetermined noncontiguous set of nucleotides collectively consists of N nucleotides, where N is an integer in the set {4, ..., 20}.

[0028] In some embodiments, the first sequence read corresponds to a subset of the test nucleic acid that is greater than 20 kbp, 30 kbp, 40 kbp, 50 kbp, 60 kbp, 70 kbp, or 80 kbp. In some embodiments, a first sequence read in the plurality of sequence reads is 2x36 bp, 2x50 bp, 2x76 bp, 2x100 bp, 2x150 bp or 2x250 bp, where the terminology 2 x N bp means that the sequence read has two reads of length N base pairs from a single piece of nucleic acid (e.g., from a text nucleic acid obtained from a biological sample) that are separated by an unspecified length. In some embodiments this unspecified length is between 200 to 1200 base pairs. In some embodiments, a first sequence read in the plurality of sequence reads represents at least 25 bp, at least 30 bp, at least 50 bp, at least 100 bp, at least 200 bp, at least 250 bp, at least 500 bp, less than 500 bp, less than 400 bp, or less than 300 bp of a single piece of nucleic acid (e.g., from a text nucleic acid obtained from a biological sample).

[0029] In some embodiments, the structural variation is a translocation 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 from the different portion of the test nucleic acid that is represented by the second set of sequence reads.

[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] Another aspect of the present disclosure is 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] Another aspect of the present disclosure provides 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.** Another aspect of the present disclosure provide s 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

more programs for execution by the one or more processors, 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.

[0035] The method further comprises 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.

[0036] The method further comprises 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. 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.

[0037] The method further comprises 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.

[0038] The method further comprises, 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. 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, where 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. 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.

[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:

$$p = 1 - P_{Binom}\left(n; {}^{n_1 n_2}/_{B}\right)$$

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.

[0044] In some embodiments, 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. In some embodiments, 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.

[0045] In some embodiments, 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. In some embodiments, 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. In some embodiments, 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.

[0046] In some embodiments, the structural variation is associated with a genetic disease. In some embodiments, 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. In some embodiments, 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.

[0047] In some embodiments, the plurality of bins comprises 10,000 or more bins, 100,000 or more bins, or 1,000,000 or more bins.

[0048] In some embodiments, 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.

[0049] In some embodiments, 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 x 10^{12}}.

[0050] 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.

[0051] In some embodiments, the contiguous set of oligonucleotides is an N-mer, wherein N is an integer selected from the set {4, ..., 20}.

[0052] 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, ..., 20}.

[0053] In some embodiments, the first sequence read corresponds to a first subset of the test nucleic acid that is greater than 10 kilobases. In some embodiments, the first sequence read corresponds to a first subset of the test nucleic acid that is greater than 20 kilobases.

[0054] In some embodiments, the structural variation is deemed to have occurred, the method further comprising treating a subject that originated the biological sample with a treatment regimen responsive to the structural variation.

[0055] 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.

[0056] In some embodiments, an identity of the first and second bin is determined by using sparse matrix multiplication. In some embodiments, the sparse matrix multiplication has the form:

$$V = A_1^T A_2,$$

where A_1 is a first B x N_1 matrix of barcodes that includes the first bin, A_2 is a second B x 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.

[0057] In some embodiments, 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.

[0058] In some embodiments, 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.

[0059] In some embodiments, a blackout list is maintained, where the blackout list comprises a plurality of blackout regions of the test nucleic acid, the method further comprising 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.

[0060] In some embodiments, the computed likelihood in the computing is computed as:

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

where, 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:

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

where 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.

[0061] In some embodiments:

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

where *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,

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

where $P_{frag}(r_1, l_1; a_b)$ 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.

[0062] In some embodiments, $P_{frag}(r_1, l_1; a_b)$ and $P_{frag}(r_2, l_2; a_b)$ are each computed as

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

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.

[0063] In some embodiments, $P(r_1, r_2, l_1, l_2, d \mid SM, \text{no } SV; a_b)$ is computed as

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

where 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.

[0064] In some embodiments,

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

where 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, $P(r_1, r_2, l_1, l_2, 2d' \mid DM, SV; a_b) = P_{frag}(r_1, l_1; a_b)P_{frag}(r_2, l_2; a_b)$, where $P_{frag}(r_1, l_1; a_b)$ 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. In some such embodiments $P_{frag}(r_1, l_1; a_b)$ 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. In some such embodiments $P(r_1, r_2, l_1, l_2, 2d' \mid SM,\ SV; a_b)$ is computed as

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

where 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. 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] Another aspect of the present disclosure provides 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

plurality of sequence reads and the second bin is represented by a second set of sequence reads in the plurality of sequence reads.

[0068] The one or more programs further comprise instructions for 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.

[0069] The one or more programs further comprise instructions for, 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) corresponds to a different barcode that is common to the first bin and the second bin and (ii) consists 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. 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. 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. 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. Each sequence read in the respective second subset of sequence reads is from the second bin. The one or more programs comprising instructions for 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.

[0070] **Part C, Phasing Methods.** Another aspect of the present disclosure provides 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_O) 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_O, 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_O 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_O, 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

standard variant calling algorithms that are relied upon to provide calls at positions $A_{i;p}$ between H_0 and H_1 may, in fact, call such positions incorrectly on occasion. For instance, consider the case in which there are twenty sequence reads from the same sequenced nucleic acid fragment f, each with the same barcode 132, for a position i of fragment f and further suppose that conventional variant calling algorithms call position i heterozygous with seventeen of the sequence reads called H_0 at position i and the remaining three called H_1 at position i. Further suppose that ground truth for position i from fragment f is, in fact homozygous H_0 meaning that the standard variant calling algorithms should have called position i H_0 for all twenty sequence reads. As such, the conventional haplotype assignment has miscalled three of the sequence reads. The disclosed alternative phasing embodiment $p \in \{0, 1, -1\}$ advantageously accounts for the possibility of this form of error. In the example of the twenty sequence reads above where three of the twenty sequence reads are miscalled at position i, this position i in all twenty of the sequence reads is assigned -1 (H_{-1}), the error state, when such error at position i is being sampled (to see if it provides a better phasing solution) by the disclosed phasing algorithms. The ability to selectively sample for this error state advantageously protects the phasing algorithm from error in the input data that arises, for instance due to error in the sequencing process, weak sequencing signal, and the like.

[0075] In the methods, a plurality of barcoded sequence reads $\vec{\mathbf{O}}$ is obtained. In some embodiments, each respective sequence read $\vec{\mathbf{O}}_q$ in the plurality of sequence reads, where q is an integer index into $\vec{\mathbf{O}}$, 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.

[0076] In some embodiments, each respective sequence read $\vec{\mathbf{O}}_q$ in the plurality of sequence reads $\vec{\mathbf{O}}$ is $\in \{0, 1, -\}^n$, where (i) each respective label "0" for the respective sequence read $\vec{\mathbf{O}}_q$ assigns a corresponding variant call in $A_{i;p}$ to H_0, (ii) each respective label "1" for the respective sequence read $\vec{\mathbf{O}}_q$ assigns a corresponding variant call in $A_{i;p}$ to H_1, and (iii) each respective label "–" for the respective sequence read $\vec{\mathbf{O}}_q$ indicates that the corresponding variant call in $A_{i;p}$ is not covered. For example, consider the case in which $\vec{\mathbf{O}}_q$ contains 5 of the 10 variant calls in $A_{i;p}$. In this example, $\vec{\mathbf{O}}_q$ will contain five variant calls with values "-" because they are not in the respective sequence read and $\vec{\mathbf{O}}_q$ will contain values for the five other variant calls in $A_{i;p}$. Each value in these five values will be a zero or a one depending on

the haplotype assigned to the respective variant call in the sequence read. In some embodiments, such haplotype assignments are obtained for the variant calls in individual sequence reads using conventional haplotype assignment algorithms.

[0077] In some embodiments, to account for possible error in the zygosity of a variant call as described above, each respective sequence read $\vec{\mathbf{O}}_q$ in the plurality of sequence reads $\vec{\mathbf{O}}$ is $\in \{0, 1, -1, -\}^n$, where (i) each respective label "0" for the respective sequence read $\vec{\mathbf{O}}_q$ assigns a corresponding variant call in $A_{i;p}$ to H_0, (ii) each respective label "1" for the respective sequence read $\vec{\mathbf{O}}_q$ assigns a corresponding variant call in $A_{i;p}$ to H_1, (iii) each respective label "-1" for the respective sequence read $\vec{\mathbf{O}}_q$ assigns a corresponding variant call in $A_{i;p}$ to the zygosity error condition (present, but neither H_0 nor H_1), and (iv) each respective label "–" for the respective sequence read $\vec{\mathbf{O}}_q$ indicates that the corresponding variant call in $A_{i;p}$ is not covered.

[0078] In the disclosed methods, a phasing result $\vec{\mathbf{X}}$ is obtained by optimization of haplotype assignments at individual positions i in $A_{i;p}$. In embodiments where each respective sequence read $\vec{\mathbf{O}}_q$ in the plurality of sequence reads $\vec{\mathbf{O}}$ is $\in \{0, 1, -\}^n$, these haplotype assignments are each between H_0 and H_1 at individual positions i for the plurality of sequence reads. In the alternative embodiments where possible error in the zygosity of the position i is to be additionally sampled in the phasing algorithm, each respective sequence read $\vec{\mathbf{O}}_q$ in the plurality of sequence reads $\vec{\mathbf{O}}$ is $\in \{0, 1, -1, -\}^n$, these haplotype assignments are each between H_0, H_1 and H_{-1} at individual positions i for the plurality of sequence reads, where H_{-1} denotes the zygosity error condition above.

[0079] Sequence reads are aligned to a reference genome. Furthermore, sequence reads that have the same barcode are grouped together. In this way, sequence reads with a common barcode are partitioned into groups that are likely to have originated from a single genomic input fragment f, and thus provide evidence that the alleles covered by the sequence read came from the same haplotype.

[0080] In embodiments where each respective sequence read $\vec{\boldsymbol{O}}_q$ in the plurality of sequence reads $\vec{\boldsymbol{O}}$ is $\in \{0, 1, -\}^n$, the probability of the observed sequence reads covering variant i from fragment f is computed as:

$$\log P(O_{i,f}|A_{i,p}) = \sum_r \mathbf{1}(S_r = A_{i,p})\left(1 - 10^{-Q_r/10}\right) + \mathbf{1}(S_r \neq A_{i,p})\left(10^{-Q_r/10}\right)$$

where,

r sums over all sequence reads for fragment f,

$\mathbf{1}(S_r = A_{i,p})$ is the indicator function that has the value "1" when the rth sequence read S_r from fragment f matches $A_{i,p}$ and is "0" otherwise,

$\mathbf{1}(S_r \neq A_{i,p})$ is the indicator function that has the value "1" when the rth sequence read S_r from fragment f does not matches $A_{i,p}$ and is "0" otherwise, and

Q_r is a relevant quality value associated with the rth sequence read.

[0081] In embodiments where each respective sequence read $\vec{\boldsymbol{O}}_q$ in the plurality of sequence reads $\vec{\boldsymbol{O}}$ is $\in$ {0, 1, -1, -}^{n}, that is possible zygosity error at position i is additionally sampled, the probability of the observed sequence reads covering variant i from fragment f is computed as:

$$\log P(O_{i,f}|A_{i,p}) = \sum_r \mathbf{1}(S_r = A_{i,p})\left(1 - 10^{-Q_r/10}\right) + \mathbf{1}(S_r \neq A_{i,p})\left(10^{-Q_r/10}\right)$$

$$+ \mathbf{1}(A_{i,p} = X^-)^{0.5}$$

where,

X^- is H_{-1}, and

$\mathbf{1}(A_{i,p} = X^-)$ is the indicator function that has the value "1" when $A_{i,p}$ is equal to -1 (H_{-1}) and is "0" otherwise.

[0082] In embodiments where $\vec{\mathbf{O}}$ is $\in$ {0, 1, -}^{n}, the phasing result can be obtained by optimizing an objective function expressed as a maximum likelihood phasing parity vector:

$$\hat{\mathbf{X}} = \underset{\vec{\mathbf{X}}}{\arg\max}\, P(\vec{\boldsymbol{O}}|\vec{\boldsymbol{X}})$$

Here, $\hat{\mathbf{X}}$ is the refined phasing vector while $\vec{\mathbf{X}}$ is the phasing vector result to be inferred and

$$P(\vec{\boldsymbol{O}}|\vec{\boldsymbol{X}}) = \prod_f P(O_{1,f}, \ldots, O_{N,f}).$$

[0083] In embodiments where $\vec{\mathbf{O}}$ is $\in$ {0, 1, -1, -}^{n}, the phasing vector can be found by optimizing an overall objective function:

$$\underset{\vec{\boldsymbol{X}}}{\text{argmax}}\, P(\vec{\boldsymbol{X}}|\vec{\boldsymbol{O}})\,| = \frac{P(\vec{\boldsymbol{X}})P(\vec{\boldsymbol{O}}|\vec{\boldsymbol{X}})}{C}$$

where,

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

H_{-1} is the condition of zygosity error at position i,

ε_i is an estimate of incurring this form of error at position i, and

$$P(\vec{O}|\vec{X}) = \prod_f P(O_{1,f}, \dots, O_{N,f}).$$

[0084] In some embodiments, ε_i is a function of the type of variant at position i. For instance ε_i is given a first value if the variant at position i arises through genetic insertions or deletion, and another value if the variant at position i arises by other means (e.g., single nucleotide polymorphisms).

[0085] Each $O_{i,f}$ in $(O_{1,f}, \dots, O_{N,f})$ is the respective subset of barcoded sequence reads for fragment f (e.g., contain the same barcode sequence). Moreover,

$$P(O_{1,f}, \dots, O_{N,f}|\mathbf{X}, H_f = 0) = \prod_i P(O_{i,f}|A_{i,X_i}),$$

$$P(O_{1,f}, \dots, O_{N,f}|\mathbf{X}, H_f = 1) = \prod_i P(O_{i,f}|A_{i,1-X_i}), \text{ and}$$

$$P(O_{1,f}, \dots, O_{N,f}|\mathbf{X}, H_f = M) = \prod_i 0.5.$$

Here, M indicates a mixture of $H_f = 0$ and $H_f = 1$ for fragment f. In other words, $H_f = 0$ represents fragment f mapping to H_0 and $H_f = 1$ represents fragment f mapping to H_1. The above three equations provide for the three possibilities for the sequence reads having a common barcode: either the sequence reads are haplotype 0 (from the first haplotype set), haplotype 1 (from the second haplotype set), or they are M, which arises in the rare instance where both maternal and paternal genomic material for the region of the reference sequence covered by the partition associated with sequence barcode f is present in the partition giving rise to a sequence read $O_{q,f}$.

[0086] In some embodiments, ten or more sequence reads have the same barcode, twenty or more sequence reads have the same barcode, thirty or more sequence reads have the same barcode, one hundred or more sequence reads have the same barcode, or one thousand or more sequence reads have the same barcode.

[0087] In some embodiments, the three possible haplotype assignments for each $O_{q,f}$ in $(O_{1,f}, \dots, O_{N,f})$ are scored as:

$$P(O_{i,f}, \ldots, O_{N,f}|\mathbf{X}) = \frac{(1-\propto)}{2}\left(\prod_i P(O_{i,f}|A_i, X_i) + \prod_i P(O_{i,f}|A_i, 1 - X_i)\right) + \propto \prod_i 0.5.$$

Here, $\propto$ is a predetermined fractional value representing a likelihood or probability that $H_f = M$ arises (prior probability that $H_f = M$), and $\log P(O_{i,f}|A_{i,p})$ is defined as above for either the embodiment where $\vec{\boldsymbol{O}}$ is $\in \{0, 1, -\}^n$ or the embodiment where $\vec{\boldsymbol{O}}$ is $\in \{0, 1, -1, -\}^n$.

[0088] In some embodiments, the set of variant calls comprises a plurality of heterozygous single nucleotide polymorphisms, heterozygous inserts, or heterozygous deletions in the test nucleic acid.

[0089] In some embodiments, the first set of haplotypes ($H = 0$) consists of maternal haplotypes for the single organism, and the second set of haplotypes ($H = 1$) consists of paternal haplotypes for the single organism.

[0090] In some embodiments, the plurality of barcodes comprises 1000 or more barcodes, 10,000 or more barcodes, 100,000 or more barcodes, or 1 x 10^6 or more barcodes. In some embodiments, the species is human.

[0091] In some embodiments, the plurality of variant calls $A_{i;p}$ comprises 1000 or more variant calls, or 10,000 or more variant calls. In some embodiments, the plurality of sequence reads comprises 10,000 or more sequence reads, 100,000 or more sequence reads, or 1 x 10^6 or more sequence reads.

[0092] In some embodiments $X\ is\ (x)$ where x is a binary string of length n, each value of "0" in x indicates origination of the corresponding variant call in the first set of haplotypes ($H = 0$), and each value of "1" in x indicates origination of the corresponding variant call in the second set of haplotypes $H = 1$).

[0093] In some embodiments, the first set of haplotypes ($H = 0$) consists of a single maternal haplotype and the second set of haplotypes ($H = 1$) consists of a single paternal haplotype. In some embodiments, the first set of haplotypes ($H = 0$) comprises five or more maternal haplotypes corresponding to five or more maternal chromosomes and the second set of haplotypes ($H = 1$) comprises five or more paternal haplotypes corresponding to five or more paternal chromosomes.

[0094] In some embodiments, the subset of sequence reads that originate from the same fragment f (and include the same respective sequence barcode) comprises 10 or more sequence reads, 30 or more sequence reads, or 100 or more sequence reads.

[0095] In some embodiments, a subset of sequence reads that include the same respective sequence read represent a fragment f of the reference consensus sequence that is at least 30 kilobases in length, at least 40 kilobases in length, or between 20 kilobases and 60 kilobases length. In some such embodiments, each such sequence read in the subset of sequence reads is 2x36 bp, 2x50 bp, 2x76 bp, 2x100 bp, 2x150 bp or 2x250 bp, where the terminology 2 x N bp means that the sequence read has two reads of length N base pairs from the reference consensus sequence that are separated by an unspecified length. In some embodiments this unspecified length is between 200 to 1200 base pairs. In some embodiments, each sequence read in the subset of sequence reads represents at least 25 bp, at least 30 bp, at least 50 bp, at least 100 bp, at least 200 bp, at least 250 bp, at least 500 bp, less than 500 bp, less than 400 bp, or less than 300 bp of the reference consensus sequence.

[0096] In some embodiments, one of the overall objective function is optimized. In some embodiments the overall objective function is optimized. In some embodiments, the hierarchical search comprises, for each respective local block of variant calls in $A_{i;p}$ that are localized to a corresponding subset of the reference consensus sequence, using a beam search over the assignments of $X_k, X_{k+1}, \ldots, X_{k+j}$ 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 and where assignments of $X_k, X_{k+1}, \ldots, X_{k+j}$ are found by computing one of the objective functions described above in which the phasing vector of the objective function in respective computations of the objective function for assignments X_k, $X_{k+1}, \ldots, X_{k+j}$ is limited to $X_k, X_{k+1}, \ldots, X_{k+j}$, thereby finding an optimal phasing solution for each respective local block of variant calls. Further, in some embodiments, neighboring local blocks of variant calls in $A_{i;p}$ are greedily joined using the optimal phasing solution for each respective local block of variant calls thereby obtaining an estimate of the optimal phasing configuration $\hat{\mathbf{X}}$. In some embodiments, neighboring local blocks of variant calls in $A_{i;p}$ are joined using *Monte Carlo* algorithms, or other stochastic searches such as simulated annealing or Boltzmann learning, *etc.* using the optimal phasing solution for each respective local block of variant calls. See for example Duda *et al.*, 2001, *Pattern Classification*, Second Edition, John Wiley & Sons, Inc., New York, which is hereby incorporated by reference herein for such purpose of disclosing, for instance, stochastic search methods.

[0097] In some embodiments, refining the phase result further comprises iteratively swapping the phase result of individual x_i in the estimate of the optimal phasing configuration $\hat{\mathbf{X}}$ and recomputing the objective function, thereby obtaining $\hat{\mathbf{X}}$.

[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}, ..., X_{k+j} discards all but a predetermined number of solutions for $\hat{\mathbf{X}}$. In some embodiments, the predetermined number of solutions for $\hat{\mathbf{X}}$ is 1000 or less. In some embodiments, predetermined number of solutions for $\hat{\mathbf{X}}$ 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, ..., 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}}. 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, ..., 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, ..., 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] Another aspect of the present disclosure provides 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] Another aspect of the present disclosure provides 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 $X_k, X_{k+1}, \ldots, X_{k+j}$ 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 $X_k, X_{k+1}, \ldots, X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k, X_{k+1}, \ldots, X_{k+j}$, 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 $\widehat{\mathbf{X}}$ for the single organism of the species. In some embodiments, the method further comprises iteratively swapping the phase result of individual x_i in $\widehat{\mathbf{X}}$ and recomputing the objective function, thereby obtaining $\widehat{\mathbf{X}}$. 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 $X_k, X_{k+1}, \ldots, X_{k+j}$ discards all but a predetermined number of solutions for $\widehat{\mathbf{X}}$ (e.g., 1000 or less, 5000 or less, etc.).

[00104] Another aspect of the present disclosure provides a method of addressing error in the zygosity 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 zygosity error condition H_{-1}. In the method, a plurality of sequence reads $\vec{\boldsymbol{O}}$ for the biological sample is obtained. Each respective sequence read $\vec{\boldsymbol{O}}_i$ 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 $\vec{\boldsymbol{O}}i$ 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 $\vec{\boldsymbol{O}}i$ assigns a corresponding variant call in $A_{i;p}$ to H_0, (iii) each respective label 1 for the respective sequence read $\vec{\boldsymbol{O}}i$ assigns a corresponding variant call in $A_{i;p}$ to H_0, (iv) each respective label -1 for the respective sequence read $\vec{\boldsymbol{O}}i$ assigns a corresponding variant call in $A_{i;p}$ to H_{-1}, and (v) each respective label – for the respective sequence read $\vec{\boldsymbol{O}}i$ indicates that the corresponding variant call in $A_{i;p}$ is not covered. In the method, a phasing vector result $\vec{\boldsymbol{X}}$ 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] $$\hat{\mathbf{X}} = \underset{\vec{\boldsymbol{X}}}{\operatorname{argmax}} P(\vec{\boldsymbol{X}}|\vec{\boldsymbol{O}}) \mid = \frac{P(\vec{X})P(\vec{O}|\vec{X})}{C}$$

[00106] where

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

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

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

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

[00111] C is a constant,

[00112] $\vec{\mathbf{X}}$ is the phasing vector result to be inferred, and

[00113] $(O_{1,f}, \ldots, O_{N,f})$ is the respective subset of N variant calls in the plurality of variant calls $A_{i;p}$ observed in the subset of sequence reads that include the same respective barcode from the plurality of barcodes.

[00114] In some embodiments,

[00115] $(O_{1,f}, \ldots, O_{N,f} | \vec{\mathbf{X}},\ H_f = 0) = \prod_i P(O_{i,f} | A_{i,X_i})$,

[00116] $P(O_{1,f}, \ldots, O_{N,f} | \vec{\mathbf{X}},\ H_f = 1) = \prod_i P(O_{i,f} | A_{i,1-X_i})$,

[00117] $P(O_{1,f}, \ldots, O_{N,f} | \vec{\mathbf{X}},\ H_f = M) = \prod_i 0.5$,

[00118] M indicates a mixture of $H_f = 0$ and $H_f = 1$ for the respective barcode f,

[00119] $P(O_{i,f}, \ldots, O_{N,f} | \vec{\mathbf{X}}) = \frac{(1-\propto)}{2}(\prod_i P(O_{i,f} | A_i, X_i) + \prod_i P(O_{i,f} | A_i, 1 - X_i)) + \propto \prod_i 0.5$,

[00120] $\log P(O_{i,f} | A_{i,p}) =$

$\sum_r \mathbf{1}(S_r = A_{i,p})\left(1 - 10^{-Q_r/10}\right) + \mathbf{1}(S_r \neq A_{i,p})\left(10^{-Q_r/10}\right)$

[00121] $+ \mathbf{1}(A_{i,p} = H_{-1})^{0.5}$

[00122] $\propto$ is a predetermined fractional value representing a likelihood or probability that $H_f = M$ arises,

[00123] i is the ith variant in the respective subset of N variant calls observed for the subset of sequence reads that include the same respective barcode,

[00124] r sums over the subset of sequence reads that include the same respective barcode,

[00125] $\mathbf{1}(S_r = A_{i,p})$ is an indicator function testing if the base assignment at position i in the rth sequence read S_r in the subset of sequence reads that include the same respective barcode matches Ai,p, wherein when they match $\mathbf{1}(S_r = A_{i,p})$ has a value of 1 and when they do not match $\mathbf{1}(S_r = A_{i,p})$ has a value of zero,

[00126] $\mathbf{1}(S_r \neq A_{i,p})$ is an indicator function testing if the base assignment at position i in the rth sequence read S_r in the subset of sequence reads that include the same respective barcode does not match Ai,p, wherein when they do not match $\mathbf{1}(S_r \neq A_{i,p})$ has a value of 1 and when they do match $\mathbf{1}(S_r = A_{i,p})$ has a value of zero,

[00127] $\mathbf{1}(A_{i,p} = H_{-1})$ is an indicator function that has a value of 1 when $A_{i,p}$ is equal to H_{-1} and is a value of zero otherwise, and

[00128] Q_r is a quality value for S_r for the read base at the position of i in the reference consensus sequence.

[00129] In some embodiments, the plurality of variant calls comprises a plurality of heterozygous single nucleotide polymorphisms, heterozygous inserts, or heterozygous deletions in the test nucleic acid. In some embodiments, the first set of haplotypes (H_0) consists of maternal haplotypes for the single organism, and the second set of haplotypes (H_1) consists of paternal haplotypes for the single organism.

