



(51) International Patent Classification:
C12Q 1/68 (2006.01)

(21) International Application Number:
PCT/AU2014/050446

(22) International Filing Date:
24 December 2014 (24.12.2014)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
2014900005 2 January 2014 (02.01.2014) AU

(71) Applicant: **THE UNIVERSITY OF QUEENSLAND**
[AU/AU]; The University of Queensland, St Lucia, 4072 (AU).

(72) Inventors: **TYSON, Gene**; 7/27 Elizabeth Street, Toowong, Queensland 4066 (AU). **HUGENHOLTZ, Philip**; 35 Central Avenue, Indooroopilly, Queensland 4068 (AU). **IMELFORT, Mike**; The University of Queensland, St Lucia, Queensland 4072 (AU). **BUTLER, Margaret**; The University of Queensland, St Lucia, Queensland 4072 (AU).

(74) Agent: **DAVIES COLLISON CAVE**; Level 10, 301 Coronation Drive, Milton, Queensland 4064 (AU).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— with international search report (Art. 21(3))

(54) Title: SEQUENCING METHOD AND APPARATUS

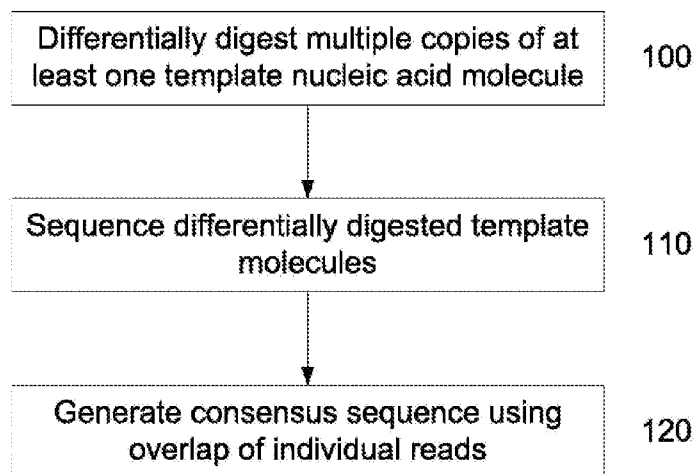


Fig. 1

(57) Abstract: A method of sequencing at least one template nucleic acid molecule, the method including differentially digesting a number of copies of the at least one template molecule to thereby generate digested template molecules, at least some of the digested template molecules having different numbers of nucleotides removed from at least one end of the template molecule, sequencing at least part of the digested template molecules to generate reads representing the sequence content of at least an end portion of the digested template molecules and generating a consensus sequence for the at least one template molecule using overlaps between the reads.

SEQUENCING METHOD AND APPARATUS

Background of the Invention

[0001] The present invention relates to a method and apparatus for sequencing a nucleic acid molecule, and in one example, to a method and apparatus for improving the sequencing read length of a sequencing device or sequencing technique.

Description of the Prior Art

[0002] The reference in this specification to any prior publication (or information derived from it), or to any matter which is known, is not, and should not be taken as an acknowledgment or admission or any form of suggestion that the prior publication (or information derived from it) or known matter forms part of the common general knowledge in the field of endeavour to which this specification relates.

[0003] Nucleic acid sequencing is the process of determining a sequence representing the order of nucleotides or bases within a nucleic acid molecule, such as DNA, RNA or the like. A number of different sequencing techniques are known, and examples of these and the associated implementing technologies are described in “*Landscape of Next-Generation Sequencing Technologies*” by Thomas P. Niedringhaus, Denitsa Milanova, Matthew B. Kerby, Michael P. Snyder, and Annelise E. Barron in Anal. Chem. 2011, 83, 4327–4341.

[0004] A limitation of most current generation sequencing machines is that of a limited read length, corresponding to the number of bases of a nucleic acid molecule that can be sequenced. For example, the Illumina HiSeq 2000 can only sequence approximately 150 bases, and whilst devices capable of sequencing greater read lengths exist, these tend to be significantly more complex, expensive and time consuming to run.

Summary of the Present Invention

[0005] In a first broad form the present invention seeks to provide a method of sequencing at least one template nucleic acid molecule, the method including:

- a) differentially digesting a number of copies of the at least one template molecule to thereby generate digested template molecules, at least some of the digested

- 2 -

template molecules having different numbers of nucleotides removed from at least one end of the template molecule;

- b) sequencing at least part of the digested template molecules to generate reads representing the sequence content of at least an end portion of the digested template molecules; and,
- c) generating a consensus sequence for the at least one template molecule using overlaps between the reads.

[0006] Typically at least some of the digested template molecules have nucleotides removed from each end and wherein the method includes sequencing at least part of either end of the digested template molecules.

[0007] Typically the method includes differentially digesting copies of the at least one template molecule using at least one nuclease enzyme.

[0008] Typically the enzyme is at least one of:

- a) Exonuclease III;
- b) Mung Bean Nuclease;
- c) S1 Nuclease; and,
- d) Bal31.

[0009] Typically the method includes differentially digesting copies of the at least one template molecule by controlling at least one of:

- a) a digestion time;
- b) a digestion rate;
- c) a digestion temperature;
- d) a nuclease enzyme concentration; and,
- e) a nuclease enzyme type.

[0010] Typically the method includes creating a plurality of collections of digested template molecules, the copies of the template molecules in a given collection being digested under similar conditions so that the digested template molecules in a collection have a similar number of nucleotides removed from at least one end, whilst digested template molecules in

- 3 -

different ones of the plurality of collections have substantially different numbers of nucleotides removed from at least one end.

[0011] Typically the method includes:

- a) amplifying the at least one template molecule to generate copies of the at least one template molecule;
- b) creating a plurality of collections, each collection including copies of the at least one template molecule; and,
- c) digesting the copies of the at least one template molecule in each collection differentially, so that each collection contains respective digested template molecules.

[0012] Typically the method includes sequencing at least part of at least one end of digested template molecules in each collection to determine at least one read for each collection.

[0013] Typically the method includes sequencing at least part of each end of the digested template molecules in each collection.

[0014] Typically the method includes:

- a) identifying overlaps between reads of different collections; and,
- b) using the reads and overlaps to generate the consensus sequence.

[0015] Typically the method includes:

- a) creating each collection by digesting copies of multiple different template molecules;
- b) sequencing at least some of the digested template molecules in each collection; and,
- c) generating a respective consensus sequence for at least some of the multiple different template molecules.

[0016] Typically the method includes creating each collection from a desired distribution of copies of multiple different template molecules.

[0017] Typically the method includes:

- a) obtaining a desired distribution of the multiple different template molecules;

- 4 -

- b) amplifying the desired distribution of multiple different template molecules to generate a number of copies of each of the multiple different template molecules; and,
- c) creating a plurality of collections, each collection including a subset of the number of copies of each of the multiple different template molecules.

[0018] Typically the method includes:

- a) adding a respective label to the digested template molecules in each collection; and,
- b) determining the consensus sequence at least partially in accordance with the respective labels.

[0019] Typically the method includes determining overlaps between the reads in different collections using the labels.

[0020] Typically each collection represents a respective tier in a sequence hierarchy, the template molecules in adjacent tiers being progressively digested.

[0021] Typically, for multiple different template molecules, the method includes:

- a) selecting reads for a next tier using the labels;
- b) determining overlapping k-mers for each read;
- c) comparing the k-mers to k-mers of reads from previous tiers to determine inter-tier overlapping reads; and,
- d) using the inter-tier overlapping reads to generate a consensus sequence for each of the multiple different template molecules.

[0022] Typically the method includes:

- a) dividing the reads from a tier into multiple streams; and,
- b) processing each stream independently.

[0023] Typically the method includes:

- a) comparing the k-mers to k-mers of previous tiers to determine inter-tier overlapping k-mers;

- 5 -

- b) using the inter-tier overlapping k-mers to determine potential inter-tier overlapping reads; and,
- c) using the potential inter-tier overlapping reads to determine the inter-tier overlapping reads.

[0024] Typically, the method includes:

- a) calculating a distribution of insert lengths of reads in each tier; and,
- b) filtering inter-tier overlapping reads using the distribution of insert lengths.

[0025] Typically, the method includes:

- a) determining subgroups of inter-tier overlapping reads; and,
- b) generating consensus sequences of at least some of the multiple different template molecules using the subgroups of inter-tier overlapping reads.

[0026] Typically the method includes, in an electronic processing device:

- a) determining read data indicative of the reads; and,
- b) using the read data to generate the consensus sequence.

[0027] Typically the method includes, in the electronic processing device, receiving the read data from a sequencing device.

[0028] Typically the method includes, in the electronic processing device:

- a) generating an indication of the consensus sequence; and,
- b) causing an indication of the consensus sequence to be displayed.

[0029] In a second broad form the present invention seeks to provide apparatus for generating a consensus sequence representing a sequence of at least one template nucleic acid molecule, the apparatus including an electronic processing device that:

- a) determines read data indicative of reads, the reads representing the sequence content of at least an end portion of digested template molecules, the digested template molecules being obtained by differentially digesting a number of copies of at least one template molecule so that at least some of the digested template molecules have different numbers of nucleotides removed from at least one end of the at least one template molecule; and,

- 6 -

- b) uses the read data to generate consensus sequence data representing the consensus sequence.

