NUCLEIC ACID SEQUENCE ASSEMBLY

CROSS-REFERENCE

[0001] This application claims the benefit of U.S. Provisional Application No. 62/117,256, filed February 17, 2015, which is hereby explicitly incorporated by reference in its entirety, and this application also claims the benefit of U.S. Provisional Application No. 62/294,208, filed February 11, 2016, which is hereby explicitly incorporated by reference in its entirety.

BACKGROUND

[0002] Currently accessible and affordable high-throughput sequencing methods are best suited to the characterization of short-range sequence contiguity and genomic variation. Achieving long-range linkage and haplotype phasing requires either the ability to directly and accurately read long (*e.g.*, tens of kilobases) sequences, or the capture of linkage and phase relationships through paired or grouped sequence reads. However, grouping sequencing information and generating an assembly of the sequence information necessary to achieve long-range linkage and haplotype phasing is computationally intensive and time consuming. Disclosed herein are computationally efficient methods and systems to obtain assemblies with chromosome-scale contiguity from sequence information informed by paired or grouped sequence reads.

SUMMARY

[0003] Disclosed herein are methods, compositions, algorithms, and systems related to the scaffolding of nucleic acid data. Approaches herein utilize read pairs to infer information regarding the phase or physical linkage information of contigs to which the reads of a read pair map in a dataset. Contigs in a nucleic acid dataset are ordered, oriented or merged end to end or in some cases inserted one into another (collectively "scaffolded") in light of the impact of such activity on a score or parameter relating to their relative positioning.

[0003a] In an aspect, the present disclosure provides a method for nucleic acid sequence data assembly, comprising: (a) obtaining a set of paired-end reads obtained by digesting sample DNA to generate internal double strand breaks within the DNA, allowing the double strand breaks to re-ligate randomly to form a plurality of junctions, and sequencing across the plurality of junctions; (b) obtaining standard paired-end read distance frequency data; (c) obtaining grouped contig sequences; and (d) scaffolding the grouped contig sequences such that read-pair distance frequency data for read pairs that map to separate contigs approximates the standard paired-end read distance frequency data, thereby assembling nucleic acid sequence data of the nucleic acid.

[0004] In some cases the score or parameter is a measure of the resulting impact of contig repositioning on aggregate read pair separation for a read pair dataset of one or another contig configuration. Depending on the approach used to generate it, a dataset of read pairs may have a particular read pair separation distribution curve. Mapped as read pair separation as a function of frequency, one can determine an expected read pair distance distribution for a given read pair dataset. One may then map the read pairs to a set of contigs, and position the contigs (in order, orientation or otherwise) so as to make the read pair distance distribution for the data set match, approximate, or more closely approximate the read pair distance distribution expected given a nucleic acid sample and a method of read pair generation.

common contig. Alternately or in combination, standard paired-end read frequency is obtained from previously generated curves. The initial configuration is a random configuration, or is preconfigured. In preferred examples, read pair distance frequency data for read pairs that map to separate contigs more closely approximates the paired-end read distance frequency data when read pair distance likelihood increases. In many cases, read-pair distance likelihood is maximized. Read pair distance frequency data for read pairs that map to separate contigs more closely approximates the paired-end read distance frequency data when a statistical measure of difference between the read pair distance frequency data and the standard paired-end read distance frequency data decreases. A number of statistical measures are available. For example, statistical measure of distance between the read pair distance frequency data and the standard paired-end read distance frequency data comprises at least one of ANOVA, a t-test, and a X-squared test in various cases. Read pair distance for read pairs that map to separate contigs more closely matches the paired-end read distance frequency data when deviation of read pair distance distribution among ordered contigs obtained as compared to standard paired-end read distance frequency decreases. Alternately or in combination, deviation of read pair distance distribution among ordered contigs obtained as compared to standard paired-end read distance frequency is minimized. In some scaffold assessment, a contig that shares sequence in a paired end read associated with a first cluster and a second cluster is assigned to a cluster having a greater number of shared end reads. Clustering often comprises placing contigs into a number of groups that is greater than or equal to the number of chromosomes in the organism. Often, a contig sharing only a single paired end read with one contig of a cluster is not included in that cluster. A contig sharing with a cluster only at least one paired end read comprising repetitive sequence is often not included in that cluster. Similarly, a contig sharing with a cluster only at least one paired end read comprising low quality sequence is often not included in that cluster. In some methods the set of paired-end reads are obtained by digesting sample DNA to generate internal double strand breaks within the nucleic acid, allowing the double strand breaks to re-ligate to form at least one religation junction, and sequencing across at least one religation junction. The DNA is crosslinked to at least one DNA binding agent, such as a nuclear protein or a nanoparticle, in some approaches to paired read generation. The DNA is isolated naked DNA that is reassembled into reconstituted chromatin, although DNA having binding proteins is suitable under some circumstances, particularly if DNA molecules are not bound to one another. Often, the reconstituted chromatin is crosslinked. The reconstituted chromatin comprises a DNA binding protein. Alternately or in combination, the reconstituted chromatin comprises a nanoparticle. Preferably in some cases, clustering of contigs is independent of the number or chromosomes for the organism. A contig that shares sequence in a paired end read associated