[00130] In some embodiments, the plurality of barcodes comprises 1000 or more barcodes. In some embodiments, the plurality of barcodes comprises 10,000 or more barcodes. In some embodiments, the plurality of barcodes comprises 100,000 or more barcodes. In some embodiments, the plurality of barcodes comprises 1 x 10^6 or more barcodes. In some embodiments, the species is human. In some embodiments, the plurality of variant calls $A_{i,p}$ comprises 1000 or more variant calls or 10,000 or more variant calls. In some embodiments, the plurality of sequence reads comprises 10,000 or more sequence reads, 100,000 or more sequence reads, or 1 x 10^6 or more sequence reads.

[00131] In some embodiments, $\vec{X}$ *is* (x), where x is a binary string of length n, each value of 0 in x indicates origination of the corresponding variant call in the first set of haplotypes (H_0), and each value of 1 in x indicates origination of the corresponding variant call in the second set of haplotypes (H_1).

[00132] In some embodiments, the first set of haplotypes (H_0) consists of a single maternal haplotype and the second set of haplotypes (H_1) consists of a single paternal haplotype.

[00133] In some embodiments, the first set of haplotypes (H_0) comprises five or more maternal haplotypes corresponding to five or more maternal chromosomes and the second set

of haplotypes (H_1) comprises five or more paternal haplotypes corresponding to five or more paternal chromosomes.

[00134] In some embodiments, the subset of sequence reads that include the same respective barcode f comprises 10 or more sequence reads. In some embodiments, the subset of sequence reads that include the same respective barcode f comprises 30 or more sequence reads. In some embodiments, the subset of sequence reads that include the same respective barcode f comprises 100 or more sequence reads. In some embodiments, a subset of sequence reads that include the same respective barcode f represent at least 30 kilobases of the reference consensus sequence or at least 40 kilobases of the reference consensus sequence.

[00135] In some embodiments, the refining optimizes the overall objective function using a hierarchical search over $\vec{X}$. In some embodiments, the hierarchical search comprises, for each respective local block of variant calls in $A_{i;p}$ that are localized to a corresponding subset of the reference consensus sequence, using a beam search over the assignments of X_k, X_{k+1}, ..., 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 and wherein assignments of X_k, X_{k+1}, ..., X_{k+j} are found by computing the objective function in which the phasing vector of the objective function in respective computations is limited to X_k, X_{k+1}, ..., X_{k+j}, thereby finding an optimal phasing solution for each respective local block of variant calls, and greedily joining neighboring local blocks of variant calls in $A_{i;p}$ using the optimal phasing solution for each respective local block of variant calls thereby obtaining an estimate of the optimal phasing configuration $\hat{\mathbf{X}}$. In some embodiments, the refining the phase result further comprises iteratively swapping the phase result of individual x_i in the estimate of the optimal phasing configuration $\hat{\mathbf{X}}$ and recomputing the objective function, thereby obtaining $\hat{\mathbf{X}}$.

[00136] 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}, ..., X_{k+j} discards all but a predetermined number of solutions for $\hat{\mathbf{X}}$. In some embodiments, the predetermined number of solutions for $\hat{\mathbf{X}}$ is 1000 or less. In some embodiments, the predetermined number of solutions for $\hat{\mathbf{X}}$ is 5000 or less.

[00137] 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.

[00138] In some embodiments, 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, ..., 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}}.

[00139] 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, wherein N is an integer selected from the set {4, ..., 20}.

[00140] 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, ..., 20}.

[00141] In some embodiments, a sequence read in the plurality of sequence reads corresponds to a portion of the reference consensus sequence that is greater than 10 kilobases or greater than 20 kilobases.

[00142] In some embodiments, the plurality of variant calls is obtained from the plurality of sequence reads. In some embodiments, the plurality of sequence reads is obtained from a plurality of barcoded-oligo coated gel-beads and wherein the test nucleic acid sample is 50 ng or less. In some embodiments, the plurality of barcoded-oligo coated gel-beads comprises 10,000 beads. In some embodiments, the plurality of barcoded-oligo coated gel-beads comprises 50,000 beads. In some embodiments, is 25 ng or less, 10 ng or less, 5 ng or less, or 2.5 ng or less.

[00143] In some embodiments, the plurality of sequencing reads $\vec{O}$ is obtained within ten minutes of exposure to the plurality of barcodes. In some embodiments, the plurality of sequencing reads $\vec{O}$ is obtained within twenty minutes of exposure to the plurality of barcodes.

[00144] The present disclosure further provides a non-transitory computer readable storage medium storing one or more programs configured for execution by a computer, the one or more programs comprising instructions for carrying out any of the disclosed methods. Thus, these methods, systems, and non-transitory computer readable storage medium provide improved methods for detecting a structural variant in sequencing data of a test nucleic acid obtained from a biological sample.

[00145] Thus, these methods, systems, and non-transitory computer readable storage medium provide improved methods for detecting structural variants in sequencing data of a test nucleic acid obtained from a biological sample and for phasing such data.

INCORPORATION BY REFERENCE

[00146] All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference in their entireties to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference.

BRIEF DESCRIPTION OF THE DRAWINGS

[00147] The implementations disclosed herein are illustrated by way of example, and not by way of limitation, in the figures of the accompanying drawings. Like reference numerals refer to corresponding parts throughout the drawings. In the figures that include method flowcharts, boxes that are dashed indicate example embodiments.

[00148] Figure 1 is an example block diagram illustrating a computing device in accordance with some implementations.

[00149] Figures 2A, 2B, 2C, 2D, and 2E illustrate a method of detecting a structural variant in sequencing data of a test nucleic acid obtained from a biological sample in accordance with some implementations.

[00150] Figures 3A, 3B, 3C, 3D, 3E, 3F, 3G and 3H illustrate exemplary constructs in accordance with some embodiments.

[00151] Figure 4 illustrates a method of identifying bins of sequence reads in accordance with some embodiments of the present disclosure.

[00152] Figure 5 illustrates bin information in accordance with some embodiments of the present disclosure.

[00153] Figure 6 illustrates the relationship between the test nucleic acid (e.g., chromosomal DNA), the different fragments of the larger test nucleic acid, and sequence reads of fragments in accordance with some embodiments.

[00154] Figures 7A, 7B, 7C, 7D, 7E, 7F, 7G, and 7H illustrate a method of detecting a structural variant in sequencing data of a test nucleic acid obtained from a biological sample in accordance with some implementations.

[00155] Figure 8 illustrates an overview of a bar-coded library creation process in accordance with an embodiment of the present disclosure.

[00156] Figure 9 illustrates linked sequence reads in accordance with the present disclosure in which each dot in the figure represents a read-pair and group of such read-pairs that are joined by a horizontal line share a common barcode from a pool more than 100,000 barcodes.

[00157] Figure 10 provides metrics on two separate sequencing runs (from two different sources of target nucleic acid) using the systems and methods of the present disclosure.

[00158] Figure 11 illustrates cis versus trans mutations in accordance with the prior art.

[00159] Figure 12 illustrates a phasing objective function in accordance with an embodiment of the systems and methods of the present disclosure in which $O_{q,f}$ are the observations of variant i, from molecule f, $A_{i,p}$ is the allele on phase p at variant i , X_i is the phasing of variant i, $Sr = A_{i,p}$ means that sequence read r matches allele $A_{i,p}$, and α is the allele collision probability.

[00160] Figure 13 illustrates a beam search in accordance with an embodiment of the systems and methods of the present disclosure.

[00161] Figure 14 illustrates phase metrics for the phasing in accordance with an embodiment of the systems and methods of the present disclosure in which an example of the phasing of the present disclosure (column "NA12878 WES") is compared with conventional phasing (columns NA12878 WGS, HuRef 1 WGS, and NA20847 WGS).

[00162] Figure 15 illustrates a phased structural variant call in accordance with an embodiment of the systems and methods of the present disclosure in which linked reads split by

the haplotype assignment of each input molecule reveals the phased structure of the structural variant. Vertical bars indicate the breakpoints of the structural variant call.

[00163] Figure 16 illustrates called deletions in NA12878 using an embodiment of the systems and methods of the present disclosure.

[00164] Figure 17 compares PCR duplication rate between the disclosed systems and methods (10X) and that of TruSeq + AGILENT SS data at various input quantities in accordance with an embodiment of the present disclosure.

[00165] Figure 18 compares variant calling performance between the disclosed systems and methods (10X) and that of TruSeq/SS at various input quantities.

[00166] Figure 19 provides sequencing metrics for various samples run using the disclosed systems and methods.

[00167] Figure 20 illustrates detection of annotated gene fusions using the systems and methods of the present disclosure.

[00168] Figure 21 illustrates high confidence detection of a tumor-specific gene fusion in HCC38 triple negative breast cancer cell line in accordance with embodiments of the present disclosure.

[00169] Figures 22 and 23 illustrate resolving complex rearrangements using the systems and methods of the present disclosure.

[00170] Figure 24 illustrates haplotype phasing of gene fusion events in accordance with some embodiments of the present disclosure.

[00171] Figures 25 and 26 illustrate how whole-genome sequencing linked reads and phasing reveal a complex deletion and allele loss event in HCC1143 triple negative breast cancer.

[00172] Figure 27 illustrates a schematic for detecting BCR-ABL with whole exome sequencing in which there is extra baiting of long introns in accordance with embodiments of the present disclosure.

[00173] Figures 28 and 29 illustrate how a 10X GemCode library was generated from an input of ~1ng of KU812 gDNA using the systems and methods of the present disclosure. Hybrid capture was done with the standard IDT Exome panel, with (Exome +) or without

(Exome) additional Ultramer DNA baits that map to the >100 kb intronic region between exon 1 and exon 2 of ABL1 (average bait spacing ~2 kb).

[00174] Figure 30 summarizes structural variant statistics for 0.2 fmol of intronic bait versus no intronic bait runs described in Figures 27 through 29.

DETAILED DESCRIPTION

[00175] The present disclosure generally provides methods, processes, and particularly computer implemented processes and non-transistory computer program products for use in the analysis of genetic sequence data, and in particular, for detecting structural variations (*e.g.*, deletions, duplications, copy-number variants, insertions, inversions, translocations, long term repeats (LTRs), short term repeats (STRs), and a variety of other useful characterizations), as well as for haplotype phasing, in sequencing data of a test nucleic acid obtained from a biological sample. Details of implementations are now described in relation to the Figures.

[00176] Figure 1 is a block diagram illustrating a structural variant detection and phasing system 100 in accordance with some implementations. The device 100 in some implementations includes one or more processing units CPU(s) 102 (also referred to as processors), one or more network interfaces 104, a user interface 106, a memory 112, and one or more communication buses 114 for interconnecting these components. The communication buses 114 optionally include circuitry (sometimes called a chipset) that interconnects and controls communications between system components. The memory 112 typically includes high-speed random access memory, such as DRAM, SRAM, DDR RAM, ROM, EEPROM, flash memory, CD-ROM, digital versatile disks (DVD) or other optical storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, other random access solid state memory devices, or any other medium which can be used to store desired information; and optionally includes non-volatile memory, such as one or more magnetic disk storage devices, optical disk storage devices, flash memory devices, or other non-volatile solid state storage devices. The memory 112 optionally includes one or more storage devices remotely located from the CPU(s) 102. The memory 112, or alternatively the non-volatile memory device(s) within the memory 112, comprises a non-transitory computer readable storage medium. In some implementations, the memory 112 or alternatively the non-transitory computer readable storage medium stores the following programs, modules and data structures, or a subset thereof:

- an optional operating system 116, which includes procedures for handling various basic system services and for performing hardware dependent tasks;
- an optional network communication module (or instructions) 118 for connecting the device 100 with other devices, or a communication network;
- an optional sequence read processing module 120 for processing sequence reads, including a structural variation determination sub-module 122 for identifying structural variations in a genetic sample from a single organism of a species and a phasing sub-module 124 for identifying the haplotype of each sequence read of the genetic sample;
- one or more nucleic acid sequencing datasets 126, each such dataset obtained using a genetic sample from a single organism of a species;
- a plurality of sequence reads 128, each respective sequence read in the plurality of sequence reads comprising at least a first portion 130 that corresponds to a subset of the test nucleic acid 602 and a second portion 132 that encodes a barcode for the respective sequence read;
- a plurality of bins, each respective bin 140 in the plurality of bins representing a different portion 142 of the test nucleic acid 602 and further associated with a set of sequence reads 144 of the test nucleic acid; and
- a haplotype visualization tool 148 for to visualizing structural variation and phasing information in nucleic acid sequencing data, including a summarization module 150, a phase visualization module 152, a structural variants (visualization) module 154, and a read visualization module 156.

[00177] In some implementations, the user interface 106 includes an input device (e.g., a keyboard, a mouse, a touchpad, a track pad, and/or a touch screen) 100 for a user to interact with the system 100 and a display 108.

[00178] In some implementations, one or more of the above identified elements are stored in one or more of the previously mentioned memory devices, and correspond to a set of instructions for performing a function described above. The above identified modules or programs (e.g., sets of instructions) need not be implemented as separate software programs, procedures or modules, and thus various subsets of these modules may be combined or otherwise re-arranged in various implementations. In some implementations, the memory 112

optionally stores a subset of the modules and data structures identified above. Furthermore, in some embodiments, the memory stores additional modules and data structures not described above. In some embodiments, one or more of the above identified elements is stored in a computer system, other than that of system 100, that is addressable by system 100 so that system 100 may retrieve all or a portion of such data when needed.

[00179] Although Figure 1 shows a "structural variation detection and phasing system 100," the figure is intended more as functional description of the various features which may be present in computer systems than as a structural schematic of the implementations described herein. In practice, and as recognized by those of ordinary skill in the art, items shown separately could be combined and some items could be separated.

[00180] **Part A, Structural Variation.** Figure 2 is a flow chart illustrating a method of determining a structural variation occurring in a test nucleic acid obtained from a single biological sample (202). In some embodiments, the method takes place at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors in accordance with some embodiments (204).

[00181] *Obtaining a plurality of sequence reads.* In accordance with the disclosed systems and methods, a plurality of sequence reads 128 is obtained (206) using a test nucleic acid 602. Such sequence reads ultimately form the basis of a nucleic acid sequencing dataset 126. Each respective sequence read 128 in the plurality of sequence reads comprises a first portion 130 that corresponds to a subset of the test nucleic acid and a second portion 132 that encodes an barcode for the respective sequence read. The barcode is independent of the sequencing data of the test nucleic acid. In other words, the barcode is not derived from, or a function of the sequencing data of the test nucleic acid. In some instances a sequence read is referred to herein as a next generation sequencing (NGS) read-pair.

[00182] In some embodiments, a first sequence read in the plurality of sequence reads corresponds to a subset of the test nucleic acid that is is 2x36 bp, 2x50 bp, 2x76 bp, 2x100 bp, 2x150 bp or 2x250 bp, where the terminology 2 x N bp means that the sequence read has two reads of length N base pairs from a single piece of nucleic acid (e.g., from a text nucleic acid obtained from a biological sample) that are separated by an unspecified length. In some embodiments this unspecified length is between 200 to 1200 base pairs. In some embodiments, a first sequence read in the plurality of sequence reads represents at least 25 bp, at least 30 bp, at least 50 bp, at least 100 bp, at least 200 bp, at least 250 bp, at least 500 bp, less than 500 bp,

less than 400 bp, or less than 300 bp of a single piece of nucleic acid (e.g., from a text nucleic acid obtained from a biological sample). More generally, sequence read 128 obtained in some embodiments, are assembled into contigs with an N50 of at least about 10 kbp, at least about 20 kbp, or at least about 50 kbp. In more preferred aspects, sequence reads are assembled into contigs of at least about 100 kbp, at least about 150 kbp, at least about 200 kbp, and in many cases, at least about 250 kbp, at least about 300 kbp, at least about 350 kbp, at least about 400 kbp, and in some cases, or at least about 500 kbp or more. In still other embodiments, sequence reads are phased into contigs with an N50 in excess of 200 kbp, in excess of 300 kbp, in excess of 400 kbp, in excess of 500 kbp, in excess of 1 Mb, or even in excess of 2 Mb are obtained in accordance with the present disclosure. See Miller et al., 2010, "Assembly algorithms for next generation sequencing data," Genomics 95, pp. 315-327, which is hereby incorporated by reference for a definition on N50 and conventional contig assembly algorithms.

[00183] In some embodiments, as illustrated in Figure 6, to obtain the plurality of sequence reads 128, a larger contiguous nucleic acid 602 (the test nucleic acid, e.g., chromosomal DNA) is fragmented to form fragments 306 and these fragments are compartmentalized, or partitioned into discrete compartments or partitions (referred to interchangeably herein as partitions). In some embodiments, the test nucleic acid 602 is the genome of a multi-chromosomal organism such as a human. In some embodiments, more than 10, more than 100, more than 1000, more than 10,000, more than 100,000, more than 1×10^6, or more than 5×10^6 sets of sequence reads are obtained, corresponding more than 10, more than 100, more than 1000, more than 10,000, more than 100,000, more than 1×10^6, or more than 5×10^6 partitions. Figure 6 thus illustrates the relationship between the larger contiguous nucleic acid 602, the different fragments 306 of the larger contiguous nucleic acid, and sequence reads 128 of fragments. Typically, between 1 and 250 fragments 604, between 5 and 500 fragments 604 or between 10 and 1000 fragments 604 are each partitioned into a separate partition. In any event, sufficiently few of the fragments 804 are partitioned into the same partition such that the chance that the fragments 804 in a single partition have any appreciable overlapping sequences is unlikely. Sequence reads 728 of each fragment 804 are made. In typical embodiments, sequence reads 128 are short in length (e.g., less than 1000 bases) so that they can be sequenced in automated sequencers. Each sequence read 128 in a partition includes a common second portion 132 that forms a barcode that is independent of the sequence of the larger contiguous nucleic 602 acid nucleic acid and that identifies the partition, in a plurality of partitions, in which the respective sequence read was formed.

[00184] In some embodiments, the test nucleic acid is the genome of a multi-chromosomal organism such as a human. In some embodiments, 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 from the multi-chromosomal species (208).

[00185] Each partition maintains separation of its own contents from the contents of other partitions. As used herein, the partitions refer to containers or vessels that may include a variety of different forms, e.g., wells, tubes, micro or nanowells, through holes, or the like. In preferred aspects, however, the partitions are flowable within fluid streams. In some embodiments, these vessels are comprised of, e.g., microcapsules or micro-vesicles that have an outer barrier surrounding an inner fluid center or core, or have a porous matrix that is capable of entraining and/or retaining materials within its matrix. In a preferred aspect, however, these partitions comprise droplets of aqueous fluid within a non-aqueous continuous phase, e.g., an oil phase. A variety of different suitable vessels are described in, for example, U.S. Patent Application No. 13/966,150, filed August 13, 2013, which is hereby incorporated by reference herein in its entirety. Likewise, emulsion systems for creating stable droplets in non-aqueous or oil continuous phases are described in detail in, e.g., Published U.S. Patent Application No. 2010-0105112, which is hereby incorporated by reference herein in its entirety. In certain embodiments, microfluidic channel networks are particularly suited for generating partitions. Examples of such microfluidic devices include those described in detail in Provisional U.S. Patent Application No. 61/977,804, filed April 4, 2014, the full disclosure of which is incorporated herein by reference in its entirety for all purposes. Alternative mechanisms may also be employed in the partitioning of individual cells, including porous membranes through which aqueous mixtures of cells are extruded into non-aqueous fluids. Such systems are generally available from, e.g., Nanomi, Inc.

[00186] In the case of droplets in an emulsion, partitioning of the test nucleic acid fragments into discrete partitions may generally be accomplished by flowing an aqueous, sample containing stream, into a junction into which is also flowing a non-aqueous stream of partitioning fluid, e.g., a fluorinated oil, such that aqueous droplets are created within the flowing stream partitioning fluid, where such droplets include the sample materials. As described below, the partitions, e.g., droplets, also typically include co-partitioned barcode oligonucleotides.

[00187] The relative amount of sample materials within any particular partition may be adjusted by controlling a variety of different parameters of the system, including, for example, the concentration of test nucleic acid fragments in the aqueous stream, the flow rate of the aqueous stream and/or the non-aqueous stream, and the like. The partitions described herein are often characterized by having overall volumes that are less than 1000 pL, less than 900 pL, less than 800 pL, less than 700 pL, less than 600 pL, less than 500 pL, less than 400pL, less than 300 pL, less than 200 pL, less than 100pL, less than 50 pL, less than 20 pL, less than 10 pL, or even less than 1 pL. Where co-partitioned with beads, it will be appreciated that the sample fluid volume within the partitions may be less than 90% of the above described volumes, less than 80%, less than 70%, less than 60%, less than 50%, less than 40%, less than 30%, less than 20%, or even less than 10% of the above described volumes. In some cases, the use of low reaction volume partitions is particularly advantageous in performing reactions with small amounts of starting reagents, e.g., input test nucleic acid fragments. Methods and systems for analyzing samples with low input nucleic acids are presented in U.S. Provisional Patent Application No. 62/017,580, filed June 26, 2014, the full disclosure of which is hereby incorporated by reference in its entirety.

[00188] Once the test nucleic acid fragments 306 are introduced into their respective partitions, the test nucleic acid fragments 306 within partitions are generally provided with unique barcodes such that, upon characterization of those test nucleic acid fragments 306, they may be attributed as having been derived from their respective partitions. In some embodiments, such unique barcodes are previously, subsequently or concurrently delivered to the partitions that hold the compartmentalized or partitioned test nucleic acid fragments, in order to allow for the later attribution of the characteristics, e.g., nucleic acid sequence information, to the sample nucleic acids included within a particular compartment, and particularly to relatively long stretches of contiguous sample nucleic acids that may be originally deposited into the partitions.

[00189] Accordingly, the fragments 604 are typically co-partitioned with the unique barcodes (e.g., barcode sequences). In particularly preferred aspects, the unique barcodes are provided in the form of oligonucleotides that comprise nucleic acid barcode sequences that is attached to test nucleic acid fragments in the partitions. The oligonucleotides are partitioned such that as between oligonucleotides in a given partition, the nucleic acid barcode sequences contained therein are the same, but as between different partitions, the oligonucleotides can, and preferably have differing barcode sequences. In preferred embodiments, only one nucleic

acid barcode sequence is associated with a given partition, although in some embodiments, two or more different barcode sequences are present in a given partition.

[00190] The nucleic acid barcode sequences will typically include from 6 to about 20 or more nucleotides within the sequence of the oligonucleotides. In some embodiments, these nucleotides are completely contiguous, i.e., in a single stretch of adjacent nucleotides. In alternative embodiments, they are separated into two or more separate subsequences that are separated by one or more nucleotides. Typically, separated subsequences are separated by about 4 to about 16 intervening nucleotides.

[00191] The test nucleic acid 602 is typically partitioned such that the nucleic acids are present in the partitions in relatively long fragments or stretches of contiguous nucleic acid molecules 306 of the original test nucleic acid 602. As illustrated in Figure 6, these fragments 306 typically represent a number of overlapping fragments of the overall test nucleic acid to be analyzed, e.g., an entire chromosome, exome, or other large genomic fragment. In some embodiments, the test nucleic acid 602 includes whole genomes, individual chromosomes, exomes, amplicons, or any of a variety of different nucleic acids of interest. Typically, the fragments 306 of the test nucleic acid 602 that are partitioned are longer than 1 kbp, longer than 5 kbp, longer than 10 kbp, longer than 15 kbp, longer than 20 kbp, longer than 30 kbp, longer than 40 kbp, longer than 50 kbp, longer than 60 kbp, longer than 70 kbp, longer than 80 kbp, longer than 90 kbp or even longer than 100 kbp.

[00192] The test nucleic acid 602 is also typically partitioned at a level whereby a given partition has a very low probability of including two overlapping fragments 306 of the starting test nucleic acid 602. This is typically accomplished by providing the test nucleic acid 602 at a low input amount and/or concentration during the partitioning process. As a result, in preferred cases, a given partition includes a number of long, but non-overlapping fragments 306 of the starting test nucleic acid 602. The nucleic acid fragments 306 in the different partitions are then associated with unique barcodes where, for any given partition, nucleic acids contained therein possess the same unique barcode, but where different partitions include different unique barcodes. Moreover, because the partitioning step allocates the sample components into very small volume partitions or droplets, it will be appreciated that in order to achieve the desired allocation as set forth above, one need not conduct substantial dilution of the sample, as would be required in higher volume processes, e.g., in tubes, or wells of a multiwell plate. Further, because the systems described herein employ such high levels of barcode diversity, one can allocate diverse barcodes among higher numbers of genomic equivalents, as provided

above. In some embodiments, in excess of 10,000, 100,000, 500,000, etc. diverse barcode types are used to achieve genome:(barcode type) ratios that are on the order of 1:50 or less, 1:100 or less, 1:1000 or less, or even smaller ratios, while also allowing for loading higher numbers of genomes (e.g., on the order of greater than 100 genomes per assay, greater than 500 genomes per assay, 1000 genomes per assay, or even more) while still providing for far improved barcode diversity per genome. Here, each such genome is an example of a test nucleic acid.