[0030] Typically the electronic processing device:

- a) determines overlapping k-mers for each read;
- b) compares the k-mers to k-mers of reads from previous tiers to determine inter-tier overlapping reads; and,
- c) uses the inter-tier overlapping reads to generate a consensus sequence for each of the multiple different template molecules.

[0031] Typically the apparatus includes multiple electronic processing devices, and wherein reads for a next tier are dividing into multiple streams and each electronic processing device processes a respective stream by:

- a) determining overlapping k-mers for each read in the stream; and,
- b) comparing the k-mers to k-mers of reads from previous tiers to determine potential inter-tier overlapping reads.

[0032] In a third broad form the present invention seeks to provide a method of sequencing at least one template nucleic acid molecule, the method including:

- a) creating a plurality of collections, each collection including a desired distribution of copies of multiple different template molecules;
- b) differentially digesting the copies of the multiple different template molecules in each collection to thereby generate digested template molecules, at least some of the digested template molecules having different numbers of nucleotides removed from at least one end of the multiple template molecules;
- c) sequencing at least part of the digested template molecules to generate reads representing the sequence content of at least an end portion of the digested template molecules; and,
- d) generating a consensus sequence for at least some of the multiple different template molecules.

[0033] In a fourth broad form the present invention seeks to provide apparatus for generating a consensus sequence representing a sequence of at least some of multiple different template nucleic acid molecules, the apparatus including an electronic processing device that:

- a) determines read data indicative of reads, the reads representing the sequence content of at least an end portion of digested template molecules, the digested template molecules being obtained by differentially digesting a number of copies of multiple template molecules so that at least some of the digested template molecules have different numbers of nucleotides removed from at least one end of the multiple template molecules; and,
- b) uses the read data to generate consensus sequence data representing a consensus sequence for at least some of the multiple different template molecules.

Brief Description of the Drawings

[0034] An example of the present invention will now be described with reference to the accompanying drawings, in which: -

[0035] Figure 1 is a flow chart of an example of a method for sequencing a template nucleic acid molecule;

[0036] Figures 2A to 2E are schematic diagrams of example nucleic acid sequences for the method of Figure 1;

[0037] Figure 3 is a flow chart of a second example of a method for sequencing a template nucleic acid molecule;

[0038] Figures 4A to 4C are a flow chart of a specific example of a method for sequencing multiple template DNA fragments;

[0039] Figure 4D shows graphs of examples of the sequence lengths of the digested template molecules;

[0040] Figures 5A and 5B are a flow chart of a specific example of a method for using an overlap graph to construct a consensus sequence for the method of Figures 4A and 4B;

[0041] Figure 5C is a schematic diagram illustrating the comparison of k-mers in the process of Figures 5A and 5B;

[0042] Figure 5D is a schematic diagram illustrating the calculation of a tier insert length in the process of Figures 5A and 5B;

[0043] Figure 5E is a schematic diagram illustrating the determination of paired inter-tier overlaps in the process of Figures 5A and 5B; and,

[0044] Figure 6 is a schematic diagram of an example of apparatus for use in a method sequencing a nucleic acid molecule.

Detailed Description of the Preferred Embodiments

[0045] For the purpose of illustration in the following examples, reference is made to the following terms. The term "nucleic acid molecule" or "polynucleotide" as used herein includes, but is not limited to mRNA, RNA, cRNA, cDNA or DNA, or the like. The term typically refers to a polymeric form of nucleotides of at least 10 bases in length, either ribonucleotides or deoxynucleotides or a modified form of either type of nucleotide. The term includes single and double stranded forms of DNA or RNA.

[0046] A "template" or "template molecule" refers to a nucleic acid molecule and/or fragment of a nucleic acid molecule whose sequence is of interest and is to be determined. The term is intended to cover nucleic acid molecules with or without added adaptors or labels, which may be required for sequencing, amplification or other process steps.

[0047] The term "digested template molecule" is intended to cover a template molecule that has undergone a digestion step to remove nucleotides therefrom. In this regard, the term is also intended to cover template molecules that have undergone null digestion and remain intact, as will be apparent from the following description.

[0048] A "collection" refers to a group of nucleic acid molecules corresponding to one or more template molecules, which undergo a common digestion process (including no digestion). A "tier" refers to a group or collection of nucleic acid molecules forming part of a hierarchy in which adjacent tiers have undergone progressive (increasing or decreasing) amounts of digestion.

[0049] A "read" is the output from a sequencing machine representing a nucleotide sequence of a sequenced molecule. "Insert size" or "insert length" refers to a length of a nucleic acid molecule that has been sequenced and "consensus sequence" is a sequence determined for a respective template molecule.

[0050] The use of the above terms is for the purpose of illustration only and is not intended to be limiting.

[0051] An example of a process for sequencing a nucleic acid molecule will now be described reference to Figures 1 and 2A to 2E.

[0052] In this example, at step 100, the method includes differentially digesting a number of copies of at least one template molecule to thereby generate digested template molecules, at least some of the digested template molecules having different numbers of nucleotides removed from at least one end of the template molecule. This can be achieved in any suitable manner, such as digesting template molecules using a nuclease enzyme or the like, as will be described in more detail below.

[0053] This step is used to create digested template molecules having different lengths representing overlapping portions of the original nucleic acid molecule. In one example, the digested template molecules of different lengths represent respective tiers in a hierarchy, with sequences obtained from different tiers being used in establishing a consensus sequence, as will be described in more detail below.

[0054] By way of an example, three copies of a template nucleic acid molecule 201, 202, 203 are shown in Figure 2A, with examples of corresponding differentially digested template molecules being shown in Figure 2B. In this example, the template molecule 201 undergoes no digestion, whilst four and eight nucleotides are digested from each end of the template molecules 202, 203, respectively. The removal of four and eight nucleotides is for the purpose of illustration only and is not intended to be limiting. In practice the number of nucleotides that are removed will depend on a range of factors, such as the length of the nucleic acid molecule, the read length of the sequencing device used to perform the sequencing, the number of collections or tiers, or the like, as will be described in more detail below. Typically the number of nucleotides digested will be in the region of tens or hundreds depending on the preferred implementation. Furthermore, although digestion of both ends of the molecules is shown, this is not essential and alternatively digestion can be performed on the 3' or 5' end only, and also can be performed asymmetrically, depending on the manner in which digestion is performed.

[0055] At step 110, the method includes sequencing the digested template molecules to generate reads representing the sequence content of at least an end portion of the digested template molecules. This may be achieved in any suitable manner and typically involves using existing sequencing techniques optionally implemented using existing sequencing devices, to thereby read the ends of the differentially digested template molecules.

[0056] Examples of the generated reads are shown in Figure 2C, with the digested template molecules being read from each end to generate reads shown in bold lettering at 201.1, 201.2, 202.1, 202.2, 203.1, 203.2. Thus, it will be appreciated that typically a plurality of digested template molecules for each tier in the hierarchy are sequenced, with different digested template molecules being sequenced from different ends, as will be appreciated by persons skilled in the art.

[0057] It will further be appreciated that the number of nucleotides that can be sequenced will depend on the read length of the sequencing device. Accordingly, whilst eight nucleotides from the end of each digested nucleic acid molecule are shown as being sequenced, this is for illustration only, and in practice the number of nucleotides would more typically be in the region of tens or hundreds of bases, depending on the nature of the sequencing device and/or sequencing technique used.

[0058] At step 120, the method includes generating a consensus sequence of the template molecule using overlaps between the reads. In this regard, it will be appreciated that the reads 201.1, 201.2, 202.1, 202.2, 203.1, 203.2 can be aligned as shown in Figure 2D, to thereby generate the consensus sequence 204 shown in Figure 2E. This is typically achieved by performing a mathematical analysis of the reads obtained from the sequencing device at step 110, for example by using a modified overlap-layout-consensus assembly algorithm, or other suitable technique.

[0059] Accordingly, the above described process describes an approach in which a number of copies of a nucleic acid molecule are differentially digested to create digested template molecules having different lengths, and representing different overlapping portions of a template molecule to be sequenced. In particular, differentially digesting ends of the copies of the digested template molecules allows ends of the digested template molecules to be

- 11 -

sequenced to generate reads that correspond to a variety of different locations within the original template molecule. These reads can be assembled into a consensus sequence so that relatively short length reads can be used to assemble a consensus sequence of the template molecule, even if this has a length that is significantly greater than the read length of the sequencing device.

[0060] Additionally, the above described process can be implemented relatively easily, for example through appropriate sample preparation involving digestion of copies of a template molecule, which can be performed by relatively minor modification of the sample preparation protocol for a sequencing device. Post-process of read data generated by the sequencing device can then be performed, for example using a suitable software application executed by a remote computer system, or by firmware forming part of the sequencing device. It will therefore be appreciated that this process can be used with existing sequencing equipment, without equipment modification, simply by performing modified pre-sequencing sample preparation and post-sequencing read data analysis. This can therefore be used to improve the effective read length of existing sequencing techniques and processes.

[0061] Whilst the above described technique is described with respect to at least one template molecule, it will be appreciated that in practice the technique is typically performed on multiple different template molecules simultaneously. Thus, for example, the process can be performed for a sample containing multiple different nucleic acid molecules, with the process establishing a respective consensus sequence for each of the different template molecules. Thus, in one specific example, a DNA molecule or the like can be fragmented into multiple template molecules, each of which can be sequenced simultaneously using the above described technique. The process can enhance the read length and quality of the sequence obtained for each of the template molecules, in turn making accurate reconstruction of the sequence of longer DNA molecule more feasible.

