

implemented system comprising a first data structure comprising a plurality of sequence reads derived from a larger contiguous nucleic acid. Sequence reads derived from a common fragment of the larger contiguous nucleic acid comprise a common barcode sequence. The first subset of sequence reads is aligned to provide a contiguous linear nucleic acid sequence

[0010] Also provided are methods of assembling nucleic acid sequence reads into larger contiguous sequences. Such methods identify a first subsequence from a set of overlapping sequence reads in a computer implemented system comprising a first data structure comprising a plurality of sequence reads derived from a larger contiguous nucleic acid. The first subsequence is then extended to one or more adjacent or overlapping sequences based upon the presence of a barcode sequence on the adjacent sequence that is common to the first subsequence. A linear nucleic acid sequence is then provided that comprises the first subsequence and the one or more adjacent sequences.

[0011] In one aspect, the present disclosure provides a sequencing method of assembling nucleic acid sequence reads 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: obtaining a plurality of sequence reads derived from a larger contiguous nucleic acid, wherein two or more sequence reads derived from a common fragment of the larger contiguous nucleic acid comprise a common barcode sequence, identifying a first subset of sequence reads in the plurality of sequence reads that comprise both overlapping sequences and a common barcode sequence; and aligning the first subset of sequence reads to provide a contiguous linear nucleic acid sequence.

[0011a] In another aspect, the present disclosure provides a method of assembling nucleic acid sequence reads into larger contiguous sequences, 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: obtaining a plurality of sequence reads derived from a larger contiguous nucleic acid, identifying a first subsequence from a set of overlapping sequence reads in the plurality of sequence reads; extending the first subsequence to one or more adjacent or overlapping sequences based upon the presence of a barcode sequence on the adjacent sequence that is common to the first subsequence; and providing a linear nucleic acid sequence that comprises the first subsequence and the one or more adjacent sequences.

[0011b] In another aspect, the present disclosure provides a sequencing method comprising, at a computer system having one or more processors, and memory storing one or more programs for

execution by the one or more processors: (A) obtaining a plurality of sequence reads, wherein the plurality of sequence reads comprises a plurality of sets of sequence reads, each respective sequence read in a set of sequence reads includes (i) a first portion that corresponds to a subset of a larger contiguous nucleic acid and (ii) a common second portion that forms an identifier that is independent of the sequence of the larger contiguous nucleic acid and that identifies a partition, in a plurality of partitions, in which the respective sequence read was formed, and each respective set of sequence reads in the plurality of sets of sequence reads is formed in a partition in the plurality of partitions and each partition includes one or more fragments of the larger contiguous nucleic acid that is used as the template for each respective sequence read in the partition; (B) creating a respective set of k-mers for each sequence read in the plurality of sequence reads, wherein the sets of k-mers collectively comprise a plurality of k-mers, the identifiers of the sequence reads for each k-mer in the plurality of k-mers is retained, k is less than the average length of the sequence reads in the plurality of sequence reads, and each respective set of k-mers includes at least eighty percent of the possible k-mers of length k of the first portion of the corresponding sequence read; (C) tracking, for each respective k-mer in the plurality of k-mers, an identity of each sequence read in the plurality of sequence reads that contains the respective k-mer and the identifier of the set of sequence reads that contains the sequence read; (D) graphing the plurality k-mers as a graph comprising a plurality of nodes connected by a plurality of directed arcs, wherein each node comprises an uninterrupted set of k-mers in the plurality of k-mers of length k with $k-1$ overlap, each arc connects an origin node to a destination node in the plurality of nodes, a final k-mer of an origin node has $k-1$ overlap with an initial k-mer of a destination node, and a first origin node has a first directed arc with both a first destination node and a second destination node in the plurality of nodes; and (E) determining whether to merge the origin node with the first destination node or the second destination node in order to derive a contig sequence that is more likely to be representative of a portion of the larger contiguous nucleic acid, wherein the contig sequence comprises (i) the origin node and (ii) one of the first destination node and the second destination node, wherein the determining uses at least the identifiers of the sequence reads for k-mers in the first origin node, the first destination node, and the second destination node.