aggregate measure of the read-pair distance. This is followed in many cases by iteratively selecting at least one additional set until a majority of said clustered contigs are ordered. Selecting often involves iteratively selecting at least one additional set until each of said clustered contigs are ordered. The nucleic acid sequence is derived from a sample such as a genome, or in some cases, a plurality of genomes. In some examples, the methods are practiced on a computer-implemented system comprising a processor configured to receive contig and read pair data as discussed herein, process the data as discussed above, and output scaffolded contig data having the improved parameters as discussed above.

[0008] Provided herein are methods for scaffolding contigs in a cluster, comprising: assigning a log-likelihood ratio score for each pair of contigs; sorting connections by ratio score; and accepting or rejecting contig connections in decreasing order of ratio score so as to increase the total score of the assembly. In some methods, the scaffolding comprises ordering the set of contigs, and/or orienting the set of contigs, and/or merging at least two contigs end to end, and/or inserting one contig into a second contig, and/or cleaving a contig into at least two constituent contigs. In many cases, the contigs comprise a genome, or a plurality of genomes. In some examples, the methods are practiced on a computer-implemented system comprising a processor configured to receive contig and read pair data as discussed herein, process the data as discussed above, and output scaffolded contig data having the improved parameters as discussed above.

[0009] Provided herein are methods for determining locally optimal contig configuration of a plurality of contigs within a cluster. Some such methods comprise a) identifying a sequence window of size *w* contigs starting at position i along a cluster of contigs; b) considering $w!\ 2^w$ ordering and orienting options for the contigs of window *w* by examining the scores of compatible orders and orientations in each position i in the window; c) orienting and ordering said *w* contigs in said window to obtain an optimal score; d) shifting the window to position $i+1$; and e) repeating steps (a), (b) and (c) for said window at position $i+1$ using the orienting and ordering of said w contigs to determine and optimal score; thereby orienting and ordering said plurality of contigs in a locally optimal configuration relative to the score. In some methods, read pair data mapping to the plurality of contigs in the cluster is obtained, a standard paired-end read frequency data set is obtained, and the score for an orienting and ordering of said *w* contigs is a measure of how closely a read pair distance data set for the read pair data mapping to the plurality of contigs in the cluster matches the standard paired-end read frequency data set. In some methods, read pair data mapping to the plurality of contigs in the cluster is obtained, the score is total read pair distance, and the score is optimized when total read pair distance is minimized. The window size *w* is 3, or alternately *w* is 4, or alternately *w* is 5, or alternately *w* is

6. In some cases *w* has a first value for a first cluster and *w* has a second value at a second cluster. *w* is selected in some methods to comprise 1% of the contigs of the set, or alternately 5% of the contigs of the set, or alternately 10% of the contigs of the set. In many methods the score is a read pair distance likelihood score, and the score is optimal when the score is maximized for a given window size. The score is calculated using formula 1 in some exemplary examples. The score is a deviation from an expected read pair distribution, and is optimal when the score is minimized for a given window size in some cases. The plurality of contigs comprises a genome, or a plurality of genomes, or a non-genomic nucleic acid source. In some examples, the methods are practiced on a computer-implemented system comprising a processor configured to receive contig and read pair data as discussed herein, process the data as discussed above, and output scaffolded contig data having the improved parameters as discussed above.