[0062] The number of different template molecules that can be analysed simultaneously will depend on a range of factors, such as the length of the template molecules, the read length of the sequencing device used, the number of nucleotides digested from the template molecules, the quality of the resulting sequences, or the like. In one example, the different template

- 12 -

molecules can include several million different templates that can be sequenced simultaneously, as will be described in more detail below.

[0063] A number of further features will now be described.

[0064] In one example, at least some of the digested template molecules have nucleotides removed from each end and wherein the method includes sequencing at least part of either end of the digested template molecules. Thus, it will be appreciated that this approach can allow reads taken from either end of the digested template molecules to be used in generating the consensus sequence.

[0065] In general, the method involves differentially digesting a number of copies of the nucleic acid molecule using at least one nuclease enzyme, such as one or more of Exonuclease III, Mung Bean Nuclease, S1 Nuclease Bal31, Exonuclease T. In one specific example, Exonuclease III and Mung Bean Nuclease are used in sequence as will be described in more detail below. Although any other suitable techniques could be used, the use of a nuclease enzyme is particularly beneficial as this digests ends of the template molecules progressively, allowing the digestion process to be highly controlled. For example, the number of nucleotides removed from each template molecule can be controlled by controlling a digestion rate, for example by controlling at least one of a digestion time, a digestion temperature, or ratio of the nuclease enzyme to template copies and a nuclease enzyme type. However, any suitable technique for controllably removing nucleotides from ends of the copies of the templates can be used, such as using restriction enzymes, cleaving, shearing or the like. Accordingly, it will be appreciated that use of the term digestion is not intended to be limiting, but rather is intended to cover any mechanism for removing nucleotides in a controlled manner.

[0066] The method may further include creating collections, such as "libraries" of digested template molecules, the copies of the template molecules in a given collection being digested under similar conditions so that the digested template molecules in a collection have a similar number of nucleotides removed from at least one end, whilst digested template molecules in different ones of the plurality of collections have substantially different numbers of nucleotides removed from at least one end. In this regard, it will be appreciated that not

every molecule undergoing a common digestion process will necessarily have the same number of nucleotides removed from each end, simply due to inherent variations in the digestion process. Nevertheless, such variations can be accounted for during subsequent construction of the consensus sequence, so this process still allows reads obtained from digested nucleic acids in different collections to be used to assemble the consensus sequence.

[0067] Thus, in one example, the sequencing method uses a plurality of collections containing copies of one or more different template molecules to be sequenced. In this example, the method typically includes amplifying the at least one template molecule to generate the copies of the at least one template molecule, which are then used in creating a plurality of collections, each collection including copies of the at least one template molecule, and digesting the copies of the at least one template molecule in each collection differentially, so that each collection contains respective digested template molecules.

[0068] It will be appreciated that creation of collections or libraries is typically performed as part of the sample preparation protocol for nucleic acid sequencing, and this may also include additional steps such as repairing ends of the digested template molecules, adding adaptors to ends of the digested template molecules, or the like, as known in the art.

[0069] In one example, the method includes sequencing at least part of at least one end, and more typically both ends, of the digested template molecules in each collection to determine at least one read for each collection. Once completed, overlaps between reads of different collections can be identified, with the reads and overlaps being used to generate the consensus sequence.

[0070] As mentioned above, whilst this can be performed for a single template molecule, more typically this is performed on multiple different template molecules. In this case, the method involves creating each collection by digesting copies of multiple different template molecules, sequencing at least some of the digested template molecules in each collection and generating a respective consensus sequence for at least some of the multiple different template molecules. Thus, this allows a respective consensus sequence to be generated for at least some of, and optionally all of, the multiple different template molecules.

[0071] In general, when sequencing multiple different template molecules, the method includes creating each collection from a desired distribution of copies of the multiple different template molecules. In this regard, the term desired distribution is intended to cover having a desired number of copies of each of the multiple different template molecules, in each collection. The desired number is typically the same number for each of the different template molecules and each of the different collections. Additionally, the desired number is selected based on aspects of the sequencing process to ensure that reads are obtained for each of the different template molecules, in each of the different collections.

[0072] In this regard, in an ideal scenario in which every molecule in a sample is perfectly sequenced, then each collection could contain one copy of each respective different template molecule. So for t different collections, t copies of each different template molecule could be prepared, with one copy being placed in each collection. However, in practice, this is not a feasible approach for a number of reasons.

[0073] For example, it is not generally practical to manipulate individual molecules, so instead copies of each different template molecule are prepared in solution and then split into different collections by splitting or sampling the solution. Consequently, the solution must contain a sufficiently large number of copies of each different template, to ensure that when the solution is divided into different collections, each collection contains copies of each different template.

[0074] A further issue is that sequencing platforms typically have a throughput limit defined by $N \times L$, where N is the number of reads on any given run and L is the length of the reads. In order to ensure that the necessary chemical processes for initial library preparation work, the libraries must contain a large number of template molecules, far in excess of the number of reads N , typically by several orders of magnitude. This means that a large number of the templates molecules in each library are not sequenced. As the libraries contain a large number of copies F of each different template molecule ($F \gg N$), and as the template molecules that are ultimately sequenced are selected randomly, it is possible that the sequenced template molecules do not include any of some of the template molecules.

[0075] Consequently, it is important to ensure that the number of copies F of each template molecule is significantly smaller than the number of reads N ($F \ll N$), to ensure reads are obtained across all of the multiple different template molecules. As a result, it is important to ensure that each collection includes a relatively small number of copies F of each of the different template molecules, to ensure that the sequencing process is not randomly saturated by copies of only some of the multiple different template molecules.

[0076] However, if there are too few copies F of each template molecule, this can also result in a reduction in the likelihood of the template molecule being sequenced successfully, as the overall throughput will be lower than required to establish a consensus sequence.

[0077] Accordingly, if the number of copies F of each different template molecule in each collection is too small, there will not be sufficient reads to establish a consensus sequence for that template molecule, whereas if the number of copies F of each template molecule is too high, there is a risk that some template molecules will saturate the reads, so again some template molecules will not be sequenced.

[0078] Accordingly, in the above described process, a dilution step can be performed once a library has been prepared for amplifying the template molecules to generate the copies, to thereby ensure a desired a concentration of each template molecule in the library. Once amplification has been completed, obtaining a specific sample volume of solution containing the copies of each of the template molecules can be used to ensure each collection contains a desired number of copies F of each different template molecule.

[0079] The process of diluting the library prior to amplification is known as bottlenecking and is performed based on a number of factors, including, but not limited to, one or more of:

- the number of tiers t ;
- the read length L of the sequencing machine;
- the length of the assembled consensus sequence;
- the number of PCR cycles used in the various stages of creating each collection; and,
- the type of sample preparation performed.

[0080] Accordingly, the bottlenecking process is used to prevent uneven distributions of template molecules, which could result in certain template molecules being over or under

- 16 -

represented in the collections. For example, if too many copies of certain template molecules exist, these can saturate the sequencing process, so sequences for other ones of the template molecules cannot be accurately determined.

[0081] In one particular example, the method includes obtaining a desired distribution of the multiple different template molecules by amplifying the desired distribution of multiple different template molecules to generate a number of copies of each of the multiple different template molecules and creating a plurality of collections, each collection including a subset of the number of copies of each of the multiple different template molecules. However, other techniques could be used.

[0082] To facilitate the process of generating the consensus sequence, the method can include adding a respective label to the digested template molecules in each collection and determining the consensus sequence at least partially in accordance with the respective labels. In this instance, the method can include determining overlaps between the reads in different collections using the labels, for example by identifying a collection associated with each read using the labels. However, it will be appreciated that this is only required if the digested template molecules in each collection are sequenced collectively, and in the event that the molecules from different collections are sequenced separately, for example by using a respective sequencing run for each collection, then this step may not be required.

[0083] Thus, the process can involve determining reads for each collection using the labels, generating a collection sequence for each collection using the reads for the respective collection and generating the consensus sequence using the collection sequences. However, the use of labels is not required in the event that the different collections are sequenced in different sequencing runs, as mentioned above.

[0084] Irrespective or not of whether labels are used, the method can involve comparing reads obtained from digested template molecules within a given collection and using these to determine a collection sequence, indicative of idealized sequence for the ends of digested template molecules within the collection, with the collection sequences of different collections being compared to determine the consensus sequence.

[0085] Alternatively, reads for the different libraries can be used collectively to generate the consensus sequence directly. The method may include sequencing a fragment of at least one end of the digested nucleic acid molecule in each collection to determine at least one read for each collection. The method may then further include identifying overlaps between reads of different collections and using the reads and overlaps to generate the consensus sequence.

[0086] In one example, the consensus sequence is obtained by generating a hash indicative of an overlap between and sequence content for each read, using the hash to construct an overlap graph, and constructing a consensus sequence using the overlap graph, as will be described in more detail below.

[0087] In one example, each collection represents a respective tier in a sequence hierarchy, the template molecules in adjacent tiers being progressively digested. In this example, when multiple different template molecules are being analysed the method can include selecting reads for a next tier using the labels, determining overlapping k-mers for each read, comparing the k-mers to k-mers of reads from previous tiers to determine inter-tier overlapping reads and then using the inter-tier overlapping reads to generate a consensus sequence for each of the multiple different template molecules. In one example, this involves using the inter-tier overlapping reads to generate an overlap graph and using the overlap graph to construct the consensus sequence.

