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 disclosure in order to provide a basic understanding of some of the aspects of the disclosure. This summary is not an extensive overview of the disclosure. It is not intended to identify key/critical elements of the disclosure or to delineate the scope of the disclosure. Its sole purpose is to present some of the concepts of the disclosure 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

**[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 embodiment 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 embodiment 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.

**[0033a]** In a first aspect, there is provided 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.

**[0034]** **Part B, Additional Embodiment for Structural Variation.** Another embodiment 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

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 $m$. 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 embodiment 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

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 embodiment 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 ($H_0$) and a second set of haplotypes ($H_1$). In other words, the biological sample is diploid and inherits maternal and paternal haplotypes. For instance, some portions of the genome of the biological sample are paternally inherited while other portions of the genome are maternally inherited. If the portions that are maternally inherited were to arbitrarily be designated the haplotype $H_0$, the portions that are paternally inherited are designated $H_1$.

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

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

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

**[0098]** In some embodiments, a respective local block of variant calls consists of between 20 and 60 variants in $A_{i;p}$. In some embodiments, a respective local block of variant calls consists of between 30 and 80 variants in $A_{i;p}$. In some embodiments, an iteration of the beam search for the assignments of one of $X_k$, $X_{k+1}$, …, $X_{k+j}$ discards all but a predetermined number of solutions for $\widehat{\mathbf{X}}$. In some embodiments, the predetermined number of solutions for $\widehat{\mathbf{X}}$ is 1000 or less. In some embodiments, predetermined number of solutions for $\widehat{\mathbf{X}}$ 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 x $10^{12}$}. 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 embodiment 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.

**[00102a]** In another aspect, there is provided a computing system, comprising:

one or more processors;

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

the one or more programs comprising instructions for phasing sequencing data of a test anucleic 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:

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

wherein,

$\widehat{\mathbf{X}}$ 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*,

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

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

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

*i* is the i[th] 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 *r*[th] 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 *r*[th] 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.

**[00102b]** In yet another aspect, there is provided a computing system, comprising:

one or more processors;

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

the one or more programs comprising instructions 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 $\vec{\boldsymbol{O}}$ 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:

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

wherein,

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

$\varepsilon_i$ is an estimate of incurring $H_{-1}$ at position $i$, and

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

$\widehat{\mathbf{X}}$ 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.

**[00102c]** In a further aspect, there is provided a computing system, comprising:

one or more processors;

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

the one or more programs comprising instructions for 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k$, $X_{k+1}$, …, $X_{k+j}$, 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 $\widehat{\mathbf{X}}$ for the single organism of the species.

**[00102d]** In yet another aspect, there is provided a computing system, comprising:

one or more processors;

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

the one or more programs comprising instructions for 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k$, $X_{k+1}$, …, $X_{k+j}$, 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 $\widehat{\mathbf{X}}$ for the single organism of the species.

**[00103]** Another embodiment 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k$, $X_{k+1}$, …, $X_{k+j}$, 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ discards all but a predetermined number of solutions for $\widehat{\mathbf{X}}$ (e.g., 1000 or less, 5000 or less, etc.).

**[00103a]** In another aspect, there is provided 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 $H_O$ 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:

$$\widehat{\mathbf{X}} = \underset{\vec{\mathbf{X}}}{\arg\max}\, P\left(\vec{\boldsymbol{O}}\middle|\vec{\mathbf{X}}\right)$$

wherein,

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

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

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

$(O_{1,f}, \ldots, O_{N,f})$ 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,

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

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

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

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

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

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

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

*i* is the i[th] 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,

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

$\mathbf{1}\left(S_r \neq A_{i,p}\right)$ is an indicator function testing if the base assignment at position *i* in the *r*[th] 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 $\mathbf{1}\left(S_r \neq A_{i,p}\right)$ has a value of 1 and when they do match $\mathbf{1}\left(S_r = A_{i,p}\right)$ 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.