[0010] Provided herein are methods for nucleic acid sequence assembly, comprising: obtaining purified DNA; binding the purified DNA with a DNA binding agent to form DNA/chromatin complexes; incubating the DNA-chromatin complexes with restriction enzymes to leave sticky ends; performing ligation to join ends of DNA; sequencing across ligated DNA junctions to generate paired end reads; and using the paired end reads to scaffold a nucleic acid data set comprising contigs representing sequence of the purified DNA. In some methods the purified DNA is derived from a genome, or from a plurality of genomes. In some examples, the methods are practiced on a computer-implemented system comprising a processor configured to receive contig and read pair data as discussed herein, process the data as discussed above, and output scaffolded contig data having the improved parameters as discussed above.

[0011] Provided herein are methods for identifying a read-paired sequence read mapping to a repetitive contig region, comprising: obtaining a contig dataset for a nucleic acid sample; obtaining at least one read-paired sequence read corresponding to nonadjacent physically linked sequence information; and excluding the read-paired sequence read if at least one read of the read paired sequence read maps to two distinct loci of a contig data set. In some methods the repetitive region comprises sequence having a shotgun read depth exceeding a first threshold. In some methods the repetitive region comprises a base position having a read depth exceeding a second threshold. Often, the first threshold and second threshold are fixed relative to the overall distribution of read depth. The first threshold is 3 times the overall distribution of read depth in many cases. Alternately, the first threshold is 2, 2.5, 3.5, 4, 4.5, 5, 5.5, 6, or a non-integer value within or adjacent to this set. The second threshold is often 3.5 times the overall distribution of read depth. Alternately, the second threshold is 2, 2.5, 3, 4, 4.5, 5, 5.5, 6, or a non-integer value within or adjacent to this set. In some methods the purified DNA is derived from a genome, or from a plurality of genomes. In some examples, the methods are practiced on a computer-

implemented system comprising a processor configured to receive contig and read pair data as discussed herein, process the data as discussed above, and output scaffolded contig data having the improved parameters as discussed above.

[0012] Provided herein are methods for guiding contig assembly decisions, comprising the step of determining the probability of observing the number and implied separations of spanning read-paired sequence between a first contig and a second contig, wherein the contigs have relative orientations of o within the set [++,+-,-+,--] and are separated by a gap length. Some methods further comprise normalizing the probability of distribution of read-pair sequence over separation distances, wherein normalizing includes comparing the read-pair sequence to noise pairs which sample the nucleic acid sample independently. In some cases the nucleic acid sample comprises a genome. Alternately, the nucleic acid sample comprises a plurality of genomes, or a non-genomic source. Often, the total number of noise pair is determined by tabulating the densities of links for a sample of contig pairs. Further provided herein are methods wherein the highest and lowest 1% of densities are excluded. In alternatives thereto, the highest .5%, .6%, .7%, .8%, .9%, 1.1%, 1.2%, 1.3%, 1.4%, 1.5%, 1.6%, 1.7%, 1.8%, 1.9%, 2%, 3%, 4%, 5%, or greater than 5% are excluded, and similarly the lowest .5%, .6%, .7%, .8%, .9%, 1.1%, 1.2%, 1.3%, 1.4%, 1.5%, 1.6%, 1.7%, 1.8%, 1.9%, 2%, 3%, 4%, 5%, or greater than 5% are excluded. Some methods comprise determining contig order. Some methods comprise determining contig orientation. Some methods comprise determining both contig order and contig orientation. In some examples, the methods are practiced on a computer-implemented system comprising a processor configured to receive contig and read pair data as discussed herein, process the data as discussed above, and output scaffolded contig data having the improved parameters as discussed above.

[0013] Provided herein are methods for contig misjoin correction comprising obtaining a set of contig sequences having an initial configuration; obtaining a set of paired end reads; obtaining standard paired-end read distance frequency data; grouping contig pairs sharing sequence that coexists in at least one paired end read; comparing read pair frequency data for the grouping of the contigs to the standard paired-end read distance frequency data; determining whether introducing a break in a contig of the grouping causes the read pair frequency data for the grouping of the contigs to more closely approximate the standard paired-end read distance frequency data; and, if the read pair frequency data for the grouping of the contigs to more closely approximate the standard paired-end read distance frequency data, then introducing a break into the contig. In some methods the first position is merged with at least one adjacent second position having said log likelihood below said threshold prior to introducing the break. The second adjacent position is no greater than 300 bases pairs from the first position.