[0088] In one example, this process can be at least partially parallelised by dividing the reads from a tier into multiple streams and processing each stream independently, as will be described in more detail below.

[0089] The method of identifying the inter-tier overlapping reads generally involves comparing the k-mers from reads in one tier to k-mers of previous tiers to determine inter-tier overlapping k-mers, using the inter-tier overlapping k-mers to determine potential inter-tier overlapping reads and using the potential inter-tier overlapping reads to determine the inter-tier overlapping reads.

[0090] The potential inter-tier overlapping reads are typically filtered to identify the inter-tier overlapping reads. Such filtering may be achieved in any manner, but typically involves

calculating a distribution of insert lengths of reads in each tier and filtering inter-tier overlapping reads using the distribution of insert lengths.

[0091] Once inter-tier overlapping reads have been determined, these can be divided into subgroups, each one corresponding to one of the multiple different template molecules, which are then used in generating consensus sequences of at least some of the multiple different template molecules.

[0092] The method is typically performed at least in part using an electronic processing device, such as a processing system or the like, which operates to determine read data indicative of the reads and use the read data to generate the consensus sequence. It will be appreciated that this could be performed by a stand alone or networked computer system coupled to or forming part of a sequencing device. Thus, this could be performed by a computer system coupled to an existing sequencing device, for example via a computer network, or by custom firmware or software installed on an existing sequencing device.

[0093] In either case, the electronic processing device is typically adapted to receive read data, performing an analysis of the read data and generate an indication of the consensus sequence, which may then optionally be displayed or otherwise provided to an operator. Thus, it will be appreciated that any suitable electronic processing device could be used, as will be described in more detail below.

[0094] In one example, the apparatus includes multiple electronic processing devices, and wherein the method includes dividing reads for a next tier into multiple streams with each electronic processing device processing a respective stream by determining overlapping k-mers for each read in the stream and comparing the k-mers to k-mers of reads from previous tiers to determine potential inter-tier overlapping reads. Once this has been completed information from each of the streams can be consolidated to enable inter-tier overlapping reads to be determined.

[0095] It will be appreciated that the length of nucleic acid molecules that can be sequenced using the above described techniques will vary depending on the preferred implementation and for example will depend on factors such as the read length of the sequencing apparatus, the number of collections (ie the number of tiers), the number of nucleotides removed from

each molecule, or the like. In one example, the process can be used to easily allow sequencing of molecules having a length of at least two to three times the read length of the sequencing device, even when only using minimal number of tiers, such as three tiers. However, it will be appreciated that increasing the number of tiers allows a corresponding increase in the length that can be sequenced.

[0096] It will also be appreciated that the number of nucleotides digested, and hence the degree of overlap between the reads in adjacent tiers can also be adjusted. This will typically be set based on the read length of the sequencing device and must be at least less than the read length, to ensure overlap between the reads of adjacent tiers. In one example, the number of nucleotides digested is set to be less than half of the read length of the sequencing device. Thus ensures a degree of overlap between non-adjacent tiers, which in turn can assist with error correction, as well as ensuring the process can be continued even if the sequence for any one tier cannot be successfully read for any reason.

[0097] In one specific example, when used with an Illumina HiSeq 2000 and using three tiers, the method can be used to sequence a nucleic acid molecule having at least 250 bases, and more typically a double stranded nucleic acid molecule having up to 1200 base pairs, thereby representing a significant increase in the effective read length of the device. It will also be appreciated that this can be further increased by using further tiers. Furthermore, this technique could easily be applied to multiple different template nucleic acid molecules simultaneously, such as over two and a half million templates, depending on the preferred implementation. Thus, it will be appreciated that this can be used to significantly increase the effective read length of existing sequencing devices without requiring modification of the device itself.

[0098] An example of a method for sequencing nucleic acid molecules, and in particular to sequencing a combination of multiple different template nucleic acid molecules will now be described in more detail with reference to Figure 3.

[0099] In this example, at step 300 the multiple template molecules are amplified utilising an appropriate amplification technique, such as PCR (Polymerase Chain Reaction), or the like.

- 20 -

This is performed so as to generate copies of each of the multiple template molecules, in turn allowing these to be differentially digested.

[0100] At step 310, a plurality of collections is created, each collection containing copies of each of the multiple template molecules in solution. This may be achieved in any suitable manner and will typically involve forming a solution containing the copies of the multiple template molecules and then dividing this between multiple sample preparation wells.

[0101] At step 320 the template molecules in each collection are digested under different conditions. Usually, this is achieved by adding a specific amount and concentration of a nuclease enzyme, and in particular an exonuclease enzyme, to the wells, and then controlling a digestion time for each well, for example by inhibiting the digestion after a respective time period.

[0102] Different digestion durations can be achieved by adding the enzyme to different wells at different times, and then inhibiting digestion simultaneously across all wells, for example by adding a nuclease enzyme inhibitor, or heating the well to deactivate the enzyme. Alternatively, this can be achieved by simultaneously adding the nuclease enzyme and then selectively inhibiting the digestion after a set time for each well, and this will largely depend on the preferred sample preparation protocol.

[0103] Alternatively differential digestion could be achieved by altering different parameters for each well, such as the enzyme concentration, type of enzyme used, a temperature or the like. The use of different digestion times is not therefore intended to be limiting, although it will be noted that controlling digestion times is generally preferred as this is typically more straightforward than controlling other parameters. Additionally, at least one of the wells is typically not exposed to the enzyme so the nucleic acid molecules therein remain undigested.

[0104] Differentially digesting the template molecules in each collection results in digested template molecules of different lengths in the different collections so that each collection can define a respective tier in the hierarchy used to establish the consensus sequence, as will be described in more detail below. It will also be appreciated that each tier typically contains digested template molecules corresponding to each of the multiple template molecules.

[0105] Following optional standard library preparation techniques, such as addition of adaptors or the like to the template molecules, at step 330 ends of the template molecules in each collection are sequenced utilising an existing sequencing device, such as the Illumina Hi-Seq 2000 sequencing device, or the like.

[0106] Read data indicative of the sequence content of each the sequences read by the sequencing device are then provided to the electronic processing device, which compares the reads at step 340, using the comparison to establish overlaps between reads from different collections, and hence tiers, at step 350. These overlaps, together with the reads, can be used to identify reads from the same template molecule, allowing the sequences from different tiers to be grouped and to generate a consensus sequence for each template molecule at step 360. This can be achieved in any suitable manner, using a modified overlap-layout-consensus graph approach, or the like, and an example will be described in more detail below.

[0107] An example process for sequencing multiple DNA fragments will now be described in more detail with reference to Figures 4A to 4D and 5A to 5E.

[0108] In this example, at step 400 multiple double stranded DNA fragments, acting as multiple templates, are acquired, for example from a biological sample or the like. In one example, this can involve fragmenting a larger piece of DNA using known techniques, as will be understood by those skilled in the art. At step 405 a library is created containing the multiple template DNA fragments, with adaptors being added to the DNA fragments as part of this process. The adaptors allow PCR amplification and sequencing to be performed, and may also act as a unique label for undigested fragments (corresponding to tier "0"), as will be described in more detail below.

[0109] Prior to performing PCR, bottlenecking is performed at step 410 by diluting the library to a desired concentration so as to obtain a desired distribution of DNA fragments. In this regard, a desired distribution corresponds to a given number of each of the DNA fragments. This is performed so that each of the DNA fragments is evenly represented in the collections which are sequenced as described below, thereby ensuring equal representation in

the sequencing results, thereby helping to ensure accurate identification of consensus sequences.

[0110] A sample is then obtained from the library, with PCR being performed on the sample at step 415 to create copies of the multiple DNA fragments. It will be appreciated that this will utilise standard techniques and this will not therefore be described in any further detail.

[0111] At step 420 t wells are prepared, each well representing a respective tier or collection and each containing copies of the multiple different DNA fragments in solution. At step 425 Exonuclease III is added to $t-1$ wells to digest the DNA fragments therein. In this regard, the double stranded DNA fragments are incubated with Exonuclease III, which digests one strand (from the 3' end) of the double strand DNA, resulting in a single stranded, 5' 'tail' on the double strand DNA fragments. Each of the $t-1$ wells undergo digestion for respective different amounts of time at step 430, with the times being selected to differentially digest the DNA fragments, and in particular to remove a respective number of nucleotides from each DNA fragment in the well. The time selected will depend on a number of factors such as the digestion temperature, concentration of the DNA fragments in the well, concentration of the enzyme, number of nucleotides to be removed, or the like. This is therefore typically performed in accordance with a set protocol which defines specific values for these parameters so that the operator preparing the examples need only follow the set instructions and add predefined concentrations of the nuclease enzyme to the wells at the defined times.

[0112] At step 435, digestion is halted by heating the well to 70°C for 20min to inactivate the Exonuclease III. At step 440, Mung Bean or S1 Nuclease is added to the $t-1$ wells to degrade the single strand DNA, thus removing the 5' 'tail' from the double strand DNA fragments created by the Exonuclease III digestion.

[0113] Results of an example digestion process are shown in Figure 4D. In this example, three collections are prepared, with the collection 401 representing tier "0" containing undigested DNA fragments, whilst the collections 402, 403 represent tiers "1" and "2" respectively, and contain progressively digested DNA fragments. The graphs illustrate a distribution of the number of nucleotides in each tier, showing that tier "0" contains undigested DNA fragments most of which have approximately 600 nucleotides, with

- 23 -

variations arising during the amplification process. The strands in tiers "1" and "2" are approximately symmetrically digested at each end, with most of tier "1" fragments losing about 100 nucleotides at each end and most tier "2" fragments losing about 150 nucleotides at each end.