**[00103b]** In yet another aspect, there is provided 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 ($H_O$) 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k$, $X_{k+1}$, …, $X_{k+j}$, 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 $\widehat{\mathbf{X}}$ for the single organism of the species.

**[00103c]** In a further aspect, there is provided 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k$, $X_{k+1}$, …, $X_{k+j}$, 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 $\widehat{\mathbf{X}}$ for the single organism of the species.

**[00104]** Another embodiment 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 $H_O$, label 1 assigns the respective variant call to $H_1$, and label -1 assigns the respective variant call to the zygosity error condition $H_{-1}$. 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 $H_O$, (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. 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 $H_O$, $H_1$ and $H_{-1}$ for the plurality of sequence reads using an overall objective function:

**[00105]** $$\widehat{\mathbf{X}} = \underset{\vec{\boldsymbol{X}}}{\operatorname{argmax}} P\left(\vec{\boldsymbol{X}}|\vec{\boldsymbol{O}}\right) \Big| = \frac{P(\vec{\boldsymbol{X}})P(\vec{\boldsymbol{O}}|\vec{\boldsymbol{X}})}{C}$$

**[00106]** where

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

**[00108]** $\varepsilon_i$ is an estimate of incurring $H_{-1}$ at position $i$, and

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

**[00110]** $\widehat{\mathbf{X}}$ 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.

**[00113a]** In another aspect, there is provided 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 $\vec{\boldsymbol{X}}$ 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:

$$\widehat{\mathbf{X}} = \underset{\vec{\boldsymbol{X}}}{\operatorname{argmax}} P(\vec{\boldsymbol{X}}|\vec{\boldsymbol{O}}) \Big| = \frac{P(\vec{\boldsymbol{X}})P(\vec{\boldsymbol{O}}|\vec{\boldsymbol{X}})}{C}$$

wherein,

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

$\varepsilon_i$ is an estimate of incurring $H_{-1}$ at position $i$, and

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

$\widehat{\mathbf{X}}$ 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.

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

**[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 i[th] 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 $r^{th}$ 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,

**[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.

**[00144a]** In yet another aspect, there is provided 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,

$\widehat{\mathbf{X}}$ 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*,

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

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

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

*i* is the i[th] 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 *r*[th] 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 *r*[th] 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.

**[00144b]** In yet another aspect, there is provided 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 $H_O$, 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 $\vec{\boldsymbol{X}}$ by optimization of haplotype assignments at individual positions $i$ in $A_{i;p}$ between $H_O$, $H_1$ and $H_{-1}$ for the plurality of sequence reads using an overall objective function:

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

wherein,

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

$\varepsilon_i$ is an estimate of incurring $H_1$ at position $i$, and

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

$\widehat{\mathbf{X}}$ 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.

**[00144c]** In a further aspect, there is provided 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 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k$, $X_{k+1}$, …, $X_{k+j}$, 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 $\widehat{\mathbf{X}}$ for the single organism of the species.

**[00144d]** In another aspect, there is provided 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k$, $X_{k+1}$, …, $X_{k+j}$, 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 $\widehat{\mathbf{X}}$ for the single organism of the species.

**[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.

**What is claimed is:**

1. 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.

2. The non-transitory computer readable storage medium of claim 1, wherein the plurality of barcodes comprises 1000 or more barcodes.

3. The non-transitory computer readable storage medium of claim 1, wherein the plurality of barcodes comprises 10,000 or more barcodes.

4. The non-transitory computer readable storage medium of claim 1, wherein the plurality of barcodes comprises 100,000 or more barcodes.

5. The non-transitory computer readable storage medium of claim 1, wherein the plurality of barcodes comprises 1 x $10^6$ or more barcodes.