[00193] Referring to Figure 3A, panel B, often the above-described partitioning is performed by combining the sample containing the test nucleic acid with a set of oligonucleotide tags 302 (containing the barcodes 132) that are releasably-attached to beads 308 prior to the partitioning step. The oligonucleotides 302 may comprise at least a primer region 316 and a barcode 132 region. Between oligonucleotides 302 within a given partition, the barcode region 132 is substantially the same barcode sequence, but as between different partitions, the barcode region in most cases is a different barcode sequence. In some embodiments, the primer region 316 is an N-mer (either a random N-mer or an N-mer designed to target a particular sequence) that is used to prime the fragments 306 within the sample within the partitions. In some cases, where the N-mer is designed to target a particular sequence, the primer region 316 is designed to target a particular chromosome (e.g., human chromosome 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X, or Y), or region of a chromosome, e.g., an exome or other targeted region. In some cases, the N-mer is designed to target a particular gene or genetic region, such as a gene or region associated with a disease or disorder (e.g., cancer). In some cases, the N-mer is designed to target a particular structural variation. Within the partitions, an amplification reaction is conducted using the primer sequence 316 (e.g. N-mer) to prime the test nucleic acid fragments 306 at different places along the length of the fragment. As a result of the amplification, each partition contains amplified products of the nucleic acid 602 that are attached to an identical or near-identical barcode, and that represent overlapping, smaller fragments of the nucleic acids in each partition. The barcode 132 therefore serves as a marker that signifies that a set of nucleic acids originated from the same partition, and thus potentially also originated from the same test nucleic acid fragment 306. It will be appreciated that there are typically several fragments 306 in any given partition. Nevertheless, in typical embodiments, fragments 306 that are in the same partition typically do not have any significant overlap and so it is possible to localize the amplified sequence reads to the correct fragment 304 in any given partition. Following

amplification, the amplified nucleic acids are pooled, sequenced to form sequence reads, and aligned using a sequencing algorithm. Because shorter sequence reads may, by virtue of their associated barcode sequences, be aligned and attributed to a single, long originating fragment of the test nucleic acid 602, all of the identified variants on that sequence can be attributed to a single originating fragment 306 and single originating chromosome of the test nucleic acid 602. Further, by aligning multiple co-located variants across multiple long fragments 306, one can further characterize that chromosomal contribution. Accordingly, conclusions regarding the phasing of particular genetic variants may then be drawn. Such information may be useful for identifying haplotypes, which are generally a specified set of genetic variants that reside on the same nucleic acid strand or on different nucleic acid strands. Moreover, additionally or alternatively, structural variants are identified.

[00194] In some embodiments, referring to Figure 3A, the co-partitioned oligonucleotide tags 302 also comprise functional sequences in addition to the barcode sequence 132 and the primer region sequence 316. For instance, in some embodiments, the co-partitioned oligonucleotide tags 302 also comprise other functional sequences useful in the processing of the partitioned nucleic acids such as targeted or random/universal amplification primer sequences for amplifying test nucleic acid fragments 306 within the partitions 304 while attaching the associated barcode sequences, sequencing primers, hybridization or probing sequences, e.g., for identification of presence of the sequences, or for pulling down barcoded nucleic acids, or any of a number of other potential functional sequences. See, for example, the disclosure on co-partitioning of oligonucleotides and associated barcodes and other functional sequences, along with sample materials as described in, for example, U.S. Patent Application Nos. 61/940,318, filed February 7, 2014, 61/991,018, Filed May 9, 2014, and U.S. Patent Application No. 14/316,383, (Attorney Docket No. 43487-708.201) filed on June 26, 2014, as well as U.S. Patent Application No. 14/175,935, filed February 7, 2014, the full disclosures of which is hereby incorporated by reference in their entireties.

[00195] In one exemplary process, beads are provided, where each such bead includes large numbers of the above described oligonucleotides releasably attached to the beads. In such embodiments, all of the oligonucleotides attached to a particular bead include the same nucleic acid barcode sequence, but a large number of diverse barcode sequences are represented across the population of beads used. Typically, the population of beads provides a diverse barcode sequence library that includes at least 1000 different barcode sequences, at least 10,000 different barcode sequences, at least 100,000 different barcode sequences, or in

some cases, at least 1,000,000 different barcode sequences. Additionally, each bead typically is provided with large numbers of oligonucleotide molecules attached. In particular, the number of molecules of oligonucleotides including the barcode sequence on an individual bead may be at least about 10,000 oligonucleotides, at least 100,000 oligonucleotide molecules, at least 1,000,000 oligonucleotide molecules, at least 100,000,000 oligonucleotide molecules, and in some cases at least 1 billion oligonucleotide molecules.

[00196] In some embodiments, the oligonucleotides are releasable from the beads upon the application of a particular stimulus to the beads. In some cases, the stimulus is a photo-stimulus, e.g., through cleavage of a photo-labile linkage that may release the oligonucleotides. In some cases, a thermal stimulus is used, where elevation of the temperature of the beads environment results in cleavage of a linkage or other release of the oligonucleotides form the beads. In some cases, a chemical stimulus is used that cleaves a linkage of the oligonucleotides to the beads, or otherwise results in release of the oligonucleotides from the beads.

[00197] In some embodiments, the beads including the attached oligonucleotide tags 302 are co-partitioned with the individual samples, such that a single bead and a single sample are contained within an individual partition. In some cases, where single bead partitions are desired, it may be desirable to control the relative flow rates of the fluids such that, on average, the partitions contain less than one bead per partition, in order to ensure that those partitions that are occupied, are primarily singly occupied. Likewise, in some embodiments, the flow rate is controlled to provide that a higher percentage of partitions are occupied, e.g., allowing for only a small percentage of unoccupied partitions. In preferred aspects, the flows and channel architectures are controlled as to ensure a desired number of singly occupied partitions, less than a certain level of unoccupied partitions and less than a certain level of multiply occupied partitions.

[00198] Figure 3 of United States Patent Application No. 62/072,214, filed October 29, 2014, entitled "Analysis of Nucleic Acid Sequences," which is hereby incorporated by reference and the portions of the specification describing Figure 3 provide a detailed example of one method for barcoding and subsequently sequencing a test nucleic acid (referred to in the reference as a "sample nucleic acid") in accordance with one embodiment of the present disclosure. As noted above, while single bead occupancy may be the most desired state, it will be appreciated that multiply occupied partitions, or unoccupied partitions may often be present. Figure 4 of United States Patent Application No. 62/072,214, filed October 29, 2014, entitled

"Analysis of Nucleic Acid Sequences," which is hereby incorporated by reference and the portions of the specification describing Figure 4 provide a detailed example of a microfluidic channel structure for co-partitioning samples and beads comprising barcode oligonucleotides in accordance with one embodiment of the present disclosure.

[00199] Once co-partitioned, the oligonucleotide tags 302 disposed upon the bead are used to barcode and amplify the partitioned samples. One process for use of these barcode oligonucleotides in amplifying and barcoding samples is described in detail in U.S. Patent Application Nos. 61/940,318, filed February 7, 2014, 61/991,018, filed May 9, 2014, and 14/316,383, (Attorney Docket No. 43487-708.201) filed on June 26, 2014, the full disclosures of which are hereby incorporated by reference in their entireties. Briefly, in one aspect, the oligonucleotides present on the beads that are co-partitioned with the samples are released from their beads into the partition with the samples. The oligonucleotides typically include, along with the barcode sequence 132, a primer sequence at its 5' end 316. In some embodiments, this primer sequence is a random oligonucleotide sequence intended to randomly prime numerous different regions of the samples. In some embodiments the primer sequence 316 is a specific primer sequence targeted to prime upstream of a specific targeted region of the sample.

[00200] Once released, the primer portion of the oligonucleotide anneals to a complementary region of test nucleic acid fragments 306 in the partition. Extension reaction reagents, e.g., DNA polymerase, nucleoside triphosphates, co-factors (e.g., Mg^{2+} or Mn^{2+} etc.), that are also co-partitioned with the fragments 306 and beads 304, extend the primer sequence using the fragments 306 as a template, to produce a complementary sequence to the strand of the test nucleic fragment 306 to which the primer annealed, and this complementary sequence includes the oligonucleotide 302 and its associated barcode sequence 132. Annealing and extension of multiple primers to different portions of the fragments 306 in the partition 304 may result in a large pool of overlapping complementary portions of the test nucleic acid fragments 306, each possessing its own barcode sequence 132 indicative of the partition 304 in which it was created. In some cases, these complementary fragments may themselves be used as a template primed by the oligonucleotides present in the partition 304 to produce a complement of the complement that again, includes the barcode sequence 132. In some cases, this replication process is configured such that when the first complement is duplicated, it produces two complementary sequences at or near its termini, to allow the formation of a hairpin structure or partial hairpin structure that reduces the ability of the molecule to be the

basis for producing further iterative copies. A schematic illustration of one example of this is shown in Figure 3F.

[00201] As Figure 3A shows, oligonucleotides 302 that include a barcode sequence 132 are co-partitioned in, e.g., a droplet 304 in an emulsion, along with a sample test nucleic acid fragment 306. In some embodiments, the oligonucleotides 302 are provided on a bead 308 that is co-partitioned with the test nucleic acid fragment 306. The oligonucleotides 302 are preferably releasable from the bead 308, as shown in Figure 3A, panel (A). As shown in Figure 3A panel (B), the oligonucleotides 302 includes a barcode sequence 132, in addition to one or more functional sequences, e.g., sequences 312, 132 and 316. For example, oligonucleotide 302 is shown as further comprising attachment sequence 312 that may function as an attachment or immobilization sequence for a given sequencing system, e.g., a P5 sequence used for attachment in flow cells of an ILLUMINA, HISEQ or MISEQ system. In other words, attachment sequence 312 is used to reversibly attach oligonucleotide 302 to a bead 308 in some embodiments. As shown in Figure 3A, panel B, the oligonucleotide 302 also includes a primer sequence 316, which may include a random or targeted N-mer (discussed above) for priming replication of portions of the sample test nucleic acid fragment 306. Also included within exemplary oligonucleotide 302 of Figure 3A, panel B, is a sequence 310 which may provide a sequencing priming region, such as a "read1" or R1 priming region, that is used to prime polymerase mediated, template directed sequencing by synthesis reactions in sequencing systems. In many cases, the barcode sequence 132, immobilization sequence 312 and exemplary R1 sequence 310 may be common to all of the oligonucleotides 302 attached to a given bead. The primer sequence 316 may vary for random N-mer primers, or may be common to the oligonucleotides on a given bead for certain targeted applications.

[00202] Referring to Figure 3B, based upon the presence of primer sequence 316, the oligonucleotides 302a and 302b are able to prime the test nucleic acid fragment 306, which allows for extension of the oligonucleotides 302a and 302b using polymerase enzymes and other extension reagents also co-portioned with the bead 304 and sample test nucleic acid fragment 306.

[00203] As shown in Figure 3C, following extension of the oligonucleotides that, for random N-mer primers, would anneal to multiple different regions of the sample test nucleic acid fragment 306; multiple overlapping complements or fragments of the test nucleic acid fragment 306 are created, e.g., fragments 130-1 and 130-2. As such, Figure 3C illustrates (A) obtaining a plurality of sequence reads, wherein each respective sequence read 128 in the

plurality of sequence reads comprises a first portion 130 that corresponds to a subset of the test nucleic acid 602 and a common second portion 132 that forms an barcode that is independent of the sequence of the larger contiguous nucleic acid 602 and that identifies a partition 304, in a plurality of partitions, in which the respective sequence read was formed (e.g., barcode sequence 132).

[00204] Although including sequence portions that are complementary to portions of test nucleic acid, e.g., sequences 306-1 and 306-2, these constructs are generally referred to herein as comprising fragments of the sample test nucleic acid 602, having the attached barcode sequences. As will be appreciated, the replicated portions of the template sequences as described above are often referred to herein as "fragments" of that template sequence. Notwithstanding the foregoing, however, the term "fragment" encompasses any representation of a portion of the originating test nucleic acid sequence, e.g., a template or sample nucleic acid, including those created by other mechanisms of providing portions of the template sequence, such as actual fragmentation of a given molecule of sequence, e.g., through enzymatic, chemical or mechanical fragmentation. In preferred aspects, however, fragments of a test nucleic acid sequence will denote replicated portions of the underlying sequence or complements thereof.

[00205] The barcoded nucleic acid fragments of Figure 3B may then be subjected to characterization, e.g., through sequence analysis, or they may be further amplified in the process, as shown in Figure 3D. For example, additional oligonucleotides, e.g., oligonucleotide 302c, also released from bead 308, may prime the fragment 302b. In particular, again, based upon the presence of the random N-mer primer 316 in oligonucleotide 302c (which in many cases will be different from other random N-mers in a given partition) the oligonucleotide anneals with the fragment 302b, and is extended to create a complement 130-3 to at least a portion of fragment 302b which comprises a duplicate of a portion of the test nucleic acid sequence. Extension of the oligonucleotide 302b continues until it has replicated through the oligonucleotide portion 130 of fragment 302b. As noted elsewhere herein, and as illustrated in Figure 3D, the oligonucleotides may be configured to promptly stop in the replication by the polymerase at a desired point, e.g., after replicating through sequences 316 and 310 of oligonucleotide 302b that is included within fragment. As described herein, this may be accomplished by different methods, including, for example, the incorporation of different nucleotides and/or nucleotide analogues that are not capable of being processed by the polymerase enzyme used. For example, this may include the inclusion of uracil containing

nucleotides within the sequence region 310 to prevent a non-uracil tolerant polymerase to cease replication of that region. As a result, referring to Figure 3E, a sequence read 128-3 is created that includes the full-length oligonucleotide 302b at one end, including the barcode sequence 132, the attachment sequence 312, the R1 primer region 310, and the random N-mer sequence 316. At the other end of the sequence is included the complement 316' to the random N-mer of the first oligonucleotide 302, as well as a complement to all or a portion of the R1 sequence, shown as sequence 310'. The R1 sequence 310 and its complement 310' are then able to hybridize together to form a partial hairpin structure 360. As will be appreciated, because the random N-mers differ among different oligonucleotides, these sequences and their complements would not be expected to participate in hairpin formation, e.g., sequence 316', which is the complement to random N-mer 316, would not be expected to be complementary to random N-mer sequence 316b. This would not be the case for other applications, e.g., targeted primers, where the N-mers would be common among oligonucleotides within a given partition.

[00206] By forming these partial hairpin structures, it allows for the removal of first level duplicates of the sample sequence from further replication, e.g., preventing iterative copying of copies. The partial hairpin structure also provides a useful structure for subsequent processing of the created fragments, e.g., fragment 130-3.

[00207] All of the sequence reads 128 from multiple different partitions may then be pooled for sequencing on high throughput sequencers as described herein. Because each sequence read 128 is coded as to its partition of origin, the sequence of that sequence read may be attributed back to its origin based upon the presence of the barcode 132. Such sequence reads, and analysis of such sequence reads, form the basis of the disclosed nucleic acid sequencing dataset 126.

[00208] This is schematically illustrated in Figure 4. As shown in one example, a test nucleic acid fragment 306-1 and a test nucleic acid fragment 306-2 are each partitioned along with their own sets of barcode oligonucleotides 132 as described above. Within each partition, each fragment (306-1 and 306-2) is then processed to separately provide overlapping sequence reads 128 of the fragments 306-1 and 306-2 to form a respective set of sequence reads 414 and 416. This processing provides sequence reads 414 with a barcode sequence 132 that is the same for each of the sequence reads 414 derived from a particular first fragment 306-1. As shown, the set of sequence reads 414 is denoted by "1" while the set of sequence reads 416 is denoted by "2". A diverse library of barcodes may be used to differentially barcode large numbers of different sets of fragment reads. However, it is not necessary for every sequence

read in a given partition to be barcoded with different barcode sequence. In fact, in many cases, multiple different first fragments may be processed concurrently to include the same barcode sequence. Diverse barcode libraries are described in detail elsewhere herein.

[00209] The sets of sequence reads may then be pooled for sequencing using, for example, sequence by synthesis technologies available from Illumina or Ion Torrent division of Thermo Fisher, Inc. Once sequenced, the sequence reads 128 can be attributed to their respective fragment set, e.g., as shown in aggregated reads, at least in part based upon the included barcodes, and optionally, and preferably, in part based upon the sequence of the fragment itself. The attributed sequence reads for each fragment set are then assembled to provide the assembled sequence for each sample fragment, e.g., sequences 418 and 420, which in turn, may be further attributed back to their respective original fragments (306-1 and 302-2). Methods and systems for assembling genomic sequences are described in, for example, U.S. Provisional Patent Application No. 62/017,589 (Attorney Docket No. 43487-729.101), filed June 26, 2014, the full disclosure of which is hereby incorporated by reference in its entirety.

[00210] In some embodiments, the biological sample is from a multi-chromosomal species and the test nucleic acid 602 comprises a plurality of nucleic acids collectively representing a plurality of chromosomes from the multi-chromosomal species (208). In some embodiments, the barcode in the second portion 132 of each respective sequence read in the plurality of sequence reads encodes a unique predetermined value selected from the set {1, ..., 1024}, {1, ..., 4096}, {1, ..., 16384}, {1, ..., 65536}, {1, ..., 262144}, {1, ..., 1048576}, {1, ..., 4194304}, {1, ..., 16777216}, {1, ..., 67108864}, or {1, ..., 1 x 1012} (210). For instance, consider the case in which the barcode sequence 132 is represented by a set of five nucleotide positions. In this instance, each nucleotide position contributes four possibilities (A, T, C or G), giving rise, when all five positions are considered, to 4 x 4 x 4 x 4 x 4 = 1024 possibilities. As such, the five nucleotide positions form the basis of the set {1,..., 1024}. In other words, when the barcode sequence 132 is a 5-mer, the second portion 132 of each sequence read 128 encodes a unique predetermined value selected from the set {1,..., 1024}. Likewise, when the barcode sequence 132 is represented by a set of six nucleotide positions, the six nucleotide positions collectively contribute 4 x 4 x 4 x 4 x 4 x 4 = 4096 possibilities. As such, the six nucleotide positions form the basis of the set {1,..., 4096}. In other words, when the barcode sequence 132 is a 6-mer, the second portion 132 of each sequence read 128 encodes a unique predetermined value selected from the set {1,..., 4096}.

[00211] In some embodiments, the barcode in the second portion 132 of a sequence read in the plurality of sequence reads is localized to a contiguous set of oligonucleotides within the sequence read (212). In one such exemplary embodiment, the contiguous set of oligonucleotides is an N-mer, where N is an integer selected from the set {4, ..., 20} (214). In other words, in some embodiments, the barcode 132 in, for instance Figure 3B, panel B, is a contiguous set of nucleotide positions (e.g., 4 contiguous nucleotide positions, 5 contiguous nucleotide positions, 6 contiguous nucleotide positions, 7 contiguous nucleotide positions, 8 contiguous nucleotide positions, 9 contiguous nucleotide positions, 10 contiguous nucleotide positions, 11 contiguous nucleotide positions, 12 contiguous nucleotide positions, 13 contiguous nucleotide positions, 14 contiguous nucleotide positions, 15 contiguous nucleotide positions, 16 contiguous nucleotide positions, 17 contiguous nucleotide positions, 18 contiguous nucleotide positions, 19 contiguous nucleotide positions, or 20 contiguous nucleotide positions) within oligonucleotide tag 302 which ultimately becomes second portion 132 upon transcription of the test nucleic acid.

[00212] By contrast, in some embodiments, the barcode in the second portion of a sequence read in the plurality of sequence reads is localized to a noncontiguous set of oligonucleotides within the sequence read (216). In one such exemplary embodiment, the predetermined noncontiguous set of nucleotides collectively consists of N nucleotides, where N is an integer in the set {4, ..., 20} (218). As an example, in some embodiments, referring to Figure 3A, panel B, barcode sequence 132 comprises a first set of contiguous nucleotide positions at a first position in oligonucleotide tag 302 and a second set of contiguous nucleotide positions at a second position in oligonucleotide tag 302, that is displaced from the first set of contiguous nucleotide positions by a spacer. In one specific example, the barcode sequence 132 comprises $(X1)_nY_z(X2)_m$, where X1 is n contiguous nucleotide positions, Y is a constant predetermined set of z contiguous nucleotide positions, and X2 is m contiguous nucleotide positions. In this example, the barcode in the second portion of the sequence read 128 produced by a schema invoking this exemplary barcode is localized to a noncontiguous set of oligonucleotides, namely $(X1)_n$ and $(X2)_m$. This is just one of many examples of noncontiguous formats for barcode sequence 132.

[00213] In some embodiments, the first sequence read in the plurality of sequence reads corresponds to a subset of the test nucleic acid that is 2x36 bp, 2x50 bp, 2x76 bp, 2x100 bp, 2x150 bp or 2x250 bp, where the terminology 2 x N bp means that the sequence read has two reads of length N base pairs from a single piece of nucleic acid (e.g., from a text nucleic acid

obtained from a biological sample) that are separated by an unspecified length. In some embodiments this unspecified length is between 200 to 1200 base pairs. In some embodiments, a first sequence read in the plurality of sequence reads represents at least 25 bp, at least 30 bp, at least 50 bp, at least 100 bp, at least 200 bp, at least 250 bp, at least 500 bp, less than 500 bp, less than 400 bp, or less than 300 bp of a single piece of nucleic acid (e.g., from a text nucleic acid obtained from a biological sample). (220).

[00214] *Obtaining bin information.* In accordance with the disclosed systems and methods, bin information for a plurality of bins is obtained (222). Each respective bin 140 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. Figure 5 illustrates. In Figure 5, a test nucleic acid 602 is depicted. Further shown in Figure 5, each respective bin 140 in a plurality of bins represents a different portion 142 of the test nucleic acid 602. Further, the bin information identifies, for each respective bin 140 in the plurality of bins, a set 144 of sequence reads 128 corresponding to the respective bin.

[00215] Continuing to refer to Figure 5, each sequence read 128 in each set 144 of sequence reads in the plurality of sets of sequence reads is in the plurality of sequence reads. That is, each sequence read 128 corresponds to a portion 142 of the test nucleic acid 602 that has been binned.

[00216] Each respective sequence read 128 in each respective set 144 of sequence reads in the plurality of sets of sequence reads has a respective first portion 130 that corresponds to a subset of the test nucleic acid that at least partially overlaps the different portion 142 of the test nucleic acid 602 that is represented by the bin 140 corresponding to the respective set of sequence reads. For example, referring to set 144-1 illustrated in Figure 5, each sequence read 128-1-1, ..., 128-1-K, includes a first portion 130 that corresponds to at least a subset of the region 142-1 of the test nucleic acid 140-1 corresponding to bin 140-1. It will be appreciated from the schema and constructs of Figure 3, that in typical embodiments, the first portions 130 of sequence reads 128-1-1, ..., 128-1-K of set 144-1 are of different lengths and are displaced, yet overlapping, with respect to each other. In some embodiments, such first portions 130 may include some nucleotide positions from adjacent regions 142 of the test nucleic acid 602. In some embodiments the first portions 130 only represent a subset of the region 142 of the test nucleic acid 602 of the bin 140.

[00217] In some embodiments, the region 142 of the test nucleic acid corresponding to a bin 140- is substantially longer than any given first portion 130 in the set 144 of sequence reads 128 within the bin. Moreover, in some embodiments, a given set 144 of sequence reads 128 includes five or more different barcode sequences 132, ten or more different barcode sequences 132, fifteen or more different barcode sequences 132, twenty or more different barcode sequences 132, twenty-five or more different barcode sequences 132, thirty or more different barcode sequences 132, thirty-five or more different barcode sequences 132, forty or more different barcode sequences 132, forty-five or more different barcode sequences 132, or fifty or more different barcode sequences 132.

[00218] In some embodiments, two or more, three or more, four or more, five or more, six or more, seven or more, eight or more, nine or more, ten or more, or eleven or more sequence reads 128 in a bin 144 have the same barcode sequences 132. In some embodiments, each sequence read 128 in a bin 144 has a different barcode sequence 132.

[00219] In some embodiments, each bin 140 in the plurality of bins represents at least 20 kbp, at least 50 kbp, at least 100 kbp, at least 250 kbp, or at least 500 kbp of the test nucleic acid (224). In other words, referring to Figure 5, in some embodiments, the portion 142-1 of test nucleic acid 602 represented by bin 140-1 is at least 20 kbp, at least 50 kbp, at least 100 kbp, at least 250 kbp, or at least 500 kbp.

[00220] In some embodiments, there is no overlap between each different portion 406 of the test nucleic acid represented by each respective bin in the plurality of bins (226). Such an embodiment is illustrated in Figure 5. In such embodiments it is possible for overlap in the first portions 130 of the sequence reads 128, in the corresponding bin 144, with neighboring regions 142. In some embodiments, there is no overlap between each different portion 406 of the test nucleic acid represented by each respective bin in the plurality of bins (226), but there is some overlap in the first portions 130 of the sequence reads 128 in the corresponding bin 144 with neighboring portions 406.

[00221] In some embodiments, each respective sequence read 128 in each respective set 144 of sequence reads in the plurality of sequence reads has a respective first portion 130 that corresponds to a subset of the test nucleic acid 602 that fully overlaps the different portion 142 of the test nucleic acid 602 that is represented by the bin 144 corresponding to the respective set 144 of sequence reads. In other words, using bin 144-1 of Figure 5 as an example, in some

embodiments, each first portion 130 of each sequence read 128-1-1,... 128-1-k, includes a full copy of regions 142-1.

[00222] In some embodiments, the plurality of bins comprises 10,000 or more bins, 100,000 or more bins, or 1,000,000 or more bins (230).

[00223] *Determining a number of unique* barcodes *common to two sets of sequence reads.* The method continues with a determination of the number of unique barcodes that are found in both the first and second set of sequence reads in the plurality of sequence reads (232). An example of such a determination is the determination of how many sequence reads 128-1-X in set 144-1 have a barcode sequence 132 that is the same as a barcode sequence 132 in a sequence read 128-2-Y in set 144-2. This determination results in a number of unique sequence barcodes. For example, consider the case in which there are five unique barcodes that are found in the sequence reads 128-1-X in set 144-1 as well as in the sequence reads 128-2-Y in set 144-2. In this instance, the number of distinct common barcodes is five.

[00224] *Determine a probability or likelihood that the number is attributable to chance.* The method continues by determining a probability or likelihood that the above-identified number determined at (232) is attributable to chance. This is done by comparison of a metric based upon the number to a threshold criterion (234). 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.

