

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] According to a first aspect, there is provided a computer implemented system for scaffolding contigs of nucleic acid sequence information comprising a processor configured to: receive a set of contig sequences having an initial configuration and having initial configuration read pair frequency data; receive a set of paired end reads; receive standard paired-end read distance frequency data; process contig pairs such that contig pairs sharing sequence that coexists in at least one paired end read are grouped to form grouped contig sequences; scaffold the grouped contig sequences such that read pair distance frequency data for read pairs that map to separate contigs more closely approximates the standard paired-end read distance frequency data relative to the initial configuration read pair frequency data; and output processed contig scaffolds to a network, screen, or server.

[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.

CLAIMS

What is claimed is:

1. A computer implemented system for scaffolding contigs of nucleic acid sequence information comprising a processor configured to:
receive a set of contig sequences having an initial configuration and having initial configuration read pair frequency data;
receive a set of paired end reads;
receive standard paired-end read distance frequency data;
process contig pairs such that contig pairs sharing sequence that coexists in at least one paired end read are grouped to form grouped contig sequences;
scaffold the grouped contig sequences such that read pair distance frequency data for read pairs that map to separate contigs more closely approximates the standard paired-end read distance frequency data relative to the initial configuration read pair frequency data; and
output processed contig scaffolds to a network, screen, or server.
2. The computer system of claim 1, wherein the set of paired end reads is obtained by digesting sample DNA to generate internal double strand breaks within the nucleic acid, allowing the double strand breaks to re-ligate randomly to form a plurality of re-ligation junctions, and sequencing across the plurality of re-ligation junctions.
3. The computer system of claim 1, wherein the standard paired-end read distance frequency data is obtained from paired-end reads where both reads map to a common contig.
4. The computer system of claim 1, wherein the standard paired-end read distance frequency data is obtained from previously generated curves.
5. The computer system of claim 1, wherein the initial configuration is a random configuration.
6. The computer system 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.
7. The computer system of claim 6, wherein read-pair distance likelihood is maximized.
8. The computer system 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.
9. The computer system of claim 8, 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.

10. The computer system of claim 1, wherein read pair distance for read pairs that map to separate contigs more closely matches the standard 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.

11. The computer system of claim 10, wherein deviation of read pair distance distribution among ordered contigs obtained as compared to standard paired-end read distance frequency is minimized.

12. The computer system of claim 1, wherein 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.

13. The computer system of claim 12, wherein the DNA is crosslinked to at least one DNA binding agent.

14. The computer system of claim 12, wherein the DNA is isolated naked DNA.

15. The computer system of claim 14, wherein the isolated DNA is reassembled into reconstituted chromatin.

16. The computer system of claim 15, wherein the reconstituted chromatin is crosslinked.