(19) DANMARK

(10) DK/EP 3146070 T5

Patent- og Varemærkestyrelsen

(12) Rettet oversættelse af europæisk patentskrift

(51) Int.Cl.: *C 12 Q 1/68 (2006.01)*

(45) Oversættelsen bekendtgjort den: **2019-01-21**

(80) Dato for Den Europæiske Patentmyndigheds bekendtgørelse om meddelelse af patentet: **2018-09-26**

(86) Europæisk ansøgning nr.: **15727042.2**

(86) Europæisk indleveringsdag: **2015-05-22**

(87) Den europæiske ansøgnings publiceringsdag: **2017-03-29**

(86) International ansøgning nr.: **GB2015051518**

(87) Internationalt publikationsnr.: **WO2015177570**

(30) Prioritet: **2014-05-23 GB 201409282**

(84) Designerede stater: **AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR**

(73) Patenthaver: **Longas Technologies Pty Ltd, Level 7, 210 George Street, Sydney NSW 2000, Australien**

(72) Opfinder: **BURKE, Catherine Maree, c/o University of Technology, 1 Broadway, Broadway, Sydney, New South Wales 2007, Australien**
DARLING, Aaron Earl, c/o University of Technology, 1 Broadway, Broadway, Sydney, New South Wales 2007, Australien

(74) Fuldmægtig i Danmark: **Zacco Denmark A/S, Arne Jacobsens Allé 15, 2300 København S, Danmark**

(54) Benævnelse: **Sekvenseringsfremgangsmåde**

(56) Fremdragne publikationer:
WO-A1-2006/137734
WO-A2-2012/162267
AARON E DARLING ET AL: "The genome of Clostridium difficile 5.3", GUT PATHOGENS, BIOMED CENTRAL LTD, LONDON, UK, vol. 6, no. 1, 24 February 2014 (2014-02-24), page 4, XP021182419, ISSN: 1757-4749, DOI: 10.1186/1757-4749-6-4
Catherine Burke ET AL: Resolving microbial microdiversity with high accuracy full length 16S rRNA Illumina sequencing, 6 November 2014 (2014-11-06), XP055207693, DOI: 10.1101/010967 Retrieved from the Internet: URL:http://biorxiv.org/content/biorxiv/ear ly/2014/11/06/010967.full.pdf [retrieved on 2015-08-13]

DESCRIPTION

TECHNICAL FIELD

[0001] The present invention relates to methods for generating sequences of template nucleic acid molecules, computer-implemented methods for determining sequences of at least two template nucleic acid molecules, computer programs adapted to perform the methods and computer readable media storing the computer programs.

BACKGROUND

[0002] In general there are difficulties with sequencing long nucleic acid sequences (for example those greater than 1 Kbp) effectively and quickly. Presently sequencing technology can produce either large volumes of short sequence reads (i.e. sequences of short nucleic acid molecules) or small numbers of long sequence reads. It is, at present, difficult to sequence large numbers of long sequence reads.

[0003] The 16S rRNA gene is used for phylogenetic studies as it is highly conserved between different species of bacteria and archaea. In addition to highly conserved primer binding sites, 16S rRNA gene sequences contain hypervariable regions that can provide species-specific signature sequences useful for bacterial identification. As a result, 16S rRNA gene sequencing has become prevalent in medical microbiology as a rapid and cheap alternative to phenotypic methods of bacterial identification. In addition, although it was originally used to identify bacteria, 16S sequencing was subsequently found to be capable of reclassifying bacteria into completely new species, or even genera. It has also become one of the primary criteria used to identify and describe new species of bacteria, both in laboratory culture and in uncultured environmental samples. However, the use of 16S rRNA sequence analysis is hampered due to the difficulties associated with sequencing large numbers of nucleic acid molecules greater than 1 Kbp. This has meant that, in general, most researchers performing 16S sequence analysis tend to focus on short, up to 500bp, regions of the 16S gene. Sequencing such short regions results in a lack of taxonomic resolution.

[0004] In addition general sequencing methods tend to lack accuracy due to recombination events that can occur during the sequencing process. Sequencing involves steps of amplifying the nucleic acid molecules to be sequenced. During these amplification steps recombination events can occur. This can mean that when samples of nucleic acid molecules contain genes of similar sequences, sequencing methods will generate, not only the sequences of the original genes, but also the sequences of nucleic acid molecules produced via recombination between these similar genes. Since 16S rRNA genes tend to be similar across different species, a nucleic acid template molecule within a sample of nucleic acid molecules comprising nucleic acid molecules from multiple different 16S rRNA genes may recombine during sequencing.

Such recombination events become increasingly frequent as the amount of amplification required to analyse the sample grows, especially at levels required to analyse certain host-associated microbiota and forensic samples. Thus it is beneficial, when sequencing samples of DNA having 16S rRNA genes, to be able to identify and remove sequences of nucleic acids produced via recombination.

[0005] Computational methods for detecting recombination are limited, however, because they can only detect recombination events that occur between two parental molecules that are substantially different in sequence. Recombination among highly similar sequences (e.g. >97% identity) remains difficult to discriminate from true biological diversity using computational methods. Molecular approaches to boost the accuracy of computational recombination detection do not currently exist.

[0006] Approaches to boost the read length of high-throughput sequencing instruments have been described previously. Among these are the complexity reduction approaches such as Illumina's Moleculo which assigns unique barcodes to pools of 100s of DNA molecules, and molecular tagging methods, which add a unique barcode to each single molecule in a sample. Both approaches reconstruct the original template molecules by analysing a collection of short reads belonging to each barcode, computationally reconstructing a consensus sequence of the original templates. Both approaches depend on amplification to create many copies of the barcoded pools or tagged single molecules. However, none of these previous approaches employ a molecular system to detect in-vitro recombination error introduced by the amplification.

SUMMARY OF THE INVENTION

[0007] The present inventors have developed a technique which allows for sequencing of long sequences of nucleic acids quickly and accurately. This technique can be used in many different applications but is particularly advantageous for use in 16S rRNA gene sequencing since it can be used to generate large volumes of long reads spanning the entire length of the 1.5 Kbp gene. Thus this technique can be used to sequence the entire 16S rRNA gene providing greater taxonomic resolution than previous methods which involved sequencing shorter regions of the 16S rRNA gene.

[0008] In addition the present inventors have developed a technique which allows for sequences of recombination products generated during the sequencing process to be identified and disregarded. This improves the sensitivity and accuracy of sequencing in general and such accuracy improves the taxonomic resolution when the technique is used for phylogenetic studies using 16S sequencing.

[0009] In a first aspect of the present invention there is provided a method for generating sequences of at least one individual template nucleic acid molecule which is greater than 1 Kbp in size comprising:

1. a) providing at least one sample of nucleic acid molecules comprising at least two template nucleic acid molecules which are greater than 1 Kbp in size;
2. 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 wherein 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;
3. 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;
4. 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;
5. 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;
6. f) sequencing the multiple fragmented template nucleic acid molecules; and
7. g) reconstructing a consensus sequence for at least one of the at least two template nucleic acid molecules from sequences comprising at least a subset of the sequences produced in step f).

[0010] In a second aspect of the present invention there is provided a computer-implemented method for determining sequences of at least one individual target template nucleic acid molecule comprising the following steps:

1. (a) obtaining data comprising sequences of regions of multiple copies of at least two tagged template nucleic acid molecules wherein 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, wherein each target template nucleic acid molecule is tagged with a unique first molecular tag and a unique second molecular tag and wherein the regions comprise the first molecular tag and the second molecular tag;
2. (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;
3. (c) obtaining data comprising sequences of multiple fragments of the at least two tagged template nucleic acid molecules wherein each of the fragments comprise either the first molecular tag or the second molecular tag;
4. (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;

5. (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
6. (f) performing steps (c) to (e) in respect of a second and/or further template nucleic acid molecule.

[0011] In a third aspect of the invention there is provided a computer-implemented method for determining sequences of at least one target template nucleic acid molecule comprising the following steps:

1. (a) obtaining data comprising clusters of sequences wherein:
    1. (i) each cluster comprises sequences of regions of multiple copies of at least two tagged template nucleic acid molecules wherein 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, wherein each of the at least two target template nucleic acids is tagged with a unique first molecular tag and a unique second molecular tag and wherein the regions comprise the first molecular tag and the second molecular tag;
    2. (ii) each cluster comprises sequences of multiple fragments of the at least two tagged template nucleic acid molecules wherein each of the fragments comprises either the first molecular tag or the second molecular tag;
    3. (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;
    4. (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 target 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;
2. (b) reconstructing the sequence of a first 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
3. (c) performing step (b) in respect of a second and/or further template nucleic acid molecule.

[0012] In a fourth aspect of the invention there is provided a method generating sequences of at least one individual target template nucleic acid molecule comprising:

1. a) providing at least one sample of nucleic acid molecules comprising at least two template nucleic acid molecules;
2. 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 wherein 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;
3. c) amplifying the at least two tagged template nucleic acid molecules provide multiple copies of the at least two tagged template nucleic acid molecules;
4. d) sequencing regions of the at least two tagged template nucleic acid molecules comprising the first molecular tag and the second molecular tag; and
5. e) reconstructing a consensus sequence for at least one of the at least two target template nucleic acid molecules wherein step e) comprises
    1. (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;
    2. (ii) selecting at least one cluster of sequences wherein 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;
    3. (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
    4. (iv) performing steps (ii) to (iii) in respect of a second and/or further template nucleic acid molecule.

[0013] In a fifth aspect of the invention there is provided a computer-implemented method for determining sequences of at least one individual target template nucleic acid molecule comprising the following steps:

1. (a) obtaining data comprising sequences of regions of multiple copies of at least two tagged template nucleic acid molecules wherein 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, wherein each target template nucleic acid molecule is

tagged with a unique first molecular tag and a unique second molecular tag and wherein the regions comprise the first molecular tag and the second molecular tag;

2. (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 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;
3. (c) selecting at least one cluster of sequences wherein 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;
4. (d) reconstructing a consensus sequence of a first 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
5. (e) performing steps (c) to (d) in respect of a second and/or further template nucleic acid molecule.

[0014] In a sixth aspect of the invention there is provided a computer-implemented method for determining sequences of at least one target template nucleic acid molecule comprising

1. (a) obtaining data comprising a cluster of sequences;
2. (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;
   wherein the sequences in the selected cluster comprise sequences of regions of multiple copies of at least two tagged template nucleic acid molecules wherein 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, wherein each target template nucleic acid molecule is tagged with a unique first molecular tag and a unique second molecular tag and wherein the regions comprise the first molecular tag and the second molecular tag; and each sequence in the selected cluster
   1. (i) comprises first molecular tag which is homologous to the first molecular tag of the other sequences in that and the second molecular tag which is
   2. (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.

[0015] In a seventh aspect of the invention there is provided a computer program adapted to perform the methods or method steps of the invention when said program is run on an electronic device.

[0016] In an eighth aspect of the invention there is provided a computer readable medium storing the computer program of the invention.

[0017] There is disclosed a kit comprising:

1. (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;
2. (ii) instructions describing how to perform the method of the invention.

[0018] There is disclosed a kit comprising

1. (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;
2. (ii) the computer readable medium storing the computer program of the invention.

DESCRIPTION OF FIGURES

[0019]

Figure 1. Fully automated computational workflow used for processing reads from a single MiSeq nano run to sequence full length 16S rRNA templates. A pool of molecules containing both full length templates and "fill-in" fragments was sequenced on the instrument and processed computationally using the steps shown.

Figure 2. Description of sequences of primers used for 16S gene amplification and sequencing.

Figure 3. Abundance of barcode clusters identified as putatively recombinant (left column), along with abundances of the progenitor molecules producing recombinant forms (right). Parental templates are on average 28-35x more abundant than the putatively recombinant forms.

Figure 4. Graph illustrating the length distribution of the assembled 16S sequences.

Figure 5. Graph comparing phyla level taxonomic assignments of OTUs using long and short sequences. The clear bar represents the average value across all 12 samples for the short sequencing method. The black bar represents the average value across all 12 samples for the "long" method. The grey bar represents the average value across all 12 samples from the assembled V4 region from the "long" sequencing method.

Figure 6. Graph comparing genus level taxonomic assignments of OTUs using long and short sequences. The clear bar represents the average value across all 12 samples for the short sequencing method. The black bar represents the average value across all 12 samples for the "long" method. The grey bar represents the average value across all 12 samples from the assembled V4 region from the long sequencing method.

Figure 7. Accumulation curves showing the number of random barcodes observed versus the number of template molecules sequenced. The 50x and 100x dilutions were predicted to have an appropriate level of redundancy in templates to permit reconstruction of the full length template molecule by fill-in sequencing on an Illumina MiSeq. The top line represents a 1 in 10 dilution, the second line from the top represents a 1 in 50 dilution. The third line from the top represents a 1 in 100 dilution. The fourth line from the top represents a 1 in 500 dilution and the bottom line represents a 1 in 1000 dilution.

Figures 8-11. Flow charts depicting methods of the invention

DETAILED DESCRIPTION OF THE INVENTION

GENERATING OR DETERMINING SEQUENCES OF AT LEAST ONE INDIVIDUAL TARGET TEMPLATE NUCLEIC ACID MOLECULE

[0020] The present invention relates to a method for generating or a computer implemented method for determining the sequences of at least one individual target template nucleic acid molecules.

[0021] The term 'target template nucleic acid molecule' refers to a nucleic acid molecule which the operator of the method intends to sequence. A 'template nucleic acid molecule' may comprise part of a larger nucleic acid molecule such as a chromosome. A 'template nucleic acid molecule' may comprise a gene, multiple genes or a fragment of a gene. A 'template nucleic acid molecule' may be isolated using primers which are capable of hybridising to the template nucleic acid molecule.

[0022] There are at least two target template nucleic acid molecules in the sample of nucleic acid molecules. In the case of 16S sequencing the at least two target template nucleic acid molecules could include multiple molecules each encoding a different 16S rRNA. For example, the at least two target template nucleic acid molecules could include nucleic acids encoding 16S rRNA from different bacteria, nucleic acids encoding different molecules of 16s rRNA from the same bacterium or both. Alternatively the at least two target template nucleic acid molecules may comprise multiple copies of the same gene. The 'target template nucleic acid molecules' may comprise a fragment of the 16s rRNA, however it is preferable that the

fragment is at least 1 Kbp in length. This is because the inventors have demonstrated that when 16S sequencing is used for phylogenetic studies, the longer the strand of 16s rRNA that is sequenced the higher the level of taxonomic resolution that can be obtained.

[0023] In one embodiment of the invention the at least one target template nucleic acid molecule is greater than 1 Kbp, greater than 1.2 Kbp, greater than 1.3 Kbp or greater than 1.5 Kbp in size. In a further embodiment of the invention the at least one target template nucleic acid molecule is less than 100 Kbp, less than 50 Kbp, less than 25 Kbp, less than 15 Kbp, less than 10 Kbp, less than 5 Kbp, less than 3 Kbp or less than 2 Kbp in size.

[0024] In a further embodiment of the invention the method is a high throughput method for generating sequences of at least one target template nucleic acid molecule.