[0114] Thus, these data highlight the ability of the digestion process to controllably digest a number of nucleotides from each end of the DNA fragments by varying a digestion time. It will be appreciated from this that the amount of digestion is not necessarily exactly equal for each DNA fragments in a given well, but that the amount of digestion will be substantially similar for DNA fragments in a well and substantially different for DNA fragments in different wells. As a result, differences in the amount of digestion within a tier is within limits that can be accommodated using suitable statistical analysis techniques when analyzing the resulting read data, as will be described in more detail below.

[0115] In any event, the result of the digestion process is the production of t wells each including digested DNA fragments for each of the multiple DNA fragment templates.

[0116] At step 445, libraries are created for the each of the $t-1$ wells, which typically includes adding adaptors to the digested DNA fragments in each of the $t-1$ wells, for example by adding an adaptor oligo mix, ligase and buffer solutions as required by the particular sequencing device and/or protocol. In this regard, it will be appreciated that the undigested molecules corresponding to tier "0" already have adaptors attached, and so this step may not be required for that well. Creation of the library can also include adding a unique label, which may form part of the adaptor added above. In this regard, a unique label is used for each well, so that the digested template molecules from each well can be identified so as to allow the analysis process to ascertain with which tier resulting reads should be associated. However, the tier "0" fragments may already include a label in which case a further label may not be needed for this well. The nature of the unique label may vary depending on the preferred implementation, but typically this includes a predefined sequence of nucleotides which are bound to the end of the DNA fragments in each collection.

[0117] At step 450, samples from the different wells are combined and provided to the sequencing device, which is used to perform sequencing of the digested DNA fragments.

- 24 -

This is performed in accordance with normal device operation and will typically involve the sequencing device generating reads representing sequence content of the unique label and an end portion of each digested DNA fragment at step 455. Alternatively however the DNA fragments from different wells can be sequenced separately, in which case labels may not be required.

[0118] It will be appreciated from the above description that the digested DNA fragments in a given tier are not identical, for example as they are obtained from different template DNA fragments, as well as due to copy errors during the initial amplification, and variations in the digestion process. Accordingly, the read data is statistically analysed to identify a consensus sequence for each DNA fragment.

[0119] To achieve this, at step 460 the electronic processing device, which has received read data representing the reads generated by the sequencer, groups reads into tiers using the corresponding label, if this has not already been done by the sequencing device itself. At step 465 the electronic processing device compares reads between different tiers to identify inter-tier overlaps. In particular, this process typically identifies multiple potential inter-tier overlaps for different reads, which are then filtered at step 480, based on insert length distributions and other factors.

[0120] At step 475 the electronic processing device sorts the overlapping inter-tier reads across all tiers into subgroups, based on the overlaps, allowing the subgroups of overlaps and reads to be used to assemble a consensus sequence for each template DNA fragment at step 480.

[0121] In more detail, the process of analysing the reads can be broadly understood to include the following steps:

1. Labels added to the reads during sequencing preparation are used to sort paired reads into tier specific groups, if this has not already been done by the sequencing device.
2. A search is performed with the aim of identifying all inter-tier read overlaps.
3. The distributions of insert sizes for each tier are calculated and stored.

- 25 -

4. The complete set of inter-tier overlaps is filtered such that only paired inter-tier (PIT) overlaps remain. Furthermore, the insert sizes inferred by these overlaps must match the distributions calculated at step 3.
5. Paired reads spanning all tiers are sorted into subgroups.
6. The overlap and tier-specific information is used to reconstruct the consensus sequence of each subgroup.

[0122] In practice, the algorithm used to identify all inter-tier overlaps is performed iteratively for each tier. At each iteration a storage hash is constructed which encodes information about the reads in that tier. At the same time, the reads are checked against a comparison hash generated during the previous iteration and potential overlaps are identified. When all the reads in a tier have been parsed the comparison hash can be destroyed and the newly constructed storage hash becomes the comparison hash for the next iteration.

[0123] This process will now be described in more detail with reference to Figures 5A to 5E.

[0124] In this example, in step 500 the electronic processing device selects a next tier in the hierarchy, typically starting with the lowest tier, in this case tier "0", and optionally divides reads within the tier into multiple streams at step 505. This is performed to allow each stream to be processed separately, for example by a respective electronic processing device, allowing a massively parallel architecture to be employed. This significantly reduces the time required to perform an analysis of each tier, in turn allowing this to be used for large datasets without an undue time delay.

[0125] For each stream, paired reads are then divided into overlapping k-mers at step 510. In this regard, the term k-mer refers to small sub-strings that commonly include a small number of nucleotides, such as between three and a number of bases shorter than the read being analysed. The length of the k-mer and the amount of overlap between each k-mer can vary depending on the length and quality of the input reads. As each k-mer is produced, information is generated about each k-mer including but not limited to:

- a description of the read it came from;
- it's position within that read; and,
- it's orientation (with respect to strandedness) at that position.

[0126] At step 515 information regarding each k-mer is added to a storage hash. In this regard, the term storage hash refers to a data store, such as a database, containing information regarding the k-mers for the current stream, and which usually stores the information mapped to a predetermined format, which typically but not necessarily reduces overall storage requirements. It will therefore be appreciated that the use of a hash, whilst advantageous, is not intended to be limiting, and that any appropriate data store and structure could be used.

[0127] As individual k-mers will be encountered multiple times, the k-mers and associated information are stored using the sequence of the k-mer as a key. In such cases the k-mer sequence key will point to a list of information packets, one for each occurrence of the k-mer within the stream. It should be noted that it is not necessary to construct a storage hash for the final iteration, corresponding to the highest tier in the hierarchy, however k-mers must still be generated for comparison with the hash produced in the previous iteration, as will become apparent from the following description.

[0128] At step 520, each identified k-mer is compared to k-mers in a comparison hash, to identify the same k-mers within the different tiers. In this regard, the comparison hash is a hash resulting from identification of k-mers in previous tiers in the hierarchy, and which is created by combining storage hashes from the previous tier. This process therefore effectively determines inter-tier overlapping k-mers. It will also be appreciated that this step is not performed for tier "0" in the hierarchy.

[0129] As each k-mer is produced, the algorithm checks to see whether it exists in the comparison hash. Co-occurrence of k-mers between two reads (common k-mers) indicates a potential overlapping read between the tiers. Accordingly, if the k-mer exists in the comparison hash, it is used to generate a list of reads from the previous tier which contain this k-mer. These reads are called query sequences or "queries" and the read currently being parsed is called the subject read or "subject".

[0130] Following this at step 530, the subjects and queries are analysed to identify potential inter-tier overlapping reads.

[0131] For example, as shown in Figure 5C, the first common k-mer found for a pair of reads is denoted K_1 , the second common k-mer is denoted K_2 and so forth, with the final common

k-mer is denoted K_N . Using information about the positions and orientations of K_1 in each read, it is possible to predict the positions and orientations of $K_1, K_2, \dots, K_N$ for the subject and also for each query. Thus, for each query it is possible to produce a collection of k-mers, positions and orientations which, if found in the subject, identify a potential overlap between the subject and the query.

[0132] When such k-mers are produced the potential overlap is confirmed and information describing it is stored. This information includes but is not limited to: a description of the reads which are overlapped, the relative orientation of the subject with respect to the query and visa versa and the amount of overlap shared between the subject and the query.

[0133] It will be appreciated that the process of comparing all k-mers across a possible overlap and requiring these to be equal will only find overlapping reads that match perfectly. Whilst this is fine in theory, in practice errors may arise in sequences, for example due to sequencing errors, or the like. Accordingly, in one example, the process is modified to allow identification of imperfect overlaps. In this example, instead of identifying a set of k-mers and corresponding positions and overlaps that confirm an inter-tier read overlap, the existence of common k-mers is used to indicate a possible overlap. Following this, more generalised string matching algorithms can be used to confirm the overlap and calculate the number of mismatches etc. This extra information can be stored alongside the information which would normally be stored for each overlap.

[0134] Once the streams for a respective tier have been analysed, it is determined if all the tiers are complete at step 535, and if not the process moves on to step 540 to replace the comparison hash with the storage hash to form a new comparison hash, which is then used in analysing the next tier in the hierarchy. In this regard, it will be appreciated that if a parallel processing approach is used at step 505, then a respective storage hash will be produced for each stream and these must be combined when forming the new comparison hash.

[0135] It will be appreciated that parallelisation can be used as the process does not attempt to examine intra-tier overlaps, but rather only looks for inter-tier overlaps. In this regard, each stream is typically analysed by a respective processor which can achieve this by having read only access to the comparison hash generated for the previous tier. The processor then

creates its own storage hash, with the individual storage hashes being merged upon completion of the analysis of each stream.

[0136] In any event, having identified all potential inter-tier overlapping reads, these are then filtered to determine the most-likely inter-tier overlapping reads.

[0137] Filtering may be achieved in any suitable manner. In one example, this is done based on the distribution of insert lengths within each tier, and accordingly, at step 545 a distribution of insert lengths is determined for each tier. In this regard, as shown in Figure 5D, each paired read has two ends denoted R^1 and R^2 . If both ends of the paired read have been mapped onto a reference sequence or template, then the regions lying directly under each read are known as read regions, the points that mark the inside edges of the read regions the inner boundaries and the points that mark the outside edges of the read regions the outer boundaries. In this case, the distance separating the outer boundaries is referred to as the insert length.