[00225] For example, continuing to use the example of the comparison of sets 144-1 and 144-2 of Figure 5 from above, a probability or likelihood that the number determined at (232) is attributable to chance is made. This is done by comparison of a metric based upon (i) the five unique barcodes (barcode sequences 132) common to set 144-1 and set 144-2 (ii) to a threshold criterion (234). When the metric satisfies the threshold criterion, a structural variation is deemed to have occurred in (i) portion 142-1 of the test nucleic acid 602 and/or (ii) the different portion 142-2 of the test nucleic acid.

[00226] In some embodiments, the metric is computed as:

$$p = \prod_{i}^{n} f_{b_i} \qquad \text{(A)}$$

where, $\{b_1, b_2, \ldots, b_n\}$ is the set of n unique barcodes (e.g., barcode sequences 132) that is found in both the first and second sets of sequence reads in the plurality of sequence reads , i is

an integer index to n, and f_{b_i} is the fraction of the plurality of bins in which the first portion of barcode b_i appears (236). In other words, the product of formula A is taken over the distinct common barcodes, not over the common reads.

[00227] For instance, in the comparison of sets 144-1 to 144-2, the metric is computed as:

$$p = \prod_{i}^{n} f_{b_i} \qquad (1)$$

where, $\{b_1, b_2, \ldots, b_n\}$ is the set of n barcodes in set 144-1 having unique (that is, unique across the set $\{b_1, b_2, \ldots, b_n\}$) barcode sequences 132 that are also found in the set 144-2, i is an integer index to n, and f_{b_i} is the fraction of the plurality of bins in which barcode b_i appears (236). More specifically, in embodiments in which the first portion of the sequence read 128 includes a respective barcode sequence 132, f_{b_i} is the fraction of the plurality of bins in which the respective barcode sequence 132 appears.

[00228] The probability in (1) is the probability of observing all n barcodes at the same time, assuming that they occur independently of each other. If this probability is small, the hypothesis that the overlap between the first bin and the second bin is random.

[00229] In some embodiments, the metric is computed as:

$$p = \prod_{i}^{n} \left(1 - P_{Binom}(0; max(n_1, n_2), f_{b_i})\right) \qquad (2)$$

where, $\{b_1, b_2, \ldots, b_n\}$ is the set of n unique barcodes that is found in both 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, i is an integer index to n, f_{b_i} is the fraction of the plurality of bins in which the barcode b_i appears, and $P_{Binom}(0; max(n_1, n_2) f_{b_i})$ is a cumulative distribution function of a Binomial distribution. $P_{Binom}(0; max(n_1, n_2) f_{b_i})$ is equated to a probability that, given $max(n_1, n_2)$ barcodes, each of which has a f_{b_i} probability of carrying barcode b_i, at least one occurrence of b_i in $\{b_1, b_2, \ldots, b_n\}$ is achieved. This assumes that the probability that a sequence read has barcode b_i is equal to the frequency of the barcode in the plurality of barcodes.

[00230] Computing the probability in (2) or (1) requires iteration over all common barcodes between each pair of bins. Since it is typically the case that most pairs of bins will not have a significant sequence barcode overlap, an initial set of candidate pairs of bins can be

obtained by performing a less computationally intense test that assumes that all barcodes have the same frequency. In some such embodiments, the metric is computed as:

$$p = 1 - P_{Binom}\left(n; {}^{n_1, n_2}/_{B}\right) \qquad (3)$$

where n is the number of unique barcodes that is found in both 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. Equation (3) is the probability of observing more than n overlapping barcodes assuming that all barcodes have the same frequency of appearance in the plurality of bins and are chosen at random.

[00231] Iterating over all pairs of bins genome-wide (e.g., across the plurality of bins) can be time consuming. For example, if the human genome (which is roughly 3×10^9 base pairs long) is binned in non-overlapping windows of size 10 kb, this would result in 3×10^5 bins. In other words, there would be 3×10^5 bins in the plurality of bins. In this case, the number of pairwise bin comparisons would be on the order of 10^{10}. In some embodiments, to be able to perform this comparison efficiently, matrix operations are used. In particular, for each chromosome, a $B \times N$ matrix, A, is defined, where B is the number of unique barcodes present in the experiment and N is the number of bins in that chromosome. This matrix will tend to be very sparse, since only a small subset of the barcodes will appear in any given bin. To find the overlapping bins between two chromosomes (or a chromosome and itself), with respective barcode matrices A_1 and A_2 (defined as above for A), the product $V = A_1^T A_2$ is computed, where T denotes the transpose of the matrix A_1. If N_1 and N_2 are the number of bins in A_1 and A_2 respectively, then V will be a $N_1 \times N_2$ matrix. V_{ij} will be non-zero if and only if there is barcode overlap between the ith bin of A_1 and the jth bin of A_2. The time and memory required to perform this sparse matrix multiplication depends on the number of non-zero elements in the resulting matrix, rather than on the total size of the matrix. Thus, advantageously, the time required to detect all bins that share barcodes increases with the number of such bins and not with the total number of bins.

[00232] Accordingly, in some embodiments, an identity of the first and second bin is determined using sparse matrix multiplication. In some embodiments, the sparse matrix multiplication comprises computing:

$$V = A_1^T A_2,$$

where, A_1 is a first B x N_1 matrix of barcodes that includes those of the first bin, A_2 is a second B x N_2 matrix of barcodes that includes those of the second bin, B is the number of unique barcodes in 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.

[00233] In some embodiments, the first and second bin are in different chromosomes. Accordingly, in some such embodiments, 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.

[00234] In alternative embodiments, the first and second bins are in the same chromosome. Accordingly, in some such embodiments, the first and second bins are 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.

[00235] As an example of the computations of the present disclosure, with reference to Figure 6, consider the EML4-ALK fusion, a gene fusion commonly found in the lung cancer cell line NCI-H2228. In this cell line, the intron between exons 6 and 7 of EML4 is fused to the intron between exons 20 and 21 of ALK. This creates a novel fused gene which consists of parts of both EML4 and ALK. In a sample that harbors the fusion, there will be fragments spanning the fused parts of EML4 and ALK. Therefore, the sequence reads 128 mapping to EML4 and ALK will tend to have common barcodes 132. Such barcode sharing would be much more unlikely in the absence of a fusion, since, normally, EML4 and ALK are too far apart (more than 10 million base-pairs) to be spanned by the same fragment. In a sample from the NCI-H2228 cell line, 12 barcodes 132 in common between EML4 and ALK were identified, suggesting that there were at least 12 fragments spanning the fusion region. Given the empirical frequencies of these barcodes across the plurality of bins, the probability of such an overlap occurring purely by chance, computed using metric (1) from above, is less than 10^{-8}.

[00236] In some embodiments, the metric (e.g, metric (1), (2) or (3) from above) is deemed to satisfy the threshold criterion when p is 10^{-2} or less, 10^{-3} or less, 10^{-4} or less, or 10^{-5} or less (238).

[00237] In some embodiments, the structural variation is an insertion or deletion of 50 consecutive bases or more, 500 consecutive bases or more, or 5000 consecutive bases or more,

into the different portion of the test nucleic acid that is represented by the first set of sequence reads (240).

[00238] In some embodiments, the different portion of the test nucleic acid represented by the first bin overlaps the different portion of the test nucleic acid represented by the second bin (242). In other words, referring to Figure 5, in some embodiments sections 142 overlap each other rather than abut each other as depicted. In some such embodiments, at least 50 percent, at least 80 percent, or at least 95 percent of the different portion of the test nucleic acid 602 represented by the first bin (e.g. portion 142-1 corresponding to bin 144-1) overlaps the different portion of the test nucleic acid 602 represented by the second bin (e.g. portion 142-2 corresponding to bin 144-2) (244).

[00239] In some embodiments, the structural variation is a translocation of 50 consecutive bases or more, of 50 consecutive bases or more, 100 consecutive bases or more, 250 consecutive bases or more, 500 consecutive bases or more, 1000 consecutive bases or more, 10,000 consecutive bases or more, 20,000 consecutive bases or more, 40 kb consecutive bases or more, 100 kb consecutive bases or more, or 250 kb consecutive bases or more, into the different portion of the test nucleic acid 602 that is represented by the first set (e.g., set 144-1) of sequence reads from the different portion of the test nucleic acid that is represented by the second set (e.g., set 144-2) of sequence reads (246).

[00240] 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 (248). In some such embodiments, the first chromosome is a paternal chromosome and the second chromosome is a maternal chromosome (250). In some such embodiments, the biological sample is human and the first chromosome is chromosome 21, 18, or 13 (252).

[00241] In some embodiments in which the structural variation is deemed to have occurred, the method further comprises treating a subject that originated the biological sample with a treatment regimen responsive to the identified structural variation (254). In some embodiments, this treatment regimen comprises a diet modification (256). For instance, in some embodiments the structural variation is associated with cancer and the diet modification includes avoidance of desserts and other foods that have high sugar content. In some

embodiments, the treatment regimen comprises application of a pharmaceutical composition that inhibits or augments a biological pathway associated with the structural variation (258). Nonlimiting examples of such biological pathways are found in the KEGG database. *See* Kanehisa and Goto, 2000, "KEGG: Kyoto Encyclopedia of Genes and Genomes," Nucleic Acids Research 28, 27-30, which is hereby incorporated by reference herein in its entirety.

[00242] Accurately detecting structural variation relies on the ability of sequence aligners to map reads uniquely and confidently on the genome. In some instances this is not the case, even with state-of-the-art aligners such as BWA. *See* Heng and Durbin, 2010, "Fast and accurate long-read alignment with Burrows-Wheeler transform, "Bioinformatics, 26(5):589-95, which is hereby incorporated by reference for disclosure on BWA. For example, consider two highly homologous regions of the test nucleic acid 602, i.e. two regions with a large degree of sequence similarity. Sequence reads 128 coming from either region would align reasonably well to both regions of the test nucleic acid 602. Sequencing errors, combined with natural sequence variation, might make sequence reads 128 from such a region look like they came from its homologous partner. This can lead to a spurious barcode 132 sharing between homologous regions. Algorithms that rely on read-pair information suffer from the same problem. To be able to detect such false positive structural variants, in some embodiments a step of local realignment of sequence reads 128 around the detected breakpoints is performed. Sequence reads 128 that align with similar quality to both breakpoints of a called structural variant are discarded as spurious.

[00243] Accordingly, in some embodiments in which the metric is deemed to satisfy the threshold criterion, the method further comprises aligning each respective sequence read 128 in the number of respective sequence reads to the subset 142 (e.g., subset 142-1) of the test nucleic acid 602 corresponding to the first set of sequence reads (e.g., set 144-1) using a local realignment procedure that is more accurate than genome-wide alignment with algorithms such as BWA. Then, a first alignment quality is determined for each respective sequence read in the number of respective sequence reads against the subset of the test nucleic acid corresponding to the first set of sequence reads (e.g., subset 142-1) based on the aligning. In such embodiments, the method further comprises aligning each respective sequence read 128 in the number of respective sequence reads with the subset of the test nucleic acid 142 (e.g., subset 142-2) corresponding to the second set of sequence reads. Moreover, a second alignment quality is determined for each respective sequence read in the number of respective sequence reads against the subset of the test nucleic acid corresponding to the second set of sequence reads

(e.g, subset 142-2) based on the aligning. The purpose of such alignments is to eliminate from the number of respective sequence reads those sequence reads having alignment scores that do not discriminate between the first subset (e.g., subset 141-1) and the second subset (e.g., subset 142-2). In other words, those sequence reads in which first alignment quality and the second alignment are similar are eliminated thereby resulting in a reduced number of sequence reads. With this reduced number of sequence reads, process 234 is repeated (260). A probability or likelihood that the number of sequence reads (now reduced) is attributable to chance is determined by comparison of a metric based upon the number of sequence reads 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. In some embodiments, this metric is defined in equation (1) above.

[00244] In some embodiments, a blackout list is maintained. The blackout list comprises a plurality of blackout regions of the test nucleic acid 602. In some such embodiments, the determining process 234 further comprises eliminating a sequence read 128 from the number of respective sequence reads when the first portion 130 of the sequence read overlaps a blackout region in the plurality of blackout regions (262). In some such embodiments, the determining process 234 further comprises eliminating a sequence read 128 from the number of respective sequence reads only when the first portion 130 of the sequence read is completely within a blackout region in the plurality of blackout regions.

[00245] In some embodiments, a white list is maintained. The white list comprises a plurality of regions of the test nucleic acid 602. In some such embodiments, the determining process 234 further comprises eliminating sequence reads 128 from the number of respective sequence reads when the first portion 130 of the sequence read does not overlap a white list region in the plurality of white list regions). In some such embodiments, the determining process 234 further comprises eliminating a sequence read 128 from the number of respective sequence reads only when the first portion 130 of the sequence read is completely outside all white list regions in the plurality of white list regions.

[00246] **Part B. Additional Embodiment for Structural Variation.** Referring to Figure 7, there is disclosed additional methods for detecting structural variation in sequencing data obtained from a single biological sample. Sequence reads are obtained, each comprising a portion corresponding to a subset of the test nucleic acid and a portion encoding an barcode that

is independent of the sequencing data. Bin information is obtained. Each bin represents a different portion of the sample nucleic acid. Each bin corresponds to a set of sequence reads in a plurality of sets of sequence reads formed from the obtained sequence reads such that each sequence read in a respective set of sequence reads corresponds to a subset of the test nucleic acid represented by the bin corresponding to the respective set. Binomial tests identify bin pairs having more sequence reads with the same barcode in common than expected by chance. Probabilistic models determine structural variation likelihood from the sequence reads of these bin pairs.

[00247] Figure 7 is a flow chart illustrating a method of determining a likelihood of a structural variation occurring in a test nucleic acid obtained from a single biological sample (702). In some embodiments, the method takes place at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors in accordance with some embodiments (704).

[00248] *Obtaining a plurality of sequence reads.* In accordance with the disclosed systems and methods, a plurality of sequence reads 128 is obtained (706). Methods for obtaining sequence reads are disclosed in Figure 7 at elements 706 through 720. Furthermore, any of the methods of obtaining sequence reads disclosed in Part A, Structural Variations, (e.g., elements 206 through 220 of Figure 2) as described above may be used and for brevity are not repeated here.

[00249] *Obtaining bin information.* In accordance with the disclosed systems and methods, a plurality of bins 140 is obtained (722). Methods for obtaining bins are disclosed in Figure 7 at elements 722 through 730. Furthermore, any of the methods of obtaining binds disclosed in Part A, Structural Variations, (e.g., elements 222 through 230 of Figure 2) as described above may be used and for brevity are not repeated here.

[00250] In some embodiments, to call large-scale structural variants, the test nucleic acid (e.g., the genome) is binned into 10 kb windows (hereinafter referred to as bins 140) and the unique barcodes 132 of Q60 reads within each bin 140 is counted. In some embodiments, to call large-scale structural variants, the test nucleic acid (e.g., the genome) is binned into bins 140 of a predetermined size (e.g., 5 kb, 10 kb, 20 kb, 40 kb and the unique barcodes 132 of Q40 reads, Q50 reads, Q60 reads or Q70 reads within each bin 140 are counted.

[00251] In some embodiments, each respective bin 140 in the plurality of bins represents a different portion of the test nucleic acid 602. The bin information identifies, for each

respective bin in the plurality of bins, a set of sequence reads 128 in a plurality of sets of sequence reads. Figure 5 illustrates. In Figure 5, a test nucleic acid 602 is depicted. Further shown in Figure 5, each respective bin 140 in a plurality of bins represents a different portion 142 of the test nucleic acid 602. Further, the bin information identifies, for each respective bin 140 in the plurality of bins, a set 144 of sequence reads 128 corresponding to the respective bin.

[00252] Continuing to refer to Figure 5, each sequence read 128 in each set 144 of sequence reads in the plurality of sets of sequence reads is in the plurality of sequence reads. That is, each sequence read 128 corresponds to a portion 142 of the test nucleic acid 602 that has been binned.

[00253] Each respective sequence read in each respective set 144 of sequence reads in the plurality of sets of sequence reads has a respective first portion 130 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 (722). For example, referring to set 144-1 illustrated in Figure 5, each sequence read 128-1-1, ..., 128-1-K, includes a first portion 130 that corresponds to at least a subset of the region 142-1 of the test nucleic acid 140-1 corresponding to bin 140-1. It will be appreciated from the schema and constructs of Figure 3, that in typical embodiments, the first portions 130 of sequence reads 128-1-1, ..., 128-1-K of set 144-1 are of different lengths and are displaced, yet overlapping, with respect to each other. In some embodiments, such first portions 130 may include some nucleotide positions from adjacent regions 142 of the test nucleic acid 602. In some embodiments the first portions 130 only represent a subset of the region 142 of the test nucleic acid 602 of the bin 140.

[00254] In some embodiments, the region 142 of the test nucleic acid corresponding to a bin 140- is substantially longer than any given first portion 130 in the set 144 of sequence reads 128 within the bin. Moreover, in some embodiments, a given set 144 of sequence reads 128 includes five or more different barcode sequences 132, ten or more different barcode sequences 132, fifteen or more different barcode sequences 132, twenty or more different barcode sequences 132, twenty-five or more different barcode sequences 132, thirty or more different barcode sequences 132, thirty-five or more different barcode sequences 132, forty or more different barcode sequences 132, forty-five or more different barcode sequences 132, or fifty or more different barcode sequences 132.

[00255] In some embodiments, two or more, three or more, four or more, five or more, six or more, seven or more, eight or more, nine or more, ten or more, or eleven or more sequence reads 128 in a bin 144 have the same barcode sequences 132. In some embodiments, each sequence read 128 in a bin 144 has a different barcode sequence 132.

[00256] In some embodiments, each bin 140 in the plurality of bins represents at least 20 kbp, at least 50 kbp, at least 100 kbp, at least 250 kbp, or at least 500 kbp of the test nucleic acid (724). In other words, referring to Figure 5, in some embodiments, the portion 142-1 of test nucleic acid 602 represented by bin 140-1 is at least 20 kbp, at least 50 kbp, at least 100 kbp, at least 250 kbp, or at least 500 kbp.

[00257] In some embodiments, there is no overlap between each different portion of the test nucleic acid represented by each respective bin in the plurality of bins (726). Such an embodiment is illustrated in Figure 5. In such embodiments it is possible for overlap in the first portions 130 of the sequence reads 128, in the corresponding bin 144, with neighboring regions 142. In some embodiments, there is no overlap between each different portion 406 of the test nucleic acid represented by each respective bin in the plurality of bins (726), but there is some overlap in the first portions 130 of the sequence reads 128 in the corresponding bin 144 with neighboring portions 406.

[00258] In some embodiments, each respective sequence read in each respective set of sequence reads in the plurality of sequence reads has a respective first portion 130 that corresponds to a subset of the test nucleic 602 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 (728). In other words, using bin 144-1 of Figure 5 as an example, in some embodiments, each first portion 130 of each sequence read 128-1-1,... 128-1-k, includes a full copy of regions 142-1.

[00259] In some embodiments, the plurality of bins comprises 20 or more bins, 100 or more bins, 1000 or more bins, 10,000 or more binds, 100,000 or more bins or 1,000,000 or more bins (730).

[00260] *Identify bins that have common barcodes.* In the disclosed systems and methods, there is identified, from among the plurality of bins 140, a first bin and a second bin that correspond to portions of the test nucleic acid that are nonoverlapping. The first bin is represented by a first set of sequence reads 144 in the plurality of sequence reads and the second bin is represented by a second set of sequence reads 144 in the plurality of sequence

reads (732). In some embodiments, each such set of sequence reads 144 comprises 10 or more sequence reads 128, 20 or more sequence reads 128, 30 or more sequence reads 128, or 100 or more sequence reads 128.

[00261] In some embodiments, the first bin (140) and the second bin (140) are at least a predetermined number of kilobases apart on the test nucleic acid (734). In some embodiments, the first bin and the second bin are at least 5 kilobases, at least 25 kilobases, at least 50 kilobases, at least 50 kilobases, or at least 100 kilobases apart on the test nucleic acid (736).

[00262] In some embodiments, a first value that represents a numeric probability or likelihood that the number of barcodes 132 common to the first set 144 and the second set 144 is attributable to chance is computed (738). In some embodiments, a binomial test to compute the first value (740). For instance, in some embodiments a binomial test is used to find all pairs of bins that are at least a predetermined distance (e.g., 50 kb or on different chromosomes) apart and share more barcodes 132 than what would be expected by chance (e.g., using a p-value cutoff of 10^{-15} without any multiple hypothesis correction). Advantageously, it has been discovered that such a cutoff is loose enough to include all interesting regions of potential structural variation. In some embodiments, the binomial test has the form:

$$p = 1 - P_{Binom}\left(n; {}^{n_1 n_2}/_{B}\right)$$

where, 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 (742). In some embodiments, the single biological sample is human, the test nucleic acid 602 is the genome of the biological sample, and the first value satisfies the predetermined cut-off value when the first value is 10^{-12} or less, when the first value is 10^{-13} or less, when the first value is 10^{-14} or less, or when the first value is 10^{-15} or less (746).

[00263] In some embodiments, the number of pairs of bins that are compared here is roughly in the order of 10^{10}. In order to perform these comparisons efficiently, in some embodiments the set of barcodes 132 in each bin 140 is coded as non-zero entries in a (very sparse) matrix and sparse matrix multiplications are used to identify regions that overlap (748). This allows for the quick identification of candidate bins 140 for further structural variation study. However, the disclosed binomial tests generate a very large number of false positives in

some instances since it does not account for many aspects of the system, such as the length distribution of the test nucleic acid fragments 306 and the variation in the amplification rate across GEMs. In some embodiments, an identity of the first and second bin is determined using sparse matrix multiplication of the form:

[00264]

$$V = A_1^T A_2,$$

where A_1 is a first B x N_1 matrix of barcodes that includes the first bin, A_2 is a second B x 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 is the transpose of matrix A_1 (750). In some embodiments, 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 (752). In some embodiments, the first bin and the second bin are each 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 (754).

[00265] In a second pass, a probabilistic approach is used to clean up the initial candidate list of bins 140 (744). In some embodiments, an estimate of the set of library molecules is obtained by joining nearby sequence reads 128 (e.g. closer than 30 kb) having the same barcode 132. In the following discussion, the term "fragment" is used to refer to a span of nearby sequence reads 128 with the same barcode 132. Fragments originate from some unobserved molecules (that may be longer than the observed fragments). Based on the set of fragments, quantities such as read generation rate (sequenced reads per bp) of individual GEMs, the number of molecules inside each partition 304, and the molecule length distribution are estimated.

[00266] Given a pair of candidate bins W_1, W_2, the sets of fragments that overlap them are identified and then pairs of fragments in bins W_1 and W_2 with the same barcode 132 are identified. Such pairs of bins are potentially evidence for structural variation, since they suggest that the same molecule might have spanned two relatively distant loci of the genome. Accordingly, in some embodiments, responsive to a determination that the first value (described above) satisfies a predetermined cut-off value, for each barcode that is common to a first bin and a second bin 140, there is obtain a fragment pair thereby obtaining one or more fragment pairs. Each fragment pair in the one or more fragment pairs (i) corresponds to a

different barcode that is common to the first bin and the second bin and (ii) consists of a different first calculated fragment and a different second calculated fragment. In some embodiments the one or more fragment pairs is a single fragment pair. In some embodiments the one or more fragment pairs is 2 or more fragment pairs, 5 or more fragment pairs, 10 or more fragment pairs or 100 or more fragment pairs.

[00267] 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 128 in the plurality of sequence reads having the barcode corresponding to the respective fragment pair. Each sequence read 128 in the respective first subset of sequence reads is within a predefined genetic distance (e.g., 30 kb) of another sequence read 128 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. 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. Each sequence read in the respective second subset of sequence reads is within a predefined genetic distance (e.g., 30 kb) 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 (744).

[00268] To quantify the above identified evidence that a pair of bins exhibit structural variation a likelihood ratio score is computed in some embodiments (756). In some embodiments, a respective likelihood is computed 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 602. 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. 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 (756).

[00269] In some embodiments, the computed likelihood is computed as a ratio score between the probability of occurrence of the first model and the probability of occurrence of the second model (758). Referring to element 760 of Figure 7F, in some embodiments, the computed likelihood is computed as:

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

[00270] Since fragments with different barcodes are independent, this score decomposes to a product of terms with one term for each of the pairs of fragments with the same barcode b:

$$\frac{P(r_1, r_2, l_1, l_2, d \mid SV; a_b)}{P(r_1, r_2, l_1, l_2, d \mid no\ SV; a_b)} \quad (1)$$

[00271] where:

[00272] r_1, r_2 are the number of sequence reads (128) on each of the two fragments,

[00273] l_1, l_2 are the observed lengths of the two fragments,

[00274] d is the distance between the two fragments, and

[00275] a_b is the rate (reads/bp) of the GEM/barcode b.

[00276] In other words, LR is equal to a product of one or more terms (e.g., a plurality of terms), where each respective term (i) represents a respective fragment pair in the one or more fragment pairs and (ii) has the form:

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

where 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 (760).

[00277] Referring to element 762 of Figure 7G, in some embodiments, the two candidate fragments might have originated from the same molecule or from different molecules, therefore:

$$P(r_1, r_2, l_1, l_2, d \mid no\ SV; a_b) = \\ P(r_1, r_2, l_1, l_2, d \mid same\ molecule, no\ SV; a_b) P(same\ molecule \mid no\ SV) + \\ P(r_1, r_2, l_1, l_2, d \mid different\ molecules, no\ SV; a_b) P(different\ molecules \mid no\ SV) \quad (2)$$

[00278] The probability given that the fragments originated from different molecules is:

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

where $P_{frag}(r, l; a_b)$ is the probability of observing r sequence reads 128 from a molecule of unknown length such that the reads span an observed length of l. Referring to element 764 of Figure 7G, assuming that the sequence reads 128 are generated from a Poisson process with constant rate across the genome, we get:

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

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 is m.

