

What is claimed is:

1. 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 barcode 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;

(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 of 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.

2. The sequencing method of claim 1, wherein,

the first origin node and the first destination node are part of a first path in the graph that includes one or more additional nodes other than the origin node and the first destination node,

the first origin node and the second destination node are part of a second path in the graph that includes one or more additional nodes other than the origin node and the second destination node, and

the determining (E) comprises determining whether the first path is more likely representative of the larger contiguous nucleic acid than the second path by evaluating a number of identifiers shared between the k-mers of the nodes of a first portion of the first path and the k-mers of the nodes of a second portion of the first path versus a number of identifiers shared between the

k-mers of the nodes of a first portion of the second path and the k-mers of the nodes of a second portion of the second path.

3. The sequencing method of claim 1, wherein

the first origin node and the first destination node are part of a first path in the graph that includes one or more additional nodes other than the origin node and the first destination node,

the first origin node and the second destination node are part of a second path in the graph that includes one or more additional nodes other than the origin node and the second destination node, and

the determining (E) up-weights the first path relative to the second path when the first path has higher average coverage than the second path.

4. The sequencing method of claim 1, wherein the

the first origin node and the first destination node are part of a first path in the graph that includes one or more additional nodes other than the origin node and the first destination node,

the first origin node and the second destination node are part of a second path in the graph that includes one or more additional nodes other than the origin node and the second destination node, and

the determining (E) up-weights the first path relative to the second path when the first path represents a longer contiguous portion of the larger contiguous nucleic acid sequence than the second path.

5. The sequencing method of any one of claims 1-4, wherein a first k-mer in the first node is present in a sub-plurality of the plurality of sequence reads and the identity of each sequence read in the sub-plurality of sequence reads is retained for the first k-mer and used by the determining (E) to determine whether the first path is more likely representative of the larger contiguous nucleic acid sequence than the second path.

6. The sequencing method of any one of claims 1-5, wherein

a partition in the plurality of partitions comprises at least 1000 molecules with the barcode, and

each molecule in the at least 1000 molecules further comprises a primer sequence complementary to at least a portion of the larger contiguous nucleic acid.

7. The sequencing method of any one of claims 1-5, wherein
a partition in the plurality of partitions comprises at least 1000 molecules with the barcode, and

each molecule in the at least 1000 molecules further comprises a primer site and a semi-random N-mer priming sequence that is complementary to part of the larger contiguous nucleic acid.

8. The sequencing method of any one of claims 1-7, wherein the one or more fragments of the larger contiguous nucleic acid in a partition in the plurality of partitions are greater than 50 kilobases in length.

9. The sequencing method of any one of claims 1-7, wherein the one or more fragments of the larger contiguous nucleic acid in a partition in the plurality of partitions are between 20 kilobases and 200 kilobases in length.

10. The sequencing method of any one of claims 1-9, wherein the one or more fragments of the larger contiguous nucleic acid in a partition in the plurality of partitions consists of between 1 and 500 different fragments of the larger contiguous nucleic acid.

11. The sequencing method of any one of claims 1-9, wherein the one or more fragments of the larger contiguous nucleic acid in a partition in the plurality of partitions consists of between 5 and 100 fragments of the larger contiguous nucleic acid.

12. The sequencing method of any one of claims 1-11, wherein the plurality of sequence reads is obtained from less than 5 nanograms of nucleic or ribonucleic acid.

13. The sequencing method of any one of claims 1-12, wherein the identifier in the barcode of each respective sequence read in the set of sequence reads encodes a common value selected from the set {1, ..., 1024}, the set {1, ..., 4096}, the set {1, ..., 16384}, the set {1, ..., 65536}, the

set {1, ..., 262144}, the set {1, ..., 1048576}, the set {1, ..., 4194304}, the set {1, ..., 16777216}, the set {1, ..., 67108864}, or the set {1, ..., 1×10^{12} }.

14. The sequencing method of any one of claims 1-12, wherein the identifier is an N-mer, and N is an integer selected from the set {4, ..., 20}.

15. The sequencing method of any one of claims 1-14, wherein an average sequence read length of the plurality of sequence reads is between 40 bases and 200 bases.

16. The sequencing method of any one of claims 1-14, wherein an average sequence read length of the plurality of sequence reads is between 60 bases and 140 bases.