6. 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 $\vec{O}$ 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 $\in \{0, 1, -\}^n$, 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 $H_O$, 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:

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

wherein,

$\widehat{\mathbf{X}}$ 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*,

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

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

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

*i* is the i$^{\text{th}}$ 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 $r^{\text{th}}$ 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 $r^{th}$ 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.

7. 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 $H_0$, 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:

$$\widehat{\mathbf{X}} = \underset{\vec{\boldsymbol{X}}}{\operatorname{argmax}} P(\vec{\boldsymbol{X}}|\vec{\boldsymbol{O}}) \Big| = \frac{P(\vec{\boldsymbol{X}})P(\vec{\boldsymbol{O}}|\vec{\boldsymbol{X}})}{C}$$

wherein,

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

$\varepsilon_i$ is an estimate of incurring $H_{-1}$ at position *i*, and

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

$\widehat{\mathbf{X}}$ 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.

8. The method of claim 7, 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*,

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

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

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

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

*i* is the i[th] 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 $r^{th}$ 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 $r^{th}$ 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.

9. The method of any one of claims 6-8, wherein the plurality of variant calls comprises a plurality of heterozygous single nucleotide polymorphisms, heterozygous inserts, or heterozygous deletions in the test nucleic acid.

10. The method of any one of claims 6-9, 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.

11. The method of any one of claims 6-10, wherein the plurality of barcodes comprises 1000 or more barcodes.

12. The method of any one of claims 6-10, wherein the plurality of barcodes comprises 10,000 or more barcodes.

13. The method of any one of claims 6-10, wherein the plurality of barcodes comprises 100,000 or more barcodes.

14. The method of any one of claims 6-10, wherein the plurality of barcodes comprises 1 x $10^6$ or more barcodes.

15. The method of any one of claims 6-14, wherein the species is human.

16. The method of any one of claims 6-15, wherein the plurality of variant calls $A_{i;p}$ comprises 1000 or more variant calls.

17. The method of any one of claims 6-15, wherein the plurality of variant calls $A_{i;p}$ comprises 10,000 or more variant calls.

18. The method of any one of claims 6-15, wherein the plurality of sequence reads comprises 10,000 or more sequence reads.

19. The method of any one of claims 6-18, wherein the plurality of sequence reads comprises 100,000 or more sequence reads.

20. The method of any one of claims 6-19, wherein the plurality of sequence reads comprises 1 x $10^6$ or more sequence reads.

21. The method of any one of claims 6-20, wherein $\vec{\boldsymbol{X}}\ is\ (x)$, 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 ($H_1$).

22. The method of any one of claims 6-21, 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.

23. The method of any one of claims 6-21, 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.

24. The method of any one of claims 6-23, wherein the subset of sequence reads that include the same respective barcode *f* comprises 10 or more sequence reads.

25. The method of any one of claims 6-23, wherein the subset of sequence reads that include the same respective barcode *f* comprises 30 or more sequence reads.

26. The method of any one of claims 6-23, wherein the subset of sequence reads that include the same respective barcode *f* comprises 100 or more sequence reads.

27. The method of any one of claims 6-26, wherein a subset of sequence reads that include the same respective barcode *f* represent at least 30 kilobases of the reference consensus sequence.

28. The method of any one of claims 6-26, wherein a subset of sequence reads that include the same respective barcode *f* represent at least 40 kilobases of the reference consensus sequence.

29. The method of any one of claims 6-28, wherein the refining (D) optimizes the overall objective function using a hierarchical search over $\vec{\boldsymbol{X}}$.

30. The method of claim 29, 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ 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 $\widehat{\mathbf{X}}$.

31. The method of claim 30, wherein the refining the phase result further comprises iteratively swapping the phase result of individual $x_{\mathrm{i}}$ in the estimate of the optimal phasing configuration $\widehat{\mathbf{X}}$ and recomputing the objective function, thereby obtaining $\widehat{\mathbf{X}}$.

32. The method of claim 30, wherein a respective local block of variant calls consists of between 20 and 60 variants in $A_{i;p}$.