PROVIDING AT LEAST ONE SAMPLE OF NUCLEIC ACIDS

[0025] Some aspects of the invention require a step of providing at least one sample of nucleic acids comprising at least two target template nucleic acid molecules. Optionally the at least two target template nucleic acid molecules are greater than 1 Kbp in size.

[0026] In general the term 'comprising' is intended to mean including but not limited to, for example the phrase 'comprising the following steps' indicates that the method includes those steps but that additional steps may also be performed. In some embodiments of the invention the word 'comprising' can be replaced by the word 'consisting'. The term 'consisting' is intended to be limiting, for example if a method is 'consisting the following steps' the method includes those steps and no others.

[0027] The sample may be any sample of nucleic acids. The sample of nucleic acids may be a sample of nucleic acids derived from a human, for example a sample extracted from a skin swab of a human patient. Alternatively the sample of nucleic acids may be derived from other sources such as a sample from a water supply. Such a sample could contain billions of template nucleic acid molecules. It would be possible to sequence each of these billions of template nucleic acid molecules simultaneously using the method of the invention thus there is no upper limit on the template nucleic acid molecules which could be used in the method of the invention.

[0028] In a further embodiment of the invention the method comprises providing multiple samples of nucleic acids, for example 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 15, 20, 25, 50, 75 or 100 samples. Optionally less than 100, 75, 50, 25, 20, 15, 11, 10, 9, 8, 7, 6, 5 or 4 samples of nucleic acids provided. In a further embodiment between 2 and 100, 2 and 75, 2 and 50, 2 and 25, 5 and 15 or 7 and 15 samples are provided.

INTRODUCING A FIRST MOLECULAR TAG AND A SECOND MOLECULAR TAG AND AMPLIFYING THE AT LEAST TWO TAGGED TEMPLATE NUCLEIC ACID MOLECULES

[0029] Some of the methods of the invention involve introducing a first molecular tag into one end of each of the at least two template nucleic acid molecules and a second molecular tag into the other end of each of the at least two template nucleic acid molecules to provide at least two tagged template nucleic acid molecules. Some of the methods of the invention involve amplifying the at least two tagged template nucleic acid molecules to provide multiple copies of the at least two template nucleic acid molecules.

[0030] In order to allow the template nucleic acid molecules in the sample to be sequenced the template nucleic acid molecules should be amplified optionally by PCR in order to provide multiple copies of each template nucleic acid molecule (*i.e.* to ensure that the template nucleic acid molecules are at sufficient concentration for the sequencing reaction). In one embodiment the amplification is carried out by polymerase chain reaction (PCR). The amplification step also ensures that the target template nucleic acid molecules are enriched relative to the other nucleic acids in the sample. The amplification step uses primers that hybridise to the target template nucleic acid molecules thus amplifying the target template nucleic acid only and increasing the concentration of the target template nucleic acid molecules relative to the other nucleic acids in the sample (enrichment). However, since samples generally contain multiple target template nucleic acid molecules this amplification step may amplify multiple target template nucleic acid molecules. For example in 16S sequencing, a sample may contain 16S DNA templates from multiple bacteria, the primers used in the amplification step may hybridise to all of these 16S gene sequences and thus all of these DNA templates will be amplified. This can be achieved by using degenerate primers which may vary slightly in sequence such that a group of degenerate primers can hybridise to (or are complementary to) similar but not identical target template nucleic acid sequences.

[0031] It is advantageous to be able to determine which of the sequences generated in the sequencing steps originated from the same original template nucleic acid molecule. Accordingly the term 'tagged template nucleic acid molecule' refers to a molecule comprising a 'target template nucleic acid molecule' and a tag at each end. This allows consensus sequences for each molecule of original template nucleic acids to be determined. This can be achieved by adding molecular tags into both ends (the 5' and 3' ends) of each of the original template nucleic acid molecules (Lundberg et al; Nature Methods 10: 999-1002) to produce tagged template nucleic acid molecules. The first and/or second molecular tag will be considered to have been introduced into the ends of the template DNA molecules as long as they are near, in sequence, to the terminal nucleotides (the first or last nucleotide in the sequence) of the template DNA molecules. In one embodiment there are less than 50, 40, 30, 25, 20, 15, 10 or 5 nucleotides between a terminal nucleotide and the first molecular tag. In a further embodiment there are less than 50, 40, 30, 25, 20, 15, 10 or 5 nucleotides between a terminal nucleotide and the second molecular tag.

[0032] The methods of the invention require that the first molecular tag and the second molecular tag are unique. In this case the term 'unique' refers to molecular tags which

comprise a random sequence of base pairs, assuming that there are enough random nucleotide sequences used each first molecular tag and each second molecular tag will be have a different sequence from every other tag that is generated. However in some embodiments the same tag sequence may occur more than once, in this embodiment the first molecular tag and the second molecular tag will still be considered to be 'unique'. In a further embodiment each first molecular tag and each second molecular tag comprise nucleotide sequences which are different to the nucleotide sequences of every other first molecular tag and second molecular tag. In a further embodiment at least 90% of the first molecular tags and the second molecular tags comprise nucleotide sequences which are different to the nucleotide sequences of every other first molecular tag and second molecular tag. This means that sequences of nucleic acid molecules sharing the same pair of first and second unique molecular tags are likely to have originated from the same original template nucleic acid molecule (birthday paradox). In addition sequences of fragments of nucleic acid comprising either the first molecular tag or the second molecular tag associated with a target template nucleic acid molecule are also likely to have originated from that target template DNA molecule. The use of two unique molecular tags also allows for sequences that are generated by recombination during the methods of the invention to be identified and disregarded.

[0033] The first molecular tag and the second molecular tag sequences may also comprise a few nucleotides from the target template nucleic acid sequence, for example less than 50, 40, 35, 30, 25, 20, 15 or 10 base pairs of the target template nucleic acid molecule sequence.

[0034] In one embodiment the first molecular tag and the second molecular tag are greater than 5 bp, greater than 6 bp or greater than 7 bp in size. In a further embodiment the first molecular tag and the second molecular tag are less than 20 bp, less than 18 bp, less than 15 bp, or less than 10 bp in size.

[0035] Such unique molecular tags can be introduced using a variety of techniques including PCR, tagmentation and physical shearing or restriction digestion of target nucleic acids combined with subsequent adapter ligation (optionally sticky-end ligation). For example PCR can be carried out on the at least two target template nucleic acid molecules using a first set of primers capable of hybridising to (optionally complementary to) the at least two target template nucleic acid molecules. In one embodiment of the invention the first molecular tag and the second molecular tag are introduced into each of the at least two template nucleic acid molecules by PCR using primers comprising a portion (a 5' end portion) comprising the first molecular tag or the second molecular tag and a portion (a 3' end portion) having a sequence that is capable of hybridising to (optionally complementary to) the at least two target template nucleic acid molecules. Such primers will hybridise to target template nucleic acid molecule, PCR primer extension will then provide a nucleic acid molecule which comprises either the first molecular tag or the second molecular tag. A further round of PCR with these primers will provide tagged template nucleic acid molecules comprising a first molecular tag at one end and a second molecular tag at the other end. In a further embodiment the primers are degenerate, i.e. the 3' end portion of the primers are similar but not identical to one another. For example, if the method of the invention is used for 16S ribosomal sequencing the 3' end

portion of the primers may vary slightly primer to primer but each 3' end portion will be complementary to the 16S sequence in at least one organism. This allows sequencing of a 16S sequence whose origin is unknown, thus enabling sequencing of any 16S rRNA sequence irrespective of its origin (for example the bacterium from which it is derived). Such sequences can then be used in phylogenetic studies. In an embodiment where the at least two target template nucleic acid molecules are 16S rRNA genes, suitable primers may have a 3' end portion comprising the 27F (Weisberg et al, J Bacteriol. 1991 Jan; 173(2): 697-703) or 1391R (Turner *et al*, 1999) bacterial primer sequences.

[0036] In a separate embodiment of the invention the first molecular tag and the second molecular tag can be introduced using tagmentation. In an embodiment wherein the first molecular tag and the second molecular tag are introduced using tagmentation they can be introduced using direct tagmentation, or by introducing a defined sequence by tagmentation followed by two rounds of PCR using primers that comprise a portion capable of hybridising to the defined sequence and a portion comprising the first molecular tag or the second molecular tag. In a further embodiment of the invention the first molecular tag and the second molecular tag can be introduced by restriction digestion of the original nucleic acids followed by ligation of nucleic acids comprising the first or second molecular tag. The restriction digestion of the original nucleic acids should be performed such that the digestion results in a molecule comprising the region to be sequenced (the at least one target template nucleic acid molecule).

[0037] In an embodiment where the first molecular tag and the second molecular tag are introduced into the at least two target template nucleic acid molecules by PCR, the primers used may comprise a further portion comprising a constant 'stub sequence'. This constant stub sequence is preferably 5' of the unique molecular tag. In this embodiment the tagged template nucleic acid molecules provided will further comprise a stub sequence.

[0038] In an embodiment where multiple samples of nucleic acids are provided, the method comprises a further step of introducing a sample barcode into one of the ends of the target template nucleic acid molecules in each sample. This further step occurs before or during the step of introducing a first molecular tag into one end and a second molecular tag into the other end of each of the at least two target template nucleic acid molecules in the methods of the present invention. These sample barcodes may be introduced in a similar way to introducing the first molecular tag and the second molecular tag, for example a round of PCR may be carried out on each sample separately in which the primers used hybridise to (or are complementary to) the at least two target template nucleic acid molecules and comprise a portion (optionally a 3' portion) which comprises the sample barcode. Optionally in an embodiment where the first molecular tag and the second molecular tag are introduced into the at least two template nucleic acid molecules by PCR, the primers used to introduce the tags may comprise a further portion comprising a sample specific barcode. In this embodiment a first round of PCR is carried out on each sample of nucleic acids separately. The first round of PCR may use primers comprising the first molecular tag or the second molecular tag, a sample specific barcode which is identical for every nucleic acid template molecule in the

sample, a region which hybridises to the template nucleic acid molecules and optionally a stub region. The samples of nucleic acids can then be pooled and subject to further rounds of PCR using primers (optionally which are capable of hybridising to or are complementary to the 'stub' region) which do not comprise a sample specific barcode. Optionally a second round of PCR is carried out using a primer which comprises a second sample specific barcode, in this embodiment the samples of nucleic acids are not pooled until after the second round of PCR.

[0039] The step of amplifying the at least two tagged template nucleic acid may involve PCR using a second set of primers which are capable of hybridising with the ends of the tagged template nucleic acid molecules in such a way that primer extension will result in multiple copies of the tagged template nucleic acid molecules and will maintain the first molecular tag and the second molecular tag. In an embodiment where the first set of primers comprises a stub sequence, the second set of primers may comprise a region which is capable of hybridising to the stub sequence of the tagged template nucleic acid molecules.

ISOLATING A FRACTION OF THE AMPLIFIED TEMPLATE NUCLEIC ACID MOLECULES AND FRAGMENTING THE AMPLIFIED TEMPLATE NUCLEIC ACID MOLECULES IN THE FRACTION

[0040] The method may comprise isolating a fraction of the amplified template nucleic acid molecules and fragmenting the amplified template nucleic acid molecules in the fraction to provide multiple fragmented template nucleic acid molecules.

[0041] By the term 'fragment' we are referring to a short segment of a nucleic acid molecule i.e. to a string of nucleotides which form part of a 'full length' sequence. Fragments according to the invention will be at least 10, 15, 20, 50, 100, 200, 250 or 500 base pairs long. Optionally fragments according to the invention will be less than 2500, 2200, 2000 or 1500 base pairs long.

[0042] Fragmentation can be carried out using any appropriate method. For example, fragmentation can be carried out using restriction digestion or using PCR with primers complementary to at least one internal region of the tagged template nucleic acid molecules. Preferably fragmentation is carried out using a method that produces arbitrary fragments. The term "arbitrary fragment" refers to a randomly generated fragment, for example a fragment generated by tagmentation. Fragments generated using restriction enzymes are not "arbitrary" as restriction digestion occurs at specific DNA sequences defined by the restriction enzyme that is used. Even more preferably fragmentation is carried out by tagmentation. If fragmentation is carried out by tagmentation, the tagmentation reaction optionally introduces an adapter region into the fragmented template nucleic acid molecules. This adapter region is a short DNA sequence which may encode, for example adapters to allow the fragmented template nucleic acid molecules to be sequenced using Illumina MiSeq technology.

[0043] In a typical embodiment this step may comprise a further 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. In this preferred embodiment the step of enriching the multiple fragmented template nucleic acid molecules is preferably carried out by PCR. Preferably the PCR is carried out using primers which are capable of hybridising to (optionally complementary to) either the first or second molecular tag and primers which are capable of hybridising to (optionally complementary to) internal regions of the at least two tagged template nucleic acid molecules. Such a PCR step will increase the concentration of fragments comprising the first molecular tag or the second molecular tag.

[0044] In an embodiment where fragmentation is carried out by tagmentation and the tagmentation introduces an adapter region into the fragmented template nucleic acid molecules, enrichment may be carried out by PCR using primers that are capable of hybridising to (optionally complementary to) either the first or second molecular tag and primers that are capable of hybridising to (optionally complementary to) the adapter sequence.

SEQUENCING REGIONS OF THE AT LEAST TWO TAGGED TEMPLATE NUCLEIC ACID MOLECULES AND/OR SEQUENCING THE MULTIPLE FRAGMENTED TEMPLATE NUCLEIC ACID MOLECULES

[0045] In general sequencing steps can be carried out using any method of sequencing. Examples of possible sequencing methods include Maxam Gilbert Sequencing, Sanger Sequencing, or sequencing comprising bridge PCR. In a typical embodiment the sequencing steps involve bridge PCR, optionally the bridge PCR step is carried out using an extension time of greater than 5, 10, 15 or 20 seconds. An example of the use of bridge PCR is in Illumina Genome Analyzer Sequencers.

[0046] The method of the invention may comprise a step of sequencing regions of the at least two tagged template nucleic acid molecules. As described above, the method of the invention requires that a first and a second molecular tag are introduced into the at least two target template nucleic acid molecules and that each of the at least two template nucleic acid molecules is tagged with a unique tag. Since each of the tagged at least two template nucleic acid molecules comprises a unique tag then, even though multiple copies of the at least two template nucleic acid molecules are produced after the amplification step, it is possible to see which sequences correspond to which individual target template nucleic acid molecule. In order to achieve this the operator must be able to determine the sequence of the first and second unique molecular tag associated with each original target template nucleic acid molecule. This is achieved by sequencing regions of the at least two tagged template nucleic acid molecules wherein the regions comprise the first molecular tag and the second molecular tag. This step may comprise sequencing the entire length of the at least two tagged template nucleic acid molecules or typically comprises sequencing only the ends of the at least two tagged template nucleic acid molecules.