17. The sequencing method of any one of claims 1-16, wherein

the plurality of sequence reads collectively provide at least 15X coverage for the larger contiguous nucleic acid,

more than ten percent of the k-mers in the plurality of k-mers are from more than one sequence read in the plurality sequence reads, and

the identifier of each sequence read for each k-mer represented by more than one sequence read is retained.

18. The sequencing method of any one of claims 1-16, wherein

the plurality of sequence reads collectively provide at least 25X coverage for the larger contiguous nucleic acid,

more than thirty percent of the plurality of k-mers are from more than one sequence read in the plurality sequence reads, and

the identifier of each sequence read for each k-mer represented by more than one source sequence read is retained.

19. The sequencing method of any one of claims 1-18, wherein the sequence reads in the plurality of sequence reads encode between 75 and 125 bases of the larger contiguous nucleic acid and the value k is an odd integer between 5 and 73.

20. The sequencing method of any one of claims 1-19, wherein a set of sequence reads in the plurality of sequence reads comprises more than 100 sequence reads, and each sequence read of the more than 100 sequence reads includes the same barcode.
21. The sequencing method of any one of claims 1-20, wherein the larger contiguous nucleic acid is a chromosome.
22. The sequencing method of any one of claims 1-21, wherein the larger contiguous nucleic acid is greater than 40 million base pairs in length.
23. The sequencing method of any one of claims 1-22, wherein creating the respective set of k-mers comprises hashing each sequence read in the plurality of sequence reads according to a predetermined k-mer length thereby creating a respective set of k-mers for each sequence read in the plurality of sequence reads.
24. 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 barcode 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;
 - (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 of 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.

25. 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 barcode 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;

(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.

26. The sequencing method of claim 25, wherein the using (F) comprises evaluating a number of identifiers shared between each possible branch segment pair in the set of branch segments.

27. The sequencing method of claim 25 or 26, wherein
a partition in the plurality of partitions comprises at least 1000 molecules with the barcode,
and
each molecule in the at least 1000 molecules includes a primer sequence complementary to
at least a portion of the larger contiguous nucleic acid.
28. The sequencing method of any one of claims 25-27, wherein the identifier in the barcode of
each respective sequence read in the set of sequence reads encodes a common value selected from
the set {1, ..., 1024}, the set {1, ..., 4096}, the set {1, ..., 16384}, the set {1, ..., 65536}, the set
{1, ..., 262144}, the set {1, ..., 1048576}, the set {1, ..., 4194304}, the set {1, ..., 16777216},
the set {1, ..., 67108864}, or the set {1, ..., 1×10^{12} }.
29. The sequencing method of any one of claims 25-28, wherein the identifier is an N-mer, wherein
N is an integer selected from the set {4, ..., 20}.
30. The sequencing method of any one of claims 25-29, wherein an average sequence read length
of the plurality of sequence reads is between 40 bases and 200 bases.
31. The sequencing method of any one of claims 25-29, wherein an average sequence read length
of the plurality of sequence reads is between 60 bases and 140 bases.
32. The sequencing method of any one of claims 25-31, wherein
the plurality of sequence reads collectively provide at least 15X coverage for the larger
contiguous nucleic acid,
more than ten percent of the k-mers in the plurality of k-mers are from more than one source
sequence read in the plurality sequence reads, and
the identifier of each such source sequence read for each k-mer represented by more than
one source sequence read is retained.
33. The sequencing method of any one of claims 25-31, wherein
the plurality of sequence reads collectively provide at least 25X coverage for the larger
contiguous nucleic acid,

more than thirty percent of the plurality of k-mers are from more than one source sequence read in the plurality sequence reads, and

the identifier of each such source sequence read for each k-mer represented by more than one source sequence read is retained.

34. The sequencing method of any one of claims 25-33, wherein the sequence reads in the plurality of sequence reads encode between 75 and 125 bases of the larger contiguous nucleic acid and k is an odd integer between 5 and 73.

35. The sequencing method of any one of claims 25-34, wherein a set of sequence reads in the plurality of sequence reads comprises more than 100 sequence reads, each with the same barcode.

36. The sequencing method of any one of claims 25-35, wherein the larger contiguous nucleic acid is a chromosome.

37. The sequencing method of any one of claims 25-36, wherein the larger contiguous nucleic acid is greater than 40 million base pairs in length.

38. 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;

(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.