

CLAIMS

1. A method for generating sequences of at least one individual target template nucleic acid molecule comprising:

(a) providing at least one sample of nucleic acid molecules comprising at least two target template nucleic acid molecules;

(b) introducing a first molecular tag into one end of each of the at least two target template nucleic acid molecules and a second molecular tag into the other end of each of the at least two target template nucleic acid molecules to provide at least two tagged template nucleic acid molecules in which each tagged template nucleic acid molecule is tagged with a unique first molecular tag and a unique second molecular tag;

(c) amplifying the at least two tagged template nucleic acid molecules to provide multiple copies of the at least two tagged template nucleic acid molecules;

(d) sequencing regions of the at least two tagged template nucleic acid molecules comprising the first molecular tag and the second molecular tag; and

(e) reconstructing a consensus sequence for at least one of the at least two target template nucleic acid molecules

in which step (e) comprises

(i) identifying clusters of sequences of the regions of the multiple copies of the at least two tagged template nucleic acid molecules which are likely to correspond to the same target template nucleic acid molecule by assigning sequences comprising first molecular tag sequences which are homologous to one another and second molecular tag sequences which are homologous to one another to the same cluster;

(ii) selecting at least one cluster of sequences in which the sequences within the selected clusters comprise a first molecular tag and a second molecular tag which are more commonly associated with one another than with a different first molecular tag or second molecular tag;

(iii) reconstructing a consensus sequence of a first target template nucleic acid molecule by aligning sequences of the at least two template nucleic acid molecules

in the cluster selected in step (ii) and defining a consensus sequence from these sequences; and

(iv) performing steps (ii) to (iii) in respect of a second and/or further template nucleic acid molecule.

5 2. A method for generating sequences of at least one individual target template nucleic acid molecule which is greater than 1 Kbp in size comprising:

(a) providing at least one sample of nucleic acid molecules comprising at least two target template nucleic acid molecules which are greater than 1 Kbp in size;

10 (b) introducing a first molecular tag into one end of each of the at least two target template nucleic acid molecules and a second molecular tag into the other end of each of the at least two target template nucleic acid molecules to provide at least two tagged template nucleic acid molecules in which each of the at least two tagged template nucleic acid molecules is tagged with a unique first molecular tag and a unique second molecular tag;

15 (c) amplifying the at least two tagged template nucleic acid molecules to provide multiple copies of the at least two tagged template nucleic acid molecules;

(d) isolating a fraction of the multiple copies of the at least two tagged template nucleic acid molecules and fragmenting the tagged template nucleic acid molecules in the fraction to provide multiple fragmented template nucleic acid molecules;

20 (e) sequencing regions of the multiple copies of the at least two tagged template nucleic acid molecules comprising the first molecular tag and the second molecular tag;

(f) sequencing the multiple fragmented template nucleic acid molecules; and

25 (g) reconstructing a consensus sequence for at least one of the at least two target template nucleic acid molecules from sequences comprising at least a subset of the sequences produced in step (f).

3. A method according to Claim 2 in which the method further comprises a step of enriching the multiple fragmented template molecules to increase the proportion of the multiple fragmented template nucleic acid molecules comprising the first molecular tag or the second molecular tag and in which this step is before step (f).

30 4. A method according to Claim 2 or 3 in which step (g) comprises:

(i) identifying clusters of sequences of regions of the multiple copies of the at least two tagged template nucleic acid molecules which are likely to correspond to the same individual target template nucleic acid molecule by assigning sequences comprising first molecular tag sequences which are homologous to one another and second molecular tag sequences which are homologous to one another to the same cluster;

(ii) analysing the sequences of the multiple fragmented template nucleic acid molecules to identify sequences of the multiple fragmented template nucleic acid molecules which comprise a first molecular tag which is homologous to the first molecular tag of the sequences of a first cluster or a second molecular tag which is homologous to the second molecular tag of the sequences of the first cluster;

(iii) reconstructing the sequence of a first template nucleic acid molecule by aligning sequences comprising at least a subset of the sequences of the multiple fragmented template nucleic acid molecules identified in step (ii) and defining a consensus sequence from these sequences; and

(iv) performing steps (i) to (iii) in respect of a second and/or further template nucleic acid molecule.