Alternately, the second position does not include positions greater than 1000 base pairs from the first position. Alternately, the second adjacent position is no greater than 50, 100, 150, 200, 250, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850, 900, 950, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800 1900, or 2000, or an integer value within the range spanned by the values recited. Further provided herein are methods wherein determining the log likelihood change comprises identifying an average paired end mapping density for a contig, identifying segments of the contig having a paired end mapping density of at least 3x that of the average paired end mapping density, and excluding segments of the contig having a paired end mapping density of at least 3x that of the average paired end mapping density. Alternately, a threshold of 1.5x, 1.6x, 1.7x, 1.8x, 1.9x, 2.0x, 2.1x, 2.2x, 2.3x, 2.4x, 2.5x, 2.6x, 2.7x, 2.8x, 2.9x, 3.1x, 3.2x, 3.3x, 3.4x, 3.5x, 3.6x, 3.7x, 3.8x, 3.9x, 4x, 4.1x, 4.2x, 4.3x, 4.4x, 4.5x, 4.6x, 4.7, 4.8x, 4.9x, 5x or greater than 5x is used. Further provided herein are methods wherein the set of contig sequences is derived from a genome. Further provided herein are methods wherein the set of contig sequences is derived from a plurality of genomes. In some examples, the methods are practiced on a computer-implemented system comprising a processor configured to receive contig and read pair data as discussed herein, process the data as discussed above, and output scaffolded contig data having the improved parameters as discussed above.

[0014] Provided herein are methods for contig assembly, comprising: indicating broken contigs of the starting assembly, wherein broken contigs are nodes and edges of the broken contigs are labeled with a list of ordered pairs of integers and wherein the edges of the breaks correspond to mapped read-paired sequence; and excluding edges with fewer than a threshold number of mapped connections. In some methods the threshold number is less than 5%. Alternately, the number is less than 20%, 15%, 14%, 13%, 12%, 11%, 10%, 9%, 8%, 7%, 6%, 4%, 3%, 2%, 1% or lower. In some cases the threshold number is fewer than t_L links. In some methods contigs comprise edges where the ratio of the degree in the graph of a corresponding node to contig length is base pairs exceeds about 5% of high end of the distribution all values. In some methods the contigs are derived from a genome. In some methods the contigs are derived from a plurality of genomes. In some examples, the methods are practiced on a computer-implemented system comprising a processor configured to receive contig and read pair data as discussed herein, process the data as discussed above, and output scaffolded contig data having the improved parameters as discussed above.

[0015] Provided herein are methods of assembling contig sequence information into at least one scaffold, comprising obtaining sequence information corresponding to a plurality of contigs, obtaining paired-end read information from a nucleic acid sample represented by the plurality of contigs, and configuring the plurality of contigs such that deviation of a read pair distance

parameter from a predicted read pair distance data set is minimized, wherein the configuring occurs in less than 8 hours. The predicted read pair distance data set comprises a read pair distance likelihood curve in many preferred examples. In some cases the read pair distance parameter is maximum distance likelihood relative to a read pair distance likelihood curve. Alternately, the read pair distance parameter is minimum variation relative to a read pair distance likelihood curve. The locally adjacent set of contigs comprises 2 contigs. Alternately, the locally adjacent set of contigs comprises 3 contigs. Alternately, the locally adjacent set of contigs comprises 4 contigs. Alternately, the locally adjacent set of contigs comprises 5 contigs. Alternately, the locally adjacent set of contigs comprises 6 contigs. Preferably, the configuring occurs in less than 7 hours. Alternately, the configuring occurs in less than 6 hours, 5 hours, 4 hours, 3 hours, 2 hours, 1 hour, or less than 1 hour. The contig information is derived from a genome in many cases. Alternately, the contig sequence information is derived from a plurality of genomes. In some examples, the methods are practiced on a computer-implemented system comprising a processor configured to receive contig and read pair data as discussed herein, process the data as discussed above, and output scaffolded contig data having the improved parameters as discussed above.