[00279] Referring to element 766 of Figure 7G, the probability given that the fragments came from the same molecule can be computed in a similar way as:

$$P(r_1, r_2, l_1, l_2, d \mid same\ molecule, no\ SV; a_b) = \\ \sum_{m:m \geq l_1 + l_2 + d} (m - l_1 - l_2 - d)\, P_p(r_1 - 2; a_b l_1) P_p(r_2 - 2; a_b l_2) P_p\big(0; a_b(m - l_1 - l_2)\big) a_b^4 P_L(m)$$

where m is 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, $P_p(r_2 - 2; a_b l_2)$ and r_2, $P_p\big(0; a_b(m - l_1 - l_2)\big)$ are each probability mass functions of a Poisson distribution with parameter b, and $P_L(m)$ is a pre-estimated probability that the true common molecule length is m.

[00280] In the presence of a structural variation, the likelihood is similar to equation (2). However, in this case, there is an additional unknown, namely the exact position of the

breakpoints with respect to the observed fragments. For example, assume that there was a deletion between positions 100,000 and 200,000 of chr1 and that the observed fragments span the regions 85,000-90,000 and 210,000-230,000. If the exact breakpoints were known, the previous calculations could be used with d set to 10 kb + 10kb = 20 kb. Since the position of the true breakpoints (and therefore the true distance between the observed fragments) is unknown, it is integrate out in some embodiment. Rather, in some embodiments, to simplify calculations, a rough estimate of d is obtained by computing the maximum extent d' such that $P_p(0; a_b d') \geq 0.75$. Then, d is set to 2d' in the equations above in order to compute $P(r_1, r_2, l_1, l_2, d \mid SV; a_b)$. In other words, in some embodiments $P(r_1, r_2, l_1, l_2, d \mid SV; a_b) =$

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

where

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,

$P(r_1, r_2, l_1, l_2, 2d' \mid DM, SV; a_b) = P_{frag}(r_1, l_1; a_b)P_{frag}(r_2, l_2; a_b)$, 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. Here $P_{frag}(r_1, l_1; a_b)$ and $P_{frag}(r_2, l_2; a_b)$ are each computed as

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

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

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

where 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, $P_p\big(0; a_b(m - l_1 - l_2)\big)$ 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.

[00281] In some embodiments, to get the priors in equation (2) above, it is assumed that the probability of having a second molecule at any given position of the genome is M/G where M is the total number of input molecules and G is the size of the genome. These priors are actually independent of whether there is a structural variation or not so P(same molecule j no SV) = P(same molecule) and P(different molecules) = 1 – P(same molecule).

[00282] The above discussion for equations (1) and (2) assumes whole-genome sequencing data. In the case of targeted sequencing, the composition of the target set is taken into consideration. In some embodiments, the assumption is made that the off-target regions generate reads following a similar Poisson process as the target regions, but with a different rate. In particular, let b_t be the fraction of reads on target and g_t be the fraction of the genome that is covered by the target regions. If a_b the Poisson rate of target regions, then the rate of off-target regions is:

$$\tilde{a}_b = \frac{1 - b_t}{1 - g_t} \frac{g_t}{b_t} a_b$$

[00283] The probability of observing r reads from a region that contains l_t bp of targets and l_n bp of off-target regions is:

$$P_{mixture}(r, l_1, l_2; a_b, \tilde{a}_b) = \sum_{n=0}^{n=r} P_p(r - n; a_b l_t) P_p(n; \tilde{a}_b l_n)$$

[00284] The probability of observing r reads from a molecule of unknown length that spanned an observed length of $l = l_t + l_n$ is:

$$\sum_{m:m\geq l} P_{mixture}(r-2, l_t, l_n; a_b, \tilde{a}_b) a_b^2 P_L(m) \sum_{f\in offsets} P_{mixture}(0, d_{f_t}, d_{f_n}; a_b, \tilde{a}_b)$$

where the inner sum is taken over all $m - l$ offsets of the unobserved molecule with respect to the observed fragment, and d_{f_t} and d_{f_n} are the bases on and off-target for the corresponding offset. To simplify calculations, for a given value of m, the average fraction of bases on and off-target is computed across all offsets and the assumption is made that all offsets have the same target composition in some embodiments. The rest of the probabilities needed to compute (1) are adjusted in a similar way from the WGS case.

[00285] In some embodiments, all probabilities were computed in log-space to avoid underflows. In some embodiments, a log-likelihood ratio cutoff of 200 is used. This cutout has been empirically found to result in high-quality calls with very low false positive rates (after the filtering steps described below).

[00286] In some embodiments, the structural variation is an insertion or deletion of 50 consecutive bases or more, 500 consecutive bases or more, or 5000 consecutive bases or more, into the different portion of the test nucleic acid that is represented by the first set of sequence reads (768). In some embodiments, the different portion of the test nucleic acid of the first bin overlaps the different portion of the test nucleic acid represented by the second bin (770). In some embodiments, at least 50, 80 percent, or 95 percent of the different portion of the test nucleic acid of the first bin overlaps the different portion of the test nucleic acid of the second bin (772). In some embodiments, the structural variation is a translocation 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 from the different portion of the test nucleic acid that is represented by the second set of sequence reads (774).

[00287] 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 (776). In some embodiments, the first chromosome is a paternal chromosome and the second chromosome is a maternal chromosome (778). In some embodiments, the biological sample is human and the first chromosome is any of chromosome 1-21 (780).

[00288] 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 (782). In some embodiments, the treatment regimen comprises a diet modification (784). In some embodiments, the treatment regimen comprises application of a pharmaceutical composition that inhibits a biological pathway associated with the structural variation (786).

[00289] *Refining breakpoints using short read information.* In some embodiments, after obtaining breakpoint windows using the approach described above, this information from read pairs and split reads is used to further refine the breakpoint locations. For each called structural variant, all read pairs and split reads within the called breakpoint bins are selected. Then a probabilistic approach similar to Layer *et al,* 2014, "LUMPY: A probabilistic framework for structural variant discovery," *Genome Biology*, *15*(6), R84, doi:10.1186/gb-2014-15-6-r84, which his hereby incorporated by reference herein in its entirety, is used to infer the breakpoint loci based on the combined evidence from all selected read pairs and split reads. In order to avoid false positives, inference of the exact breakpoint loci was only attempted when there were at least four read pairs and split reads supporting the call.

[00290] *Filtering calls based on gaps and segmental duplications.* In some embodiments, a blackout list is maintained (788), the blackout list comprising a plurality of blackout regions of the test nucleic acid, and the method further comprises eliminating a sequence read, prior to the identifying, from the number of respective sequence reads when the first portion of the sequence read overlaps a blackout region in the plurality of blackout regions (788). For instance, in some embodiments, structural variant calls whose breakpoints overlap different copies of the same segmental duplication (using the Segmental Duplication track from the UCSC browser) were excluded. Structural variation is enriched in such regions (Mills *et al.*, 2011, "Mapping copy number variation by population-scale genome sequencing," *Nature*, *470*(7332), 59–65, doi:10.1038/nature09708, which is hereby incorporated by reference in its entirety), so some of these calls might represent true events. However, a large fraction of calls in regions of structural variation have been observed to be the result of the inability of aligners to properly resolve repetitive regions, since a small amount of variation is sufficient to make reads map uniquely and with high mapping quality to one or the other copy of the segmental duplication. Structural variant calls within a predetermined distance (e.g. 10 kb) from gaps (using the gaps track from the UCSC browser) or from new sequence introduced

in hg38 (using the hg19 diff track from the UCSC browser) were also excluded . The rationale was that such calls are not interesting since they probably represent misassemblies in hg19.

[00291] **Part C. Phasing.** In this section, exemplary methods implemented by some embodiments of phasing sub-module 124 are disclosed. Motivation for phasing the sequence reads 128 into parental haplotypes is illustrated. Phased variants are a more complete representation of the state of a diploid genome. A variety of research has found that phased variants allow for a better understanding of observed phenotypes. See, for example, Tewhey *et al.*, 2011, "The importance of phase information for human genomics," Nat Rev Genet, 12:215-223, which is hereby incorporated by reference in its entirety. For example, in the presence of compound heterozygosity, phasing is required to disambiguate the loss of one or both copies of the gene. Referring to Figure 11, for example, two loss of function mutations in cis leave a functional copy of the gene, whereas two loss of function mutations in trans leaves both copies inactivated. Thus, it is necessary to properly phase sequence reads in order to elucidate the state of a diploid genome.

[00292] *Algorithm for phasing.* In order to phase the variants (e.g., the variants called in Part A or Part B above), the likelihood of sequence read and barcode support for each allele given a phasing configuration is modeled. See, for example, Bansal *et al.*, 2008, "An MCMC algorithm for haplotype assembly from whole-genome sequence data," Genome Res, 18:1336-1346, which is hereby incorporated by reference in its entirety. Then a search for the maximum likelihood phasing configuration is conducted. This algorithm first finds near-optimal local haplotype configurations with beam-search over blocks of adjacent variants. In some embodiments, the blocks of adjacent variants comprise about 40 variants. In some embodiments, the blocks of adjacent variants consist of between 10 and 30 variants, between 20 and 50 variants, between 30 and 60 variants, or more than 60 variants. Neighboring blocks are greedily joined to form a global solution, which is iteratively refined until convergence. The confidence of each phasing decision is the likelihood-ratio between optimal and next-best solutions.

[00293] In particular, a pre-determined set of variant calls is obtained (e.g., from parts A or part B above, a plurality of single nucleotide polymorphisms, from other sources, etc.). Alleles $A_{i;p}$ are labeled, where $i \in 1; \ldots, N$ indexes the variant. In some embodiments, $p \in 0, 1$ is an arbitrary label for the two alleles of the variant. The set of alleles that come from the same parent chromosome is referred to as a haplotype, and are arbitrarily labeled H_0 and H_1. The goal of the phasing algorithm is to determine which allele from each variant came from each

parent chromosome. In some embodiments, the phasing result can be described by a binary variable for each variant $X_i \in 0, 1$ where $X_i = 0$ indicates the $A_{i,0} \in H_0$ and $A_{i,1} \in H_1$ and $X_i = 1$ indicates that $A_{i,0} \in H_1$ and $A_{i,1} \in H_0$.

[00294] 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 error condition represents an error in the zygosity of the underlying variant. This alternative embodiment takes into consideration that standard variant calling algorithms that are relied upon to provide variant calls at positions $A_{i;p}$ between H_0 and H_1 may, in fact, call the zygosity of such positions incorrectly on occasion. The disclosed alternative phasing embodiment $p \in \{0, 1, -1\}$ advantageously allows for exploration of this form of zygosity error during the phasing. The ability to selectively sample for this error state advantageously protects the phasing algorithm from error in the input data that arises, for instance due to error in the sequencing process, weak sequencing signal, and the like.

[00295] Neighboring variants on the test nucleic acid sequence 602 (e.g. genome) are often separated by distances longer than the read-pair length (e.g., sequence read length 128), causing very short phase blocks. As disclosed above, long test nucleic acid fragments 306 (e.g. input fragments) covering a small fraction (0:01 - 0:001) of the test nucleic acid sequence 602 (e.g. genome) are exposed to each barcode 132, so the probability that a barcode 132 contains sequence reads 128 from both haplotypes is small.

[00296] In some embodiments, sequence reads 128 are aligned to the genome of the target organism. Sequence reads 128 are grouped by the attached barcode 132 sequences. Sequence reads 128 with common barcode sequences 132 are partitioned into groups that are likely to have originated from a single test nucleic acid fragment 306 (e.g. genomic input fragment), and thus provide evidence that the alleles covered by the sequence reads 128 came from the same haplotype. In some embodiments, the plurality of barcoded sequence reads is denoted $\vec{\mathbf{O}}$. In some embodiments, each respective sequence read $\vec{\mathbf{O}}_q$ in the plurality of sequence reads, where q is an integer index into $\vec{\mathbf{O}}$, 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.

[00297] In some embodiments, each respective sequence read $\vec{\mathbf{O}}_q$ in the plurality of sequence reads $\vec{\mathbf{O}}$ is $\in \{0, 1, -\}^n$, where (i) each respective label "0" for the respective sequence read $\vec{\mathbf{O}}_q$ assigns a corresponding variant call in $A_{i;p}$ to H_0, (ii) each respective label "1" for the respective sequence read $\vec{\mathbf{O}}_q$ assigns a corresponding variant call in $A_{i;p}$ to H_1, and (iii) each respective label "–" for the respective sequence read $\vec{\mathbf{O}}_q$ indicates that the corresponding variant call in $A_{i;p}$ is not covered. For example, consider the case in which $\vec{\mathbf{O}}_q$ contains 5 of the 10 variant calls in $A_{i;p}$. In this example, $\vec{\mathbf{O}}_q$ will contain five variant calls with values "-" because they are not in the respective sequence read and $\vec{\mathbf{O}}_q$ will contain values for the five other variant calls in $A_{i;p}$. Each value in these five values will be a zero or a one depending on the haplotype assigned to the respective variant call in the sequence read. In some embodiments, such haplotype assignments are obtained for the variant calls in individual sequence reads using conventional haplotype assignment algorithms.

[00298] In some embodiments, to account for possible error in the zygosity of a variant call as described above, each respective sequence read $\vec{\mathbf{O}}_q$ in the plurality of sequence reads $\vec{\mathbf{O}}$ is $\in \{0, 1, -1, -\}^n$, where (i) each respective label "0" for the respective sequence read $\vec{\mathbf{O}}_q$ assigns a corresponding variant call in $A_{i;p}$ to H_0, (ii) each respective label "1" for the respective sequence read $\vec{\mathbf{O}}_q$ assigns a corresponding variant call in $A_{i;p}$ to H_1, (iii) each respective label "-1" for the respective sequence read $\vec{\mathbf{O}}_q$ assigns a corresponding variant call in $A_{i;p}$ to the zygosity error condition (present, but neither H_0 nor H_1), and (iv) each respective label "–" for the respective sequence read $\vec{\mathbf{O}}_q$ indicates that the corresponding variant call in $A_{i;p}$ is not covered.

[00299] In the disclosed methods, a phasing result $\vec{\mathbf{X}}$ is obtained by optimization of haplotype assignments at individual positions i in $A_{i;p}$. In embodiments where each respective sequence read $\vec{\mathbf{O}}_q$ in the plurality of sequence reads $\vec{\mathbf{O}}$ is $\in \{0, 1, -\}^n$, these haplotype assignments are each between H_0 and H_1 at individual positions i for the plurality of sequence reads. In the alternative embodiments where possible error in the zygosity of the position i is to be additionally sampled in the phasing algorithm, each respective sequence read $\vec{\mathbf{O}}_q$ in the plurality of sequence reads $\vec{\mathbf{O}}$ is $\in \{0, 1, -1, -\}^n$, these haplotype assignments are each between H_0, H_1 and H_{-1} at individual positions i for the plurality of sequence reads, where H_{-1} denotes the zygosity error condition above.

[00300] Sequence reads are aligned to a reference genome. Furthermore, sequence reads that have the same barcode are grouped together. In this way, sequence reads with a common barcode are partitioned into groups that are likely to have originated from a single genomic input fragment *f*, and thus provide evidence that the alleles covered by the sequence read came from the same haplotype.

[00301] In some embodiments where each respective sequence read $\vec{\boldsymbol{O}}_q$ in the plurality of sequence reads $\vec{\boldsymbol{O}}$ is $\in \{0, 1, -\}^n$, the probability of the observed sequence reads 128 covering variant *i* from test nucleic acid fragment 306 *f* is computed as:

$$\log P(O_{i,f}|A_{i,p}) = \sum_r \mathbf{1}(S_r = A_{i,p})\left(1 - 10^{-Q_r/10}\right) + \mathbf{1}(S_r \neq A_{i,p})\left(10^{-Q_r/10}\right)$$

where *r* sums over the reads and $\mathbf{1}(S_r = A_{i,p})$ is the indicator function testing whether the rth sequence read 128 S_r matches allele $A_{i,p}$. When the rth sequence read 128 S_r matches allele $A_{i,p}$, the indicator function has a first value (e.g., "1") and is a second value (e.g. "0") ortherwise. The term $\mathbf{1}(S_r \neq A_{i,p})$ is the indicator function that has a first value (e.g., "1") when the rth sequence read S_r from fragment *f* does not match $A_{i,p}$ and is a second value (e.g., "0") otherwise. In some embodiments, the probability assigned is derived from the inverse-Phred transformed quality value of relevant read base Q_r.

[00302] In embodiments where each respective sequence read $\vec{\boldsymbol{O}}_q$ in the plurality of sequence reads $\vec{\boldsymbol{O}}$ is $\in \{0, 1, -1, -\}^n$, that is possible zygosity error at position *i* is additionally sampled, the probability of the observed sequence reads covering variant *i* from fragment *f* is computed as:

$$\log P(O_{i,f}|A_{i,p}) = \sum_r \mathbf{1}(S_r = A_{i,p})\left(1 - 10^{-Q_r/10}\right) + \mathbf{1}(S_r \neq A_{i,p})\left(10^{-Q_r/10}\right)$$

$$+ \mathbf{1}(A_{i,p} = X^-)^{0.5}$$

where,

X^- is H_{-1}, and

$\mathbf{1}(A_{i,p} = X^-)$ is the indicator function that has a first value (e.g., "1") when $A_{i,p}$ is equal to -1 (H_{-1}) and is a second value (e.g., "0") otherwise.

[00303] In embodiments where $\vec{\mathbf{O}}$ is $\in \{0, 1, -\}^n$, the phasing result can be obtained by optimizing an objective function expressed as a maximum likelihood phasing parity vector:

$$\hat{\mathbf{X}} = \underset{\vec{\mathbf{X}}}{\arg\max}\, P(\vec{\boldsymbol{O}}|\vec{\boldsymbol{X}})$$

Here, $\hat{\mathbf{X}}$ is the refined phasing vector while $\vec{\mathbf{X}}$ is the phasing vector result to be inferred and

$$P(\vec{\boldsymbol{O}}|\vec{\boldsymbol{X}}) = \prod_f P(O_{1,f}, \ldots, O_{N,f}).$$

[00304] In embodiments where $\vec{\mathbf{O}}$ is $\in \{0, 1, -1, -\}^n$, the phasing vector can be found by optimizing an overall objective function:

$$\hat{\mathbf{X}} = \underset{\vec{\boldsymbol{X}}}{\text{argmax}}\, P(\vec{\boldsymbol{X}}|\vec{\boldsymbol{O}})\,| = \frac{P(\vec{\boldsymbol{X}})P(\vec{\boldsymbol{O}}|\vec{\boldsymbol{X}})}{C}$$

where,

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

H_{-1} is the condition of zygosity error at position i,

ε_i is an estimate of incurring this form of error at position i, and

$$P(\vec{\boldsymbol{O}}|\vec{\boldsymbol{X}}) = \prod_f P(O_{1,f}, \ldots, O_{N,f}).$$

[00305] In some embodiments, ε_i is a function of the type of variant at position i. For instance ε_i is given a first value if the variant at position i arises through genetic insertions or deletion, and another value if the variant at position i arises by other means (e.g., single nucleotide polymorphisms).

[00306] The data from a test nucleic acid fragment 306f come from one of three cases. The first two cases are that the alleles present are only from H_0 or only from H_1. These cases are the typical cases and have a high prior probability, governed by the fraction of the test nucleic acid fragment 306 (e.g., genome) present in each partition 304. The third case is that input DNA from both haplotypes was present at the locus, so either allele is equally likely to be observed:

[00307] $P(O_{1,f}, \ldots, O_{N,f}|\mathbf{X},\ H_f = 0) = \prod_i P(O_{i,f}|A_{i,X_i})$

[00308] $P(O_{1,f}, \ldots, O_{N,f}|\mathbf{X},\ H_f = 1) = \prod_i P(O_{i,f}|A_{i,1-X_i})$

[00309] $P(O_{1,f}, \ldots, O_{N,f} | \mathbf{X}, H_f = M) = \prod_i 0.5$

[00310] The above three equations give the probability of the observed sequence reads 128 from fragment 306 f at variant location i, X_i, and fragment haplotype H_f. Observations are independent given the variant party and fragment haplotype. The prior probability of the third case is $\propto$ - the probability that a partition contains both haplotypes at a locus. The overall likelihood can be computed by summing over the three cases:

$$P(O_{i,f}|X_i) = \frac{(1-\propto)}{2}\left(P(O_{i,f}|X_i, H_f = 0) + P(O_{i,f}|X_i, H_f = 1)\right) + \propto P(O_{i,f}|H_f = M)$$

$$P(O_{i,f}, \ldots, O_{N,f}|\mathbf{X}) = \frac{(1-\propto)}{2}\left(\prod_i P(O_{i,f}|A_i, X_i) + \prod_i P(O_{i,f}|A_i, 1 - X_i)\right) + \propto \prod_i 0.5$$

Each position in the test nucleic acid 602 (e.g., genome) with a heterozygous variant is given a variant index $v \in 1, \ldots, N$, and each of the two alleles are arbitrarily assigned an index $a \in 0, 1$.

[00311] *Optimization.* In some embodiments, one of the overall objective functions presented above is optimized using a hierarchical search over the phasing vector $\vec{\mathbf{X}}$. The objective function for the embodiment of $\vec{\mathbf{O}} \in \{0, 1, -\}^n$ is illustrated in Figure 12. In some embodiments, $\vec{\mathbf{X}}$ is broken up into local chunks of $n \approx 40$ variants and the relative phasing of the block is determined using beam search over the assignments of $X_k, X_{k+1}, \ldots, X_{k+n}$, where k is the first variant in the local block. In some embodiments local chunks consists of between 10 and 20 variants, between 20 and 30 variants, between 30 and 40 variants, between 35 and 45 variants, between 40 and 50 variants, or more than 50 variants. Beam search is described at, for example, en.wikipedia.org/wiki/Beam_search, which is hereby incorporated by reference herein in its entirety. See also Figure 13 which illustrates the beam search. In Figure 13, at each stage of the beam search, only the k highest scoring partial solutions are retained. In Figure 13, k=6. In Figure 13, nodes marked with a red X are not in the top k, so they are not explored further.

[00312] The relative phasing of neighboring blocks is found greedily, yielding a candidate phasing vector $\vec{\mathbf{X}}$. Finally $\vec{\mathbf{X}}$ is iteratively refined by swapping the phase of individual variants. Convergence of this refinement yields an estimate of the optimal phasing configuration $\vec{\mathbf{X}}$. In some embodiments, the confidence of each phasing decision is the likelihood-ratio between optimal and next-best solutions. In some embodiments, an estimate of the accuracy of the phasing configuration is determined by computing the likelihood ratio

between the optimal configuration $\vec{\mathbf{X}}$ and some alternate configuration $\mathbf{X}_{alt}$ by computing the likelihood ratio between the hypotheses. In some embodiments, the confidence is then reported as a Phred-scaled quality value:

$$Q(\mathbf{X}_{alt}) = -10\log_{10}\left(\frac{\mathrm{P}(\mathbf{O}|\mathbf{X}_{alt})}{\mathrm{P}(\mathbf{O}|\widehat{\mathbf{X}})}\right)$$

[00313] In some embodiments, there are two classes of errors considered: short switch errors and long switch errors. Short switch errors are single variants that are assigned the wrong phasing in an otherwise correctly phased region. To measure the short switch confidence of variant i, X_i is flipped to form X_{alt}. When the short switch confidence is low, the variant is marked as not phased in the output, rather than reporting a phasing call likely to be erroneous.

[00314] Long switch errors occur when two neighboring block of variants $\ldots, X_{i-2}, X_{i-1}$ and $X_i, X_{i+1}, \ldots$ are correctly phased internally, but have the wrong relative phasing between the two blocks. In this case, a long switch error is called at position i. The long switch confidence at position i is tested by inverting the phase of X_j for all $i >= j$. When the long switch confidence falls below a threshold, a new phase block is started and variants in different phase blocks are not called as phased with respect to one another.

[00315] **Example 1.**

[00316] *Sample preparation.* Figure 8 provides an example of sample preparation in accordance with an exemplary embodiment of the present disclosure. The GemCode Platform massively partitions and barcodes DNA, producing sequencing-ready libraries with >100,000 unique barcodes. Custom algorithms use this barcode information to map reads back to original, long molecules of DNA, creating linked reads that span many tens of kilobases. Long template molecules from ~1 ng of gDNA are randomly distributed across >100,000 barcoded partitions, giving <10 fg (< 0.3% of the genome) per partition. Each partition carries primers with barcodes that are constant within a partition, but distinct across partitions. An amplification reaction creates barcoded short read library fragments within each partition. The resulting library is compatible with standard exome capture, while preserving long range linkage information. In particular, the resulting libraries are sample-indexed and can be whole genome sequenced, or used as input into one of many commercially available hybrid capture platforms to generate targeted sequencing libraries. The powerful nature of this new linked

read data type means that long-range information is retained after target enrichment, even though intronic regions are not selected or sequenced.

[00317] *Linked reads.* Referring to Figure 9, after mapping, reads from the same locus with the same barcode form a set of 'linked reads' that were generated from a single input molecule. In exome data, linked reads span multiple target regions, allowing phasing and structural variant calling from exome data. In Figure 9, each dot represents NGS read-pair (e.g., a sequence read). Groups of such sequence reads joined by a horizontal line in Figure 9 share a common barcode from a pool of over 100,000 barcodes used to sequence the test nucleic acid. Moreover, the horizontal lines are arranged into respective groups 902 and 904, each such group representing the two parental haplotypes of the diploid organism from which the test nucleic acid was obtained. Further illustrated in Figure 9 are the locations 906 of exons within the genome of the test nucleic acid. Each such exon is illustrated as a vertical bar within the figure.