[0138] The manner in which the insert length can be determined will vary depending on the preferred implementation. For example, the easiest way to determine the insert length distributions is to map reads from each tier to a closely related reference sequence or collection of reference sequences. A sufficiently large number of reads from each tier should map to a given reference, allowing for direct calculation of insert lengths and the resulting distributions. Alternatively, this can be achieved by performing a rough assembly of the sequence for one of the tiers, typically tier "0", and using this for a point of comparison for determining insert lengths. As a further alternative, for experiments where the target genomes are expected to have absolutely no relation to any known reference, a suitable small reference genome can be added as a template to sequencing, such as PhiX.

[0139] At step 550 the list of potential inter-tier overlaps are filtered using the insert length distributions for each tier. To perform this, a paired inter-tier overlap (paired overlap) is defined as a set of two inter-tier overlaps O_A and O_B between paired reads R_X and R_Y that respectively belong to adjacent tiers T_N and T_{N+1} and have the following properties:

- Overlap O_A links end R_X^1 of R_X and end R_Y^1 of R_Y and overlap O_B links end R_X^2 of R_X and end R_Y^2 of R_Y .

- 29 -

- The overlaps are such that the outer boundaries of R_Y fall within the read regions of R_X .
- The distance between the outer boundaries of R_Y^1 and R_X^1 (left offset) and between the outer boundaries of R_Y^2 and R_X^2 (right offset) are what we would expect to see given our knowledge of the distributions of insert sizes obtained during the previous step.

[0140] Accordingly, the electronic processing device examines the list of potential inter-tier overlaps and removes those that do not meet the above criteria.

[0141] At step 555 the resulting overlapping paired reads are sorted into subgroups. This can be achieved using any suitable technique, such as by identifying common reads, clustering algorithms or the like. The resulting subgroups can then be used to determine a consensus sequence for a corresponding template DNA fragment, for example using known assembly techniques.

[0142] Thus, the above described process allows for assembly of the consensus sequences for each of the sequenced template molecules. Furthermore, by performing the process by initially identifying only inter-tier overlaps, achievable through the use of the labelling of the different tiers, this can advantageously allow a parallelised implementation to be performed.

[0143] An example of a processing system including an electronic process device is shown in Figure 6.

[0144] In this example, the processing system 600 includes at least one microprocessor 610, a memory 611, an optional input/output device 612, such as a keyboard and/or display, and an external interface 613, interconnected via a bus 614 as shown. In this example the external interface 613 can be utilised for connecting the processing system 600 to peripheral devices, such as a sequencing device, communications network, external databases, other storage devices, or the like. Although a single external interface 613 is shown, this is for the purpose of example only, and in practice multiple interfaces using various methods (eg. Ethernet, serial, USB, wireless or the like) may be provided.

[0145] In use, the microprocessors 610 execute instructions in the form of applications software stored in the memory 611 to allow the above described analysis process to be performed, as well as to perform any other required processes, such as communicating with the sequencing device. The applications software may include one or more software modules, and may be executed in a suitable execution environment, such as an operating system environment, or the like.

[0146] Accordingly, it will be appreciated that the processing system 600 may be formed from any suitable processing system, such as a suitably programmed computer system, PC, web server, network server, or the like. In one particular example, the processing system 600 is a standard processing system such as an Intel or AMD Architecture based processing system, which executes software applications stored on non-volatile (*e.g.*, hard disk) storage, although this is not essential. However, it will also be understood that the processing system could be any electronic processing device such as a microprocessor, microchip processor, logic gate configuration, firmware optionally associated with implementing logic such as an FPGA (Field Programmable Gate Array), or any other electronic device, system or arrangement.

[0147] It will therefore be appreciated from the above that the sequencing method and apparatus can be used to enhance the sequencing ability of existing known sequencing techniques and machines using minor changes in the protocol for preparing template molecules for sequencing and through the use of suitable post sequencing data analysis.

[0148] Throughout this specification and claims which follow, unless the context requires otherwise, the word “comprise”, and variations such as “comprises” or “comprising”, will be understood to imply the inclusion of a stated integer or group of integers or steps but not the exclusion of any other integer or group of integers.

[0149] Persons skilled in the art will appreciate that numerous variations and modifications will become apparent. All such variations and modifications which become apparent to persons skilled in the art, should be considered to fall within the spirit and scope that the invention broadly appearing before described.

THE CLAIMS DEFINING THE INVENTION ARE AS FOLLOWS:

- 1) A method of sequencing at least one template nucleic acid molecule, the method including:
 - a) differentially digesting a number of copies of the at least one template molecule to thereby generate digested template molecules, at least some of the digested template molecules having different numbers of nucleotides removed from at least one end of the template molecule;
 - b) sequencing at least part of the digested template molecules to generate reads representing the sequence content of at least an end portion of the digested template molecules; and,
 - c) generating a consensus sequence for the at least one template molecule using overlaps between the reads.
- 2) A method according to claim 1, wherein at least some of the digested template molecules have nucleotides removed from each end and wherein the method includes sequencing at least part of either end of the digested template molecules.
- 3) A method according to claim 1 or claim 2, wherein the method includes differentially digesting copies of the at least one template molecule using at least one nuclease enzyme.
- 4) A method according to claim 3, wherein the enzyme is at least one of:
 - a) Exonuclease III;
 - b) Mung Bean Nuclease;
 - c) S1 Nuclease; and,
 - d) Bal31.
- 5) A method according to any one of the claims 1 to 4, wherein the method includes differentially digesting copies of the at least one template molecule by controlling at least one of:
 - a) a digestion time;
 - b) a digestion rate;
 - c) a digestion temperature;
 - d) a nuclease enzyme concentration; and,
 - e) a nuclease enzyme type.
- 6) A method according to any one of the claims 1 to 5, wherein the method includes creating a plurality of collections of digested template molecules, the copies of the template

molecules in a given collection being digested under similar conditions so that the digested template molecules in a collection have a similar number of nucleotides removed from at least one end, whilst digested template molecules in different ones of the plurality of collections have substantially different numbers of nucleotides removed from at least one end.

- 7) A method according to any one of the claims 1 to 6, wherein the method includes:
 - a) amplifying the at least one template molecule to generate copies of the at least one template molecule;
 - b) creating a plurality of collections, each collection including copies of the at least one template molecule; and,
 - c) digesting the copies of the at least one template molecule in each collection differentially, so that each collection contains respective digested template molecules.
- 8) A method according to claim 6 or claim 7, wherein the method includes sequencing at least part of at least one end of digested template molecules in each collection to determine at least one read for each collection.
- 9) A method according to claim 8, wherein the method includes sequencing at least part of each end of the digested template molecules in each collection.
- 10) A method according to claim 8 or claim 9, wherein the method includes:
 - a) identifying overlaps between reads of different collections; and,
 - b) using the reads and overlaps to generate the consensus sequence.
- 11) A method according to any one of the claims 6 to 10, wherein the method includes:
 - a) creating each collection by digesting copies of multiple different template molecules;
 - b) sequencing at least some of the digested template molecules in each collection; and,
 - c) generating a respective consensus sequence for at least some of the multiple different template molecules.
- 12) A method according to claim 11, wherein the method includes creating each collection from a desired distribution of copies of multiple different template molecules.
- 13) A method according to claim 11 or claim 12, wherein the method includes:
 - a) obtaining a desired distribution of the multiple different template molecules;
 - b) amplifying the desired distribution of multiple different template molecules to generate a number of copies of each of the multiple different template molecules; and,

- 33 -

- c) creating a plurality of collections, each collection including a subset of the number of copies of each of the multiple different template molecules.
- 14) A method according to any one of the claims 6 to 13, wherein the method includes:
- a) adding a respective label to the digested template molecules in each collection; and,
 - b) determining the consensus sequence at least partially in accordance with the respective labels.
- 15) A method according to claim 14, wherein the method includes determining overlaps between the reads in different collections using the labels.
- 16) A method according to claim 14 or claim 15, wherein each collection represents a respective tier in a sequence hierarchy, the template molecules in adjacent tiers being progressively digested.
- 17) A method according to claim 16, wherein, for multiple different template molecules, the method includes:
- a) selecting reads for a next tier using the labels;
 - b) determining overlapping k-mers for each read;
 - c) comparing the k-mers to k-mers of reads from previous tiers to determine inter-tier overlapping reads; and,
 - d) using the inter-tier overlapping reads to generate a consensus sequence for each of the multiple different template molecules.
- 18) A method according to claim 17, wherein the method includes:
- a) dividing the reads from a tier into multiple streams; and,
 - b) processing each stream independently.
- 19) A method according to claim 17 or claim 18, wherein the method includes:
- a) comparing the k-mers to k-mers of previous tiers to determine inter-tier overlapping k-mers;
 - b) using the inter-tier overlapping k-mers to determine potential inter-tier overlapping reads; and,
 - c) using the potential inter-tier overlapping reads to determine the inter-tier overlapping reads.
- 20) A method according to claim 19, wherein, the method includes:
- a) calculating a distribution of insert lengths of reads in each tier; and,
 - b) filtering inter-tier overlapping reads using the distribution of insert lengths.