[0011c] In yet another aspect, 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: (A) obtaining a

plurality of sequence reads, wherein the plurality of sequence reads comprises a plurality of sets of sequence reads, each respective sequence read in a set of sequence reads includes (i) a unique first portion that corresponds to a subset of a larger contiguous nucleic acid and (ii) a common second portion that forms an identifier that is independent of the sequence of the larger contiguous nucleic acid and that identifies a partition, in a plurality of partitions, in which the respective sequence read was formed, and each respective set of sequence reads in the plurality of sets of sequence reads is formed in a partition in the plurality of partitions and each partition includes one or more fragments of the larger contiguous nucleic acid that is used as the template for each respective sequence read in the partition; (B) creating a respective set of k-mers for each sequence read in the plurality of sequence reads, wherein the sets of k-mers collectively comprise a plurality of k-mers, the identifier of the sequence read for each k-mer in the plurality of k-mers is retained, k is less than the average length of the sequence reads in the plurality of sequence reads, and each respective set of k-mers includes at least eighty percent of the possible k-mers of length k of the first portion of the corresponding sequence read; (C) tracking, for each respective k-mer in the plurality of k-mers, an identity of each sequence read in the plurality of sequence reads that contains the respective k-mer and the identifier of the set of sequence reads that contains the sequence read; (D) graphing the plurality k-mers as a graph comprising a plurality of nodes connected by a plurality of directed arcs, wherein each node comprises an uninterrupted set of k-mers in the plurality of k-mers of length k with $k-1$ overlap, each arc connects an origin node to a destination node in the plurality of nodes, a final k-mer of an origin node has $k-1$ overlap with an initial k-mer of a destination node, and a first origin node has a first directed arc with both a first destination node and a second destination node in the plurality of nodes; and (E) determining whether to merge the origin node with the first destination node or the second destination node in order to derive a contig sequence that is more likely to be representative of a portion of the larger contiguous nucleic acid, wherein the contig sequence comprises (i) the origin node and (ii) one of the first destination node and the second destination node, wherein the determining uses at least the identifiers of the sequence reads for k-mers in the first origin node, the first destination node, and the second destination node.

[0011d] In a further aspect, the present disclosure provides a sequencing method comprising: at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors: (A) obtaining a plurality of sequence reads, wherein the plurality of sequence reads comprises a plurality of sets of sequence reads, each respective sequence read in a set of sequence reads includes (i) a unique first portion that corresponds to a subset of a

larger contiguous nucleic acid and (ii) a common second portion that forms an identifier that is independent of the sequence of the larger contiguous nucleic acid and that identifies a partition, in a plurality of partitions, in which the respective sequence read was formed, each respective set of sequence reads in the plurality of sets of sequence reads is formed in a partition in the plurality of partitions and each partition includes one or more fragments of the larger contiguous nucleic acid that is used as the template for each respective sequence read in the partition; (B) creating a respective set of k-mers for each sequence read in the plurality of sequence reads, wherein the sets of k-mers collectively comprise a plurality of k-mers, the identifier of the sequence read for each k-mer in the plurality of k-mers is retained, k is less than the average length of the sequence reads in the plurality of sequence reads, and each respective set of k-mers includes at least eighty percent of the possible k-mers of the first portion of the corresponding sequence read; (C) tracking, for each respective k-mer in the plurality of k-mers, an identity of each sequence read in the plurality of sequence reads that contains the respective k-mer; (D) identifying a first path, wherein the first path comprises a first set of k-mers with $k-1$ overlap in the plurality of k-mers; (E) identifying a second path, wherein the second path comprises a second set of k-mers with $k-1$ overlap in the plurality of k-mers and wherein the first path intersects the second path, thereby forming the set of branch segments comprising: a left portion of the first path, a right portion of the first path, a left portion of the second path and a right portion of the second path; and (F) using the identifiers associated with the k-mers of the set of branch segments, verify the connectivity of the branch segments.