[0016] Provided herein are methods of scaffolding a set of contig sequences comprising obtaining a set of contig sequences representative of a nucleic acid sample obtaining read pair data for the nucleic acid sample, and ordering and orienting the set of contig sequences such that read pair data for the nucleic acid sample more closely approximates an expected read pair distribution, wherein 70% of the set of contig sequences are ordered and oriented so as to match the relative order and orientation of their sequences in the nucleic acid sample in no more than 8 hours. The scaffolding comprises at least one of ordering the set of contigs, orienting the set of contigs, merging at least two contigs end to end, inserting one contig into a second contig, and/or cleaving a contig into at least two constituent contigs. In some methods, 80% of the set of contig sequences are ordered and oriented so as to match the relative order and orientation of their sequences in the nucleic acid sample in no more than 8 hours. Alternately, 90% of the set of contig sequences are ordered and oriented so as to match the relative order and orientation of their sequences in the nucleic acid sample in no more than 8 hours. Alternately, 95% of the set of contig sequences are ordered and oriented so as to match the relative order and orientation of their sequences in the nucleic acid sample in no more than 8 hours. In some cases, 70% of the set of contig sequences are ordered and oriented so as to match the relative order and orientation of their sequences in the nucleic acid sample in no more than 4 hours, or alternately in no more than 2 hours. or alternately in no more than 1 hour. The contig information is derived from a

genome in many cases. Alternately, the contig sequence information is derived from a plurality of genomes.

[0017] Provided herein are methods of configuring a set of nucleic acid sequence data comprising: obtaining sequence information corresponding to a plurality of contigs which comprise the scaffold, obtaining pair-end read information, and configuring the plurality of contigs such that paired-end read distance distribution for the paired-end read information is globally optimized to approximate a reference paired-end read distance distribution, wherein the configuring occurs in less than 8 hours. The contig information is derived from a genome in many cases. Alternately, the contig sequence information is derived from a plurality of genomes. In some examples, the methods are practiced on a computer-implemented system comprising a processor configured to receive contig and read pair data as discussed herein, process the data as discussed above, and output scaffolded contig data having the improved parameters as discussed above.

[0018] Provided herein are methods of improving a scaffold assembly comprising obtaining a scaffold set comprising a plurality of joined nodepairs, wherein each node of a node pair comprises at least one contig sequence, obtaining pair-end read information mapped to the plurality of joined nodes, counting the number of read pairs shared by a joined nodepair, comparing said number to a threshold, and cleaving a node pair into unjoined nodes if said number falls below a threshold. In some cases only read pairs mapping to unique contig sequence are counted. Further provided herein are methods wherein read pairs mapping to a contig sequence segment to which a threshold number of distinct reap pair ends map are discarded. The threshold number is 3x the average number for non-repetitive sequence in many cases. Alternately, threshold values of 1.5x, 1.6x, 1.7x, 1.8x, 1.9x, 2x, 2.1x, 2.2x, 2.3x, 2.4x, 2.5x, 2.6x, 2.7x, 2.7x, 2.8x, 2.9x, 3,1x, 3.2x, 3.3x, 3.4x, 3.5x, 3.6x, 2.7x, 2.8x, 3.9x, 4x, 4.5x, 5x or greater than 5x are employed. The contig information is derived from a genome in many cases. Alternately, the contig sequence information is derived from a plurality of genomes. In some examples, the methods are practiced on a computer-implemented system comprising a processor configured to receive contig and read pair data as discussed herein, process the data as discussed above, and output scaffolded contig data having the improved parameters as discussed above.

[0019] Provided herein are methods of improving a scaffold assembly comprising obtaining a scaffold set comprising a plurality of joined nodepairs, wherein each node of a node pair comprises at least one contig sequence, obtaining pair-end read information mapped to the plurality of joined nodes, obtaining standard paired-end read distance frequency data; comparing pair-end read frequency data for the paired end read information mapped to the plurality of

joined nodes to the standard paired-end read distance frequency data; and cleaving at least one joined node if cleaving the joined node results in pair-end read frequency data for the paired end read information mapped to the plurality of joined nodes to more closely approximate the standard paired-end read distance frequency data. The contig information is derived from a genome in many cases. Alternately, the contig sequence information is derived from a plurality of genomes. In some examples, the methods are practiced on a computer-implemented system comprising a processor configured to receive contig and read pair data as discussed herein, process the data as discussed above, and output scaffolded contig data having the improved parameters as discussed above.