[0047] The method of the invention may comprise a step of sequencing multiple fragmented template nucleic acid molecules. In an embodiment wherein the method comprises a step of sequencing multiple fragmented template nucleic acid molecules, this can be performed in the same sequencing run as the sequencing run in which the at least two tagged template nucleic acid molecules are sequenced. On the other hand it can be more efficient and accurate to sequence the multiple fragmented template nucleic acid molecules in a separate sequencing run from the at least two tagged template nucleic acid molecules.

RECONSTRUCTING A CONSENSUS SEQUENCE FOR AT LEAST ONE OF THE AT LEAST TWO TEMPLATE NUCLEIC ACID MOLECULES

[0048] The methods of the invention may comprise a step of reconstructing a consensus sequence for at least one of the at least two template nucleic acid molecules.

[0049] Optionally the step of reconstructing a consensus sequence comprises a step of identifying clusters of sequences of 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 (for example step S2). For the purposes of the present invention the phrase "homologous to one another" requires that two sequences have greater than 75%, 80%, 85%, 90%, 95%, 98%, 99% or 100% sequence identity to one another over the whole length of the longest sequences. For example if the sequences comprise molecular tags of 10 bp, two molecular tags will be 90% identical to one another if the tags differ in only one base pair. This difference can be a replacement or a deletion of a base pair. This can be determined by aligning the sequences of the molecular tags and comparing them using the 'uclust' algorithm or any similar sequence clustering algorithm such as CD-HIT.

[0050] Optionally the step of reconstructing a consensus sequence comprises a step of analysing the sequences of the at least two tagged template nucleic acid molecules and/or the multiple fragmented template nucleic acid molecules to identify sequences of the at least two tagged template nucleic acid molecules and/or multiple fragmented template nucleic acid molecules which comprise a first molecular tag or a second molecular tag which is homologous to the first molecular tag or the second molecular tag of the sequences of a first cluster (for example step S4 or S7). This may comprise a step of determining a consensus sequence for the first molecular tag sequence and the second molecular tag sequence of a cluster. As described above sequences will be assigned to the same cluster if the first molecular tag sequences and the second molecular tag sequences are homologous to one another. The first molecular tag and the second molecular tag sequences may be slightly different to one another even where the sequences have originated by the same individual target template nucleic acid molecule due to errors in the sequence introduced during the method of the invention. Thus a consensus sequence from these homologous first molecular tag and second molecular tag sequences can be defined. This consensus sequence is highly likely to represent

the sequence of the tag as it was introduced into the target template nucleic acid molecule. Once a consensus sequence for the first molecular tag and the second molecular tag for a cluster have been defined sequences of the multiple fragmented template nucleic acid molecules which comprise a first molecular tag or a second molecular tag which is homologous to one of these consensus sequences can be identified. This provides greater accuracy in identifying the multiple fragmented template nucleic acid molecules which correspond to a particular original template nucleic acid molecule.

[0051] As described above each tagged template nucleic acid molecule comprises a first molecular tag and a second molecular tag. These tagged template nucleic acid molecules are copied and the copies fragmented. Each fragment will have the same sequence as a portion of the individual target template nucleic acid molecule (notwithstanding the possibility of some error in replication during PCR amplification steps) and thus can be considered to 'correspond' to a portion of the original individual target template nucleic acid molecule. A portion of these fragments will comprise the first molecular tag or the second molecular tag. Once sequenced it can, therefore, be identified which individual target template nucleic acid molecule the fragment corresponds to.

[0052] Optionally the step of reconstructing a consensus sequence comprises a step of reconstructing the sequence of a first template nucleic acid molecule by aligning at least a subset of the sequences of the multiple fragmented template nucleic acid molecules identified as comprising a first molecular tag or a second molecular tag homologous to the first molecular tag or the second molecular tag of the sequences of the first cluster and defining a consensus sequences from these sequences (for example step S4, S6 or S7).

[0053] As described above the nature of the first molecular tag or second molecular tag associated with each fragment allows the operator to determine which original template nucleic acid molecule the fragment corresponds to. There will be multiple fragments produced corresponding to the same original template nucleic acid molecule. The sequences of each one of these fragments will correspond to a different (potentially overlapping) region of the template nucleic acid molecule. The sequence of the template can be reconstructed by aligning these fragments and calculating a consensus sequence from the aligned fragments.

[0054] The term 'aligning' refers to arranging the sequences of the fragments in such a way as to align areas of the sequences sharing a common sequence. This can be carried out using software such as Clustal W2, IDBA-UD or SOAPdenovo. Once the sequences are aligned the consensus sequence can be determined. As described above, during the sequencing reaction, mutations may be introduced into the sequences, however these mutated sequences will be at a lower concentration than the accurate sequences. For this reason a 'consensus sequence' is defined. The term 'consensus sequence' can, in the context of the present invention be considered to refer to the most likely sequence for at least one individual target template nucleic acid molecule when considering the sequence of all fragments corresponding to that at least one template nucleic acid molecule.

[0055] In one embodiment each of the sequences of the multiple fragmented template nucleic acid molecules which were identified as comprising a first molecular tag or a second molecular tag homologous to the first molecular tag or the second molecular tag of the sequences of the first cluster are aligned and used to define the consensus sequence (the consensus sequence that is defined does not comprise the first molecular tag or the second molecular tag). In a further embodiment at least a subset but not all of the identified multiple fragmented template nucleic acid molecule sequences are aligned and used to define the consensus sequence. In a further embodiment 90%, 92%, 95%, 98%, 99% or 100% of the identified multiple fragmented template nucleic acid molecule sequences are aligned and used to define the consensus sequence. In a further embodiment the sequences of the full length at least one tagged template nucleic acid molecule are also included in the alignment and used to define the consensus sequence.

[0056] Optionally the method of the invention comprises performing the steps required to reconstruct a consensus sequence for a second or further template nucleic acid molecule. Generally this will involve repeating steps for a second cluster of sequences having first molecular tags that are homologous to one another and second molecular tags that are homologous to one another.

[0057] Optionally these steps of reconstructing a consensus sequence for at least one of the target template nucleic acid molecules are performed by a computer. In a further aspect of the invention there is provided a computer program capable of carrying out these steps of reconstructing a consensus sequence for at least one of the target template nucleic acid molecules optionally stored on a computer-readable medium.

DISREGARDING SEQUENCES OF RECOMBINATION PRODUCTS

[0058] In an aspect of the present invention there is provided a method for generating sequences which comprises or further comprises selecting at least one cluster of sequences wherein the sequences within the selected clusters comprise a first molecular and a second molecular tag which are more commonly associated with one another (e.g. at least 2 times, at least 5 times, at least 8 times or at least 10 times more commonly) than with a different first molecular tag or second molecular tag.

[0059] Optionally this step of selecting at least one cluster consists of identifying groups of clusters of sequences of the at least two tagged template nucleic acid molecules wherein the sequences within the clusters of each group have first molecular tags which are homologous to one another or identifying groups of clusters of sequences of the at least two tagged template nucleic acid molecules wherein the sequences within the clusters of each group have second molecular tags which are homologous to one another. Such a method may further comprise selecting a cluster from a group of clusters of sequences wherein the cluster that is selected contains the highest number of sequences; wherein the sequence of the first template nucleic acid molecule is reconstructed from the sequences in the cluster that was selected. This allows

products of recombination to be detected. Such recombination can result in nucleic acid molecules comprising a sequence corresponding to one part of an original template nucleic acid molecule and a sequence corresponding to one part of a different original template nucleic acid molecule being produced. However such recombination products can be detected if first and second unique molecular tags are introduced into the template nucleic acid molecules. If a recombination event occurs the pair of unique molecular tags will not be the same as any of the pairings of unique molecular tags on any of the original tagged template nucleic acid molecules. This means whilst one might expect one single cluster of sequences to be identified where all the sequences comprise the same first molecular or second molecular tags, if a small amount of recombination has taken place there may be more than one cluster having the same first molecular tag, but pairing this first molecular tag with at least two different second molecular tags. However, these clusters will contain fewer sequences than the cluster which has the same pair of first molecular and second molecular tags as the original template nucleic acid molecule, as a smaller number of copies of the products of recombination will tend to be present than the original template nucleic acid.

[0060] Indeed, it is possible to use the methods of the invention to determine the rate at which recombination is occurring (or the number of recombinants that are generated in a sequencing process). For example, clusters can be identified comprising sequences having a first molecular tag and a second molecular tag which are most commonly associated with one another. Other clusters comprising sequences having the same first molecular tag but a different second molecular tag or the same second molecular tag but a different first molecular tag are likely the result of recombination events and these clusters may be referred to as recombination product clusters. The numbers of sequences in these recombination product clusters may be quantified. The proportion of these sequences (that are the result of recombination) compared to the total number of sequences may be calculated.

[0061] A method of the invention may comprise steps of:

1. a) providing at least one sample of nucleic acid molecules comprising at least two target template nucleic acid molecules;
2. 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 wherein each tagged template nucleic acid molecule is tagged with a unique first molecular tag and a unique second molecular tag;
3. 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;
4. d) sequencing regions of the at least two tagged template nucleic acid molecules comprising the first molecular tag and the second molecular tag; and
5. e) identifying and disregarding sequences that are the product of recombination events.

[0062] Step e) may comprise a step of identifying clusters of sequences of 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. Step e) may further comprise selecting clusters of sequences wherein the sequences within the selected clusters comprise a first molecular and a second molecular tag which are more commonly associated with one another than with a different first molecular tag or second molecular tag. Step e) may further comprise disregarding any sequences that are not present within one of these selected clusters.

[0063] Optionally such a method further comprises a step of determining a consensus sequence from one of the selected clusters. This method may also comprise a step f) of determining the rate at which recombination occurs or the percentage of the total amount of DNA that is a result of a recombination event. In order to perform such a step f), one should determine the total number of sequences present, and the number of sequences that have been disregarded. The percentage of the total DNA that is a result of a recombination event will be equal to the number of sequences that have been disregarded/ the total number of sequences x 100. When generating a consensus sequence, the estimated recombination rate for the cluster can be applied to remove reads that diverge from the majority consensus where the divergent sequence occurs at the rate expected for recombinant fragments. Typically, a sequence that occurs at one of the following frequencies may be disregarded: less than 30%, less than 20%, less than 15%, less than 12% or less than 11%. The estimated recombination rate for a cluster can be reported as a quality metric for the sequence.

COMPUTER-IMPLEMENTED METHODS FOR DETERMINING A SEQUENCE OF AT LEAST ONE TEMPLATE NUCLEIC ACID MOLECULE

[0064] The invention further provides computer-implemented methods for determining sequences of at least two template nucleic acid molecules.

[0065] In such a method data is obtained/input (S1, S3 or S5), for example data comprising sequences of at least two template nucleic acid molecules and/or the data comprising sequences of regions of the at least two template nucleic acid molecule comprising the first molecular tag and the second molecular tag, can be obtained using the method steps described above.

[0066] This method is performed by a computer. In a further aspect there is provided a computer program adapted to perform the methods of the invention when the program is run on an electronic device. In a further aspect there is provided a computer readable medium storing the computer program of the invention.

[0067] As discussed, aspects of the methods discussed herein, including the methods illustrated in Figures 8 to 11, are implemented by a computer. It is well known that an individual computer can comprise standard hardware elements such as CPUs, RAM, storage devices,

etc. It is also well known that pluralities of computers can be connected together and can cooperate so as to perform computing tasks collectively (as a distributed processing system). It will therefore be appreciated that references to computer-implemented methods is intended to include, but not be limited to, methods that use a data processing system (computer) that can perform one or more of the steps described independently or a distributing processing system. A desktop PC working with a cloud computing system via an internet connection is an example of a distributed processing system. Referring to Figure 8, for example, the data to be input in steps S1 and S3 could be stored at a central server in a cloud computing system (this may be referred to as a cloud storage system) and accessed by a desktop computer that is configured to perform the analysis steps S2, S4 and S5. Alternatively, the data to be input in steps S1 and S3 could be provided by the desktop computer and the cloud computing system could be configured to perform the analysis steps S2, S4 and S5 and return the results to the desktop computer. It will be appreciated any other distribution of the data storage and data processing tasks between different computers could be adopted according to the needs of a particular application.

FURTHER DEVELOPMENTS

[0068] Methods of the invention can be modified for even longer sequences. For example, in a method comprising fragmenting the template nucleic acid molecules a further step of introducing further molecular tags (for example a third and a fourth molecular tag) into the fragmented template nucleic acid molecules can be performed. This allows for the fragmented template nucleic acid molecules to be fragmented further, and the further fragmented template nucleic acid molecules to be sequenced. The use of a third molecular tag and a fourth molecular tag allows for reconstructing the sequence of the full length sequence from the further fragmented template nucleic acid molecules.

[0069] The methods of the invention may be used to sequence multiple different genes within a sample of nucleic acids. For example the method of the invention could be used to sequence the whole or a large proportion of the genome of an organism of interest such as a medically relevant pathogen by using a range of primers capable of hybridising to nucleic acids comprising multiple genes. In one embodiment these primers are tethered to a solid surface or coupled to a selectable marker such as biotin.

KITS

[0070] Kits are disclosed. Optionally these kits comprise one or more of the following:

1. (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 a target template nucleic acid molecule; optionally wherein the primers comprise a 'stub region';
2. (ii) primers comprising a portion capable of hybridising to the primers of (i), for example

primers comprising a region complementary to the 'stub region';

3. (iii) a component capable of fragmenting a target template nucleic acid molecule for example a transposase, restriction enzymes or further primers which are complementary to internal regions of the target template nucleic acid molecule;
4. (iv) primers comprising a portion capable of hybridising to a fragmented target template nucleic acid molecule;
5. (v) reagents for performing amplification, for example by polymerase chain reaction;
6. (vi) instructions describing how to perform the methods of the invention; and/or
7. (vii) a computer readable medium storing a computer program of the invention.

EXAMPLE 1

Extraction of microbial DNA from foot skin

[0071] DNA was extracted from skin swabs taken from the feet of 6 different healthy individuals. 12 samples were taken in total. Skin swabs were collected by swabbing either the ball or heel area of the left or right foot with a rayon swab moistened in a solution of 0.15 M NaCl and 0.1% Tween 20. The swab was rubbed firmly over the skin for approximately 30 seconds. Swab heads were cut into bead beating tubes, and DNA was extracted from the swabs using the BiOstic Bacteriemia DNA Isolation Kit (Mo-Bio), as per the manufacturers instructions. DNA was quantified on a Qubit with a dsDNA HS assay (Life Technologies).

EXAMPLE 2

Preparation of short read 16S libraries for Illumina sequencing

[0072] A library of the V4 region of the 16S gene was prepared for Illumina sequencing from the microbial foot skin DNA samples using a previously published method (Caporaso et al, 2012, ISME 6(8)). Briefly, samples were amplified using primers based on the Caporaso design, which were modified to include 8bp rather than 12 bp sample barcodes, and include a barcode on both the forward and reverse primer (primer sequences are described in Figure 2). The V4 region was amplified from 500 pg template DNA using 10 cycles of PCR with the modified Caporaso primers (Caporaso_forward and Caporaso reverse), using different barcoded primers for each sample. After removal of excess primer via a magnetic bead clean-up (Agencourt) samples were pooled, and subjected to a further 20 cycles of PCR to enrich for amplicons containing the Illumina adaptors, using primers Illumina_E_1 and Illumina_E_2 (see Figure 2 for details of primers). PCRs were carried out with a Taq core PCR kit (Qiagen), under

the conditions described in Caporaso et al, (2012, ISME 6(8)). Amplicons were sequenced using a nano flow cell and a 500 cycle V2 kit on an Illumina MiSeq, following the method described in Caporaso et al (2012, ISME 6(8)). This method will be referred to as "short sequencing" and data produced with this method as "V4" data, from here-on in.