[00318] *Sequencing results.* Referring to Figure 10, the libraries of sequence reads from target nucleic acid obtained from two different organisms (respectively labeled NA12878 WGS and NA12878 WES) achieve low PCR duplication rates and high mapped fractions, despite requiring only 1 ng of input material. Each input molecule generates a group of tens of linked reads.

[00319] *Phasing.* The disclosed systems and methods for phasing (column NA12878 WES) were compared with phased genomes generated from pedigree phasing (Cleary *et al.*, 2014, "Joint variant and de novo mutation identification on pedigrees from high-throughput sequencing data," *J Comput Biol*, **21**:405-419, which is hereby incorporated by reference in its entirety) (NA12878 WGS), sperm direct haplotyping (Kirkness *et al.*, 2013, "Sequencing of isolated sperm cells for direct haplotyping of a human genome," *Genome Res*, **23**:826-832, which is hereby incorporated by reference in its entirety) (HuRef1 WGS), and fosmid pooling (Kitzman *et al.*, 2011, "Haplotype-resolved genome sequencing of a Gujarati Indian individual." *Nat Biotechnol*, **29**:59-63, which is hereby incorporated by reference in its entirety) (NA20847 WGS), by phasing existing variant calls. These results are summarized in Figure 14. High concordance with previous results is found and MB-scale phase blocks at modest coverage levels is consistently achieved. With exome sequencing, 96% of the genes shorter than 100 kb are contained in single phase blocks using the systems and method of the present disclosure.

[00320] *Structural Variant Determination - results of large scale structural variant detection - WGS case study CEPH trio.* In this example, all called deletions were found in previous studies. Eight deletions previously reported by three other studies, Kidd *et al.*, 2010, "A human genome structural variation sequencing resource reveals insights into mutational mechanisms," Cell, 143:837-47; Layer *et al.*, 2014, "LUMPY: a probabilistic framework for structural variant discovery," Genome Biol, 15:R84; and Mills *et al.*, 2011, "Mapping copy number variation by population-scale genome sequencing," Nature, 470:59-65, each of which is hereby incorporated by reference, were selected for further validation using OS-seq (Myllykangas *et al.*, 2011, "Efficient targeted resequencing of human germline and cancer genomes by oligonucleotide-selective sequencing," *Nat Biotechnol*, **29**:1024-1027, which is hereby incorporated by reference, phasing information, LOH, and depth of coverage. Tables 1 and 2 provide the results.

[00321] Table 1

	Detected by the SV algorithm of the present disclosure	Not detected by the SV algorithm of the present disclosure
Validated	5	0
Not Validated	0	3

[00322] All calls in the child were consistent with Mendelian inheritance (see structural variant phasing section). Trio information was not used for phasing the structural variants.

[00323] Table 2

	NA12878 (mother)	NA12877 (father)	NA12882 (child)
Number of structural variants greater than 50 kb	20	9	13
Percentage of het structural variants phased	81.2%	88.9%	100%

[00324] *Structural Variant Determination – Exome case study: H2228.* The lung cancer cell line H2228 has two validated gene fusions with breakpoints inside long introns (Choi *et al.*, 2008, "Identification of novel isoforms of the EML4-ALK transforming gene in non–small cell lung cancer," Cancer Res, 68:4971-4976; and Jung *et al.*, 2012, "Discovery of ALK-PTPN3

gene fusion from human non-small cell lung carcinoma cell line using next generation RNA sequencing," Genes Chromosomes Cancer, 51:590-597, each of which is hereby incorporated by reference. These breakpoints can be detected with linked reads despite the absence of overlapping capture targets. Table 3 provides the results.

[00325] Table 3

	Rank	Intron Lengths
ALK-EML4	3	2K, 16K
ALK-PTPN3	8	2K, 6K

[00326] *Structural Variant Phasing – Approach taken in Example 1.* During phasing the haplotype of each template molecule is computed concurrently with the variants. By associating each barcode supporting a structural variant with the haplotype it derived from, structural variants can be phased. Phasing structural variants provides a powerful confirmation method – false positive structural variants are unlikely to phase to a single haplotype. A p-value for the association of a structural variant to a haplotype is computed using a binomial test. Figure 15 illustrates a phased structural variant call. Splitting linked reads by the haplotype assignment of each input molecule reveals the phased structure of the structural variant. Vertical bars 1504 indicate the breakpoints of the structural variant call.

[00327] Figure 16 illustrates called deletions in NA12878. Barcodes supporting the structural variant calls are phased to a single haplotype. The three structural variants at the bottom of the table represent short-read false positive calls, correctly filtered by the disclosed algorithm. False positive calls do not phase uniquely or follow Mendelian inheritance patterns.

[00328] **Example 2.** This example describes technologies relating to detecting structural variants and phasing haplotypes from cancer exome sequencing. Specifically incorporated into Example 2 is United States Provisional Patent Application 62/120,330, entitled "Detecting Structural Variants and Phasing Haplotypes from Cancer Exome Sequencing Using 1ng Dna Input," filed February 24, 2015, which is hereby incorporated by reference in its entirety.

[00329] Structural changes, and particularly gene fusions, are known driving mutations in many cancers. In many cases they have also proven to be effective drug targets. However, detecting fusions is a challenge with existing short-read sequencing technologies, particularly when using exon target enrichment approaches to achieve the ultra-deep coverage required to

sensitively detect important mutations in heterogeneous cancer samples. Accordingly, what is needed in the art are improved systems and methods for detecting important mutations in heterogeneous cancer samples.

[00330] Technical solutions (e.g., computing systems, methods, and non-transitory computer readable storage mediums) for detecting important mutations in heterogeneous cancer samples are provided in this Example. In particular, a new platform from 10X Genomics is used to address multiple challenges in cancer sequencing: (i) detecting gene fusion events from exome sequencing data, (ii) obtaining complex sequencing libraries from ~1ng DNA input, (iii) phasing single nucleotide polymorphisms and structural variants, and (iv) resolving complex rearrangements.

[00331] In Example 2, sample preparation and determination of linked reads was as given in Example 1 with reference to Figures 8 and 9. Cancer cell line and matched normal DNA was acquired from ATCC and size selected for fragments ≥ 20 kb using the BluePippin from Sage Science. Sequencing libraries were prepared using ~1ng of gDNA input into the GemCode Platform. 5mg of library was used for exome capture using the AGILENT SURESELECT Human All Exon V5+UTRs with IDT xGen® Universal Blocking Oligos. Libraries were sequenced on an ILLUMINA HiSeq 2500 using a paired-end 2x98 run. Sequencing results were analyzed and visualized using the GEMCode Software Suite. Figure 17 compares the PCR duplication rate between the methodology of the disclosed systems and methods and that of TruSeq + AGILENT SS data at various input quantities. Figure 18 compares variant calling performance between the 10X methodology and that of TruSeq/SS at various input quantities. Figure 19 provides sequencing metrics for various sample runs using the 10X methodology. Figures 17 through 19 show that the disclosed systems and methods (e.g., the 10X GemCode Platform) generates high complexity sequencing libraries using inputs in the ~1ng range. This high complexity allows for deeper targeted sequencing and more sensitive variant calling when sample quantities are limited.

[00332] Referring to Figure 20, previously annotated rearrangements that occur within 30 kb of a region targeted with exome capture were compared to gene fusions detected using the systems and methods of the present disclosure (e.g., GemCode Platform with Whole Exome Sequencing). Previous annotations were found using whole genome sequencing or RNA-based analysis.

[00333] Figure 21 illustrates high confidence detection of a tumor-specific gene fusion in HCC38 triple negative breast cancer cell line using the systems and methods of the present disclosure.

[00334] Figures 22 and 23 illustrate resolving complex rearrangements using the systems and methods of the present disclosure. In particular, looking for the EML4/ALK fusion in H2228 lung cancer cell line reveals a much more complex event: whole genome sequencing BC and read counts confirm presence of a deletion within exons 2-19 of ALK.

[00335] Figure 24 illustrates haplotype phasing of gene fusion events in accordance with some embodiments of the present disclosure.

[00336] Figures 25 and 26 illustrate how whole-genome sequencing linked reads and phasing reveal a complex deletion and allele loss event in HCC1143 triple negative breast cancer. Lines 26A and 26B provide the relative positioning of Figure 26 with respect to the X axis of Figure 25.

[00337] Figure 27 illustrates a schematic for detecting BCR-ABL with whole exome sequencing in which there is extra baiting of long introns. Referring to Figures 28 and 29, a 10X GemCode library was generated from an input of ~1ng of KU812 gDNA using the disclosed systems and methods. Hybrid capture was done with the standard IDT Exome panel, with (Exome +) or without (Exome) additional Ultramer DNA baits that map to the >100 kb intronic region between exon 1 and exon 2 of ABL1 (average bait spacing ~2 kb). Figure 30 summarizes structural variant statistics for 0.2 fmol of intronic bait versus no intronic bait runs.

[00338] **CONCLUSION**

[00339] Plural instances may be provided for components, operations or structures described herein as a single instance. Finally, boundaries between various components, operations, and data stores are somewhat arbitrary, and particular operations are illustrated in the context of specific illustrative configurations. Other allocations of functionality are envisioned and may fall within the scope of the implementation(s). In general, structures and functionality presented as separate components in the example configurations may be implemented as a combined structure or component. Similarly, structures and functionality presented as a single component may be implemented as separate components. These and other variations, modifications, additions, and improvements fall within the scope of the implementation(s).

[00340] It will also be understood that, although the terms "first," "second," etc. may be used herein to describe various elements, these elements should not be limited by these terms. These terms are only used to distinguish one element from another. For example, a first object could be termed a second object, and, similarly, a second object could be termed a first object, without changing the meaning of the description, so long as all occurrences of the "first object" are renamed consistently and all occurrences of the "second object" are renamed consistently. The first object and the second object are both objects, but they are not the same object.

[00341] The terminology used herein is for the purpose of describing particular implementations only and is not intended to be limiting of the claims. As used in the description of the implementations and the appended claims, the singular forms "a", "an" and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will also be understood that the term "and/or" as used herein refers to and encompasses any and all possible combinations of one or more of the associated listed items. It will be further understood that the terms "comprises" and/or "comprising," when used in this specification, specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof.

[00342] As used herein, the term "if" may be construed to mean "when" or "upon" or "in response to determining" or "in accordance with a determination" or "in response to detecting," that a stated condition precedent is true, depending on the context. Similarly, the phrase "if it is determined (that a stated condition precedent is true)" or "if (a stated condition precedent is true)" or "when (a stated condition precedent is true)" may be construed to mean "upon determining" or "in response to determining" or "in accordance with a determination" or "upon detecting" or "in response to detecting" that the stated condition precedent is true, depending on the context.

[00343] The foregoing description included example systems, methods, techniques, instruction sequences, and computing machine program products that embody illustrative implementations. For purposes of explanation, numerous specific details were set forth in order to provide an understanding of various implementations of the inventive subject matter. It will be evident, however, to those skilled in the art that implementations of the inventive subject matter may be practiced without these specific details. In general, well-known instruction instances, protocols, structures and techniques have not been shown in detail.

[00344] The foregoing description, for purpose of explanation, has been described with reference to specific implementations. However, the illustrative discussions above are not intended to be exhaustive or to limit the implementations to the precise forms disclosed. Many modifications and variations are possible in view of the above teachings. The implementations were chosen and described in order to best explain the principles and their practical applications, to thereby enable others skilled in the art to best utilize the implementations and various implementations with various modifications as are suited to the particular use contemplated.

What is claimed is:

1. A method of detecting a structural variant in sequencing data of a test nucleic acid obtained from a 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, 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, wherein each sequence read in each set of sequence reads in the plurality of sets of sequence reads is 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) determining a number of unique barcodes in a first set of sequence reads, in the plurality of sequence reads, wherein 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, wherein the second set of sequence reads is of a second bin in the plurality of bins; and

(D) determining a probability or likelihood that the number is attributable to chance by comparison of a metric based upon the number to a threshold criterion, wherein, 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.

2. The method of claim 1, wherein the metric is computed as:

$$p = \prod_{i}^{n} f_{b_i}$$

wherein,

$\{b_1, b_2, \ldots, b_n\}$ is the set of n unique barcodes that is found in both the first and second set of sequence reads,

i is an integer index to n, and

f_{b_i} is the fraction of the plurality of bins in which the barcode b_i appears.

3. The method of claim 2, wherein the metric is deemed to satisfy the threshold criterion when p is 10^{-2} or less.

4. The method of claim 2, wherein the metric is deemed to satisfy the threshold criterion when p is 10^{-3} or less.

5. The method of claim 2, wherein the metric is deemed to satisfy the threshold criterion when p is 10^{-4} or less.

6. The method of claim 2, wherein the metric is deemed to satisfy the threshold criterion when p is 10^{-5} or less.

7. The method of claim 1, wherein the metric is computed as:

$$p = \prod_{i}^{n} \left(1 - P_{Binom}\left(0; max(n_1, n_2), f_{b_i}\right)\right)$$

wherein,

$\{b_1, b_2, \ldots, b_n\}$ is the set of n unique barcodes that is found in both 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,

i is an integer index to n, and

f_{b_i} is the fraction of the plurality of bins in which the barcode b_i appears.

8. The method of claim 7, wherein the metric is deemed to satisfy the threshold criterion when p is 10^{-2} or less.

9. The method of claim 7, wherein the metric is deemed to satisfy the threshold criterion when p is 10^{-3} or less.

10. The method of claim 7, wherein the metric is deemed to satisfy the threshold criterion when p is 10^{-4} or less.

11. The method of claim 7, wherein the metric is deemed to satisfy the threshold criterion when p is 10^{-5} or less.

12. The method of claim 1, wherein the metric is computed as:

$$p = 1 - P_{Binom}\left(n; {}^{n_1 n_2}/_{B}\right)$$

wherein,

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.

13. The method of any one of claims 1-12, 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.

14. The method of any one of claims 1-12, 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.

15. The method of any one of claims 1-12, 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.

16. The method of any one of claims 1-15, wherein the metric is deemed to satisfy the threshold criterion, and the method further comprises:

(E) aligning each respective sequence read in the number of respective sequence reads to (i) the subset of the test nucleic acid corresponding to the first set of sequence reads and (ii) the subset of the test nucleic acid corresponding to the second set of sequence reads;

(F) determining a first alignment quality for each respective sequence read in the number of respective sequence reads against the subset of the test nucleic acid corresponding to the first set of sequence reads based on the aligning (E);

(G) determining a second alignment quality for each respective sequence read in the number of respective sequence reads against the subset of the test nucleic acid corresponding to the second set of sequence reads based on the aligning (E); and

(H) eliminating from the number of respective sequence reads each sequence read having a first alignment quality and a second alignment quality that are similar; and

(I) repeating the determining (D).

17. The method of any one of claims 1-16, the method further comprising:

maintaining a blackout list, the blackout list comprising a plurality of blackout regions of the test nucleic acid, wherein the determining (C) further comprises:

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

18. The method of any one of claims 1-17, wherein each bin in the plurality of bins represents at least 20 kbp of the test nucleic acid.

19. The method of any one of claims 1-17, wherein each bin in the plurality of bins represents at least 50 kbp of the test nucleic acid.

20. The method of any one of claims 1-17, wherein each bin in the plurality of bins represents at least 100 kbp of the test nucleic acid.

21. The method of any one of claims 1-17, wherein each bin in the plurality of bins represents at least 250 kbp of the test nucleic acid.

22. The method of any one of claims 1-17, wherein each bin in the plurality of bins represents at least 500 kbp of the test nucleic acid.

23. The method of any one of claims 1-22, wherein the different portion of the test nucleic acid represented by the first bin overlaps the different portion of the test nucleic acid represented by the second bin.

24. The method of claim 23, wherein at least 50 percent of the different portion of the test nucleic acid represented by the first bin overlaps the different portion of the test nucleic acid represented by the second bin.

25. The method of claim 23, wherein at least 80 percent of the different portion of the test nucleic acid represented by the first bin overlaps the different portion of the test nucleic acid represented by the second bin.

26. The method of claim 23, wherein at least 95 percent of the different portion of the test nucleic acid represented by the first bin overlaps the different portion of the test nucleic acid represented by the second bin.

27. The method of any one of claims 1-22, wherein there is no overlap between each different portion of the test nucleic acid represented by each respective bin in the plurality of bins.

28. The method of any one of claims 1-22, 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.

29. The method of any one of claims 1-28, wherein the plurality of bins comprises 10,000 or more bins.

30. The method of any one of claims 1-28, wherein the plurality of bins comprises 100,000 or more bins.

31. The method of any one of claims 1-28, wherein the plurality of bins comprises 1,000,000 or more bins.

32. The method of any one of claims 1-31, 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 chromosome from the multi-chromosomal species.

33. The method of any one of claims 1-32, 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}.

34. The method of any one of claims 1-32, 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, ..., 4096}.

35. The method of any one of claims 1-32, 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, ..., 16384}.

36. The method of any one of claims 1-32, 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, ..., 65536}.

37. The method of any one of claims 1-32, 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, ..., 262144}.

38. The method of any one of claims 1-32, 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, ..., 1048576}.

39. The method of any one of claims 1-32, wherein the barcode in the second portion of each respective sequence read encodes a unique predetermined value selected from the set {1, ..., 4194304}.

40. The method of any one of claims 1-32, 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, ..., 16777216}.

41. The method of any one of claims 1-32, 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, ..., 67108864}.

42. The method of any one of claims 1-32, 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, ..., 1×10^{12}}.

43. The method of any one of claims 1-42, wherein the barcode in the second portion of a sequence read in the plurality of sequence reads is localized to a contiguous set of oligonucleotides within the sequence read.

44. The method of claim 43, wherein the contiguous set of oligonucleotides is an N-mer, wherein N is an integer selected from the set {4, ..., 20}.

45. The method of any one of claims 1-42, 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.

46. The method of claim 45, wherein the predetermined noncontiguous set of nucleotides collectively consists of N nucleotides, wherein N is an integer in the set {4, ..., 20}.

47. The method of any one of claims 1-46, wherein the first sequence read in the plurality of sequence reads corresponds to a subset of the test nucleic acid that is greater than 20 kbp.

48. The method of any one of claims 1-47, wherein the structural variation is a translocation 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 from the different portion of the test nucleic acid that is represented by the second set of sequence reads.

49. The method of any one of claims 1-48, wherein

the different portion of the test nucleic acid represented by the first bin is from a first chromosome of the biological sample, and

the different portion of the test nucleic acid represented by the second bin is from a second chromosome of the biological sample, wherein the second chromosome is other than the first chromosome.

50. The method of claim 49, wherein the first chromosome is a paternal chromosome and the second chromosome is a maternal chromosome.

51. The method of any one of claims 1-50, wherein the biological sample is human and the first chromosome is a chromosome in the set $\{1, \ldots, 23\}$.

52. The method of any one of claims 1-51, wherein the structural variation is deemed to have occurred, the method further comprising treating a subject that originated the biological sample with a treatment regimen responsive to the structural variation.

53. The method of claim 52, wherein the treatment regimen comprises a diet modification.

54. The method of claim 52, wherein the treatment regimen comprises application of a pharmaceutical composition that inhibits or augments a biological pathway associated with the structural variation.

55. The method of any one of claims 1-54, wherein the plurality of barcodes comprises 1000 or more barcodes.

56. The method of any one of claims 1-54, wherein the plurality of barcodes comprises 10,000 or more barcodes.

57. The method of any one of claims 1-54, wherein the plurality of barcodes comprises 100,000 or more barcodes.

58. The method of any one of claims 1-54, wherein the plurality of barcodes comprises 1×10^6 or more barcodes.

59. The method of any one of claims 1-54, wherein an identity of the first and second bin is determined using sparse matrix multiplication.

60. The method of claim 59, wherein the sparse matrix multiplication comprises computing:

$$V = A_1^T A_2,$$

wherein,

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

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

B is the number of unique barcodes in 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.

61. The method of claim 60, 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.

62. The method of claim 61, 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.

63. 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, 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, wherein each sequence read in each set of sequence reads in the plurality of sets of sequence reads is 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) determining a number of unique barcodes in a first set of sequence reads in the plurality of sequence reads, wherein 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, wherein the second set of sequence reads is of a second bin in the plurality of bins; and

(D) determining a probability or likelihood that the number is attributable to chance by comparison of a metric based upon the number to a threshold criterion, wherein, 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.

64. The computing system of claim 63, wherein the metric is computed as:

$$p = \prod_{i}^{n} f_{b_i}$$

wherein,

$\{b_1, b_2, \ldots, b_n\}$ is the set of n unique barcodes that is found in both the first and second set of sequence reads,

i is an integer index to n, and

f_{b_i} is the fraction of the plurality of bins in which the barcode b_i appears.

65. The computing system of claim 64, wherein the metric is deemed to satisfy the threshold criterion when p is 10^{-4} or less.

66. The computing system of claim 64, wherein the metric is deemed to satisfy the threshold criterion when p is 10^{-5} or less.

67. The computing system of any one of claims 63-66, 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.

68. The computing system of any one of claims 63-67, wherein the metric is deemed to satisfy the threshold criterion, and wherein the one or more programs further comprise instructions for:

(E) aligning each respective sequence read in the number of respective sequence reads to (i) the subset of the test nucleic acid corresponding to the first set of sequence reads and (ii) the subset of the test nucleic acid corresponding to the second set of sequence reads;

(F) determining a first alignment quality for each respective sequence read in the number of respective sequence reads against the subset of the test nucleic acid corresponding to the first set of sequence reads based on the aligning (E);

(G) determining a second alignment quality for each respective sequence read in the number of respective sequence reads against the subset of the test nucleic acid corresponding to the second set of sequence reads based on the aligning (E);

(H) eliminating from the number of respective sequence reads each sequence read having a first alignment quality and a second alignment quality that are similar; and

(I) repeating the determining (D).

69. The computing system of any one of claims 63-68, wherein the one or more programs further comprise instructions for:

maintaining a blackout list, the blackout list comprising a plurality of blackout regions of the test nucleic acid, wherein the determining (C) further comprises:

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

70. The computing system of any one of claims 63-69, wherein the barcode in the second portion of a sequence read in the plurality of sequence reads is localized to a contiguous set of oligonucleotides.

71. The computing system of claim 70, wherein the contiguous set of oligonucleotides is an N-mer, wherein N is an integer selected from the set {4, ..., 20}.

72. The computing system of any one of claims 63-69, 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.

73. The computing system of claim 72, wherein the predetermined noncontiguous set of nucleotides consists of N nucleotides, wherein N is an integer in the set {4, ..., 20}.

74. The computing system of any one of claim 63-73, wherein the structural variation is deemed to have occurred, and wherein the one or more programs further comprise instructions for communicating a treatment regimen to a subject that originated the biological sample, or a health practitioner that is in charge of caring for the subject, wherein the treatment regimen is responsive to the structural variation.

75. The computing system of claim 74, wherein the treatment regimen comprises a diet modification.

76. The computing system of claim 74, wherein the treatment regimen comprises application of a pharmaceutical composition that inhibits or augments a biological pathway associated with the structural variation.

77. The computing system of claim 63, wherein the metric is computed as:

$$p = \prod_{i}^{n} \left(1 - P_{Binom}(0; max(n_1, n_2), f_{b_i})\right)$$

wherein,

$\{b_1, b_2, \ldots, b_n\}$ is the set of n unique barcodes that is found in both 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,

i is an integer index to n, and

f_{b_i} is the fraction of the plurality of bins in which the barcode b_i appears.

78. The computing system of claim 63, wherein the metric is computed as:

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

wherein,

n is the number of unique barcodes that is found in both 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 a total number of unique barcodes across the plurality of bins.

79. The computing system of any one of claims 63-78, wherein the plurality of barcodes comprises 1000 or more barcodes.

80. The computing system of any one of claims 63-78, wherein the plurality of barcodes comprises 10,000 or more barcodes.

81. The computing system of any one of claims 63-78, wherein the plurality of barcodes comprises 100,000 or more barcodes.

82. The computing system of any one of claims 63-78, wherein the plurality of barcodes comprises 1×10^6 or more barcodes.

83. The computing system of any one of claims 63-82, wherein an identity of the first and second bin is determined using sparse matrix multiplication.

84. The computing system of claim 83, wherein the sparse matrix multiplication comprises computing:

$$V = A_1^T A_2,$$

wherein,

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

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

B is the number of unique barcodes in 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.

85. The computing system of claim 84, 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.

86. The computing system of claim 84, 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.

87. A non-transitory computer readable storage medium storing one or more programs configured for execution by a computer, the one or more programs comprising instructions for:

(A) obtaining a plurality of sequence reads, 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, wherein each sequence read in each set of sequence reads in the plurality of sets of sequence reads is 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) determining a number of unique barcodes in a first set of sequence reads in the plurality of sequence reads, wherein 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, wherein the second set of sequence reads is of a second bin in the plurality of bins; and

(D) determining a probability or likelihood that the number is attributable to chance by comparison of a metric based upon the number to a threshold criterion, wherein, 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.

88. The non-transitory computer readable storage medium of claim 87, wherein the metric is computed as:

$$p = \prod_{i}^{n} f_{b_i}$$

wherein,

$\{b_1, b_2, \ldots, b_n\}$ is the set of n unique barcodes that is found in both the first and second set of sequence reads,

i is an integer index to n, and

f_{b_i} is the fraction of the plurality of bins in which the barcode b_i appears.

89. The non-transitory computer readable storage medium of claim 88, wherein the metric is deemed to satisfy the threshold criterion when p is 10^{-4} or less.

90. The non-transitory computer readable storage medium of claim 88, wherein the metric is deemed to satisfy the threshold criterion when p is 10^{-5} or less.

91. The non-transitory computer readable storage medium of any one of claims 87-90, 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.