5. A method for determining sequences of at least one individual target template nucleic acid molecule comprising the following steps:

(a) obtaining data comprising sequences of regions of multiple copies of at least two tagged template nucleic acid molecules in which each of the at least two tagged template nucleic acid molecules comprises a first molecular tag at one end and a second molecular tag at the other end, in which each target template nucleic acid molecule is tagged with a unique first molecular tag and a unique second molecular tag and in which the regions comprise the first molecular tag and the second molecular tag;

(b) analysing the data comprising sequences of regions of the at least two tagged template nucleic acid molecules comprising the first molecular tag and the second molecular tag to identify clusters of sequences which are likely to correspond to the same individual target template nucleic acid molecule by assigning sequences comprising first molecular tags which are homologous to one another and second molecular tags which are homologous to one another to the same cluster;

(c) obtaining data comprising sequences of multiple fragments of the at least two tagged template nucleic acid molecules in which each of the fragments comprise either the first molecular tag or the second molecular tag;

(d) analysing the sequences of the multiple fragments of the at least two tagged template nucleic acid molecules to identify sequences of the multiple fragments of the at least two tagged template nucleic acid molecules which comprise the first molecular tag which is homologous to the first molecular tag of the sequences of a first cluster or the second molecular tag which is homologous to the second molecular tag of the sequences of the first cluster;

(e) reconstructing the sequence of a first target template nucleic acid molecule by aligning sequences comprising at least a subset of the sequences of the multiple fragments of the at least two tagged template nucleic acid molecules identified in step (d) and defining a consensus sequence from these sequences; and

(f) performing steps (c) to (e) in respect of a second and/or further target template nucleic acid molecule.

6. A method for determining sequences of at least one target template nucleic acid molecule comprising the following steps:

(a) obtaining data comprising clusters of sequences in which:

(i) each cluster comprises sequences of regions of multiple copies of at least two tagged template nucleic acid molecules in which each of the at least two tagged template nucleic acid molecules comprises a first molecular tag at one end and a second molecular tag at the other end, in which each target template nucleic acid is tagged with a unique first molecular tag and a unique second molecular tag and in which the regions comprise the first molecular tag and the second molecular tag;

(ii) each cluster comprises sequences of multiple fragments of the at least two tagged template nucleic acid molecules in which each of the fragments comprises either the first molecular tag or the second molecular tag;

(iii) the sequences of regions of multiple copies of at least two tagged template nucleic acid molecules in each cluster comprise first molecular tags and second molecular tags which are homologous to one another;

(iv) the sequences of the multiple fragments of the at least two tagged template nucleic acid molecules comprise the first molecular tag which is homologous to the first molecular tag of the sequences of regions of the multiple copies of at least two tagged template nucleic acid molecules in that cluster or the second molecular tag which is homologous to the second molecular tag of the sequences of regions of multiple copies of the at least two tagged template nucleic acid molecules in that cluster;

(b) reconstructing the sequence of a first target template nucleic acid molecule by aligning sequences comprising at least a subset of the sequences of the multiple fragments of the at least two tagged template nucleic acid molecules in a first cluster and defining a consensus sequence from these sequences; and

(c) performing step (b) in respect of a second and/or further template nucleic acid molecule.

7. A method according to Claim 4 in which step (i) further comprises determining a consensus sequence for the first molecular tag sequences and a consensus sequence for the second molecular tag sequences of a first cluster and step (ii) comprises identifying sequences of the multiple fragmented template nucleic acid molecules which comprise a first molecular tag or a second molecular tag which is homologous to the consensus sequence for the first molecular tag or the consensus sequence for the second molecular tag of the first cluster.

8. A method according to Claim 5 in which step (b) further comprises determining a consensus sequence for the first molecular tag sequences and a consensus sequence for the second molecular tag sequences of a first cluster and step (d) comprises identifying sequences of the multiple fragmented template nucleic acid molecules which comprise a first molecular tag or a second molecular tag which is homologous to the consensus sequence for the first molecular tag or the consensus sequence for the second molecular tag of the first cluster.

9. A method according to any of Claims 2 to 8 further comprising steps of:

(v) identifying clusters of sequences of regions of the multiple copies of the at least two tagged template nucleic acid molecules which are likely to correspond to the same template nucleic acid molecule by assigning sequences comprising first molecular tag sequences which are homologous to one another and second molecular tag sequences which are homologous to one another to the same cluster; and

(vi) selecting at least one cluster of sequences in which the sequences within the selected clusters comprise a first molecular tag and a second molecular tag which are more commonly associated with one another than with a different first molecular tag or second molecular tag;

5 in which the sequence of the first target template nucleic acid molecule is reconstructed from the sequences in the cluster selected in step (vi).

10. A method according to Claim 9 in which step (vi) consists of identifying groups of clusters of sequences of the at least two tagged template nucleic acid molecules in which the sequences within the clusters of each group have first molecular tags which are homologous to one another and/or identifying groups of clusters of sequences of the at least two tagged
10 template nucleic acid molecules in which the sequences within the clusters of each group have second molecular tags which are homologous to one another and selecting a cluster from the group of clusters of sequences in which the cluster that is selected contains the highest number of sequences.