EXAMPLE 3

Preparation of full length 16S libraries for Illumina sequencing with unique molecular tags

[0073] Primers for amplification of the 16S gene contained the 27F (Weisberg et al, J Bacteriol. 1991 Jan;173(2):697-703) or 1391R (Turner et al, Journal of Eukaryotic Microbiology, 1999, 46: 327-338) bacterial primer sequences, an 8bp barcode sequence, a 10bp unique molecular tag and partial Illumina PE adapter sequences. Primer sequences (Long_forward and Long reverse) are shown in Figure 2. The use of a 10bp unique molecular tag on both forward and reverse primers (10 billion possible unique tags at each end) allowed us to uniquely tag each 16S molecule in our pool, using a method similar to Lundberg et al (Nature Methods, 2013, 10: 999-1002). Template DNA was subject to one cycle of PCR with the forward primer, followed by a bead clean-up to remove excess primer, then another cycle of PCR with the reverse primer, followed by another bead clean-up. The first PCR carries out extension of the 16S gene from the forward primer, which introduces unique molecular tags into each different 16S template molecule in the reaction. The second PCR uses the extension products from the first PCR as a template, and produces molecules with unique molecular tags at both ends. While the original 16S molecules may also act as a template in the second PCR reaction, these products will only contain a partial Illumina PE adapter sequence at one end, and will therefore not be amplified in the enrichment PCR. The enrichment PCR (34 cycles) amplifies the tagged 16S molecule pool, using primers that are complementary to the partial Illumina PE adapter sequences at the ends of each tagged 16S molecule (Illumina primers PE_1 and PE_2, Figure 2).

[0074] PCRs were carried out using the Taq PCR core kit (Qiagen). Reactions were 50 µl and contained approximately 500 pg DNA template, 0.25 µM F primer, 250 µM dNTPs, 1 x PCR buffer, 1 x Q solution, and 1.25 U Taq polymerase. PCR cycle conditions were 95°C for 1 minute, 50 °C for 2 minutes then 72 °C for 3 minutes. This allows extension of the 16S gene from the forward primer, which introduces unique molecular tags into each 16S molecule in the reaction. PCR reactions were then subject to a magnetic bead clean-up using Agencourt SPRI beads as follows. PCR reactions were mixed with 0.6 volume of beads by pipetting, and incubated at room temperature for 1 minute. Tubes were placed on a magnetic rack for 3 minutes to allow the beads to concentrate on the side of the tube, and the supernatant was removed. The beads were washed with 200 µl of 85% ethanol for 30 seconds, after which the ethanol was removed and the beads allowed to air dry for 5 minutes. Once dry the tubes were removed from the magnetic rack, and the beads resuspended in 35 µl nuclease free water by

pipetting. After an incubation of one minute at room temperature, tubes were placed back on the magnetic rack for 3 minutes, followed by removal of the DNA containing solution to a new tube. The second PCR was set up as described above, except that 0.25 µM of the reverse primer was used, and the template was 31 µl of the bead-cleaned first round PCR reaction. The PCR cycle applied was 95 °C for 1 minute, 50 °C for 2 minutes and 72 °C for 3 minutes. During this second PCR, the uniquely tagged extension products from the first PCR act as the template, to produce 16S molecules with unique molecular tags on both ends. The second PCR was followed by another magnetic bead clean-up, as described above, and the output of this step was used as a template for the final PCR reaction. The final PCR reaction was set up in a 50 µl volume, and contained 0.5 µM both PE_1 and PE_2 primers (see Figure 2), 250 µM dNTPs, 1 x PCR buffer, 1 x Q solution, 31 µl template (from the second bead clean-up) and 1.25 U Taq polymerase. PCR cycling conditions were 95 °C for 2 minutes, followed by 34 cycles of 95 °C for one minute, 58 °C for 30 seconds, and 72 °C for 2 minutes. This was followed by a final extension of 72 °C for 5 minutes. PCRs were again subject to a bead clean-up as described above, before being analysed using a high-sensitivity DNA chip on a Bioanalyser (Agilent).

EXAMPLE 4

Tagmentation of full length, tagged 16S PCR products

[0075] The uniquely tagged, full length 16S PCR amplicons were subject to tagmentation. The tagmentation procedure utilises a transposase to simultaneously fragment the DNA while adding an adapter sequence for use on the Illumina platform. Tagmentation was carried out using the Nextera-XT kit as per the manufacturers instructions, with the exception of the PCR amplification step. Here, we carried out two PCRs per tagmentation reaction, each with a combination of one of the Illumina provided PCR primers with one of the primers from the extension PCR above, so as to amplify only those fragments of interest. We aimed to produce a pool of DNA fragments with either the PE_1 (5' end of the coding sequence of the 16S amplicons) or PE_2 (3' end of the coding sequence of the 16S amplicons) sequences on one end, and the i7 or i5 Illumina adaptors (added during the tagmentation reaction) at the other end, respectively (Figure 2). This provided a pool of fragments from across the 16S gene, which along with the full length 16S amplicons, can be sequenced from either end on the MiSeq. Sequences originating from the same template molecule can be identified via the unique molecular tags at either end of the molecule and re-assembled to provide full length 16S sequences. PCR products from the tagmentation reaction were initially cleaned using 1.8 V of Ampure SPRI beads according the manufacturer's instructions, and in subsequent tagmentation reactions using 0.6 V beads to remove fragments smaller than 400 bp.

EXAMPLE 5

Sequencing of full length and tagmented 16S amplicons on the Illumina MiSeq

[0076] The molarity of both full length 16S tagged amplicons and the tagmentation products was measured via a Bioanalyser High sensitivity DNA chip. During the first sequencing run, only tagmentation products (cleaned with 1.8 V Ampure SPRI beads) were loaded at an average concentration of 1.5 pM and sequenced with a MiSeq reagent kit v2 with 2 x 150 bp paired end reads, on a nano flow cell. For the second sequencing run, full length 16S tagged amplicons were combined with the tagmentation products (cleaned with 0.6 V Ampure SPRI beads to remove fragments < 400 bp) at a ratio of 1:9. The pooled sample was loaded at an average molarity of 6pM, and sequenced with a MiSeq reagent kit v2 with 2 x 250 bp paired end reads, on a nano flow cell.

[0077] When the full length 16S tagged amplicons were run, modifications were made to the running conditions of the MiSeq. The Chemistry.xml file in the Recipe folder on the Illumina MiSeq contains the protocol used by the instrument for clustering and sequencing DNA fragments. That Chemistry.xml file corresponding to the Illumina Version 2 sequencing kits was modified to increase the "WaitDuration" in the "Amplification 1" "Resyntheses" and "First extension" steps to 15 seconds. This resulted in a process that allowed the ends of individual full length 16S tagged amplicons to be sequenced.

EXAMPLE 6

Reconstructing full length 16S sequences from tagged Illumina reads

[0078] Sequencing produces data from two kinds of fragments, those which span the entire 16S gene (*end+end fragments*) and those which pair one end of the 16S gene with a region in the middle of the 16S gene (*end+internal fragments*). Sequences from end+end fragments encode a pairing of random barcodes and sample barcodes.

[0079] To assign sequences to samples, the 8 nt sample barcode region is matched against the collection of known sample barcodes with up to one mismatch tolerated. Because internal regions of the 16S sequence might match a sample barcode, all reads with a potential sample barcode match are then screened for the presence of the proximal or distal 16S primer annealing sequence downstream from the sample barcode. Reads lacking a known sample barcode or the primer annealing sequence in one end are presumed to derive from an end+internal fragment.

EXAMPLE 7

Consensus unique molecular tags and elimination of recombinants

[0080] Due to sequencing error, the reads derived from the same template molecule may have slightly different 10nt unique molecular tag sequences. To estimate the original 10nt random barcode sequences of tagged template molecules we apply the uclust (Edgar, R. C. (2010) Search and clustering orders of magnitude faster than BLAST, Bioinformatics 26(19), 2460-2461; Edgar, R.C. (2013) UPARSE: Highly accurate OTU sequences from microbial amplicon reads, Nature methods) algorithm to identify clusters of matching random barcode sequences at >89% identity (e.g. 1 out of 10 bases are allowed to mismatch), and to report the consensus sequences of these clusters. We first identify clusters of random barcodes in the end+end fragments. We then identify the highest abundance cluster with each 10nt random barcode and discard any cluster containing a 10nt random barcode that was found in a different, more abundant cluster. This step aims to identify and discard combinations of random barcodes that arose due to in-vitro recombination. Recombinant forms are likely to be at lower abundance than the parental templates (Figure 3). We note that when sequencing arbitrary 2Kbp fragments, such in vitro recombination is not expected to occur very frequently due to the diversity of the template molecule pool. Recombination detection is most important for application to amplicon sequencing protocols such as for the 16S.

[0081] The end+end fragments may not capture all random barcodes present in a sample. The remaining random barcodes might still be used to reconstruct 16S sequences even though they can not be assigned to a sample without end+end fragment information. Therefore, we apply uclust again to identify clusters of random barcodes on each end separately, and add any new consensus sequences that were not previously found in an end+end fragment.

[0082] Finally, random barcodes from entire set of reads are matched against the collection of consensus sequences and the reads are grouped into clusters for later assembly.

EXAMPLE 8

Assembly of read clusters

[0083] Read clusters contain reads that, with high probability, originate from the same template molecule. We apply a *de novo* assembly algorithm on the read cluster to reconstruct as much of the original template molecule as possible. The reads are assembled using the A5-miseq pipeline (Tritt et al (2012). An integrated pipeline for de Novo assembly of Microbial Genomes, PLoS One). A5-miseq is a revision of the original A5 pipeline, extending it to support assembly of reads up to 500nt long and to trim out adapter sequence from reads instead of discarding

reads containing adapter sequence.

[0084] This method will be referred to as "long sequencing" and data produced with this method as "long" data, from here-on in.

EXAMPLE 9

Analysis of 16S reads

[0085] 12 foot samples were sequenced with the full length protocol, 6 of which were sequence twice with the method. All 12 samples were also sequenced using the Caparoso *et al* 2012 method.

[0086] Both V4 and long reads were analysed using the software package QIIME (Caparoso et al (2010), QIME allows analysis of high-throughput community sequence data, Nature Methods 7: 335-335). V4 reads were quality filtered by removing reads less than 248 or more than 253 bp. For comparison, the corresponding V4 region was extracted from the long dataset, and only those assembled sequences that included the V4 region were included in the downstream analysis. These extracted sequences will be referred to as "long-V4" from here on in. All sequences were clustered into OTUs using the closed reference picking method, which assigns sequences to pre-clustered OTUs from a chimera free database (Greengenes). Taxonomy was assessed based on membership to the database of pre-clustered OUTs.

Short sequencing

[0087] A total of 296864 paired end V4 sequences were generated from 12 foot samples and a positive (Escherichia coli DNA only) and negative (swab only) control. Of these sequences, 11240 could not be assigned to a sample because of incorrect forward and reverse barcode combinations, indicating a recombination rate of at least 3.8%. 240938 sequences mapped to the 12 foot samples, which was reduced to 240426 after quality filtering (see Table 1 below for number of sequences assigned to each sample). OTUs clustered with the closed reference method in QIIME resulted in 1177 OTUs at 97% similarity containing 2 or more sequences. The taxonomic distribution of these OTUs was similar to what has been reported previously for skin communities, dominated by Firmicutes (79.6% ± 25.7), Actinobacteria (9.3%± 12.9), and Proteobacteria (9.9% ±22.2).

Table 1: Number of sequences analysed per sample for the different sequencing methods

	Number of sequences after quality filtering		
Sample	V4	Long	Long_V4
F1.B1	29853	69	69

	Number of sequences after quality filtering		
Sample	V4	Long	Long_V4
F1.H	10241	37	37
F2.B2	6501	30	30
F2.H	5560	80	80
F3.B2	5258	4	4
F3.H	38108	85	85
F4.LB	5647	32	32
F4.LH	3266	24	24
F5.LB	13931	505	505
F5.LH	66398	836	836
F6.LB	33714	431	431
F6.LH	21949	218	218
Total	**240426**	**2351**	**2351**

Long Sequences

[0088] 3914 16S sequences were assembled, with 2030 of these being longer than 1000 bp (Figure 4). 2957 sequences were assigned to foot samples, while 957 sequences could not be assigned to a sample because of incorrect molecular tag combinations. Only reads which contained a V4 region corresponding to that sequenced with the short sequencing method were used for downstream analysis, and these sequences were quality filtered in QIIME by removing sequences shorter than 700 bp and longer than 1500 bp. This resulted in 2351 sequences used for analysis (see Table 1 for details of how many sequences were assigned to each sample).

[0089] Long reads (2351 used for analysis) clustered into 72 OTUs, while the V4-long sequences (corresponding to the same region as the V4 dataset) clustered into 48 OTUs. These OTUs showed the same broad taxonomic distribution as the V4 sequence data (Figure 5). Although there was a small increase in the representation of Actinobacteria (13.6% ±21.6) and Proteobacteria (11.4±26.7), these differences were not significant (two tailed t-test, $p > 0.05$).

[0090] Similar taxonomic assignments were also observed at the level of genus (Figure 6), with communities dominated by Staphylococcus, followed by Corynebacterium, Enhydrobacter and Acinetobacter genera. The Corynebacterium genus had an increased representation in the long data set as compared to the short sequencing method, which likely accounts for the observed difference in representation for the Actinobacteria phyla, but as above, this difference

was not significant (two tailed t-test, $p > 0.05$). Comparison of individual samples between the short sequencing and long methods showed that Corynebacteria were not consistently over-represented in the assembled dataset, and the average was strongly influenced by one sample where Corynebacterium represented only 0.03% sequences in the V4 sample, but 46.67% of sequences in the assembled long sequencing data (sample F2_B2).

Recombination rates

Comparison at the OTU level

[0091] Assembled 16S sequences (lengths varying from 756 to 1375) were clustered in OTUs using the closed reference method in QIIME, and shared on average only 30.1%(±6.8) of OTUs with matched sample V4 data which was clustered in the same way. This may be due to comparing datasets of different lengths, and the way in which OTUs are clustered in QIIME. Sequences are assigned to OTUs by the best match against a database of sequences, which have been pre-clustered into OTUs at 97% similarity. Presumably, full length sequences from the database were used to cluster OTUs, and clusters that are 97% similar across the full 16S gene, may not be 97% similar in the V4 region only, since different regions of the 16S gene evolve at different rates (Schloss PD (2010) The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16S rRNA Gene-Based Studies. Plos Computational Biology 6). We therefore analysed OTUs clustered from the V4 region only of the long sequences (long-V4 sequences). In this case 92.2%(±12.1) of OTUs were shared with the matched Caporaso sample OTUs (Table 3). Although a lower coverage of sequencing was obtained in the long data set, and subsequently much fewer OTUs overall, this shows that the data that was obtained is broadly concurrent with that obtained using short V4 sequences. Interestingly, the long sequences clustered into ~50% more OTUs than the long-V4 sequences, demonstrating the more sensitive classification achievable with more sequence information per 16S molecule.

