Claims

1. A computer-implemented method comprising:

receiving a plurality of mutated sequence reads, each mutated sequence read corresponding to a subsequence of a sequence comprising mutations, wherein the sequence comprising mutations comprises mutations compared to a sequence not comprising mutations;

applying a common minimizer function to each mutated sequence read, thereby determining one or more respective minimizers for each mutated sequence read, wherein the one or more respective minimizers are one or more representative k-mers from a set of k-mers that exist in the respective sequence reads;

determining positions of the one or more respective minimizers in each mutated sequence read;

determining positions of one or more mutations in each mutated sequence read;

for at least two mutated sequence reads with a common minimizer, counting the number of mutations with matching position and/or with mismatching position when the respective minimizers are aligned;

determining a measure correlated to the probability that the at least two mutated sequence reads with the common minimizer derive from the same sequence comprising mutations based on the number of mutations with matching position and/or with mismatching position when the respective minimizers are aligned; and

assembling or reconstructing at least a portion of the sequence comprising mutations, by assembling the plurality of mutated sequence reads based on the measure correlated to the probability.

2. The method of claim 1, further comprising receiving a plurality of non-mutated sequence reads, each non-mutated sequence read corresponding to a subsequence of the sequence not comprising mutations.

3. The method of claim 1 or 2, wherein the step of applying a common minimizer function to each mutated sequence read comprises identifying i) the one or more k-mers in the respective mutated sequence read that is or are listed first in an ordered list of possible k-mers, or ii) the one or more k-mers that exist in a predetermined set of possible k-mers,

wherein the one or more minimizers determined for the respective mutated sequence read are the identified one or more k-mers.

4. The method of claim 3, wherein, i) in the ordered list of possible k-mers, the k-mers are ordered based on the probability that the k-mers exist in the sequence comprising mutations and do not exist in the sequence not comprising mutations, or ii) the predetermined set of possible k-mers comprises k-mers that are relatively likely to exist in the sequence comprising mutations and not in the sequence not comprising mutations, optionally wherein the predetermined set of possible k-mers does not comprise k-mers that are relatively unlikely to exist in the sequence comprising mutations.

5. The method of any one of claims 2-4, wherein the ordered list of possible k-mers or the predetermined set of possible k-mers consists of k-mers that exist more often in the plurality of mutated sequence reads than in the plurality of non-mutated sequence reads, optionally wherein k-mers that exist more often in the plurality of mutated sequence reads than in the plurality of non-mutated sequence reads are relatively likely to exist in the sequence comprising mutations.

6. The method of any one of claims 2-5, wherein the predetermined set of possible k-mers consists of k-mers that exist n or more times in the plurality of mutated sequence reads and exist less than n times in the plurality of non-mutated sequence reads, where n is an integer greater or equal to 1, optionally wherein k-mers that exist n or more times in the plurality of mutated sequence reads and exist less than n times in the plurality of non-mutated sequence reads are relatively likely to exist in the sequence comprising mutations.

7. The method of claim 6, wherein the predetermined set of possible k-mers consists of k-mers that do not exist in the plurality of non-mutated sequence reads.

8. The method of claim 6 or 7, wherein n is 2.

9. The method of any one of claims 2-8, further comprising generating the ordered list of possible k-mers or the predetermined set of possible k-mers based on a comparison of the k-mers in the plurality of mutated sequence reads and the k-mers in the plurality of non-mutated sequence reads .

10. The method of any one of claims 1-9, wherein each minimizer is a k-mer of length greater than 5.

11. The method of any one of claims 1-10, further comprising binning the mutated sequence reads in one or more minimizer bins, such that each minimizer bin contains mutated sequence reads having a common minimizer and does not contain mutated sequence reads not having a common minimizer, and

wherein the step of counting the number of mutations with matching position and/or with mismatching position is performed only on mutated sequence reads in the same minimizer bin.

12. The method of any one of claims 2-11, wherein the step of determining the positions of the one or more mutations in each mutated sequence read comprises:

obtaining a set of non-mutated seed-masked k-mers by applying each one of one or more seed patterns to k-mers in the plurality of non-mutated sequence reads;

for each mutated sequence read, applying the one or more seed patterns to k-mers in the respective mutated sequence read to obtain a plurality of mutated seed-masked k-mers, and determining the positions of the one or more mutations by identifying the one or more positions in the mutated sequence read that are masked by all of the seed patterns that correspond to the mutated seed-masked k-mers of the plurality of mutated seed-masked k-mers that exists in the set of non-mutated seed-masked k-mers.

13. The method of claim 12, wherein the one or more seed patterns are chosen such that the probability of obtaining identical seed-masked k-mers by applying at least one of the one or more seed patterns to any k-mer in the plurality of mutated sequence reads and to a corresponding k-mer in the plurality of non-mutated sequence reads is greater than 90%, preferably greater than 99%.

14. The method of claim 12 or 13, wherein the sequence comprising mutations comprises transition mutations compared to the sequence not comprising mutations; and
wherein the one or more seed patterns consist of one or more transition seed patterns.