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

34. The method of claim 30, 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 $\widehat{\mathbf{X}}$.

35. The method of claim 34, wherein the predetermined number of solutions for $\widehat{\mathbf{X}}$ is 1000 or less.

36. The method of claim 34, wherein the predetermined number of solutions for $\widehat{\mathbf{X}}$ is 5000 or less.

37. The method of any one of claims 6-36, wherein the species is human and the test nucleic acid sample comprises the genome of the biological sample.

38. The method of any one of claims 6-36, 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.

39. The method of any one of claims 6-38, 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, …, $1 \times 10^{12}$}.

40. The method of any one of claims 6-39, 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.

41. The method of claim 40, wherein the contiguous set of oligonucleotides is an N-mer, wherein N is an integer selected from the set {4, …, 20}.

42. The method of any one of claims 6-39, 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.

43. The method of claim 42, wherein the predetermined noncontiguous set of nucleotides collectively consists of N nucleotides, wherein N is an integer in the set {4, …, 20}.

44. The method of any one of claims 6-43, 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.

45. The method of any one of claims 6-43, 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.

46. The method of any one of claims 6-45, wherein the plurality of variant calls is obtained from the plurality of sequence reads.

47. The method of any one of claims 6-46, 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.

48. The method of claim 47 wherein the plurality of barcoded-oligo coated gel-beads comprises 10,000 beads.

49. The method of claim 47 wherein the plurality of barcoded-oligo coated gel-beads comprises 50,000 beads.

50. The method of claim 47 wherein the test nucleic acid sample is 25 ng or less.

51. The method of claim 47 wherein the test nucleic acid sample is 10 ng or less.

52. The method of claim 47 wherein the test nucleic acid sample is 5 ng or less.

53. The method of claim 47 wherein the test nucleic acid sample is 2.5 ng or less.

54. The method of any one of claims 6-53, wherein the plurality of sequencing reads $\vec{O}$ is obtained within ten minutes of exposure to the plurality of barcodes.

55. The method of any one of claims 6-53, wherein the plurality of sequencing reads $\vec{O}$ is obtained within twenty minutes of exposure to the plurality of barcodes.

56. 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 anucleic 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:

$$\widehat{\mathbf{X}} = \underset{\vec{\mathbf{X}}}{\arg\max}\, P\left(\vec{\boldsymbol{O}} \middle| \vec{\mathbf{X}}\right)$$

wherein,

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

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

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

$(O_{1,f}, \ldots, O_{N,f})$ 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,

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

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

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

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

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

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

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

*i* is the i[th] 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,

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

$\mathbf{1}\left(S_r \neq A_{i,p}\right)$ is an indicator function testing if the base assignment at position *i* in the *r*[th] 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 $\mathbf{1}\left(S_r \neq A_{i,p}\right)$ has a value of 1 and when they do match $\mathbf{1}\left(S_r = A_{i,p}\right)$ 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.

57. 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 $H_O$, 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 $\vec{\boldsymbol{X}}$ by optimization of haplotype assignments at individual positions $i$ in $A_{i;p}$ between $H_O$, $H_1$ and $H_{-1}$ for the plurality of sequence reads using an overall objective function:

$$\hat{\mathbf{X}} = \underset{\vec{\boldsymbol{X}}}{\operatorname{argmax}} P\left(\vec{\boldsymbol{X}}|\vec{\boldsymbol{O}}\right) \Big| = \frac{P\left(\vec{\boldsymbol{X}}\right)P\left(\vec{\boldsymbol{O}}|\vec{\boldsymbol{X}}\right)}{C}$$

wherein,

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

$\varepsilon_i$ is an estimate of incurring $H_{-1}$ at position $i$, and

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

$\widehat{\mathbf{X}}$ 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.