[0092] This data indicates that this newly developed method gives broadly concurrent community profiles with respect to taxonomy and OTU clustering, and allows for more sensitive taxonomic assignment.

EXAMPLE 10

Sequencing long fragments from *E. coli* K12 MG1655

[0093] Genomic DNA from E. coli K12 MG1655 was tagmented and fragments 1.5-3kbp were size selected using agarose gel electrophoresis. Molecular tagging was applied to these

fragments via 2 cycles PCR with random barcodes. Initial sequencing of the pool revealed an excess of diversity among template molecules, such that reconstruction of full length templates would be infeasible. A dilution series was used to determine the appropriate degree to which the population of template molecules should be bottlenecked for successful sequencing & reconstruction of full length templates (Figure 7). Both 50x and 100x dilutions were sequenced with fill-in reads.

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Non-patent literature cited in the description

- **LUNDBERG et al.**Nature Methods, vol. 10, 999-1002 [0031]
- **WEISBERG et al.**J Bacteriol., 1991, vol. 173, 2697-703 [0035] [0073]
- **CAPORASO et al.**ISME, 2012, vol. 6, 8 [0072] [0072] [0072]
- **TURNER et al.**Journal of Eukaryotic Microbiology, 1999, vol. 46, 327-338 [0073]
- **LUNDBERG et al.**Nature Methods, 2013, vol. 10, 999-1002 [0073]
- **EDGAR, R. C.**Search and clustering orders of magnitude faster than BLASTBioinformatics, 2010, vol. 26, 192460-2461 [0080]
- **EDGAR, R.C.**UPARSE: Highly accurate OTU sequences from microbial amplicon readsNature methods, 2013, [0080]
- **TRITT et al.**An integrated pipeline for de Novo assembly of Microbial GenomesPLoS One, 2012, [0083]
- **CAPAROSO et al.**QIME allows analysis of high-throughput community sequence dataNature Methods, 2010, vol. 7, 335-335 [0086]

PATENTKRAV

1. Fremgangsmåde til generering af sekvenser af mindst et individuelt mål-template-nukleinsyremolekyle, hvilken fremgangsmåde omfatter:

a) tilvejebringelse af mindst én prøve af nukleinsyremolekyler, der omfatter mindst to mål-template-nukleinsyremolekyler;

b) indføring af et første molekyl-tag i én ende af hvert af de mindst to mål-template-nukleinsyremolekyler og et andet molekyl-tag i den anden ende af hvert af de mindst to mål-template-nukleinsyremolekyler for at tilvejebringe mindst to taggede template-nukleinsyremolekyler, hvor hvert tagget template-nukleinsyremolekyle er tagget med et unikt første molekyl-tag og et unikt andet molekyl-tag;

c) amplificering af de mindst to taggede template-nukleinsyremolekyler for at tilvejebringe talrige kopier af de mindst to taggede template-nukleinsyremolekyler;

d) sekvensering af områder af de mindst to taggede template-nukleinsyremolekyler, der omfatter første molekyl-tag og det andet molekyl-tag; og

e) rekonstruktion af en konsensussekvens for mindst ét af de mindst to mål-template-nukleinsyremolekyler,

hvor trin e) omfatter

(i) identificering af klynger af sekvenser af områderne af de talrige kopier af de mindst to taggede template-nukleinsyremolekyler, der sandsynligvis svarer til det samme mål-template-nukleinsyremolekyle, ved tildeling af sekvenser, der omfatter første molekyl-tag-sekvenser, der er homologe med hinanden, og anden molekyl-tag-sekvenser, der er homologe med hinanden, til den samme klynge;

(ii) valg af mindst én klynge af sekvenser, hvor sekvenserne i de valgte klynger omfatter et første molekyl-tag og et andet molekyl-tag, der mere almindeligvis er forbundet med hinanden end med et andet første molekyl-tag eller andet molekyl-tag;

(iii) rekonstruktion af en konsensussekvens af et første mål-template-nukleinsyremolekyle ved aligning af sekvenser af de mindst to template-nukleinsyremolekyler i klyngen valgt i trin (ii) og definition af en konsensussekvens ud fra disse sekvenser; og

(iv) udførelse af trin (ii) til (iii) med hensyn til et andet og/eller yderligere template-nukleinsyremolekyle.

2. Fremgangsmåde til generering af sekvenser af mindst ét individuelt i mål-template-nukleinsyremolekyle, der er større end 1 Kbp i størrelse, hvilken fremgangsmåde omfatter:

a) tilvejebringelse af mindst én prøve af nukleinsyremolekyler, der omfatter mindst to mål-template-nukleinsyremolekyler, der er større end 1 Kbp i størrelse;

b) indføring af et første molekyl-tag i én ende af hvert af de mindst to mål-template-nukleinsyremolekyler og et andet molekyl-tag i den anden ende af hvert af de mindst to mål-template-nukleinsyremolekyler for at tilvejebringe mindst to taggede template-nukleinsyremolekyler, hvor hvert af de to taggede template-nukleinsyremolekyler er tagget med et unikt første molekyl-tag og et unikt andet molekyl-tag;

c) amplificering af de mindst to taggede template-nukleinsyremolekyler for at tilvejebringe talrige kopier af de mindst to taggede template-nukleinsyremolekyler;

d) isolering af en fraktion af de talrige kopier af de mindst to taggede template-nukleinsyremolekyler og fragmentering af de taggede template-nukleinsyremolekyler i fraktionen for at tilvejebringe talrige fragmenterede template-nukleinsyremolekyler;

e) sekvensering af områder af de talrige kopier af de mindst to taggede template-nukleinsyremolekyler, der omfatter det første molekyl-tag og det andet molekyl-tag;

f) sekvensering af de talrige fragmenterede template-nukleinsyremolekyler; og

g) rekonstruktion af en konsensussekvens for mindst ét af de mindst to mål-template-nukleinsyremolekyler ud fra sekvenser, der omfatter mindst et undersæt af sekvenserne frembragt i trin f).

3. Fremgangsmåde ifølge krav 2, hvor:

(A) fremgangsmåden endvidere omfatter et trin med berigelse af de talrige fragmenterede template-molekyler for at øge andelen af de talrige fragmenterede template-nukleinsyremolekyler, der omfatter det første molekyl-tag eller det andet molekyl-tag, og hvor dette trin er før trin f); og/eller

(B) trin g) omfatter:

(i) identificering af klynger af sekvenser af områder af de talrige kopier af de mindst to taggede template-nukleinsyremolekyler, der sandsynligvis svarer til det samme individuelle mål-template-nukleinsyremolekyle, ved tildeling af sekvenser, der omfatter første molekyl-tag-sekvenser, der er homologe med hinanden, og anden molekyl-tag-sekvenser, der er homologe med hinanden, til den samme klynge;

(ii) analyse af sekvenserne af de talrige fragmenterede template-nukleinsyremolekyler for at identificere sekvenser af de talrige fragmenterede template-nukleinsyremolekyler, der omfatter et første molekyl-tag, der er homologt med det første molekyl-tag af sekvenserne af en første klynge, eller et andet molekyl-tag, der er homologt med det andet molekyl-tag af sekvenserne af den første klynge;

(iii) rekonstruktion af sekvensen af et første template-nukleinsyremolekyle ved aligning af sekvenser, der omfatter mindst et undersæt af sekvenserne af de talrige fragmenterede template-nukleinsyremolekyler identificeret i trin (ii), og definition af en konsensussekvens ud fra disse sekvenser; og

(iv) udførelse af trin (i) til (iii) med hensyn til et andet og/eller yderligere template-nukleinsyremolekyle, eventuelt hvor trin (i) endvidere omfatter bestemmelse af en konsensussekvens for første molekyl-tag-sekvenser og en konsensussekvens for anden molekyl-tag-sekvenser af en første klynge og trin (ii) omfatter identificering af sekvenser af de talrige fragmenterede template-nukleinsyremolekyler, der omfatter et første molekyl-tag eller et andet molekyl-tag, der er homologt med konsensussekvensen for det første molekyl-tag eller konsensussekvensen for det andet molekyl-tag af den første klynge; og/eller

(C) trin g) er et computerimplementeret fremgangsmådetrin; og/eller

(D) trin e) og/eller f) udføres ved anvendelse af sekvenseringsteknologi, der omfatter et trin med bro-PCR, eventuelt hvor trinnet med bro-PCR udføres ved anvendelse af en forlængelsestid på mere end 15 sekunder; og/eller

(E) trin e) og f) udføres i forskellige sekvenseringskørsler.

4. Computerimplementeret fremgangsmåde til bestemmelse af sekvenser af mindst ét individuelt mål-template-nukleinsyremolekyle, hvilken fremgangsmåde omfatter følgende trin:

(a) opnåelse af data, der omfatter sekvenser af områder af talrige kopier af mindst to taggede template-nukleinsyremolekyler, hvor hvert af de to taggede template-nukleinsyremolekyler omfatter et første molekyl-tag ved én ende og et andet molekyl-tag ved den anden ende, hvor hvert mål-template-nukleinsyremolekyle er tagget med et unikt første molekyl-tag og et unikt andet molekyl-tag, og hvor områderne omfatter det første molekyl-tag og det andet molekyl-tag;

(b) analyse af de data, der omfatter sekvenser af områder af de mindst to taggede template-nukleinsyremolekyler, der omfatter det første molekyl-tag og det andet molekyl-tag, for at identificere klynger af sekvenser, der sandsynligvis svarer til det samme individuelle mål-template-nukleinsyremolekyle, ved tildeling af sekvenser, der omfatter første molekyl-tags, der er homologe med hinanden, og anden molekyl-tags, der er homologe med hinanden, til den samme klynge;

(c) opnåelse af data, der omfatter sekvenser af talrige fragmenter af de mindst to taggede template-nukleinsyremolekyler, hvor hvert af fragmenterne omfatter enten det første molekyl-tag eller det andet molekyl-tag;

(d) analyse af sekvenserne af de talrige fragmenter af de mindst to taggede template-nukleinsyremolekyler for at identificere sekvenser af de talrige fragmenter af de mindst to taggede template-nukleinsyremolekyler, der omfatter det første molekyl-tag, der er homologt med det første molekyl-tag af sekvenserne af en første klynge eller det andet molekyl-tag, der er homologt med det andet molekyl-tag af sekvenserne af den første klynge;

(e) rekonstruktion af sekvensen af et første mål-template-nukleinsyremolekyle ved aligning af sekvenser, der omfatter mindst et undersæt af sekvenserne af de talrige fragmenter af de mindst to taggede template-nukleinsyremolekyler identificeret i trin (d), og definition af en konsensussekvens ud fra disse sekvenser; og

(f) udførelse af trin (c) til (e) med hensyn til et andet og/eller yderligere mål-template-nukleinsyremolekyle, eventuelt

hvor trin (b) endvidere omfatter bestemmelse af en konsensussekvens for første molekyl-tag-sekvenser og en konsensussekvens for anden molekyl-tag-sekvenser af en første klynge og trin (d) omfatter identificering af sekvenser af de talrige fragmenterede template-nukleinsyremolekyler, der omfatter et første molekyl-tag eller et andet molekyl-tag, der er homologt med konsensussekvensen for det første molekyl-tag eller konsensussekvensen for det andet molekyl-tag af den første klynge.

5. Computerimplementeret fremgangsmåde til bestemmelse af sekvenser af mindst ét template-nukleinsyremolekyle, hvilken fremgangsmåde omfatter følgende trin:

(a) opnåelse af data, der omfatter klynger af sekvenser, hvor:

(i) hver klynge omfatter sekvenser af områder af talrige kopier af mindst to taggede template-nukleinsyremolekyler, hvor hvert af de to taggede template-nukleinsyremolekyler omfatter et første molekyl-tag ved én ende og et andet molekyl-tag ved den anden ende, hvor

hver mål-template-nukleinsyre er tagget med et unikt første molekyl-tag og et unikt andet molekyl-tag, og hvor områderne omfatter det første molekyl-tag og det andet molekyl-tag;

(ii) hver klynge omfatter sekvenser af talrige fragmenter af de mindst to taggede template-nukleinsyremolekyler, hvor hvert af fragmenterne omfatter enten det første molekyl-tag eller det andet molekyl-tag;

(iii) sekvenserne af områder af talrige kopier af mindst to taggede template-nukleinsyremolekyler i hver klynge omfatter første molekyl-tags og anden molekyl-tags, der er homologe med hinanden;

(iv) sekvenserne af de talrige fragmenter af de mindst to taggede template-nukleinsyremolekyler omfatter det første molekyl-tag, der er homologt med det første molekyl-tag af sekvenserne af områder af de talrige kopier af mindst to taggede template-nukleinsyremolekyler i denne klynge eller det andet molekyl-tag, der er homologt med det andet molekyl-tag af sekvenserne af områder af talrige kopier af de mindst to taggede template-nukleinsyremolekyler i denne klynge;

(b) rekonstruktion af sekvensen af et første mål-template-nukleinsyremolekyle ved aligning af sekvenser, der omfatter mindst et undersæt af sekvenserne af de talrige fragmenter af de mindst to taggede template-nukleinsyremolekyler i en første klynge, og definition af en konsensussekvens ud fra disse sekvenser; og

(c) udførelse af trin (b) med hensyn til et andet og/eller yderligere template-nukleinsyremolekyle.

6. Fremgangsmåde ifølge et hvilket som helst af kravene 2-5, der endvidere omfatter følgende trin:

(v) identificering af klynger af sekvenser af områder af de talrige kopier af de mindst to taggede template-nukleinsyremolekyler, der sandsynligvis svarer til det samme template-nukleinsyremolekyle, ved tildeling af sekvenser, der omfatter første molekyl-tag-sekvenser, der er homologe med hinanden, og anden molekyl-tag-sekvenser, der er homologe med hinanden, til den samme klynge;

(vi) valg af mindst én klynge af sekvenser, hvor sekvenserne i de valgte klynger omfatter et første molekyl-tag og et andet molekyl-tag, der mere almindeligvis er forbundet med hinanden end med et andet første molekyl-tag eller andet molekyl-tag;

hvor sekvensen af det første mål-template-nukleinsyremolekyle rekonstrueres ud fra sekvenserne i klyngen valgt i trin (vi),

eventuelt hvor trin (vi) består af identificering af grupper af klynger af sekvenser af de mindst to taggede template-nukleinsyremolekyler, hvor sekvenserne i hver gruppes klynger har første molekyl-tags, der er homologe med hinanden, og/eller identificering af grupper af klynger af sekvenser af de mindst to taggede template-nukleinsyremolekyler, hvor sekvenserne inde i hver gruppes klynger har anden molekyl-tags, der er homologe med hinanden, og valg af en klynge fra gruppen af klynger af sekvenser, hvor den valgte klynge indeholder det største antal sekvenser.