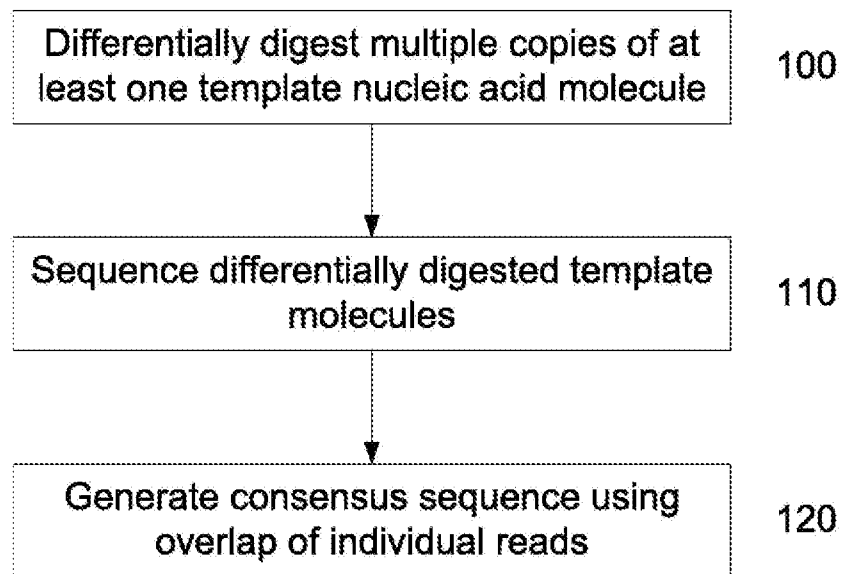
- 34 -

- 21) A method according to claim 20, wherein, the method includes:
- a) determining subgroups of inter-tier overlapping reads; and,
 - b) generating consensus sequences of at least some of the multiple different template molecules using the subgroups of inter-tier overlapping reads.
- 22) A method according to any one of the claims 1 to 21, wherein the method includes, in an electronic processing device:
- a) determining read data indicative of the reads; and,
 - b) using the read data to generate the consensus sequence.
- 23) A method according to claim 22, wherein the method includes, in the electronic processing device, receiving the read data from a sequencing device.
- 24) A method according to claim 22 or claim 23, wherein the method includes, in the electronic processing device:
- a) generating an indication of the consensus sequence; and,
 - b) causing an indication of the consensus sequence to be displayed.
- 25) Apparatus for generating a consensus sequence representing a sequence of at least one template nucleic acid molecule, the apparatus including an electronic processing device that:
- a) determines read data indicative of reads, the reads representing the sequence content of at least an end portion of digested template molecules, the digested template molecules being obtained by differentially digesting a number of copies of at least one template molecule so that at least some of the digested template molecules have different numbers of nucleotides removed from at least one end of the at least one template molecule; and,
 - b) uses the read data to generate consensus sequence data representing the consensus sequence.
- 26) Apparatus according to claim 25, wherein the electronic processing device:
- a) determines overlapping k-mers for each read;
 - b) compares the k-mers to k-mers of reads from previous tiers to determine inter-tier overlapping reads; and,
 - c) uses the inter-tier overlapping reads to generate a consensus sequence for each of the multiple different template molecules.

- 35 -

- 27) Apparatus according to claim 26, wherein the apparatus includes multiple electronic processing devices, and wherein reads for a next tier are dividing into multiple streams and each electronic processing device processes a respective stream by:
- a) determining overlapping k-mers for each read in the stream; and,
 - b) comparing the k-mers to k-mers of reads from previous tiers to determine potential inter-tier overlapping reads.
- 28) Apparatus according to any one of the claims 25 to 27, wherein the apparatus is for use in performing the method of any one of the claims 1 to 24.
- 29) A method of sequencing at least one template nucleic acid molecule, the method including:
- a) creating a plurality of collections, each collection including a desired distribution of copies of multiple different template molecules;
 - b) differentially digesting the copies of the multiple different template molecules in each collection to thereby generate digested template molecules, at least some of the digested template molecules having different numbers of nucleotides removed from at least one end of the multiple template molecules;
 - c) sequencing at least part of the digested template molecules to generate reads representing the sequence content of at least an end portion of the digested template molecules; and,
 - d) generating a consensus sequence for at least some of the multiple different template molecules.
- 30) Apparatus for generating a consensus sequence representing a sequence of at least some of multiple different template nucleic acid molecules, the apparatus including an electronic processing device that:
- a) determines read data indicative of reads, the reads representing the sequence content of at least an end portion of digested template molecules, the digested template molecules being obtained by differentially digesting a number of copies of multiple template molecules so that at least some of the digested template molecules have different numbers of nucleotides removed from at least one end of the multiple template molecules; and,
 - b) uses the read data to generate consensus sequence data representing a consensus sequence for at least some of the multiple different template molecules.

1/11

**Fig. 1**

2/11

GGATCCTCGTGAGCTGGACGTCGACTTG 201

GGATCCTCGTGAGCTGGACGTCGACTTG 202

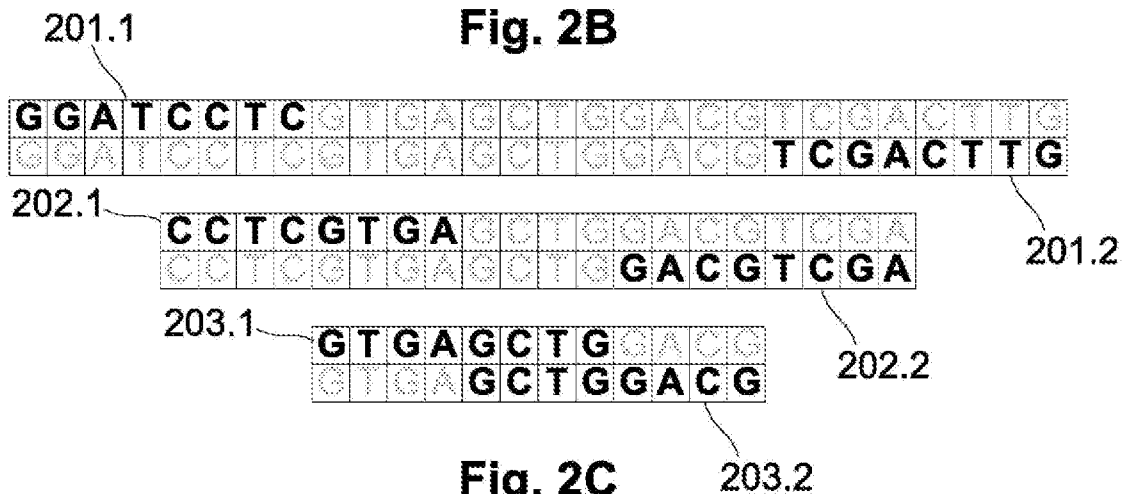
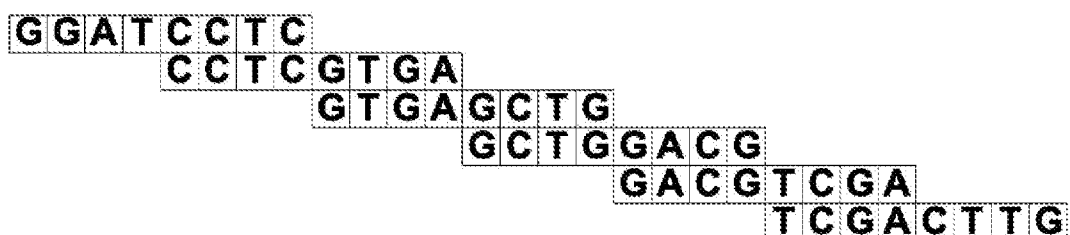
GGATCCTCGTGAGCTGGACGTCGACTTG 203