58. The computing system of claim 57, wherein

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

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

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

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

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

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

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

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

*i* is the i$^{th}$ 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,

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

$\mathbf{1}\left(S_r \neq A_{i,p}\right)$ is an indicator function testing if the base assignment at position *i* in the *r*$^{th}$ 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.

59. 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:

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

wherein,

$\widehat{\mathbf{X}}$ 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*,

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

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

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

*i* is the i[th] 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 *r*[th] 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 *r*[th] 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.

60. 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:

$$\widehat{\mathbf{X}} = \underset{\vec{\boldsymbol{X}}}{\operatorname{argmax}} P(\vec{\boldsymbol{X}}|\vec{\boldsymbol{O}}) \Big| = \frac{P(\vec{\boldsymbol{X}})P(\vec{\boldsymbol{O}}|\vec{\boldsymbol{X}})}{C}$$

wherein,

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

$\varepsilon_i$ is an estimate of incurring $H_{-1}$ at position $i$, and

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

$\widehat{\mathbf{X}}$ 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.

61. The non-transitory computer readable storage medium of claim 60, 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 i$^{th}$ 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 $r^{th}$ 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 $r^{th}$ 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.

62. 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 ($H_O$) 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k$, $X_{k+1}$, …, $X_{k+j}$, 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 $\widehat{\mathbf{X}}$ for the single organism of the species.

63. 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k$, $X_{k+1}$, …, $X_{k+j}$, 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 $\widehat{\mathbf{X}}$ for the single organism of the species.

64. The method of claim 62 or 63, the method further comprising iteratively swapping the phase result of individual $x_i$ in $\widehat{\mathbf{X}}$ and recomputing the objective function, thereby obtaining $\widehat{\mathbf{X}}$.

65. The method of any one of claims 62-64, wherein a respective local block of variant calls consists of between 20 and 60 variants in $A_{i;p}$.

66. The method of any one of claims 62-64, wherein a respective local block of variant calls consists of between 30 and 80 variants in $A_{i;p}$.

67. The method of any one of claims 62-66, 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 $\widehat{\mathbf{X}}$.

68. The method of any one of claims 62-67, wherein the predetermined number of solutions for $\widehat{\mathbf{X}}$ is 1000 or less.

69. The method of any one of claims 62-67, wherein the predetermined number of solutions for $\widehat{\mathbf{X}}$ is 5000 or less.

70. The method of any one of claims 62-69, 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.

71. The method of claim 70 wherein the plurality of barcoded-oligo coated gel-beads comprises 10,000 beads.

72. The method of claim 70 wherein the plurality of barcoded-oligo coated gel-beads comprises 50,000 beads.

73. The method of claim 70 wherein the test nucleic acid sample is 25 ng or less.

74. The method of claim 70 wherein the test nucleic acid sample is 10 ng or less.

75. The method of claim 70 wherein the test nucleic acid sample is 5 ng or less.

76. The method of claim 70 wherein the test nucleic acid sample is 2.5 ng or less.

77. The method of claim 70, wherein the plurality of sequencing reads is obtained within ten minutes of exposure to the plurality of barcodes.

78. The method of claim 70, wherein the plurality of sequencing reads is obtained within twenty minutes of exposure to the plurality of barcodes.

79. 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k$, $X_{k+1}$, …, $X_{k+j}$, 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 $\widehat{\mathbf{X}}$ for the single organism of the species.

80. 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k$, $X_{k+1}$, …, $X_{k+j}$, 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 $\widehat{\mathbf{X}}$ for the single organism of the species.

81. 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 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k$, $X_{k+1}$, …, $X_{k+j}$, 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 $\widehat{\mathbf{X}}$ for the single organism of the species.

82. 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ 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 $X_k$, $X_{k+1}$, …, $X_{k+j}$ are found by computing an objective function in which the phasing vector of the objective function in respective computations is limited to $X_k$, $X_{k+1}$, …, $X_{k+j}$, 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 $\widehat{\mathbf{X}}$ for the single organism of the species.