7. Fremgangsmåde ifølge et hvilket som helst af kravene 2-4, hvor:

(A) trin e) er et computerimplementeret fremgangsmådetrin; og/eller

(B) trin d) udføres ved anvendelse af sekvenseringsteknologi, der omfatter et trin med bro-PCR, eventuelt hvor trinnet med bro-PCR udføres ved anvendelse af en forlængelsestid på mere end 15 sekunder.

8. Computerimplementeret fremgangsmåde til bestemmelse af sekvenser af mindst ét individuelt mål-template-nukleinsyremolekyle, hvilken fremgangsmåde omfatter følgende trin:

(a) opnåelse af data, der omfatter sekvenser af områder af talrige kopier af mindst to taggede template-nukleinsyremolekyler, hvor hvert af de to taggede template-nukleinsyremolekyler omfatter et første molekyl-tag ved én ende og et andet molekyl-tag ved den anden ende, hvor hvert mål-template-nukleinsyremolekyle er tagget med et unikt første molekyl-tag og et unikt andet molekyl-tag, og hvor områderne omfatter det første molekyl-tag og det andet molekyl-tag;

(b) analyse af de data, der omfatter sekvenser af områder af de mindst to taggede template-nukleinsyremolekyler, der omfatter det første molekyl-tag og det andet molekyl-tag, for at identificere klynger af sekvenser, der sandsynligvis svarer til det samme template-nukleinsyremolekyle ved tildeling af sekvenser, der omfatter første molekyl-tags, der er homologe med hinanden, og anden molekyl-tags, der er homologe med hinanden, til den samme klynge;

(c) valg af mindst én klynge af sekvenser, hvor sekvenserne i de valgte klynger omfatter et første molekyl-tag og et andet molekyl-tag, der mere almindeligvis er forbundet med hinanden end med et andet første molekyl-tag eller andet molekyl-tag;

(d) rekonstruktion af en konsensussekvens af et første mål-template-nukleinsyremolekyle ved aligning af mindst et undersæt af sekvensmolekylerne i klyngen valgt i trin (c) og definition af en konsensussekvens ud fra disse sekvenser;

og

(e) udførelse af trin (c) til (d) med hensyn til et andet og/eller yderligere mål-template-nukleinsyremolekyle.

9. Fremgangsmådetrin ifølge krav 1 (iv) eller fremgangsmådetrin ifølge krav 8(c), der består i at identificere grupper af klynger af sekvenser af de mindst to taggede template-nukleinsyremolekyler, hvor sekvenserne inde i hver gruppes

klynger har 5'-molekyl-tags, der er homologe med hinanden, og/eller identificering af grupper af klynger af sekvenser af de mindst to taggede template-nukleinsyremolekyler, hvor sekvenserne inde i hver gruppes klynger har 3'-molekyl-tags, der er homologe med hinanden; og valg af en klynge fra en gruppe af klynger af sekvenser, hvor den valgte klynge indeholder det største antal sekvenser.

10. Computerimplementeret fremgangsmåde til bestemmelse af sekvenser af mindst ét mål-template-nukleinsyremolekyle, hvilken fremgangsmåde omfatter

(a) opnåelse af data, der omfatter en klynge af sekvenser;

(b) rekonstruktion af en konsensussekvens af et første template-nukleinsyremolekyle ved aligning af sekvenserne af mindst et undersæt af sekvenserne i den valgte klynge;

hvor sekvenserne i den valgte klynge omfatter sekvenser af områder af talrige kopier af mindst to taggede template-nukleinsyremolekyler, hvor hvert af de to taggede template-nukleinsyremolekyler omfatter et første molekyl-tag ved én ende og et andet molekyl-tag ved den anden ende, hvor hvert af de mindst to mål-template-nukleinsyremolekyler er taget med et unikt første molekyl-tag og et unikt andet molekyl-tag, og hvor områderne omfatter det første molekyl-tag og det andet molekyl-tag; og hver sekvens i den valgte klynge

(i) omfatter et første molekyl-tag, der er homologt med det første molekyl-tag af de andre sekvenser i denne klynge, og det andet molekyl-tag, der er homologt med det andet molekyl-tag af de andre sekvenser i denne klynge;

(ii) omfatter et første molekyl-tag og et andet molekyl-tag, der mere almindeligvis er forbundet med hinanden end med et andet første molekyl-tag eller andet molekyl-tag.

11. Fremgangsmåde ifølge et hvilket som helst af kravene 3(B)-10, hvor:

(A) første molekyl-tags af den samme klynges sekvenser har mindst 90 % sekvensidentitet med hinanden; og/eller

(B) anden molekyl-tags af den samme klynges sekvenser har mindst 90 % sekvensidentitet med hinanden.

12. Fremgangsmåde ifølge krav 3(B), 6 eller 7, der er en computerimplementeret fremgangsmåde.

13. Fremgangsmåde ifølge et hvilket som helst af de foregående krav, hvor:

(A) områderne omfatter mere end 25 basepar, der omfatter det første molekyl-tag eller andet molekyl-tag; og/eller

(B) områderne omfattende den fulde længde af de mindst to taggede template-nukleinsyremolekyler er sekvenseret; og/eller

(C) det første molekyl-tag og det andet molekyl-tag indføres i de mindst to template-nukleinsyremolekyler ved anvendelse af en fremgangsmåde valgt fra gruppen bestående af PCR, tagmentering og fysisk afklipning eller restriktionsdigestion af det mindst ene template-nukleinsyremolekyle efterfulgt af ligering af nukleinsyrer, der omfatter 5'-molekyl-tagget eller 3'-molekyl-tagtet, eventuelt

hvor det første molekyl-tag og det andet molekyl-tag indføres i de mindst to template-nukleinsyremolekyler ved PCR ved anvendelse af primere, der omfatter en del, der omfatter det første molekyl-tag eller det andet molekyl-tag og en del med en sekvens, der kan hybridisere til de mindst to template-nukleinsyremolekyler; og/eller

(D) de mindst to template-nukleinsyremolekyler koder for mikrobiel ribosomal 16S; og/eller

(E) mindst ét af de mindst to template-nukleinsyremolekyler er mindre end 10 Kbp i størrelse.

14. Computerprogram tilpasset til at udføre fremgangsmåden ifølge krav 4, 5, 8 eller 10, fremgangsmådetrin g) ifølge krav 2 eller fremgangsmådetrin e) ifølge krav 1, når programmet køres på en elektronisk enhed.

15. Computerlæsbart medium, hvorpå computerprogrammet ifølge krav 14 er lagret.

DRAWINGS

Figure 1

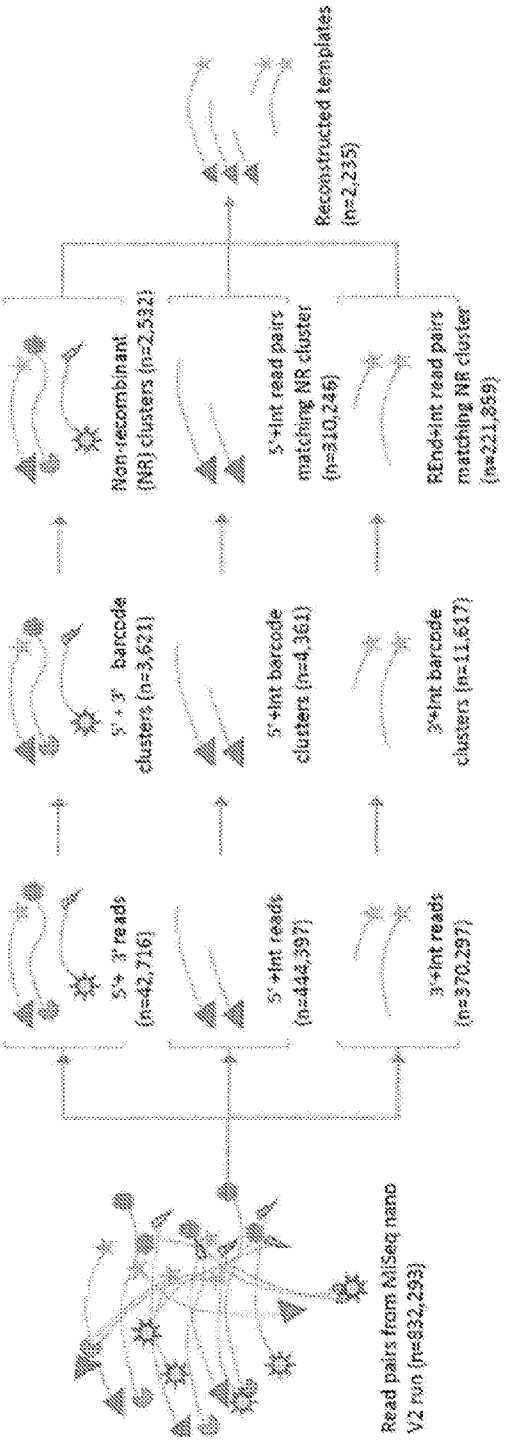


Figure 2a

Table 1: Primers used for 16S gene amplification and sequencing

Primer name	Sequence
Long_forward_1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNGTTGGCCGCGAGAGTTTGATCMTGGCTCAG
Long_forward_2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNTATTAACTNCGAGAGTTTGATCMTGGCTCAG
Long_forward_3	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNCTAATGGCNNCGAGAGTTTGATCMTGGCTCAG
Long_forward_4	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNAACCAGTCNNNCGAGAGTTTGATCMTGGCTCAG
Long_forward_5	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNGAACGGAGCGAGAGTTTGATCMTGGCTCAG
Long_forward_6	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNACTGAAGTNCGAGAGTTTGATCMTGGCTCAG
Long_forward_7	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNTTGGCTATNNCGAGAGTTTGATCMTGGCTCAG
Long_forward_8	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNTGGCGATTNNNCGAGAGTTTGATCMTGGCTCAG
Long_forward_9	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNCCTCTGATCGAGAGTTTGATCMTGGCTCAG
Long_forward_10	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNCTCATGCGNCGAGAGTTTGATCMTGGCTCAG
Long_forward_11	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNTTCAGCGANNCGAGAGTTTGATCMTGGCTCAG
Long_forward_12	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNGGATGCCANNNCGAGAGTTTGATCMTGGCTCAG
Long_forward_13	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNCGGTCGAGCGAGAGTTTGATCMTGGCTCAG
Long_forward_14	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNAAGACTACNCGAGAGTTTGATCMTGGCTCAG
Long_forward_15	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNAACGCTAANNCGAGAGTTTGATCMTGGCTCAG
Long_forward_16	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNGCCTACGCNNNCGAGAGTTTGATCMTGGCTCAG
Long_forward_17	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNTGACTGCTCGAGAGTTTGATCMTGGCTCAG
Long_forward_19	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNCAACCTTANNCGAGAGTTTGATCMTGGCTCAG
Long_forward_20	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNGGAGGCTGNNNCGAGAGTTTGATCMTGGCTCAG
Long_forward_21	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNAATCGATACGAGAGTTTGATCMTGGCTCAG
Long_forward_22	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNACCAATTGNCGAGAGTTTGATCMTGGCTCAG

Figure 2b

Long_forward_23	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNCCTAATAANNCGAGAGTTTGATCM TGGCTCAG
Long_forward_24	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNGGATTAGGNNNCGAGAGTTTGATC MTGGCTCAG
Long_forward_25	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNGCGTTACCCGAGAGTTTGATCMTG GCTCAG
Long_reverse_1	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGTTGGCCGTAGACGGGCGGTGTGT RCA
Long_reverse_2	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNTATTAACTNNNTAGACGGGCGGTG TGTRCA
Long_reverse_3	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNCTAATGGCTAGACGGGCGGTGTGT RCA
Long_reverse_4	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNAACCAGTCNNNTAGACGGGCGGTG TGTRCA
Long_reverse_5	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGAACGGAGTAGACGGGCGGTGTGT RCA
Long_reverse_6	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNACTGAAGTNNTAGACGGGCGGTGT GTRCA
Long_reverse_7	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNTTGGCTATNNNTAGACGGGCGGTG TGTRCA
Long_reverse_8	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNTGGCGATTTAGACGGGCGGTGTGT RCA
Long_reverse_9	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNCCTCTGATNTAGACGGGCGGTGTG TRCA
Long_reverse_10	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNCTCATGCGNNTAGACGGGCGGTGT GTRCA
Long_reverse_11	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNTTCAGCGANTAGACGGGCGGTGTG TRCA
Long_reverse_12	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGGATGCCANNTAGACGGGCGGTGT GTRCA
Long_reverse_13	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNCGGTCGAGNTAGACGGGCGGTGTG TRCA
Long_reverse_14	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNAAGACTACNNNTAGACGGGCGGTG TGTRCA
Long_reverse_15	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNAACGCTAATAGACGGGCGGTGTGT RCA
Long_reverse_16	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGCCTACGCNTAGACGGGCGGTGTG TRCA
Long_reverse_17	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNTGACTGCTNNTAGACGGGCGGTGT GTRCA
Long_reverse_18	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNATTGCCGCNTAGACGGGCGGTGTG TRCA
Long_reverse_19	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNCAACCTTANNTAGACGGGCGGTGT GTRCA

Figure 2c

Long_reverse_20	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGGAGGCTGNTAGACGGGCGGTGTG TRCA
Long_reverse_21	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNAATCGATANNTAGACGGGCGGTGT GTRCA
Long_reverse_22	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNACCAATTGNTAGACGGGCGGTGTG TRCA
Long_reverse_23	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNCCTAATAANTAGACGGGCGGTGTG TRCA
Long_reverse_24	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGGATTAGGNNTAGACGGGCGGTGT GTRCA
Long_reverse_25	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGCGTTACCNNNTAGACGGGCGGTG TGTRCA
PE_1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG
PE_2	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCG
Caporaso_forward _1	AATGATACGGCGACCACCGAGATCTACACAACCAGTCTATGGTAATTGTGTGCCAGCMGCCGCGGT AA
Caporaso_forward _2	AATGATACGGCGACCACCGAGATCTACACAACGCTAATATGGTAATTGTGTGCCAGCMGCCGCGGT AA
Caporaso_forward _3	AATGATACGGCGACCACCGAGATCTACACAAGACTACTATGGTAATTGTGTGCCAGCMGCCGCGGT AA
Caporaso_forward _4	AATGATACGGCGACCACCGAGATCTACACAATCGATATATGGTAATTGTGTGCCAGCMGCCGCGGT AA
Caporaso_forward _5	AATGATACGGCGACCACCGAGATCTACACACCAATTGTATGGTAATTGTGTGCCAGCMGCCGCGGTA A
Caporaso_forward _6	AATGATACGGCGACCACCGAGATCTACACACTGAAGTTATGGTAATTGTGTGCCAGCMGCCGCGGT AA
Caporaso_forward _7	AATGATACGGCGACCACCGAGATCTACACATTGCCGCTATGGTAATTGTGTGCCAGCMGCCGCGGTA A
Caporaso_forward _8	AATGATACGGCGACCACCGAGATCTACACCAACCTTATATGGTAATTGTGTGCCAGCMGCCGCGGTA A
Caporaso_forward _9	AATGATACGGCGACCACCGAGATCTACACCCTAATAATATGGTAATTGTGTGCCAGCMGCCGCGGTA A
Caporaso_forward _10	AATGATACGGCGACCACCGAGATCTACACCCTCTGATTATGGTAATTGTGTGCCAGCMGCCGCGGTA A
Caporaso_forward _14	AATGATACGGCGACCACCGAGATCTACACGAACGGAGTATGGTAATTGTGTGCCAGCMGCCGCGGT AA
Caporaso_forward _16	AATGATACGGCGACCACCGAGATCTACACGCGTTACCTATGGTAATTGTGTGCCAGCMGCCGCGGTA A
Caporaso_forward _18	AATGATACGGCGACCACCGAGATCTACACGGATGCCATATGGTAATTGTGTGCCAGCMGCCGCGGT AA
Caporaso_forward _20	AATGATACGGCGACCACCGAGATCTACACGTTGGCCGTATGGTAATTGTGTGCCAGCMGCCGCGGTA A
Caporaso_forward _22	AATGATACGGCGACCACCGAGATCTACACTGACTGCTTATGGTAATTGTGTGCCAGCMGCCGCGGTA A