[0020] Provided herein are methods of scaffold assembly comprising obtaining a set of contig sequences obtaining input data comprising a set of paired end reads, wherein at least 1% of the paired end reads comprise a read pair distance of at least 1 kb, wherein the set of paired end reads comprises paired end reads in a natural orientation, wherein the sequencing error rate for the read pairs is no greater than 0.1%, and wherein the RN50 of the input data is no greater than 20% of the assembled scaffold, and outputting a scaffold, wherein the RN50 for the scaffold is at least 2x the RN50 of the input. Optionally, the error rare is no greater than 12%, 11%, 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.9%, 0.8%, 0.7%, 0.6%, 0.5%, 0.4%, 0.3%, 0.2%, 0.1%, 0.09%, 0.08%, 0.07%, 0.06%, 0.05%, 0.04%, 0.03%, 0.02%, 0.01%, 0.001%, 0.0001%, or 0.00001% or less. In some examples, the methods are practiced on a computer-implemented system comprising a processor configured to receive contig and read pair data as discussed herein, process the data as discussed above, and output scaffolded contig data having the improved parameters as discussed above.

[0021] Provided herein are methods of scaffold assembly comprising: obtaining a set of contig sequences comprising T_0 contig sequences, obtaining a set of paired end reads, wherein at least 1% of the paired end reads comprise a read pair distance of at least 1 kb, wherein the set of paired end reads comprises paired end reads in a natural orientation, wherein the sequencing error rate for the read pairs is no more than 0.1%, and outputting a scaffold comprising T_1, wherein $T_1 < T_0$. In some cases T_1 is less than 3. Optionally, the error rare is no greater than 12%, 11%, 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.9%, 0.8%, 0.7%, 0.6%, 0.5%, 0.4%, 0.3%, 0.2%, 0.1%, 0.09%, 0.08%, 0.07%, 0.06%, 0.05%, 0.04%, 0.03%, 0.02%, 0.01%, 0.001%, 0.0001%, or 0.00001% or less. Alternately, T_1 is selected to be less than 10, 9, 8, 7, 6, 5, or 4. In some cases T_1 is two, and in some cases T_1 is a single contig. T_1 is less than 50%, 40%, 30%, 20%, 10% 5%, 3%, 2%, 1%, or less than 1% of T_1 in many cases. The contig information is derived from a genome in many cases. Alternately, the contig sequence information is derived from a plurality of genomes. In some examples, the methods are practiced on a computer-

implemented system comprising a processor configured to receive contig and read pair data as discussed herein, process the data as discussed above, and output scaffolded contig data having the improved parameters as discussed above.

[0022] Provided herein are methods of nucleic acid sequence data processing comprising: receiving an input data comprising read pairs, at least 1% of said read pairs comprising sequence data from two nucleic acid segments separated by at least 1kb and in a natural orientation, wherein an RN50 for the input data is no greater than 20% of the assembled scaffold and wherein an error rate for said input data is no greater than 0.1%; and outputting an output data comprising a scaffold, wherein RN50 for the output data is at least 2x the RN50 of the input. In some methods RN50 for the output data is at least 10x the RN50 of the input, or alternately 3x, 4x, 5x, 6x, 7x, 8x, 9x, 11x, 12x, 13x, 14x, 15x, 16x, 17x, 18x, 19x, 20x, 30x, 40x, 50x, 60x, 70x, 80x, 90x, 100x, 500x, 1000x, or greater than 1000x. Further provided herein are methods wherein the scaffold comprises at least 90% of a target genomic sample sequence in correct order and orientation. Further provided herein are methods wherein the scaffold comprises at least 99% of a target genomic sample sequence in correct order and orientation. In some examples, the methods are practiced on a computer-implemented system comprising a processor configured to receive contig and read pair data as discussed herein, process the data as discussed above, and output scaffolded contig data having the improved parameters as discussed above.