[0011e] In another aspect, 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: (A) obtaining a plurality of sequence reads, wherein the plurality of sequence reads comprises a plurality of sets of sequence reads, each respective sequence read in a set of sequence reads includes (i) a unique first portion that corresponds to a subset of a larger contiguous nucleic acid and (ii) a common second portion that forms an identifier that is independent of the sequence of the larger contiguous nucleic acid and that identifies a partition, in a plurality of partitions, in which the respective sequence read was formed, each respective set of sequence reads in the plurality of sets of sequence reads is formed in a partition in the plurality of partitions and each partition includes one or more fragments of the larger contiguous nucleic acid that is used as the template for each respective sequence read in the partition; (B) creating a respective set of k-mers for each sequence read in the plurality of sequence

reads, wherein the sets of k-mers collectively comprise a plurality of k-mers, the identifiers of the sequence reads for each k-mer in the plurality of k-mers is retained, k is less than the average length of the sequence reads in the plurality of sequence reads, and each respective set of k-mers includes at least eighty percent of the possible k-mers of the first portion of the corresponding sequence read; (C) tracking, for each respective k-mer in the plurality of k-mers, an identity of each sequence read in the plurality of sequence reads that contains the respective k-mer; (D) identifying a first path, wherein the first path comprises a first set of k-mers with $k-1$ overlap in the plurality of k-mers; (E) identifying a second path, wherein the second path comprises a second set of k-mers with $k-1$ overlap in the plurality of k-mers and wherein the first path intersects the second path, thereby forming the set of branch segments comprising: a left portion of the first path, a right portion of the first path, a left portion of the second path and a right portion of the second path; and (F) using the identifiers associated with the k-mers of the set of branch segments, verify the connectivity of the branch segments.

[0011f] The present disclosure provides a sequencing 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, implementing a method in which a plurality of sequence reads is obtained. The plurality of sequence reads comprises a plurality of sets of sequence reads. Each respective sequence read in a set of sequence reads includes (i) a first portion that corresponds to a subset of a larger contiguous nucleic acid and (ii) a common second portion that forms an identifier that is independent of the sequence of the larger contiguous nucleic acid and that identifies a partition, in a plurality of partitions, in which the respective sequence read was formed. Each respective set of sequence reads in the plurality of sets of sequence reads is formed in a partition in the plurality of partitions and each partition includes one or more fragments (e.g., two or more, three or more, ten or more) of the larger contiguous nucleic acid that is used as the template for each respective sequence read in the partition.

[0012] In the method, a respective set of k-mers is created for each sequence read in the plurality of sequence reads. The sets of k-mers collectively comprise a plurality of k-mers. The identifier of the sequence read for each k-mer in the plurality of k-mers is retained. In many instances, there are multiple such sequencing read identifiers for at least some of the k-mers. The value k is less than the average length of the sequence reads in the plurality of sequence reads. Each

respective set of k-mers includes some (e.g., at least eighty percent) of the possible k-mers of length k of the first portion of the corresponding sequence read.

[0013] In the method, there is tracked, for each respective k-mer in the plurality of k-mers, an identity of each sequence read in the plurality of sequence reads that contains the respective k-mer and the identifier of the set of sequence reads that contains the sequence read.

[0014] In the method, all or a portion (e.g., at least 1 percent, at least 5 percent, at least fifty percent) of the plurality k-mers are graphed as a graph comprising a plurality of nodes connected by a plurality of directed arcs. Each node comprises an uninterrupted set of k-mers in the plurality of k-mers of length k with k-1 overlap. Each arc connects an origin node to a destination node in the plurality of nodes. A final k-mer of an origin node has k-1 overlap with an initial k-mer of a destination node. A first origin node has a first directed arc with both a first destination node and a second destination node in the plurality of nodes.

[0015] In the method, a determination is made as to whether to merge the origin node with the first destination node or the second destination node in order to derive a contig sequence that is more likely to be representative of a portion of the larger contiguous nucleic acid. The contig sequence comprises (i) the origin node and (ii) one of the first destination node and the second destination node. The determining uses at least the identifiers of the sequence reads for k-mers in the first origin node, the first destination node, and the second destination node.

[0016] 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 executing the method described above.

[0017] The present disclosure also of the present disclosure provides a sequencing 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, obtaining a plurality of sequence reads. The plurality of sequence reads comprises a plurality of sets of sequence reads. Each respective sequence read in a set of sequence reads includes (i) a unique first portion that corresponds to a subset of a larger contiguous nucleic acid and (ii) a common second portion that forms an identifier that is independent of the sequence of the larger contiguous nucleic acid and that identifies a partition, in a plurality of partitions, in which the respective sequence read was formed. Each respective set of sequence reads in the plurality of sets of sequence reads is formed in a partition in the plurality of partitions. Each such partition includes one or more fragments of the larger contiguous nucleic acid that is used as the template for each respective sequence read in the partition.