Figure 2d

Caporaso_forward_24	AATGATACGGCGACCACCGAGATCTACACTTCAGCGATATGGTAATTGTGTGCCAGCMGCCGCGGTAA
Caporaso_reverse_1	CAAGCAGAAGACGGCATACGAGATAACCAGTCAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_reverse_7	CAAGCAGAAGACGGCATACGAGATATTGCCGCAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_reverse_8	CAAGCAGAAGACGGCATACGAGATCAACCTTAAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_reverse_9	CAAGCAGAAGACGGCATACGAGATCCTAATAAAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_reverse_15	CAAGCAGAAGACGGCATACGAGATGCCTACGCAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_reverse_16	CAAGCAGAAGACGGCATACGAGATGCGTTACCAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_reverse_17	CAAGCAGAAGACGGCATACGAGATGGAGGCTGAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_reverse_23	CAAGCAGAAGACGGCATACGAGATTGGCGATTAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_reverse_24	CAAGCAGAAGACGGCATACGAGATTTCAGCGAAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_reverse_25	CAAGCAGAAGACGGCATACGAGATTTGGCTATAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Illumina_E_1	AATGATACGGCGACCACCGA
Illumina_E_2	CAAGCAGAAGACGGCATACGA
Caporaso_read_1	TATGGTAATTGTGTGCCAGCMGCCGCGGTAA
Caporaso_read_2	AGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_index_read	ATTAGAWACCCBDGTAGTCCGGCTGACTGACT

Figure 3

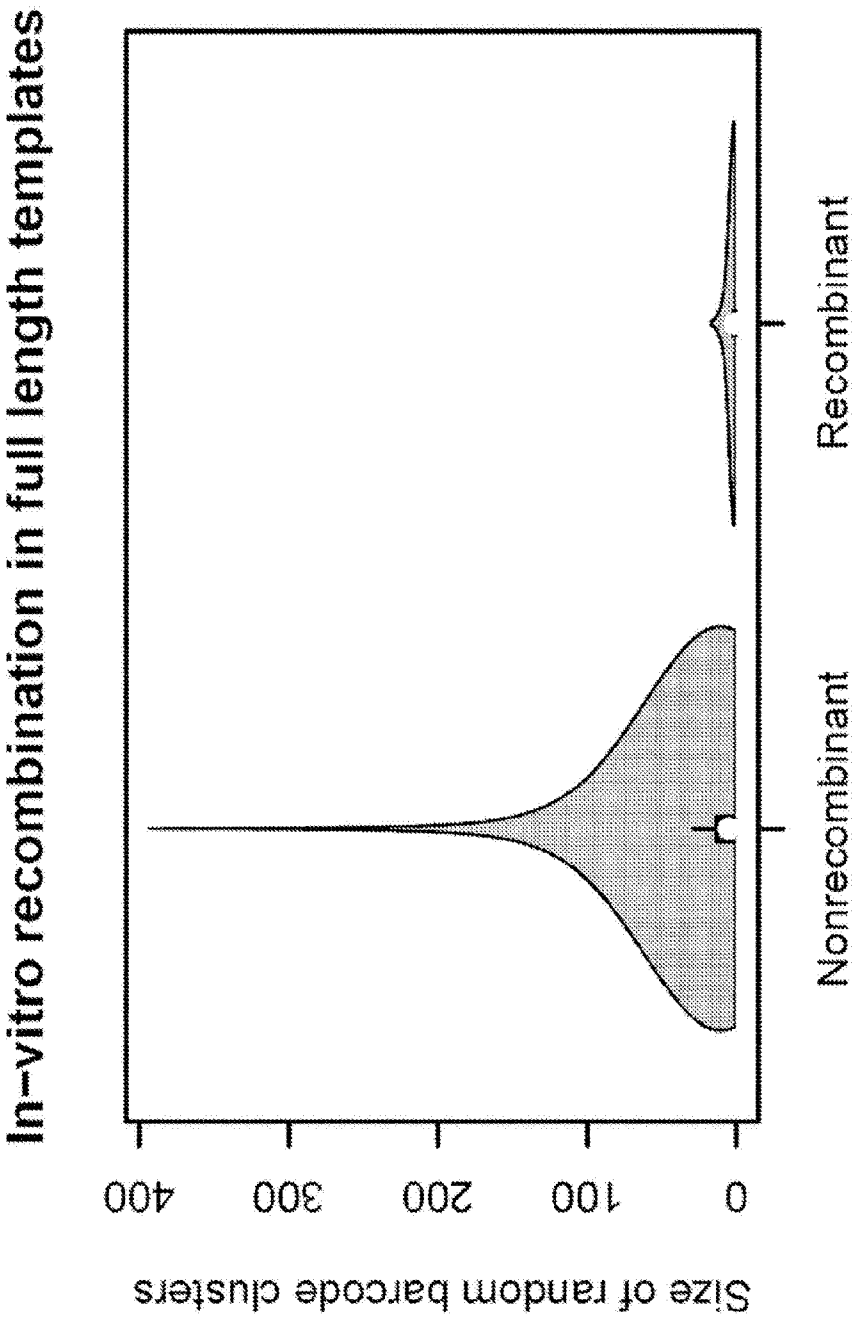

In−vitro recombination in full length templates
Recombinant
Nonrecombinant
0
100
200
300
400
Size of random barcode clusters

Figure 4

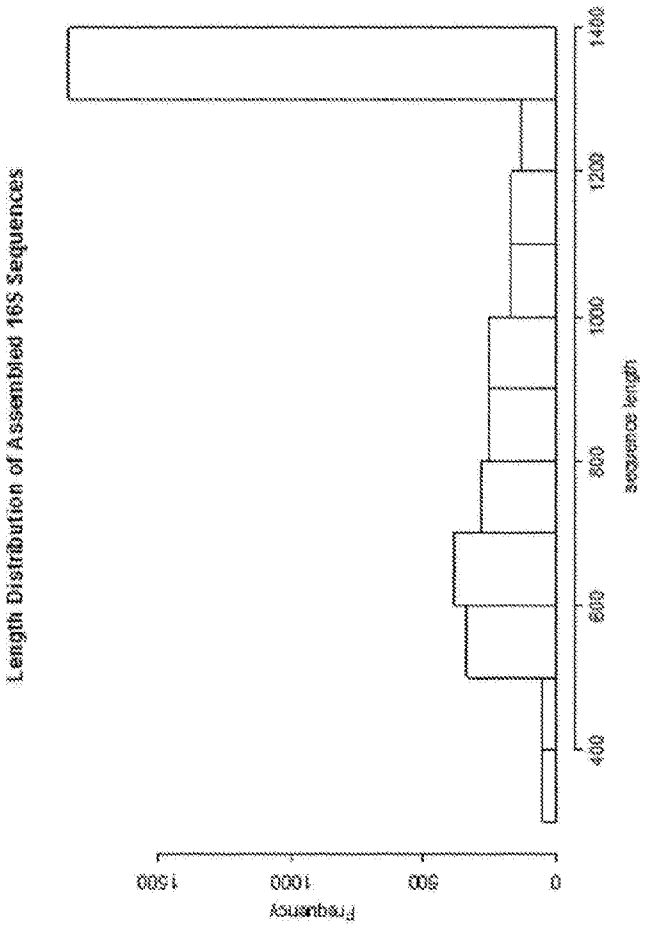

Length Distribution of Assembled 16S Sequences
sequence length
1400
1200
1000
800
600
400
1500
1000
500
0

Figure 5

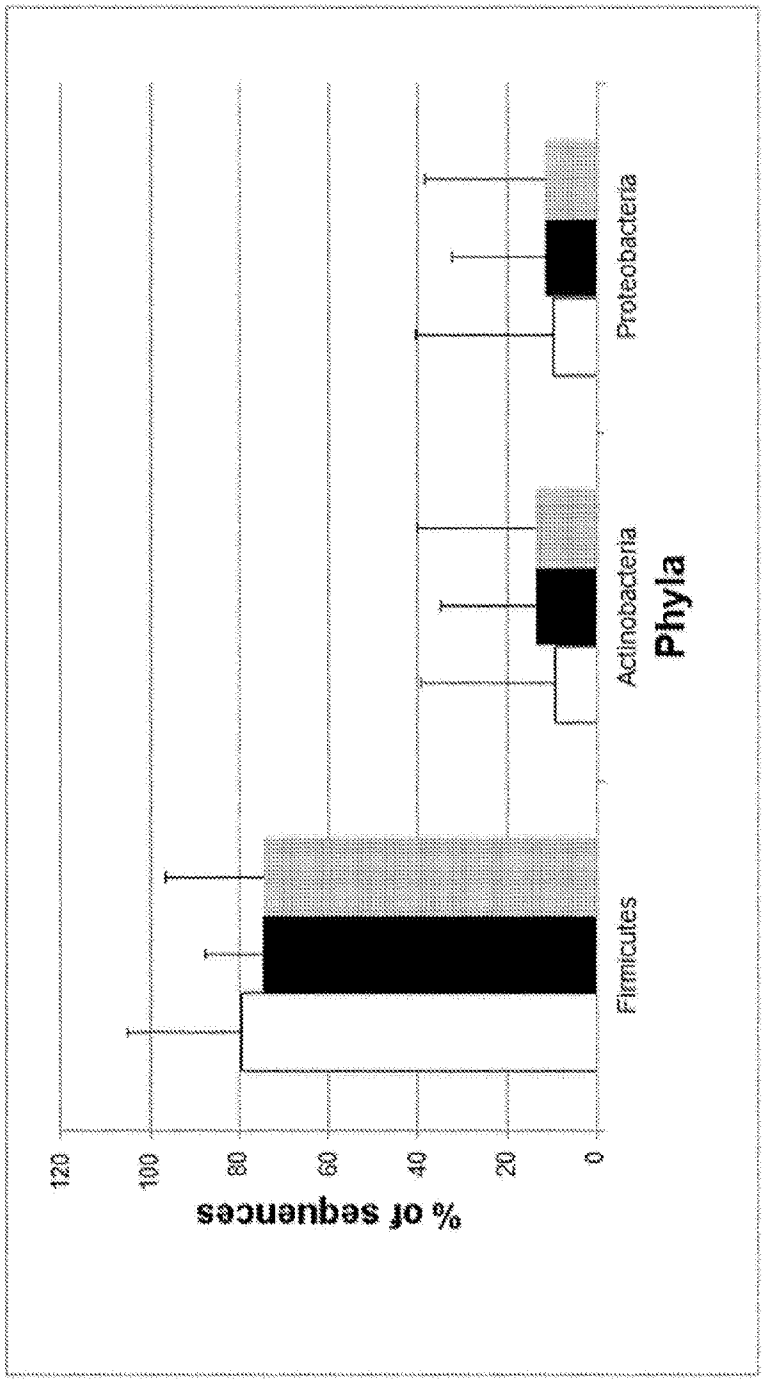

Proteobacteria
Actinobacteria
Firmicutes
Phyla
% of sequences
120
100
80
60
40
20
0