[0023] Provided herein are methods of nucleic acid sequence data processing comprising: outputting a dataset comprising read pairs, at least 1% of said read pairs comprising sequence data from two nucleic acid segments separated by at least 1kb and in a natural orientation, wherein an RN50 for the output data is no greater than 20% of the assembled scaffold and wherein an error rate for said output data is no greater than 0.1%; and receiving an dataset comprising a scaffold, wherein RN50 for the output data is at least 2x the RN50 of the input. In some methods the RN50 for the output data is at least 10x the RN50 of the input, or alternately 3x, 4x, 5x, 6x, 7x, 8x, 9x, 11x, 12x, 13x, 14x, 15x, 16x, 17x, 18x, 19x, 20x, 30x, 40x, 50x, 60x, 70x, 80x, 90x, 100x, 500x, 1000x, or greater than 1000x. Further provided herein are methods wherein the scaffold comprises at least 90% of a target genomic sample sequence in correct order and orientation. Further provided herein are methods wherein the scaffold comprises at least 99% of a target genomic sample sequence in correct order and orientation. In some examples, the methods are practiced on a computer-implemented system comprising a processor configured to receive contig and read pair data as discussed herein, process the data as discussed above, and output scaffolded contig data having the improved parameters as discussed above.

[0024] Provided herein are methods of nucleic acid sequence data processing comprising: receiving an input data comprising read pairs, at least 1% of said read pairs comprising sequence data from two nucleic acid segments separated by at least 1kb and in a natural orientation, wherein an N50 for the input data is no greater than 20% of the assembled scaffold and wherein an error rate for said output data is no greater than 0.1%; and outputting an output data comprising a scaffold, wherein N50 for the output data is at least 2x the RN50 of the input. Optionally, the error rare is no greater than 12%, 11%, 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.9%, 0.8%, 0.7%, 0.6%, 0.5%, 0.4%, 0.3%, 0.2%, 0.1%, 0.09%, 0.08%, 0.07%, 0.06%, 0.05%, 0.04%, 0.03%, 0.02%, 0.01%, 0.001%, 0.0001%, or 0.00001% or less. The contig information is derived from a genome in many cases. Alternately, the contig sequence information is derived from a plurality of genomes. In some examples, the methods are practiced on a computer-implemented system comprising a processor configured to receive contig and read pair data as discussed herein, process the data as discussed above, and output scaffolded contig data having the improved parameters as discussed above.

[0025] Provided herein are methods of nucleic acid sequence data processing comprising: outputting an output data comprising read pairs, at least 1% of said read pairs comprising sequence data from two nucleic acid segments separated by at least 1kb and in a natural orientation, wherein an N50 for the output data is no greater than 20% of the assembled scaffold and wherein an error rate for said output data is no greater than 0.1%; and receiving an input data comprising a scaffold, wherein N50 for the output data is no greater than 20% of the assembled scaffold. The contig information is derived from a genome in many cases. Optionally, the error rare is no greater than 12%, 11%, 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.9%, 0.8%, 0.7%, 0.6%, 0.5%, 0.4%, 0.3%, 0.2%, 0.1%, 0.09%, 0.08%, 0.07%, 0.06%, 0.05%, 0.04%, 0.03%, 0.02%, 0.01%, 0.001%, 0.0001%, or 0.00001% or less. Alternately, the contig sequence information is derived from a plurality of genomes. In some examples, the methods are practiced on a computer-implemented system comprising a processor configured to receive contig and read pair data as discussed herein, process the data as discussed above, and output scaffolded contig data having the improved parameters as discussed above.

[0026] Provided herein are methods of assessing the likelihood of joining two nucleic acid contigs sharing at least one paired end read, comprising: determining a density of mapped shotgun reads to the first contig, determining a density of mapped shotgun reads to the second contig, determining a likelihood score for joining the first contig and the second contig, and reducing the likelihood score when the density of mapped shotgun reads to the first contig differs significantly from the density of mapped shotgun reads to the second contig. In some methods the likelihood score is a log likelihood score. Often, the score is reduced as indicated

herein. Often, the score is reduced as a ratio of the smaller to the larger of density of mapped shotgun reads to the first contig and the density of mapped shotgun reads to the second contig. The contig information is derived from a genome in many cases. Alternately, the contig sequence information is derived from a plurality of genomes. In some examples, the methods are practiced on a computer-implemented system comprising a processor configured to receive contig and read pair data as discussed herein, process the data as discussed above, and output scaffolded contig data having the improved parameters as discussed above.

INCORPORATION BY REFERENCE

[0027] All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference 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

[0028] The novel features of the disclosure are set forth with particularity in the appended claims and in summary and detailed description herein. A better understanding of the features and advantages of the disclosure will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the disclosure are utilized, and the accompanying drawings of which.

[0029] **Figure 1A-F** depicts a flow process for generating material for *de novo* sequencing.