15 11. A method for determining sequences of at least one individual target template nucleic acid molecule comprising the following steps:

(a) obtaining data comprising sequences of regions of multiple copies of at least two tagged template nucleic acid molecules in which each of the at least two tagged template nucleic acid molecules comprises a first molecular tag at one end and a second
20 molecular tag at the other end, in which each target template nucleic acid molecule is tagged with a unique first molecular tag and a unique second molecular tag and in which the regions comprise the first molecular tag and the second molecular tag;

(b) analysing the data comprising sequences of regions of the at least two tagged template nucleic acid molecules comprising the first molecular tag and the second
25 molecular tag to identify clusters of sequences which are likely to correspond to the same template nucleic acid molecule by assigning sequences comprising first molecular tags which are homologous to one another and second molecular tags which are homologous to one another to the same cluster;

(c) selecting at least one cluster of sequences in which the sequences within the
30 selected clusters comprise a first molecular tag and a second molecular tag which are

more commonly associated with one another than with a different first molecular tag or second molecular tag;

(d) reconstructing a consensus sequence of a first target template nucleic acid molecule by aligning at least a subset of the sequences molecules in the cluster selected in step (c) and defining a consensus sequence from these sequences; and

(e) performing steps (c) to (d) in respect of a second and/or further target template nucleic acid molecule.

12. A method step according to Claim 1 (iv) or a method step according to Claim 11(c) consisting of identifying groups of clusters of sequences of the at least two tagged template nucleic acid molecules in which the sequences within the clusters of each group have 5' molecular tags which are homologous to one another and/or identifying groups of clusters of sequences of the at least two tagged template nucleic acid molecules in which the sequences within the clusters of each group have 3' molecular tags which are homologous to one another; and selecting a cluster from a group of clusters of sequences in which the cluster that is selected contains the highest number of sequences.

13. A method for determining sequences of at least one target template nucleic acid molecule comprising

(a) obtaining data comprising a cluster of sequences;

(b) reconstructing a consensus sequence of a first template nucleic acid molecule by aligning the sequences of at least a subset of the sequences in the selected cluster;

in which the sequences in the selected cluster comprise sequences of regions of multiple copies of at least two tagged template nucleic acid molecules in which each of the at least two tagged template nucleic acid molecules comprises a first molecular tag at one end and a second molecular tag at the other end, in which each of the at least two target template nucleic acid molecules is tagged with a unique first molecular tag and a unique second molecular tag and in which the regions comprise the first molecular tag and the second molecular tag; and

each sequence in the selected cluster

- (i) comprises first molecular tag which is homologous to the first molecular tag of the other sequences in that cluster and the second molecular tag which is homologous to the second molecular tag of the other sequences in that cluster; and
- (ii) comprises a first molecular tag and a second molecular tag which are more commonly associated with one another than with a different first molecular tag or second molecular tag.

14. A method according to any of Claims 4 to 13 in which the first molecular tags of the sequences of the same cluster have at least 90% sequence identity to one another.

15. A method according to any of Claims 4 to 14 in which the second molecular tags of the sequences of the same cluster have at least 90% sequence identity to one another.

16. A method according to any of Claims 2, 3 or 4 in which step g) is a computer-implemented method step or a method according to Claim 10 in which step vi) is a computer-implemented method step.

17. A method according to Claim 4 or 10 which is a computer-implemented method.

18. A method according to any preceding claim in which the regions comprise greater than 25 base pairs comprising the first molecular tag or the second molecular tag.

19. A method according to any preceding claim in which the regions comprise the entire length of the at least two tagged template nucleic acid molecules are sequenced.

20. A method according to Claim 2, 3 or 4 in which steps e) and/or f) are carried out using sequencing technology comprising a step of bridge PCR.

21. A method according to Claim 20 in which the step of bridge PCR is carried out using an extension time of greater than 15 seconds.

22. A method according to Claim 2, 3 or 4 in which steps e) and f) are carried out in different sequencing runs.

23. A method according to any preceding claim in which the first molecular tag and the second molecular tag are introduced into the at least two template nucleic acid molecules using a method selected from the group consisting of PCR, tagmentation, and physical

shearing or restriction digestion of the at least one template nucleic acid molecule followed by ligation of nucleic acids comprising the 5' molecular tag or the 3' molecular tag.

24. A method according to Claim 23 in which the first molecular tag and the second molecular tag are introduced into the at least two template nucleic acid molecules by PCR
- 5 using primers comprising a portion comprising the first molecular tag or the second molecular tag and a portion having a sequence that is capable of hybridising to the at least two template nucleic acid molecules.
25. A method according to any preceding claim in which the at least two template nucleic acid molecules encode microbial ribosomal 16S.
- 10 26. A method according to any preceding claim in which at least one of the at least two template nucleic acid molecules is less than 10 Kbp in size.
27. A computer program adapted to perform a method according to Claim 5, 6, 11 or 13, method step (g) of Claim 2 or method step (e) of Claim 1 when the program is run on an electronic device.
- 15 28. A computer readable medium storing a computer program according to Claim 27.
29. A kit comprising:
- (i) primers comprising a portion comprising a first molecular tag or a second molecular tag and a portion having a sequence that is capable of hybridising to at least two template nucleic acid molecules;
- 20 (ii) the computer readable medium storing the computer program according to Claim 28.