Figure 6

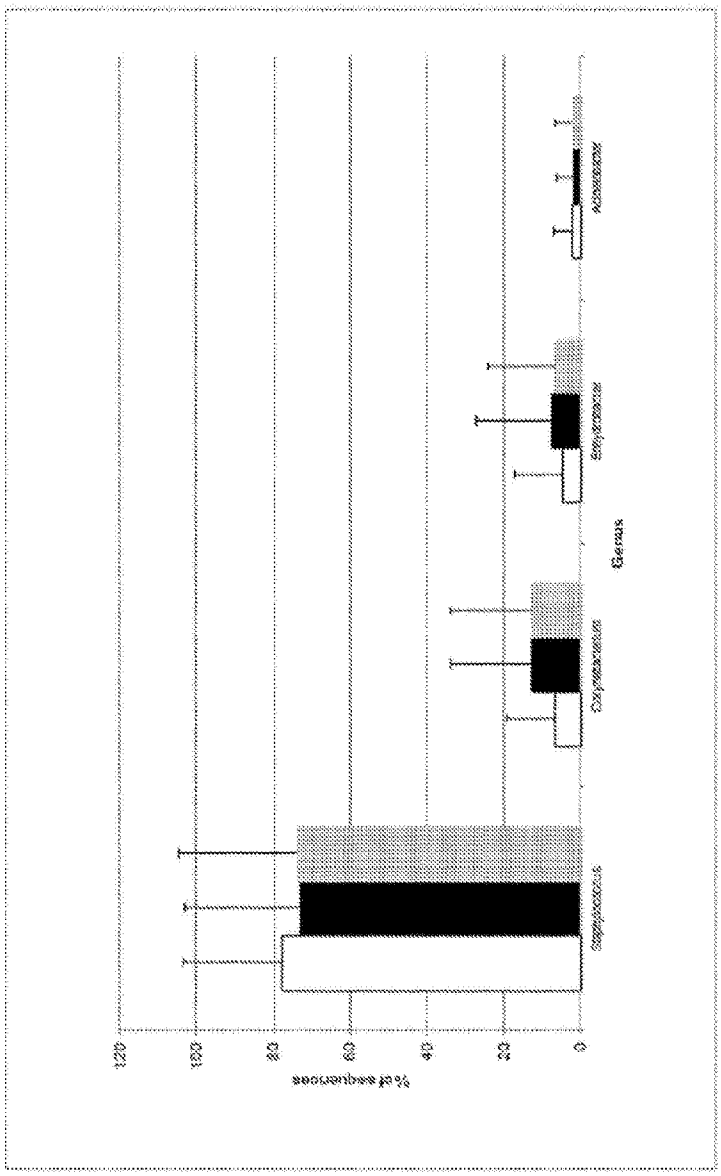


Figure 7

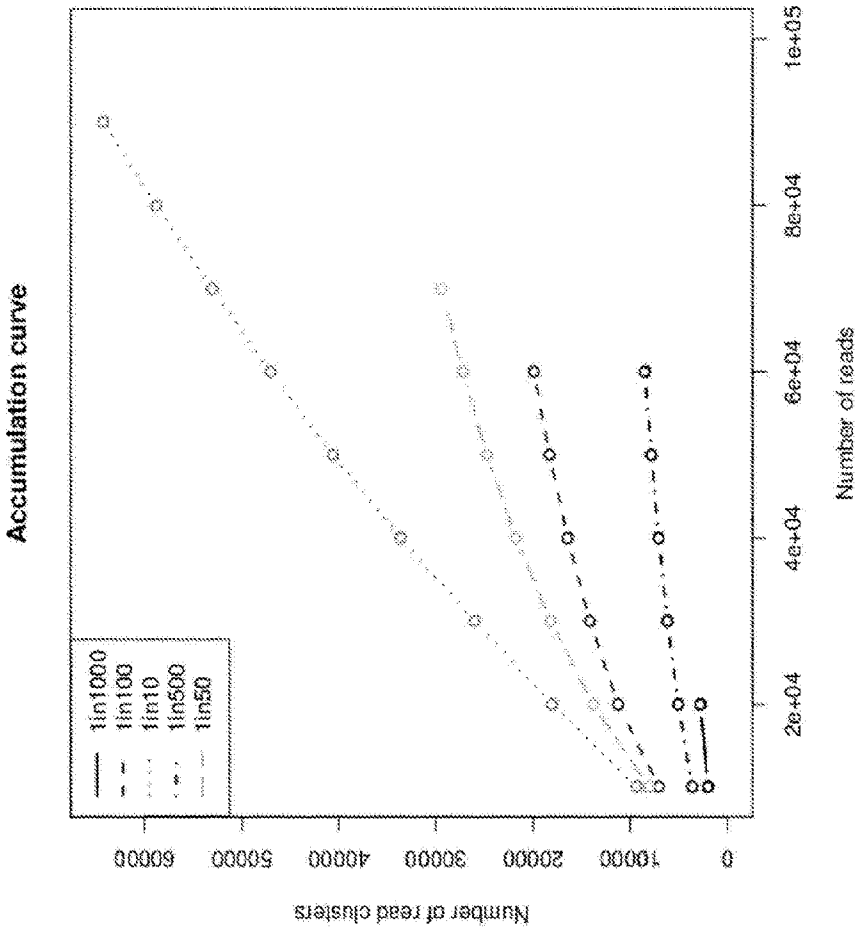

Accumulation curve
Number of read clusters
0
10000
20000
30000
40000
50000
60000
Number of reads
2e+04
4e+04
6e+04
8e+04
1e+05
1in1000
1in100
1in10
1in500
1in50

Figure 8

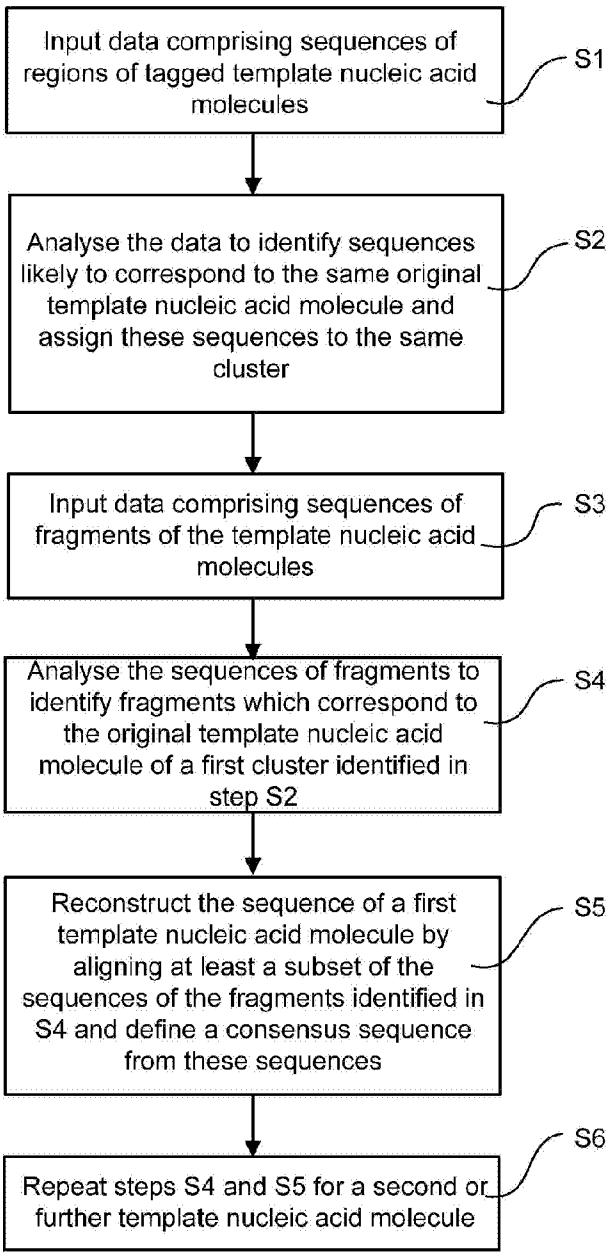

Input data comprising sequences of regions of tagged template nucleic acid molecules
S1
Analyse the data to identify sequences likely to correspond to the same original template nucleic acid molecule and assign these sequences to the same cluster
S2
Input data comprising sequences of fragments of the template nucleic acid molecules
S3
Analyse the sequences of fragments to identify fragments which correspond to the original template nucleic acid molecule of a first cluster identified in step S2
S4
Reconstruct the sequence of a first template nucleic acid molecule by aligning at least a subset of the sequences of the fragments identified in S4 and define a consensus sequence from these sequences
S5
Repeat steps S4 and S5 for a second or further template nucleic acid molecule
S6

Figure 9

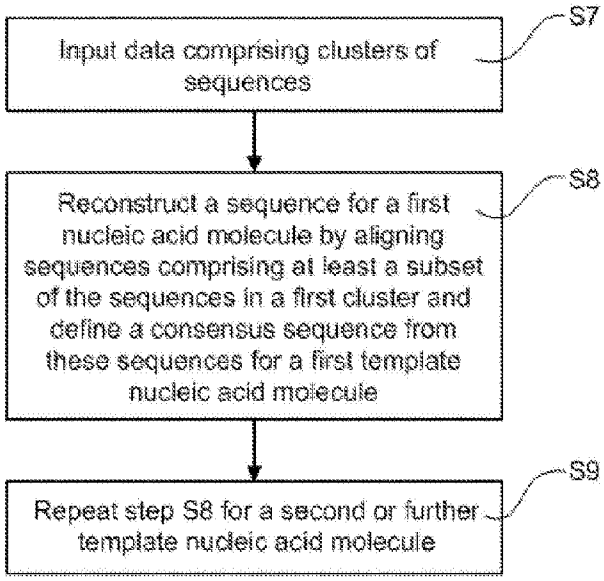

Input data comprising clusters of sequences
S7
Reconstruct a sequence for a first nucleic acid molecule by aligning sequences comprising at least a subset of the sequences in a first cluster and define a consensus sequence from these sequences for a first template nucleic acid molecule
S8
Repeat step S8 for a second or further template nucleic acid molecule
S9

Figure 10

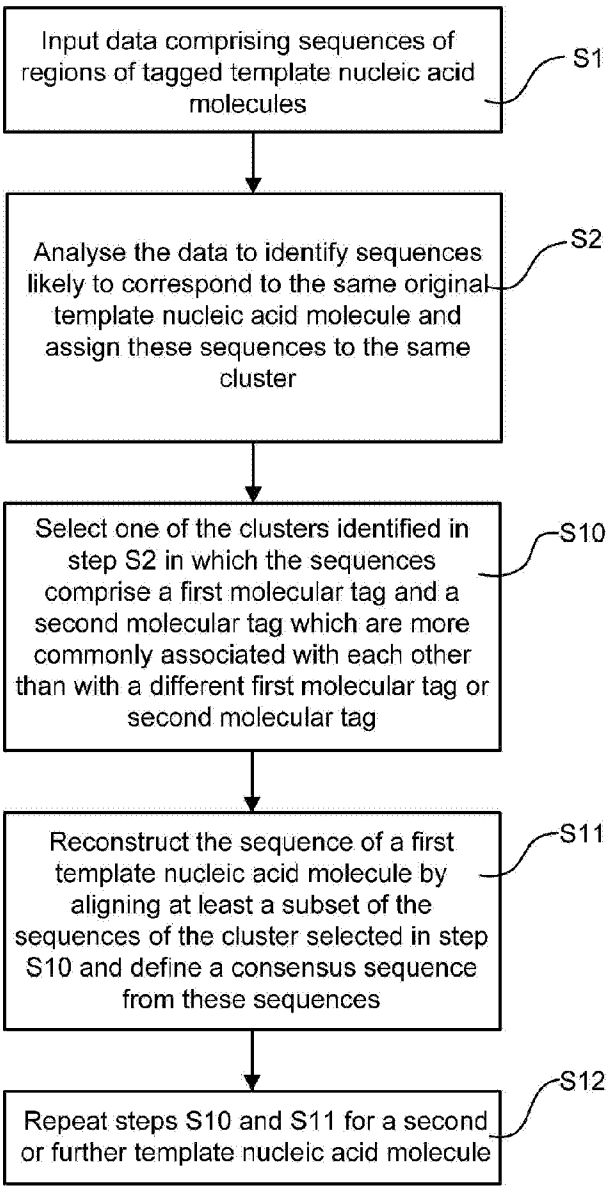
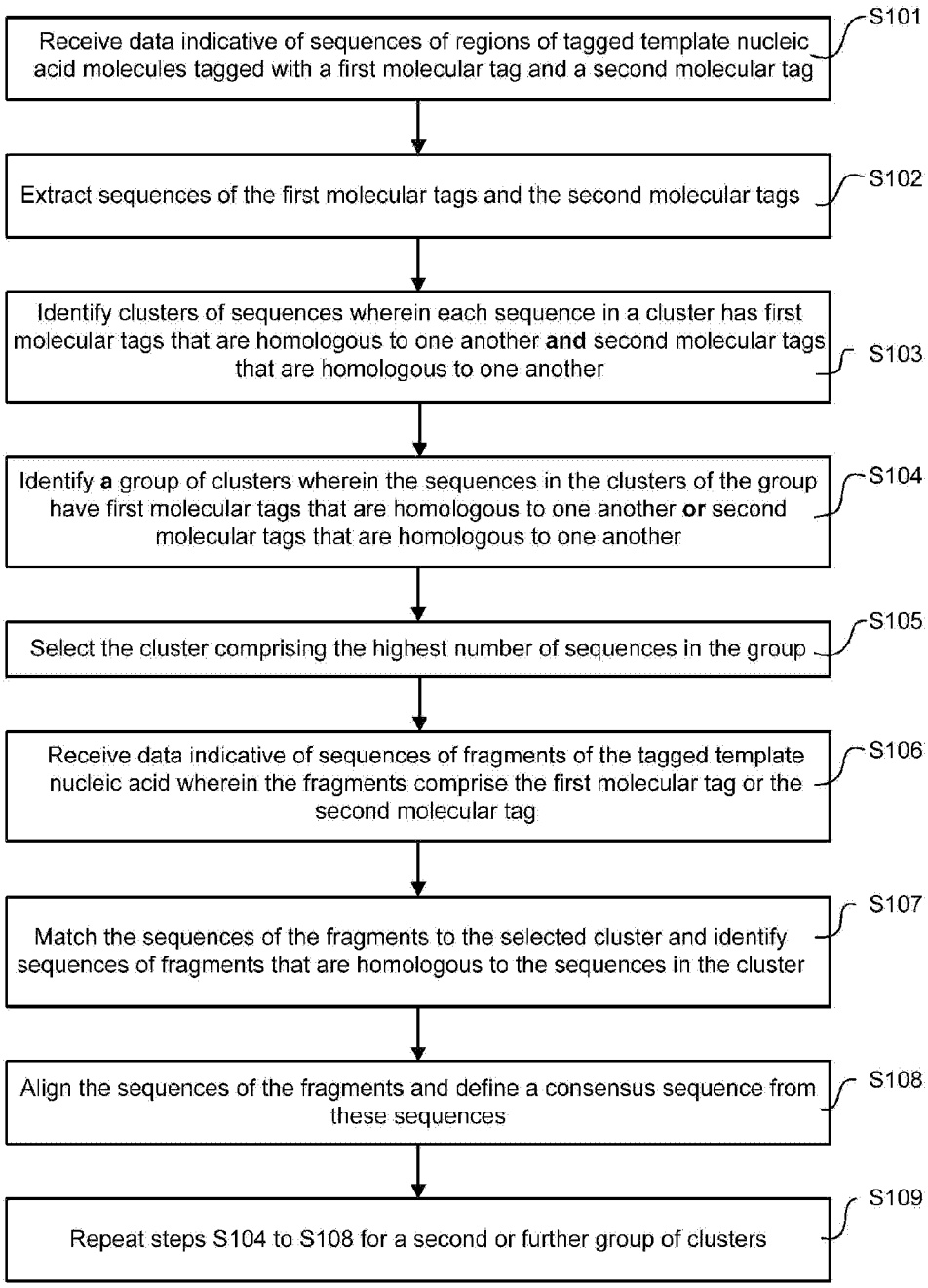

Input data comprising sequences of regions of tagged template nucleic acid molecules
S1
Analyse the data to identify sequences likely to correspond to the same original template nucleic acid molecule and assign these sequences to the same cluster
S2
Select one of the clusters identified in step S2 in which the sequences comprise a first molecular tag and a second molecular tag which are more commonly associated with each other than with a different first molecular tag or second molecular tag
S10
Reconstruct the sequence of a first template nucleic acid molecule by aligning at least a subset of the sequences of the cluster selected in step S10 and define a consensus sequence from these sequences
S11
Repeat steps S10 and S11 for a second or further template nucleic acid molecule
S12

Figure 11


Receive data indicative of sequences of regions of tagged template nucleic acid molecules tagged with a first molecular tag and a second molecular tag
S101
Extract sequences of the first molecular tags and the second molecular tags
S102
Identify clusters of sequences wherein each sequence in a cluster has first molecular tags that are homologous to one another **and** second molecular tags that are homologous to one another
S103
Identify **a** group of clusters wherein the sequences in the clusters of the group have first molecular tags that are homologous to one another **or** second molecular tags that are homologous to one another
S104
Select the cluster comprising the highest number of sequences in the group
S105
Receive data indicative of sequences of fragments of the tagged template nucleic acid wherein the fragments comprise the first molecular tag or the second molecular tag
S106
Match the sequences of the fragments to the selected cluster and identify sequences of fragments that are homologous to the sequences in the cluster
S107
Align the sequences of the fragments and define a consensus sequence from these sequences
S108
Repeat steps S104 to S108 for a second or further group of clusters
S109