[0030] **Figure 2** illustrates read pair separations for several sequencing libraries mapped to the reference human genome assembly hg19.

[0031] **Figure 3** is a chart of coverage (sum of read pair separations divided by estimated genome size) in various read pair separation bins.

[0032] **Figures 4A and 4B** illustrate the mapped locations on the GRCh38 reference sequence of read pairs are plotted in the vicinity of structural differences between GM12878 and the reference. Figure 4A depicts data for an 80Kb inversion with flanking 20kb repetitive regions. Figure 4B depicts data for a phased heterozygous deletion.

[0033] **Figure 5** depicts a flow process for generating material for *de novo* sequencing.

[0034] **Figure 6** illustrates an example of a computer system consistent with some embodiments of the present invention.

[0035] **Figure 7** is a block diagram illustrating a first example architecture of a computer system 2000 consistent with example embodiments of the present invention.

[0036] **Figure 8** is a diagram demonstrating a network 800 configured to incorporate a plurality of computer systems, a plurality of cell phones and personal data assistants, and

CLAIMS

What is claimed is:

1. A method for nucleic acid sequence data assembly, comprising:

(a) obtaining a set of paired-end reads obtained by digesting sample DNA to generate internal double strand breaks within the DNA, allowing the double strand breaks to re-ligate randomly to form a plurality of junctions, and sequencing across the plurality of junctions;

(b) obtaining standard paired-end read distance frequency data;

(c) obtaining grouped contig sequences; and

(d) scaffolding the grouped contig sequences such that read-pair distance frequency data for read pairs that map to separate contigs approximates the standard paired-end read distance frequency data,

thereby assembling nucleic acid sequence data of the nucleic acid.

2. The method of claim 1, wherein read-pair distance frequency data for read pairs that map to separate contigs more closely approximates the standard paired-end read distance frequency data when read-pair distance likelihood increases.

3. The method of claim 2, wherein read-pair distance likelihood is maximized.

4. The method of claim 1, wherein read-pair distance frequency data for read-pairs that map to separate contigs more closely approximates the standard paired-end read distance frequency data when a statistical measure of difference between the read-pair distance frequency data and the standard paired-end read distance frequency data decreases.

5. The method of claim 4, wherein the statistical measure of distance between the read-pair distance frequency data and the standard paired-end read distance frequency data comprises at least one of ANOVA, a t-test, and a X-squared test.

6. The method of claim 5, wherein read-pair distance for read-pairs that map to separate contigs more closely matches the paired-end read distance frequency data when deviation of read-pair distance distribution among ordered contigs obtained as compared to standard paired-end read distance frequency decreases.

7. The method of claim 6, wherein deviation of read-pair distance distribution among ordered contigs obtained as compared to standard paired-end read distance frequency is minimized.

8. The method of claim 1, wherein obtaining grouped contig sequences comprises extracting DNA from a biological sample and sequencing the DNA.

9. The method of claim 1, wherein the sample DNA is crosslinked to at least one DNA binding agent.

10. The method of claim 1, wherein the sample DNA comprises isolated naked DNA.

11. The method of claim 10, wherein the isolated naked DNA is reassembled into reconstituted chromatin.

12. The method of claim 11, wherein the reconstituted chromatin is crosslinked.

13. The method of claim 1, wherein standard paired-end read distance frequency data is obtained from paired-end reads where both reads map to a common contig.

14. The method of claim 1, wherein standard paired-end read distance frequency data is obtained from previously generated curves.

15. The method of claim 1, wherein said scaffolding comprises selecting a first set of putative adjacent contigs of said grouped contig sequences, determining a minimal distance order of said first set of putative adjacent contigs that reduces an aggregate measure of the read-pair distances for said read-pairs, and scaffolding said first set of putative adjacent contigs so as to reduce said aggregate measure of the read-pair distances.

16. The method of claim 15, wherein said determining a minimal distance order comprises comparing an expected read-pair distance for at least one read-pair that comprises reads mapping to two contigs of said set of putative adjacent contigs for all possible contig configurations.

17. The method of claim 16, comprising selecting a contig orientation that corresponds to a maximum likelihood read-pair distance distribution.

18. The method of claim 1, wherein the nucleic acid sequence data is obtained from a genome.

19. The method of claim 1, wherein the nucleic acid sequence data is obtained from a heterogeneous sample comprising a plurality of genomes.