92. The non-transitory computer readable storage medium of any one of claims 87-91, wherein the metric is deemed to satisfy the threshold criterion, and wherein the one or more programs further comprise instructions for:

(E) aligning each respective sequence read in the number of respective sequence reads to (i) the subset of the test nucleic acid corresponding to the first set of sequence reads and (ii) the subset of the test nucleic acid corresponding to the second set of sequence reads;

(F) determining a first alignment quality for each respective sequence read in the number of respective sequence reads against the subset of the test nucleic acid corresponding to the first set of sequence reads based on the aligning (E);

(G) determining a second alignment quality for each respective sequence read in the number of respective sequence reads against the subset of the test nucleic acid corresponding to the second set of sequence reads based on the aligning (E);

(H) eliminating from the number of respective sequence reads each sequence read having a first alignment quality and a second alignment quality that are similar; and

(I) repeating the determining (D).

93. The non-transitory computer readable storage medium of any one of claims 87-92, wherein the one or more programs further comprise instructions for:

maintaining a blackout list, the blackout list comprising a plurality of blackout regions of the test nucleic acid, wherein the determining (C) further comprises:

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

94. The non-transitory computer readable storage medium of any one of claims 87-93, wherein the barcode in the first portion of a sequence read in the plurality of sequence reads is localized to a contiguous set of oligonucleotides.

95. The non-transitory computer readable storage medium of claim 94, wherein the contiguous set of oligonucleotides is an N-mer, wherein N is an integer selected from the set $\{4, \ldots, 20\}$.

96. The non-transitory computer readable storage medium of any one of claims 87-93, wherein the barcode in the first portion of a sequence read in the plurality of sequence reads is localized to a predetermined noncontiguous set of nucleotides.

97. The non-transitory computer readable storage medium of claim 96, wherein the predetermined noncontiguous set of nucleotides consists of N nucleotides, wherein N is an integer in the set $\{4, \ldots, 20\}$.

98. The non-transitory computer readable storage medium of any one of claims 87-97, wherein the structural variation is deemed to have occurred, and wherein the one or more

programs further comprise instructions for communicating a treatment regimen to a subject that originated the biological sample, or a health practitioner that is in charge of caring for the subject, wherein the treatment regimen is responsive to the structural variation.

99. The non-transitory computer readable storage medium of claim 98, wherein the treatment regimen comprises a diet modification.

100. The non-transitory computer readable storage medium of claim 98, wherein the treatment regimen comprises application of a pharmaceutical composition that inhibits or augments a biological pathway associated with the structural variation.

101. The non-transitory computer readable storage medium of claim 87, wherein the metric is computed as:

$$p = \prod_{i}^{n} \left(1 - P_{Binom}(0; max(n_1, n_2), f_{b_i})\right)$$

wherein,

$\{b_1, b_2, \ldots, b_n\}$ is the set of n unique barcodes that is found in both 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,

i is an integer index to n, and

f_{b_i} is the fraction of the plurality of bins in which the barcode b_i appears.

102. The non-transitory computer readable storage medium of claim 87, wherein the metric is computed as:

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

wherein,

n is the number of unique barcodes that is found in both 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.

103. The non-transitory computer readable storage medium of any one of claims 87-102, wherein the plurality of barcodes comprises 1000 or more barcodes.

104. The non-transitory computer readable storage medium of any one of claims 87-102, wherein the plurality of barcodes comprises 10,000 or more barcodes.

105. The non-transitory computer readable storage medium of any one of claims 87-102, wherein the plurality of barcodes comprises 100,000 or more barcodes.

106. The non-transitory computer readable storage medium of any one of claims 87-102, wherein the plurality of barcodes comprises 1×10^6 or more barcodes.

107. The non-transitory computer readable storage medium of any one of claims 87-106, wherein an identity of the first and second bin is determined using sparse matrix multiplication.

108. The non-transitory computer readable storage medium of claim 107, wherein the sparse matrix multiplication comprises computing:

$$V = A_1^T A_2,$$

wherein,

A_1 is a first $B \times N_1$ matrix of barcodes that includes those of the first bin,
A_2 is a second $B \times N_2$ matrix of barcodes that includes those of the second bin,
B is the number of unique barcodes in 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.

109. The non-transitory computer readable storage medium of claim 108, 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.

110. The non-transitory computer readable storage medium of claim 108, 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.

111. 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.

112. The method of claim 111, 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.

113. The method of claim 111 or 112, wherein the first bin and the second bin are at least a predetermined number of kilobases apart on the test nucleic acid.

114. The method of claim 113, wherein the first bin and the second bin are at least 50 kilobases apart on the test nucleic acid.

115. The method of any one of claims 111-114, wherein the determining (D) uses a binomial test to compute the first value.

116. The method of any one of claims 111-114, wherein the determining (D) uses a binomial test to compute the first value of the form:

$$p = 1 - P_{Binom}\left(n; {}^{n_1 n_2}/_{B}\right)$$

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.

117. The method of any one of claims 111-116, 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.

118. The method of any one of claims 111-116, 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.

119. The method of any one of claims 111-118, 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.

120. The method of any one of claims 111-118, 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.

121. The method of any one of claims 111-118, 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.

122. The method of any one of claims 111-121, wherein the structural variation is associated with a genetic disease.

123. The method of any one of claims 111-122, 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.

124. The method of any one of claims 111-122, 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.

125. The method of any one of claims 111-124, wherein the plurality of bins comprises 10,000 or more bins.

126. The method of any one of claims 111-124, wherein the plurality of bins comprises 100,000 or more bins.

127. The method of any one of claims 111-124, wherein the plurality of bins comprises 1,000,000 or more bins.

128. The method of any one of claims 111-127, 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.

129. The method of any one of claims 111-128, 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}}.

130. The method of any one of claims 111-129, 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.

131. The method of claim 130, wherein the contiguous set of oligonucleotides is an N-mer, wherein N is an integer selected from the set {4, ..., 20}.

132. The method of any one of claims 111-120, 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.

133. The method of claim 132, wherein the predetermined noncontiguous set of nucleotides collectively consists of N nucleotides, wherein N is an integer in the set {4, ..., 20}.

134. The method of any one of claims 111-133, wherein the first sequence read corresponds to a first subset of the test nucleic acid that is greater than 10 kilobases.

135. The method of any one of claims 111-133, wherein the first sequence read corresponds to a first subset of the test nucleic acid that is greater than 20 kilobases.

136. The method of any one of claims 111-135, wherein the structural variation is deemed to have occurred, the method further comprising treating a subject that originated the biological sample with a treatment regimen responsive to the structural variation.

137. The method of claim 136, wherein the treatment regimen comprises a diet modification.

138. The method of claim 137, wherein the treatment regimen comprises application of a pharmaceutical composition that inhibits or augments a biological pathway associated with the structural variation.

139. The method of any one of claims 111-138, wherein an identity of the first and second bin is determined by the identifying (C) using sparse matrix multiplication.

140. The method of any one of claims 111-138, 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 x N_1 matrix of barcodes that includes the first bin,

A_2 is a second B x 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.

141. The method of claim 140, 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.

142. The method of claim 140, 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.

143. The method of any one of claims 111-142, 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.

144. The method of claim 111, wherein the computed likelihood in the computing (F) is computed as:

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

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.

145. The method of claim 144, wherein

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

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

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,

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

146. The method of claim 145, wherein $P_{frag}(r_1, l_1; a_b)$ and $P_{frag}(r_2, l_2; a_b)$ are each computed as

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

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.

147. The method of claim 145, wherein

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

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

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.

148. The method of claim 144, 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,

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

149. The method of claim 148, wherein $P_{frag}(r_1, l_1; a_b)$ and $P_{frag}(r_2, l_2; a_b)$ are each computed as

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

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.

150. The method of claim 148, wherein

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

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

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,

$P_p\big(0;a_b(m-l_1-l_2)\big)$ 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.

151. The method of claim 148, wherein 2d' is estimated by computing the maximum extent d' such that $P_p(0;a_b d')\geq 0.75$.

152. The method of any one of claims 111-151, wherein the plurality of sequence reads represents whole genome sequencing data.

153. The method of any one of claims 111-151, 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.

154. The method of any one of claims 111-154, 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.

155. The method of any one of claims 111-154, wherein the plurality of barcodes comprises 1000 or more barcodes.

156. The method of any one of claims 111-154, wherein the plurality of barcodes comprises 10,000 or more barcodes.

157. The method of any one of claims 111-154, wherein the plurality of barcodes comprises 100,000 or more barcodes.

158. The method of any one of claims 111-154, wherein the plurality of barcodes comprises 1×10^6 or more barcodes.

159. 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.

160. The computing system of claim 159, wherein the plurality of barcodes comprises 1000 or more barcodes.

161. The computing system of claim 159, wherein the plurality of barcodes comprises 10,000 or more barcodes.

162. The computing system of claim 159, wherein the plurality of barcodes comprises 100,000 or more barcodes.

163. The computing system of claim 159, wherein the plurality of barcodes comprises 1 x 10^6 or more barcodes.

164. A non-transitory computer readable storage medium storing one or more programs configured for execution by a computer, 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.

165. The non-transitory computer readable storage medium of claim 164, wherein the plurality of barcodes comprises 1000 or more barcodes.

166. The non-transitory computer readable storage medium of claim 164, wherein the plurality of barcodes comprises 10,000 or more barcodes.

167. The non-transitory computer readable storage medium of claim 164, wherein the plurality of barcodes comprises 100,000 or more barcodes.

168. The non-transitory computer readable storage medium of claim 164, wherein the plurality of barcodes comprises 1×10^6 or more barcodes.

169. A method of phasing sequencing data of a test nucleic acid sample obtained from a biological sample from a single organism of a species, wherein the test nucleic acid sample comprises a first set of haplotypes (H_0) and a second set of haplotypes (H_1), 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 reference consensus sequence for all or a portion of a genome of the species;

(B) obtaining a plurality of variant calls $A_{i;p}$ for the biological sample, wherein i is an index to a position in the reference consensus sequence, 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;

(C) obtaining a plurality of sequence reads $\vec{\boldsymbol{O}}$ for the biological sample, wherein each respective sequence read $\vec{\boldsymbol{O}}_i$ 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, and

each respective sequence read $\vec{\boldsymbol{O}}i$ in the plurality of sequence reads is $\in \{0, 1, -\}^n$, wherein (i) n is the number of variants calls in $A_{i;p}$, (ii) each respective label 0 for the respective sequence read $\vec{\boldsymbol{O}}i$ assigns a corresponding variant call in $A_{i;p}$ to H_0, (iii) each respective label 1 for the respective sequence read $\vec{\boldsymbol{O}}i$ assigns a corresponding variant call in

$A_{i;p}$ to H_O, and (iv) each respective label – for the respective sequence read $\vec{O}i$ indicates that the corresponding variant call in $A_{i;p}$ is not covered; and

(D) refining a phasing result $\vec{X}$ by optimization of haplotype assignments at individual positions i in $A_{i;p}$ between H_O and H_1 for the plurality of sequence reads using the relationship:

$$\hat{\mathbf{X}} = \underset{\vec{\mathbf{X}}}{\arg\max}\, P(\vec{O}|\vec{\mathbf{X}})$$

wherein,

$\hat{\mathbf{X}}$ is the refined phasing vector,

$\vec{\mathbf{X}}$ is the phasing vector result to be inferred,

$P(\vec{O}|\vec{\mathbf{X}})$, an overall objective function, is equal to $\prod_f P(O_{1,f}, \ldots, O_{N,f}|\vec{\mathbf{X}})$,

$(O_{1,f}, \ldots, O_{N,f})$ is the respective subset of N variant calls in the plurality of variant calls $A_{i;p}$ observed in the subset of sequence reads that include the same respective barcode from the plurality of barcodes,

$P(O_{1,f}, \ldots, O_{N,f}|\vec{\mathbf{X}},\ H_f = 0) = \prod_i P(O_{i,f}|A_{i,X_i})$,

$P(O_{1,f}, \ldots, O_{N,f}|\vec{\mathbf{X}},\ H_f = 1) = \prod_i P(O_{i,f}|A_{i,1-X_i})$,

$P(O_{1,f}, \ldots, O_{N,f}|\vec{\mathbf{X}},\ H_f = M) = \prod_i 0.5$,

M indicates a mixture of $H_f = 0$ and $H_f = 1$ for the respective barcode f,

$P(O_{i,f}, \ldots, O_{N,f}|\vec{\mathbf{X}}) = \frac{(1-\propto)}{2}\left(\prod_i P(O_{i,f}|A_i, X_i) + \prod_i P(O_{i,f}|A_i, 1 - X_i)\right) + \propto \prod_i 0.5$,

$\log P(O_{i,f}|A_{i,p}) = \sum_r \mathbf{1}(S_r = A_{i,p})\left(1 - 10^{-Q_r/10}\right) + \mathbf{1}(S_r \neq A_{i,p})\left(10^{-Q_r/10}\right)$,

$\propto$ is a predetermined fractional value representing a likelihood or probability that $H_f = M$ arises,

i is the ith variant in the respective subset of N variant calls observed for the subset of sequence reads that include the same respective barcode,

r sums over the subset of sequence reads that include the same respective barcode,

$\mathbf{1}(S_r = A_{i,p})$ is an indicator function testing if the base assignment at position i in the rth sequence read S_r in the subset of sequence reads that include the same respective barcode matches Ai,p, wherein when they match $\mathbf{1}(S_r = A_{i,p})$ has a value of 1 and when they do not match $\mathbf{1}(S_r = A_{i,p})$ has a value of zero,

$\mathbf{1}(S_r \neq A_{i,p})$ is an indicator function testing if the base assignment at position i in the rth sequence read S_r in the subset of sequence reads that include the same respective barcode does not match Ai,p, wherein when they do not match $\mathbf{1}(S_r \neq A_{i,p})$ has a value of 1 and when they do match $\mathbf{1}(S_r = A_{i,p})$ has a value of zero, and

Q_r is a quality value for S_r for the read base at the position of i in the reference consensus sequence.

170. A method of addressing error in the zygosity 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, wherein the test nucleic acid sample comprises a first set of haplotypes (H_0) and a second set of haplotypes (H_1), 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 reference consensus sequence for all or a portion of a genome of the species;

(B) obtaining a plurality of variant calls $A_{i;p}$ for the biological sample, wherein

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_O, label 1 assigns the respective variant call to H_1, and label -1 assigns the respective variant call to the zygosity error condition H_{-1};

(C) obtaining a plurality of sequence reads $\vec{\boldsymbol{O}}$ for the biological sample, wherein

each respective sequence read $\vec{\boldsymbol{O}}_i$ 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, and

each respective sequence read $\vec{\boldsymbol{O}}i$ in the plurality of sequence reads is $\in \{0, 1,-1, -\}^n$, wherein (i) n is the number of variants calls in $A_{i;p}$, (ii) each respective label 0 for the respective sequence read $\vec{\boldsymbol{O}}i$ assigns a corresponding variant call in $A_{i;p}$ to H_0, (iii) each respective label 1 for the respective sequence read $\vec{\boldsymbol{O}}i$ assigns a corresponding variant call in $A_{i;p}$ to H_O, (iv) each respective label -1 for the respective sequence read $\vec{\boldsymbol{O}}i$ assigns a corresponding variant call in $A_{i;p}$ to H_{-1}, and (v) each respective label – for the respective sequence read $\vec{\boldsymbol{O}}i$ indicates that the corresponding variant call in $A_{i;p}$ is not covered; and

(D) refining a phasing vector result $\vec{X}$ 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:

$$\hat{\mathbf{X}} = \underset{\vec{X}}{\operatorname{argmax}} P(\vec{X}|\vec{O})\,| = \frac{P(\vec{X})P(\vec{O}|\vec{X})}{C}$$

wherein,

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

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

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

$\hat{\mathbf{X}}$ is the refined phasing vector result,

C is a constant,

$\vec{\mathbf{X}}$ is the phasing vector result to be inferred, and

$(O_{1,f}, \dots, O_{N,f})$ is the respective subset of N variant calls in the plurality of variant calls $A_{i;p}$ observed in the subset of sequence reads that include the same respective barcode from the plurality of barcodes.

171. The method of claim 170, wherein

$(O_{1,f}, \dots, O_{N,f}|\vec{\mathbf{X}},\ H_f = 0) = \prod_i P(O_{i,f}|A_{i,X_i})$,

$P(O_{1,f}, \dots, O_{N,f}|\vec{\mathbf{X}},\ H_f = 1) = \prod_i P(O_{i,f}|A_{i,1-X_i})$,

$P(O_{1,f}, \dots, O_{N,f}|\vec{\mathbf{X}},\ H_f = M) = \prod_i 0.5$,

M indicates a mixture of $H_f = 0$ and $H_f = 1$ for the respective barcode f,

$P(O_{i,f}, \dots, O_{N,f}|\vec{\mathbf{X}}) = \frac{(1-\propto)}{2}(\prod_i P(O_{i,f}|A_i, X_i) + \prod_i P(O_{i,f}|A_i, 1-X_i)) + \propto \prod_i 0.5$,

$$\log P(O_{i,f}|A_{i,p}) = \sum_r \mathbf{1}(S_r = A_{i,p})\left(1 - 10^{-Q_r/10}\right) + \mathbf{1}(S_r \neq A_{i,p})\left(10^{-Q_r/10}\right)$$

$$+ \mathbf{1}(A_{i,p} = H_{-1})^{0.5}$$

$\propto$ is a predetermined fractional value representing a likelihood or probability that $H_f = M$ arises,

i is the ith variant in the respective subset of N variant calls observed for the subset of sequence reads that include the same respective barcode,

r sums over the subset of sequence reads that include the same respective barcode,

$\mathbf{1}(S_r = A_{i,p})$ is an indicator function testing if the base assignment at position i in the rth sequence read S_r in the subset of sequence reads that include the same respective barcode matches Ai,p, wherein when they match $\mathbf{1}(S_r = A_{i,p})$ has a value of 1 and when they do not match $\mathbf{1}(S_r = A_{i,p})$ has a value of zero,

$\mathbf{1}(S_r \neq A_{i,p})$ is an indicator function testing if the base assignment at position i in the rth sequence read S_r in the subset of sequence reads that include the same respective barcode does not match Ai,p, wherein when they do not match $\mathbf{1}(S_r \neq A_{i,p})$ has a value of 1 and when they do match $\mathbf{1}(S_r = A_{i,p})$ has a value of zero,

$\mathbf{1}(A_{i,p} = H_{-1})$ is an indicator function that has a value of 1 when $A_{i,p}$ is equal to H_{-1} and is a value of zero otherwise, and

Q_r is a quality value for S_r for the read base at the position of i in the reference consensus sequence.

172. The method of any one of claims 169-171, wherein the plurality of variant calls comprises a plurality of heterozygous single nucleotide polymorphisms, heterozygous inserts, or heterozygous deletions in the test nucleic acid.

173. The method of any one of claims 169-172, wherein

the first set of haplotypes (H_0) consists of maternal haplotypes for the single organism, and

the second set of haplotypes (H_1) consists of paternal haplotypes for the single organism.

174. The method of any one of claims 169-173, wherein the plurality of barcodes comprises 1000 or more barcodes.

175. The method of any one of claims 169-173, wherein the plurality of barcodes comprises 10,000 or more barcodes.

176. The method of any one of claims 169-173, wherein the plurality of barcodes comprises 100,000 or more barcodes.

177. The method of any one of claims 169-173, wherein the plurality of barcodes comprises 1×10^6 or more barcodes.

178. The method of any one of claims 169-177, wherein the species is human.

179. The method of any one of claims 169-178, wherein the plurality of variant calls $A_{i;p}$ comprises 1000 or more variant calls.

180. The method of any one of claims 169-178, wherein the plurality of variant calls $A_{i;p}$ comprises 10,000 or more variant calls.

181. The method of any one of claims 169-178, wherein the plurality of sequence reads comprises 10,000 or more sequence reads.

182. The method of any one of claims 169-181, wherein the plurality of sequence reads comprises 100,000 or more sequence reads.

183. The method of any one of claims 169-182, wherein the plurality of sequence reads comprises 1×10^6 or more sequence reads.

184. The method of any one of claims 169-183, wherein $\vec{X}$ *is* (x), wherein

x is a binary string of length n,

each value of 0 in x indicates origination of the corresponding variant call in the first set of haplotypes (H_0), and

each value of 1 in x indicates origination of the corresponding variant call in the second set of haplotypes (H_1).

185. The method of any one of claims 169-184, wherein the first set of haplotypes (H_0) consists of a single maternal haplotype and the second set of haplotypes (H_1) consists of a single paternal haplotype.

186. The method of any one of claims 169-184, wherein the first set of haplotypes (H_0) comprises five or more maternal haplotypes corresponding to five or more maternal chromosomes and the second set of haplotypes (H_1) comprises five or more paternal haplotypes corresponding to five or more paternal chromosomes.

187. The method of any one of claims 169-186, wherein the subset of sequence reads that include the same respective barcode f comprises 10 or more sequence reads.

188. The method of any one of claims 169-186, wherein the subset of sequence reads that include the same respective barcode f comprises 30 or more sequence reads.

189. The method of any one of claims 169-186, wherein the subset of sequence reads that include the same respective barcode f comprises 100 or more sequence reads.

190. The method of any one of claims 169-189, wherein a subset of sequence reads that include the same respective barcode f represent at least 30 kilobases of the reference consensus sequence.

191. The method of any one of claims 169-189, wherein a subset of sequence reads that include the same respective barcode f represent at least 40 kilobases of the reference consensus sequence.

192. The method of any one of claims 169-191, wherein the refining (D) optimizes the overall objective function using a hierarchical search over $\vec{\boldsymbol{X}}$.

193. The method of claim 192, wherein the hierarchical search comprises:

for each respective local block of variant calls in $A_{i;p}$ that are localized to a corresponding subset of the reference consensus sequence, using a beam search over the assignments of $X_k, X_{k+1}, \ldots, X_{k+j}$ in the respective local block of variant calls, wherein 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 and wherein assignments of $X_k, X_{k+1}, \ldots, X_{k+j}$ are found by computing the objective function in which the phasing vector of the objective function in respective computations is limited to $X_k, X_{k+1}, \ldots, X_{k+j}$, thereby finding an optimal phasing solution for each respective local block of variant calls, and

greedily joining neighboring local blocks of variant calls in $A_{i;p}$ using the optimal phasing solution for each respective local block of variant calls thereby obtaining an estimate of the optimal phasing configuration $\hat{\mathbf{X}}$.

194. The method of claim 193, wherein the refining the phase result further comprises iteratively swapping the phase result of individual x_i in the estimate of the optimal phasing configuration $\hat{\mathbf{X}}$ and recomputing the objective function, thereby obtaining $\hat{\mathbf{X}}$.

195. The method of claim 193, wherein a respective local block of variant calls consists of between 20 and 60 variants in $A_{i;p}$.

196. The method of claim 193, wherein a respective local block of variant calls consists of between 30 and 80 variants in $A_{i;p}$.

197. The method of claim 193, wherein an iteration of the beam search for the assignments of one of X_k, X_{k+1}, ..., X_{k+j} discards all but a predetermined number of solutions for $\widehat{\mathbf{X}}$.

198. The method of claim 197, wherein the predetermined number of solutions for $\widehat{\mathbf{X}}$ is 1000 or less.

199. The method of claim 197, wherein the predetermined number of solutions for $\widehat{\mathbf{X}}$ is 5000 or less.

200. The method of any one of claims 169-199, wherein the species is human and the test nucleic acid sample comprises the genome of the biological sample.

201. The method of any one of claims 169-199, wherein 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.

202. The method of any one of claims 169-201, wherein 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, ..., 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}}.

203. The method of any one of claims 169-202, 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.

204. The method of claim 203, wherein the contiguous set of oligonucleotides is an N-mer, wherein N is an integer selected from the set {4, ..., 20}.

205. The method of any one of claims 169-202, 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.

206. The method of claim 205, wherein the predetermined noncontiguous set of nucleotides collectively consists of N nucleotides, wherein N is an integer in the set {4, ..., 20}.

207. The method of any one of claims 169-206, wherein a sequence read in the plurality of sequence reads corresponds to a portion of the reference consensus sequence that is greater than 10 kilobases.

208. The method of any one of claims 169-206, wherein a sequence read in the plurality of sequence reads corresponds to a portion of the reference consensus sequence that is greater than 20 kilobases.

209. The method of any one of claims 169-208, wherein the plurality of variant calls is obtained from the plurality of sequence reads.

210. The method of any one of claims 169-209, wherein the plurality of sequence reads is obtained from a plurality of barcoded-oligo coated gel-beads and wherein the test nucleic acid sample is 50 ng or less.

211. The method of claim 210 wherein the plurality of barcoded-oligo coated gel-beads comprises 10,000 beads.

212. The method of claim 210 wherein the plurality of barcoded-oligo coated gel-beads comprises 50,000 beads.

213. The method of claim 210 wherein the test nucleic acid sample is 25 ng or less.

214. The method of claim 210 wherein the test nucleic acid sample is 10 ng or less.

215. The method of claim 210 wherein the test nucleic acid sample is 5 ng or less.

216. The method of claim 210 wherein the test nucleic acid sample is 2.5 ng or less.

217. The method of any one of claims 169-216, wherein the plurality of sequencing reads $\vec{O}$ is obtained within ten minutes of exposure to the plurality of barcodes.

218. The method of any one of claims 169-216, wherein the plurality of sequencing reads $\vec{O}$ is obtained within twenty minutes of exposure to the plurality of barcodes.