15. The method of any one of claims 12-14, wherein each of the plurality of mutated sequence reads corresponds to a subsequence of a sequences comprising mutations associated with one of a plurality of samples, and each of the plurality of non-mutated sequence reads corresponds to a subsequence of a sequence not comprising mutations associated with the one of the plurality of samples, wherein each sequence comprising mutations comprises mutations compared to a respective sequence not comprising mutations;

wherein obtaining a set of non-mutated seed-masked k-mers comprises obtaining a respective set of non-mutated seed-masked k-mers for each sample;

further comprising creating a set of non-mutated sample bitvectors, each non-mutated sample bitvector defining, for a respective k-mer in the set of non-mutated seed-masked k-mers, in which of the plurality of samples the respective k-mer exists; and

wherein determining the positions of the one or more mutations comprises, for each mutated sequence read, and for each set and/or each combination of sets of non-mutated seed-masked k-mers, identifying the one or more positions in the mutated sequence read that are masked by all of the seed patterns that correspond to the mutated seed-masked k-mers of the plurality of mutated seed-masked k-mers that exists in the respective set or combination of sets of non-mutated seed-masked k-mers, and associating the identified one or more positions with the one or more samples associated with the respective set or combination of sets of non-mutated seed-masked k-mers.

16. The method of any one of claims 2-15, wherein the step of determining the positions of the one or more mutations in each mutated sequence read comprises:

for one or more of the mutated sequence reads, aligning the respective mutated sequence read to a reference assembly; and

determining the positions of the one or more mutations in the respective mutated sequence read by identifying positions in the respective mutated sequence read of differences between the respective mutated sequence read and the reference assembly.

17. The method of claim 16 when dependent on any one of claims 12 to 15, wherein the step of determining the positions of the one or more mutations in each mutated sequence read comprises for each mutated sequence read:

if a position in the respective mutated sequence read is aligned to the reference assembly, determining the position in the respective mutated sequence read to be a position of a mutation in the respective mutated sequence read if the position in the respective mutated

sequence read is a position at which the respective mutated sequence read differs from the reference assembly; and

if a position in the respective mutated sequence read is not aligned to the reference assembly, determining the position in the respective mutated sequence read to be a position of a mutation in the respective mutated sequence read if the position in the respective mutated sequence read is a position that is masked by all of the seed patterns that correspond to the mutated seed-masked k-mers of the plurality of mutated seed-masked k-mers that exists in the set of non-mutated seed-masked k-mers.

18. The method of any one of claims 1-17, comprising determining the measure correlated to the probability that the at least two mutated sequence reads derive from the same sequence comprising mutations based on the number of mutations with matching position and/or with mismatching position.

19. The method of claim 18, wherein the measure correlated to the probability that the at least two mutated sequence reads derive from the same sequence comprising mutations is one of i) a probability density that the at least two mutated sequence reads derive from the same sequence comprising mutations, and ii) a score function that is correlated to the probability density that the at least two mutated sequence reads derive from the same sequence comprising mutations.

20. The method of any one of claims 1-19, further comprising creating an undirected weighted graph from the plurality of mutated sequence reads,

wherein the undirected weighted graph comprises nodes corresponding to the plurality of mutated sequence reads, and wherein the edges between the nodes are associated with respective edge weights, wherein each edge weight is determined based on the number of mutations with matching position and/or with mismatching position determined for the two mutated sequence reads corresponding to the two nodes associated with the respective edge.

21. The method of claim 20, wherein the edge weights correspond to the measure correlated to the probability that the at least two mutated sequence reads corresponding to the two nodes associated with the respective edge derive from the same sequence comprising mutations.

22. The method of claim 20 or 21, further comprising performing graph clustering on the undirected weighted graph, thereby creating clusters of mutated sequence reads that are expected to derive from the same sequence comprising mutations.

23. The method of claim 22, wherein the graph clustering comprises Markov clustering or Infomap.

24. The method of claim 22 or 23, further comprising reconstructing at least a portion of the sequence comprising mutations by assembling the mutated sequence reads in the clusters.

25. The method of claim 2 and 24, further comprising reconstructing at least a portion of the sequence not comprising mutations by inferring at least a portion of the likely sequence not comprising mutations from the reconstructed portion of the sequence comprising mutations, optionally using the plurality of non-mutated sequence reads.

26. A method for generating at least a portion of a sequence of a target template nucleic acid molecule, comprising the method of any one of claims 20-25.

27. A method for determining at least a portion of a sequence of at least one target template nucleic acid molecule, the method comprising

sequencing regions of at least one target template nucleic acid molecule comprising mutations to provide a plurality of mutated sequence reads,

carrying out the method of any one of claims 1-26 on the obtained plurality of mutated sequence reads.

28. The method of claim 27, wherein the step of sequencing comprises

(a) providing a pair of samples, each sample comprising at least one target template nucleic acid molecule;

(b) sequencing regions of the at least one target template nucleic acid molecule in a first of the pair of samples to provide the plurality of non-mutated sequence reads;

(c) introducing mutations into the at least one target template nucleic acid molecule in a second of the pair of samples to provide at least one mutated target template nucleic acid molecule;

(d) sequencing regions of the at least one mutated target template nucleic acid molecule to provide the plurality of mutated sequence reads.

29. The method of claim 28, wherein the step of introducing mutations comprises introducing transition mutations into the at least one target template nucleic acid molecule in the second of the pair of samples.