219. 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 phasing sequencing data of a test nucleic acid sample obtained from a biological sample from a single organism of a species, wherein the test nucleic acid sample comprises a first set of haplotypes (H_0) and a second set of haplotypes (H_1), by executing a method comprising:

(A) obtaining a reference consensus sequence for all or a portion of a genome of the species;

(B) obtaining a plurality of variant calls $A_{i;p}$ for the biological sample, wherein

i is an index to a position in the reference consensus sequence, 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;

(C) obtaining a plurality of sequence reads $\vec{\boldsymbol{O}}$ for the biological sample, wherein

each respective sequence read $\vec{\boldsymbol{O}}_i$ 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, and

each respective sequence read $\vec{\boldsymbol{O}}i$ in the plurality of sequence reads is $\in \{0, 1, -\}^n$, wherein (i) n is the number of variants calls in $A_{i;p}$, (ii) each respective label 0 for the respective sequence read $\vec{\boldsymbol{O}}i$ assigns a corresponding variant call in $A_{i;p}$ to H_0, (iii) each respective label 1 for the respective sequence read $\vec{\boldsymbol{O}}i$ assigns a corresponding variant call in $A_{i;p}$ to H_0, and (iv) each respective label – for the respective sequence read $\vec{\boldsymbol{O}}i$ indicates that the corresponding variant call in $A_{i;p}$ is not covered; and

(D) refining a phasing result $\vec{\boldsymbol{X}}$ by optimization of haplotype assignments at individual positions i in $A_{i;p}$ between H_0 and H_1 for the plurality of sequence reads using the relationship:

$$\hat{\mathbf{X}} = \underset{\vec{\mathbf{X}}}{\arg\max}\, P(\vec{\boldsymbol{O}}|\vec{\mathbf{X}})$$

wherein,

$\hat{\mathbf{X}}$ is the refined phasing vector,

$\vec{\mathbf{X}}$ is the phasing vector result to be inferred,

$P(\vec{\boldsymbol{O}}|\vec{\mathbf{X}})$, an overall objective function, is equal to $\prod_f P(O_{1,f}, \ldots, O_{N,f}|\vec{\mathbf{X}})$,

$(O_{1,f}, \ldots, O_{N,f})$ is the respective subset of N variant calls in the plurality of variant calls $A_{i,p}$ observed in the subset of sequence reads that include the same respective barcode from the plurality of barcodes,

$$P(O_{1,f}, \ldots, O_{N,f} | \vec{\mathbf{X}}, H_f = 0) = \prod_i P(O_{i,f} | A_{i,X_i}),$$

$$P(O_{1,f}, \ldots, O_{N,f} | \vec{\mathbf{X}}, H_f = 1) = \prod_i P(O_{i,f} | A_{i,1-X_i}),$$

$$P(O_{1,f}, \ldots, O_{N,f} | \vec{\mathbf{X}}, H_f = M) = \prod_i 0.5,$$

M indicates a mixture of $H_f = 0$ and $H_f = 1$ for the respective barcode f,

$$P(O_{i,f}, \ldots, O_{N,f} | \vec{\mathbf{X}}) = \frac{(1-\propto)}{2}\left(\prod_i P(O_{i,f} | A_i, X_i) + \prod_i P(O_{i,f} | A_i, 1 - X_i)\right) + \propto \prod_i 0.5,$$

$$\log P(O_{i,f} | A_{i,p}) = \sum_r \mathbf{1}(S_r = A_{i,p})\left(1 - 10^{-Q_r/10}\right) + \mathbf{1}(S_r \neq A_{i,p})\left(10^{-Q_r/10}\right),$$

$\propto$ is a predetermined fractional value representing a likelihood or probability that $H_f = M$ arises,

i is the ith variant in the respective subset of N variant calls observed for the subset of sequence reads that include the same respective barcode,

r sums over the subset of sequence reads that include the same respective barcode,

$\mathbf{1}(S_r = A_{i,p})$ is an indicator function testing if the base assignment at position i in the rth sequence read S_r in the subset of sequence reads that include the same respective barcode matches Ai,p, wherein when they match $\mathbf{1}(S_r = A_{i,p})$ has a value of 1 and when they do not match $\mathbf{1}(S_r = A_{i,p})$ has a value of zero,

$\mathbf{1}(S_r \neq A_{i,p})$ is an indicator function testing if the base assignment at position i in the rth sequence read S_r in the subset of sequence reads that include the same respective barcode does not match Ai,p, wherein when they do not match $\mathbf{1}(S_r \neq A_{i,p})$ has a value of 1 and when they do match $\mathbf{1}(S_r = A_{i,p})$ has a value of zero, and

Q_r is a quality value for S_r for the read base at the position of i in the reference consensus sequence.

220. 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 addressing error in the zygosity 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, wherein the test nucleic acid sample comprises a

first set of haplotypes (H_0) and a second set of haplotypes (H_1), by executing a method comprising:

A) obtaining a reference consensus sequence for all or a portion of a genome of the species;

(B) obtaining a plurality of variant calls $A_{i;p}$ for the biological sample, wherein

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 zygosity error condition H_{-1};

(C) obtaining a plurality of sequence reads $\vec{O}$ for the biological sample, wherein

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, and

each respective sequence read $\vec{O}i$ in the plurality of sequence reads is $\in \{0, 1, -1, -\}^n$, wherein (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_0, (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; and

(D) refining a phasing vector result $\vec{X}$ 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:

$$\hat{\mathbf{X}} = \underset{\vec{X}}{\operatorname{argmax}} P(\vec{X}|\vec{O})\,| = \frac{P(\vec{X})P(\vec{O}|\vec{X})}{C}$$

wherein,

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

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

$$P(\vec{O}|\vec{X}) = \prod_f P(O_{1,f}, \ldots, O_{N,f}),$$

$\hat{\mathbf{X}}$ is the refined phasing vector result,

C is a constant,

$\vec{\mathbf{X}}$ is the phasing vector result to be inferred, and

$(O_{1,f}, \ldots, O_{N,f})$ is the respective subset of N variant calls in the plurality of variant calls $A_{i;p}$ observed in the subset of sequence reads that include the same respective barcode from the plurality of barcodes.

221. The computing system of claim 220, wherein

$$(O_{1,f}, \ldots, O_{N,f}|\vec{\mathbf{X}},\ H_f = 0) = \prod_i P(O_{i,f}|A_{i,X_i}),$$

$$P(O_{1,f}, \ldots, O_{N,f}|\vec{\mathbf{X}},\ H_f = 1) = \prod_i P(O_{i,f}|A_{i,1-X_i}),$$

$$P(O_{1,f}, \ldots, O_{N,f}|\vec{\mathbf{X}},\ H_f = M) = \prod_i 0.5,$$

M indicates a mixture of $H_f = 0$ and $H_f = 1$ for the respective barcode f,

$$P(O_{i,f}, \ldots, O_{N,f}|\vec{\mathbf{X}}) = \frac{(1-\propto)}{2}\left(\prod_i P(O_{i,f}|A_i, X_i) + \prod_i P(O_{i,f}|A_i, 1 - X_i)\right) + \propto \prod_i 0.5,$$

$$\log P(O_{i,f}|A_{i,p}) = \sum_r \mathbf{1}(S_r = A_{i,p})\left(1 - 10^{-Q_r/10}\right) + \mathbf{1}(S_r \neq A_{i,p})\left(10^{-Q_r/10}\right)$$

$$+ \mathbf{1}(A_{i,p} = H_{-1})^{0.5}$$

$\propto$ is a predetermined fractional value representing a likelihood or probability that $H_f = M$ arises,

i is the ith variant in the respective subset of N variant calls observed for the subset of sequence reads that include the same respective barcode,

r sums over the subset of sequence reads that include the same respective barcode,

$\mathbf{1}(S_r = A_{i,p})$ is an indicator function testing if the base assignment at position i in the rth sequence read S_r in the subset of sequence reads that include the same respective barcode matches Ai,p, wherein when they match $\mathbf{1}(S_r = A_{i,p})$ has a value of 1 and when they do not match $\mathbf{1}(S_r = A_{i,p})$ has a value of zero,

$\mathbf{1}(S_r \neq A_{i,p})$ is an indicator function testing if the base assignment at position i in the rth sequence read S_r in the subset of sequence reads that include the same respective barcode

does not match Ai,p, wherein when they do not match $\mathbf{1}(S_r \neq A_{i,p})$ has a value of 1 and when they do match $\mathbf{1}(S_r = A_{i,p})$ has a value of zero,

$\mathbf{1}(A_{i,p} = H_{-1})$ is an indicator function that has a value of 1 when $A_{i,p}$ is equal to H_{-1} and is a value of zero otherwise, and

Q_r is a quality value for S_r for the read base at the position of i in the reference consensus sequence.

222. A non-transitory computer readable storage medium storing one or more programs configured for execution by a computer, the one or more programs comprising instructions for phasing sequencing data of a test nucleic acid sample obtained from a biological sample from a single organism of a species, wherein 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 collectively executing a method comprising:

(A) obtaining a reference consensus sequence for all or a portion of a genome of the species;

(B) obtaining a plurality of variant calls $A_{i;p}$ for the biological sample, wherein

i is an index to a position in the reference consensus sequence, and

$p \in \{0, 1\}$ in which label 0 assigns a respective variant call in $A_{i;p}$ to H_O and label 1 assigns the respective variant call to H_1;

(C) obtaining a plurality of sequence reads $\vec{\boldsymbol{O}}$ for the biological sample, wherein

each respective sequence read $\vec{\boldsymbol{O}}_i$ 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, and

each respective sequence read $\vec{\boldsymbol{O}}i$ in the plurality of sequence reads is $\in \{0, 1, -\}^n$, wherein (i) n is the number of variants calls in $A_{i;p}$, (ii) each respective label 0 for the respective sequence read $\vec{\boldsymbol{O}}i$ assigns a corresponding variant call in $A_{i;p}$ to H_0, (iii) each respective label 1 for the respective sequence read $\vec{\boldsymbol{O}}i$ assigns a corresponding variant call in $A_{i;p}$ to H_O, and (iv) each respective label – for the respective sequence read $\vec{\boldsymbol{O}}i$ indicates that the corresponding variant call in $A_{i;p}$ is not covered; and

(D) refining a phasing result $\vec{\boldsymbol{X}}$ by optimization of haplotype assignments at individual positions i in $A_{i;p}$ between H_O and H_1 for the plurality of sequence reads using the relationship:

$$\hat{\mathbf{X}} = \underset{\vec{\mathbf{X}}}{\arg\max}\, P(\vec{\boldsymbol{O}}|\vec{\mathbf{X}})$$

wherein,

$\hat{\mathbf{X}}$ is the refined phasing vector,

$\vec{\mathbf{X}}$ is the phasing vector result to be inferred,

$P(\vec{\boldsymbol{O}}|\vec{\mathbf{X}})$, an overall objective function, is equal to $\prod_f P(O_{1,f}, \ldots, O_{N,f}|\vec{\mathbf{X}})$,

$(O_{1,f}, \ldots, O_{N,f})$ is the respective subset of N variant calls in the plurality of variant calls $A_{i,p}$ observed in the subset of sequence reads that include the same respective barcode from the plurality of barcodes,

$$P(O_{1,f}, \ldots, O_{N,f}|\vec{\mathbf{X}},\ H_f = 0) = \prod_i P(O_{i,f}|A_{i,X_i}),$$

$$P(O_{1,f}, \ldots, O_{N,f}|\vec{\mathbf{X}},\ H_f = 1) = \prod_i P(O_{i,f}|A_{i,1-X_i}),$$

$$P(O_{1,f}, \ldots, O_{N,f}|\vec{\mathbf{X}},\ H_f = M) = \prod_i 0.5,$$

M indicates a mixture of $H_f = 0$ and $H_f = 1$ for the respective barcode f,

$$P(O_{i,f}, \ldots, O_{N,f}|\vec{\mathbf{X}}) = \frac{(1-\propto)}{2}\left(\prod_i P(O_{i,f}|A_i, X_i) + \prod_i P(O_{i,f}|A_i, 1 - X_i)\right) + \propto \prod_i 0.5,$$

$$\log P(O_{i,f}|A_{i,p}) = \sum_r \mathbf{1}(S_r = A_{i,p})\left(1 - 10^{-Q_r/10}\right) + \mathbf{1}(S_r \neq A_{i,p})\left(10^{-Q_r/10}\right),$$

$\propto$ is a predetermined fractional value representing a likelihood or probability that $H_f = M$ arises,

i is the ith variant in the respective subset of N variant calls observed for the subset of sequence reads that include the same respective barcode,

r sums over the subset of sequence reads that include the same respective barcode,

$\mathbf{1}(S_r = A_{i,p})$ is an indicator function testing if the base assignment at position i in the rth sequence read S_r in the subset of sequence reads that include the same respective barcode matches Ai,p, wherein when they match $\mathbf{1}(S_r = A_{i,p})$ has a value of 1 and when they do not match $\mathbf{1}(S_r = A_{i,p})$ has a value of zero,

$\mathbf{1}(S_r \neq A_{i,p})$ is an indicator function testing if the base assignment at position i in the rth sequence read S_r in the subset of sequence reads that include the same respective barcode does not match Ai,p, wherein when they do not match $\mathbf{1}(S_r \neq A_{i,p})$ has a value of 1 and when they do match $\mathbf{1}(S_r = A_{i,p})$ has a value of zero, and

Q_r is a quality value for S_r for the read base at the position of i in the reference consensus sequence.

223. A non-transitory computer readable storage medium storing one or more programs configured for execution by a computer, the one or more programs comprising instructions for addressing error in the zygosity 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, wherein 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 collectively executing a method comprising:

(A) obtaining a reference consensus sequence for all or a portion of a genome of the species;

(B) obtaining a plurality of variant calls $A_{i;p}$ for the biological sample, wherein

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 zygosity error condition H_{-1};

(C) obtaining a plurality of sequence reads $\vec{\boldsymbol{O}}$ for the biological sample, wherein

each respective sequence read $\vec{\boldsymbol{O}}_i$ 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, and

each respective sequence read $\vec{\boldsymbol{O}}i$ in the plurality of sequence reads is $\in \{0, 1, -1, -\}^n$, wherein (i) n is the number of variants calls in $A_{i;p}$, (ii) each respective label 0 for the respective sequence read $\vec{\boldsymbol{O}}i$ assigns a corresponding variant call in $A_{i;p}$ to H_0, (iii) each respective label 1 for the respective sequence read $\vec{\boldsymbol{O}}i$ assigns a corresponding variant call in $A_{i;p}$ to H_0, (iv) each respective label -1 for the respective sequence read $\vec{\boldsymbol{O}}i$ assigns a corresponding variant call in $A_{i;p}$ to H_{-1}, and (v) each respective label – for the respective sequence read $\vec{\boldsymbol{O}}i$ indicates that the corresponding variant call in $A_{i;p}$ is not covered; and

(D) refining a phasing vector result $\vec{\boldsymbol{X}}$ 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:

$$\hat{\mathbf{X}} = \underset{\vec{\boldsymbol{X}}}{\operatorname{argmax}} P(\vec{\boldsymbol{X}}|\vec{\boldsymbol{O}}) \mid = \frac{P(\vec{\boldsymbol{X}})P(\vec{\boldsymbol{O}}|\vec{\boldsymbol{X}})}{C}$$

wherein,

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

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

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

$\hat{\mathbf{X}}$ is the refined phasing vector result,

C is a constant,

$\vec{\mathbf{X}}$ is the phasing vector result to be inferred, and

$(O_{1,f}, \dots, O_{N,f})$ is the respective subset of N variant calls in the plurality of variant calls $A_{i,p}$ observed in the subset of sequence reads that include the same respective barcode from the plurality of barcodes.

224. The non-transitory computer readable storage medium of claim 223, wherein

$$(O_{1,f}, \dots, O_{N,f}|\vec{\mathbf{X}},\ H_f = 0) = \prod_i P(O_{i,f}|A_{i,X_i}),$$

$$P(O_{1,f}, \dots, O_{N,f}|\vec{\mathbf{X}},\ H_f = 1) = \prod_i P(O_{i,f}|A_{i,1-X_i}),$$

$$P(O_{1,f}, \dots, O_{N,f}|\vec{\mathbf{X}},\ H_f = M) = \prod_i 0.5,$$

M indicates a mixture of $H_f = 0$ and $H_f = 1$ for the respective barcode f,

$$P(O_{i,f}, \dots, O_{N,f}|\vec{\mathbf{X}}) = \frac{(1-\propto)}{2}\left(\prod_i P(O_{i,f}|A_i, X_i) + \prod_i P(O_{i,f}|A_i, 1-X_i)\right) + \propto \prod_i 0.5,$$

$$\log P(O_{i,f}|A_{i,p}) = \sum_r \mathbf{1}(S_r = A_{i,p})\left(1 - 10^{-Q_r/10}\right) + \mathbf{1}(S_r \neq A_{i,p})\left(10^{-Q_r/10}\right)$$

$$+ \mathbf{1}(A_{i,p} = H_{-1})^{0.5}$$

$\propto$ is a predetermined fractional value representing a likelihood or probability that $H_f = M$ arises,

i is the ith variant in the respective subset of N variant calls observed for the subset of sequence reads that include the same respective barcode,

r sums over the subset of sequence reads that include the same respective barcode,

$\mathbf{1}(S_r = A_{i,p})$ is an indicator function testing if the base assignment at position i in the rth sequence read S_r in the subset of sequence reads that include the same respective barcode matches Ai,p, wherein when they match $\mathbf{1}(S_r = A_{i,p})$ has a value of 1 and when they do not match $\mathbf{1}(S_r = A_{i,p})$ has a value of zero,

$\mathbf{1}(S_r \neq A_{i,p})$ is an indicator function testing if the base assignment at position i in the rth sequence read S_r in the subset of sequence reads that include the same respective barcode does not match Ai,p, wherein when they do not match $\mathbf{1}(S_r \neq A_{i,p})$ has a value of 1 and when they do match $\mathbf{1}(S_r = A_{i,p})$ has a value of zero,

$\mathbf{1}(A_{i,p} = H_{-1})$ is an indicator function that has a value of 1 when $A_{i,p}$ is equal to H_{-1} and is a value of zero otherwise, and

Q_r is a quality value for S_r for the read base at the position of i in the reference consensus sequence.

225. A method of phasing sequencing data of a test nucleic acid sample obtained from a biological sample from a single organism of a species, wherein the test nucleic acid sample comprises a first set of haplotypes (H_0) and a second set of haplotypes (H_1), 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 variant calls $A_{i;p}$ for the test nucleic acid sample, wherein

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 $H = 1$;

(B) for each respective local block of variant calls in $A_{i;p}$ that are localized to a corresponding subset of the reference consensus sequence, using a beam search over the haplotype assignments of local phasing vectors $X_k, X_{k+1}, \ldots, X_{k+j}$ in the respective local block of variant calls, wherein

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 $X_k, X_{k+1}, \ldots, X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k, X_{k+1}, \ldots, X_{k+j}$, 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 $A_{i;p}$; and

(C) greedily joining, 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 $A_{i;p}$ 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.

226. A method of phasing sequencing data of a test nucleic acid sample obtained from a biological sample from a single organism of a species while accounting for error in variant call zygosity, wherein the test nucleic acid sample comprises a first set of haplotypes (H_0) and a second set of haplotypes (H_1), 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 variant calls $A_{i;p}$, wherein

i is an index to a position in a reference consensus sequence for all or a portion of a genome of the species, and

$p \in \{0, 1, -1\}$ 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 a zygosity error condition H_{-1},

(B) for each respective local block of variant calls in $A_{i;p}$ that are localized to a corresponding subset of the reference consensus sequence, using a beam search over the haplotype assignments of local phasing vectors $X_k, X_{k+1}, \ldots, X_{k+j}$ in the respective local block of variant calls, wherein

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 $X_k, X_{k+1}, \ldots, X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k, X_{k+1}, \ldots, X_{k+j}$, 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 $A_{i;p}$; and

(C) greedily joining, 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 $A_{i;p}$ 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.

227. The method of claim 225 or 226, the method further comprising iteratively swapping the phase result of individual x_i in $\widehat{\mathbf{X}}$ and recomputing the objective function, thereby obtaining $\widehat{\mathbf{X}}$.

228. The method of any one of claims 225-227, wherein a respective local block of variant calls consists of between 20 and 60 variants in $A_{i;p}$.

229. The method of any one of claims 225-227, wherein a respective local block of variant calls consists of between 30 and 80 variants in $A_{i;p}$.

230. The method of any one of claims 225-229, wherein an iteration of the beam search for the assignments of one of $X_k, X_{k+1}, \ldots, X_{k+j}$ discards all but a predetermined number of solutions for $\widehat{\mathbf{X}}$.

231. The method of any one of claims 225-230, wherein the predetermined number of solutions for $\widehat{\mathbf{X}}$ is 1000 or less.

232. The method of any one of claims 225-230, wherein the predetermined number of solutions for $\widehat{\mathbf{X}}$ is 5000 or less.

233. The method of any one of claims 225-232, wherein the test nucleic acid sample is loaded onto a plurality of barcoded-oligo coated gel-beads from which a plurality of sequence reads is obtained in order to derive the plurality of variant calls $A_{i;p}$ and wherein the test nucleic acid sample is 50 ng or less.

234. The method of claim 233 wherein the plurality of barcoded-oligo coated gel-beads comprises 10,000 beads.

235. The method of claim 233 wherein the plurality of barcoded-oligo coated gel-beads comprises 50,000 beads.

236. The method of claim 233 wherein the test nucleic acid sample is 25 ng or less.

237. The method of claim 233 wherein the test nucleic acid sample is 10 ng or less.

238. The method of claim 233 wherein the test nucleic acid sample is 5 ng or less.

239. The method of claim 233 wherein the test nucleic acid sample is 2.5 ng or less.

240. The method of claim 233, wherein the plurality of sequencing reads is obtained within ten minutes of exposure to the plurality of barcodes.

241. The method of claim 233, wherein the plurality of sequencing reads is obtained within twenty minutes of exposure to the plurality of barcodes.

242. 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 phasing sequencing data of a test nucleic acid sample obtained from a biological sample from a single organism of a species, wherein the test nucleic acid sample comprises a first set of haplotypes (H_0) and a second set of haplotypes (H_1) by executing a method comprising:

(A) obtaining a plurality of variant calls $A_{i;p}$, wherein

i is an index to a position in a reference consensus sequence for all or a portion of a genome of the species, and

$p \in \{0, 1, -1\}$ 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 a zygosity error condition H_{-1};

(B) for each respective local block of variant calls in $A_{i;p}$ that are localized to a corresponding subset of the reference consensus sequence, using a beam search over the haplotype assignments of local phasing vectors $X_k, X_{k+1}, \ldots, X_{k+j}$ in the respective local block of variant calls, wherein

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 $X_k, X_{k+1}, \ldots, X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k, X_{k+1}, \ldots, X_{k+j}$, 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 $A_{i;p}$; and

(C) greedily joining, 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 $A_{i;p}$ 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.

243. 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 phasing sequencing data of a test nucleic acid sample obtained from a biological sample from a single organism of a species while accounting for error in variant call zygosity, wherein 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 executing a method comprising:

(A) obtaining a plurality of variant calls $A_{i;p}$, wherein

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;

(B) for each respective local block of variant calls in $A_{i;p}$ that are localized to a corresponding subset of the reference consensus sequence, using a beam search over the haplotype assignments of local phasing vectors $X_k, X_{k+1}, \ldots, X_{k+j}$ in the respective local block of variant calls, wherein

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 $X_k, X_{k+1}, \ldots, X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k, X_{k+1}, \ldots, X_{k+j}$, 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 $A_{i;p}$; and

(C) greedily joining, 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 $A_{i;p}$ 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.

244. A non-transitory computer readable storage medium storing one or more programs configured for execution by a computer, the one or more programs comprising instructions for phasing sequencing data of a test nucleic acid sample obtained from a biological sample from a single organism of a species, wherein the test nucleic acid sample comprises a first set of haplotypes (H_O) and a second set of haplotypes (H_1), the one or more programs collectively executing a method comprising:

A) obtaining a plurality of variant calls $A_{i;p}$, wherein

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_O and label 1 assigns the respective variant call to H_1;

(B) for each respective local block of variant calls in $A_{i;p}$ that are localized to a corresponding subset of the reference consensus sequence, using a beam search over the haplotype assignments of local phasing vectors $X_k, X_{k+1}, \ldots, X_{k+j}$ in the respective local block of variant calls, wherein

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 $X_k, X_{k+1}, \ldots, X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k, X_{k+1}, \ldots, X_{k+j}$, 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 $A_{i;p}$; and

(C) greedily joining, 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 $A_{i;p}$ 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.

245. A non-transitory computer readable storage medium storing one or more programs configured for execution by a computer, the one or more programs comprising instructions for phasing sequencing data of a test nucleic acid sample obtained from a biological sample from a single organism of a species while accounting for error in variant call zygosity, wherein 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 collectively executing a method comprising:

(A) obtaining a plurality of variant calls $A_{i;p}$, wherein

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 $H = 1$;

(B) for each respective local block of variant calls in $A_{i;p}$ that are localized to a corresponding subset of the reference consensus sequence, using a beam search over the haplotype assignments of local phasing vectors $X_k, X_{k+1}, \ldots, X_{k+j}$ in the respective local block of variant calls, wherein

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 $X_k, X_{k+1}, \ldots, X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k, X_{k+1}, \ldots, X_{k+j}$, 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 $A_{i;p}$; and

(C) greedily joining, 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 $A_{i;p}$ 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.

246. The method of any one of claims 111-158, wherein the test nucleic acid is loaded onto a plurality of barcoded-oligo coated gel-beads from which the plurality of sequence reads is obtained and wherein the test nucleic acid is 50 ng or less.

247. The method of claim 246 wherein the plurality of barcoded-oligo coated gel-beads comprises 10,000 beads.

248. The method of claim 246 wherein the plurality of barcoded-oligo coated gel-beads comprises 50,000 beads.

249. The method of claim 246 wherein the test nucleic acid is 25 ng or less.

250. The method of claim 246 wherein the test nucleic acid is 10 ng or less.

251. The method of claim 246 wherein the test nucleic acid is 5 ng or less.

252. The method of claim 246 wherein the test nucleic acid is 2.5 ng or less.

253. The method of claim 246 wherein the plurality of sequencing reads is obtained within ten minutes of exposure to a plurality of barcodes.

254. The method of claim 246 wherein the plurality of sequencing reads is obtained within twenty minutes of exposure to a plurality of barcodes.