Fig. 2A

GGATCCTCGTGAGCTGGACGTCGACTTG 201

CCTCGTGAGCTGGACGTCGA 202

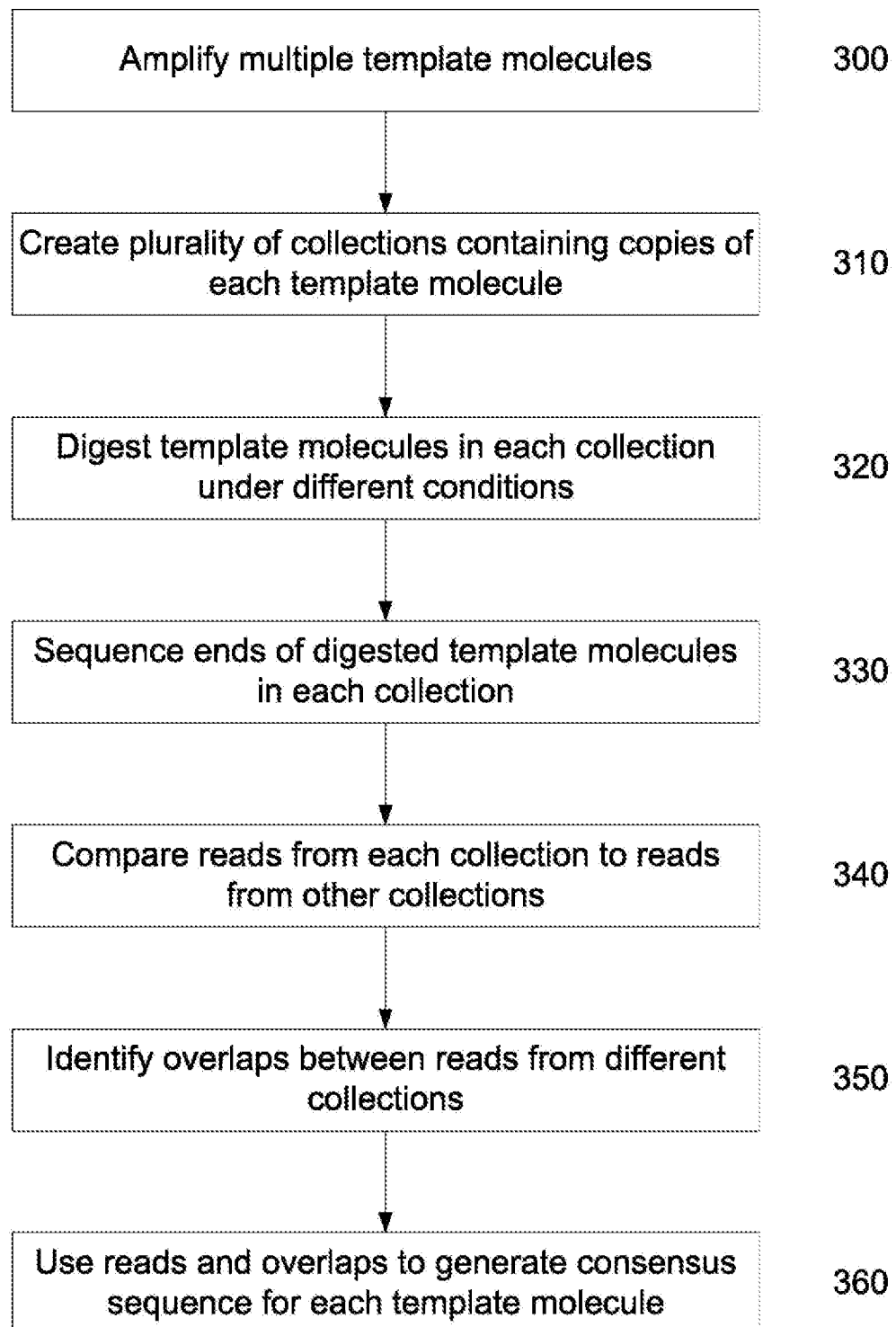
GTGAGCTGGACG 203

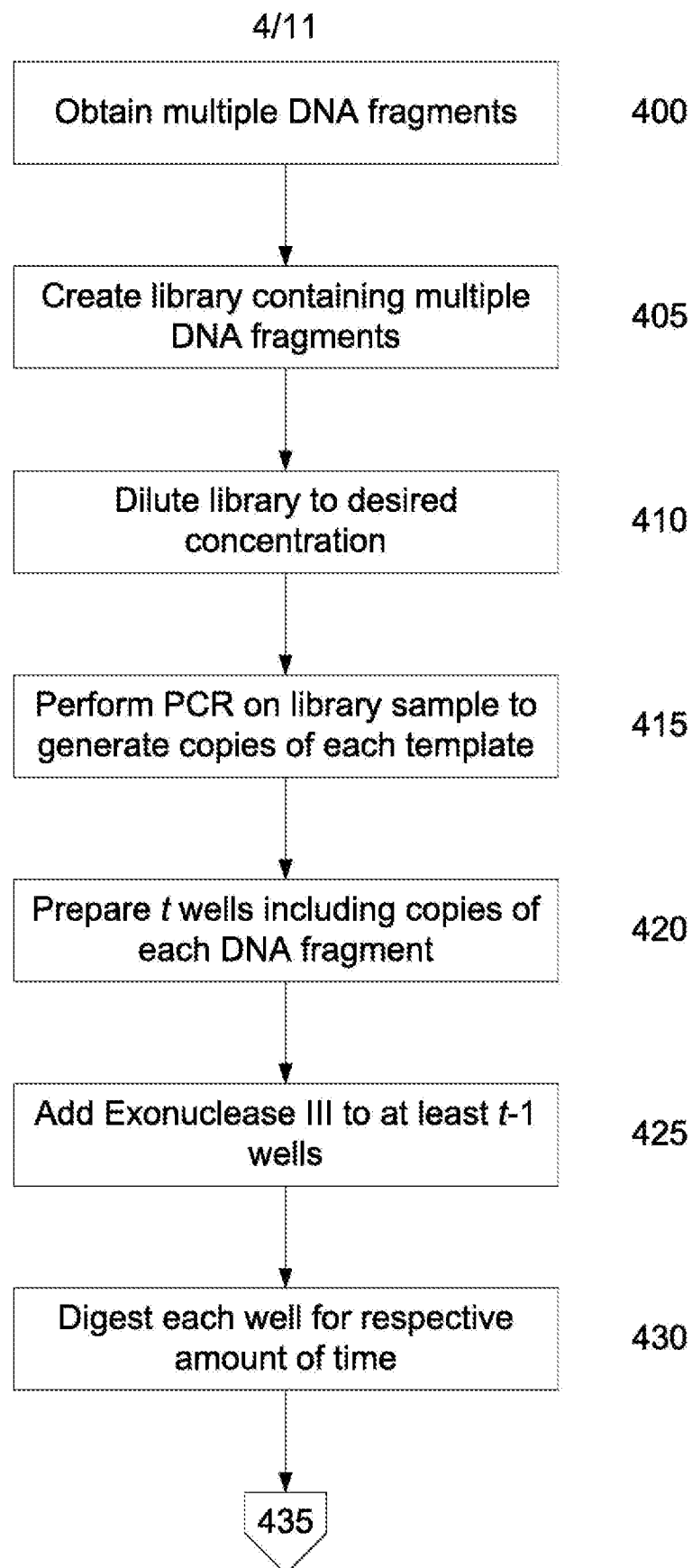
Fig. 2B**Fig. 2C****Fig. 2D**

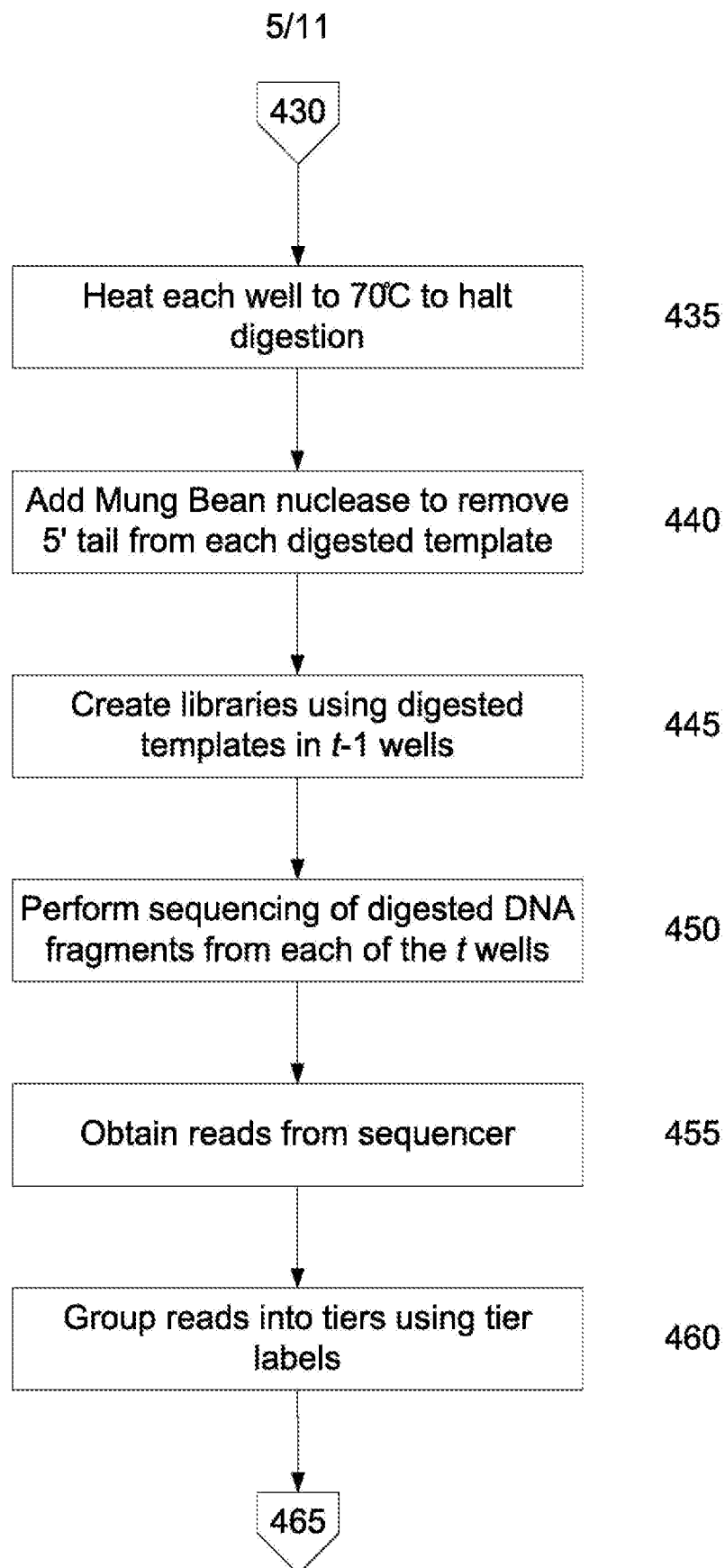
GGATCCTCGTGAGCTGGACGTCGACTTG 204

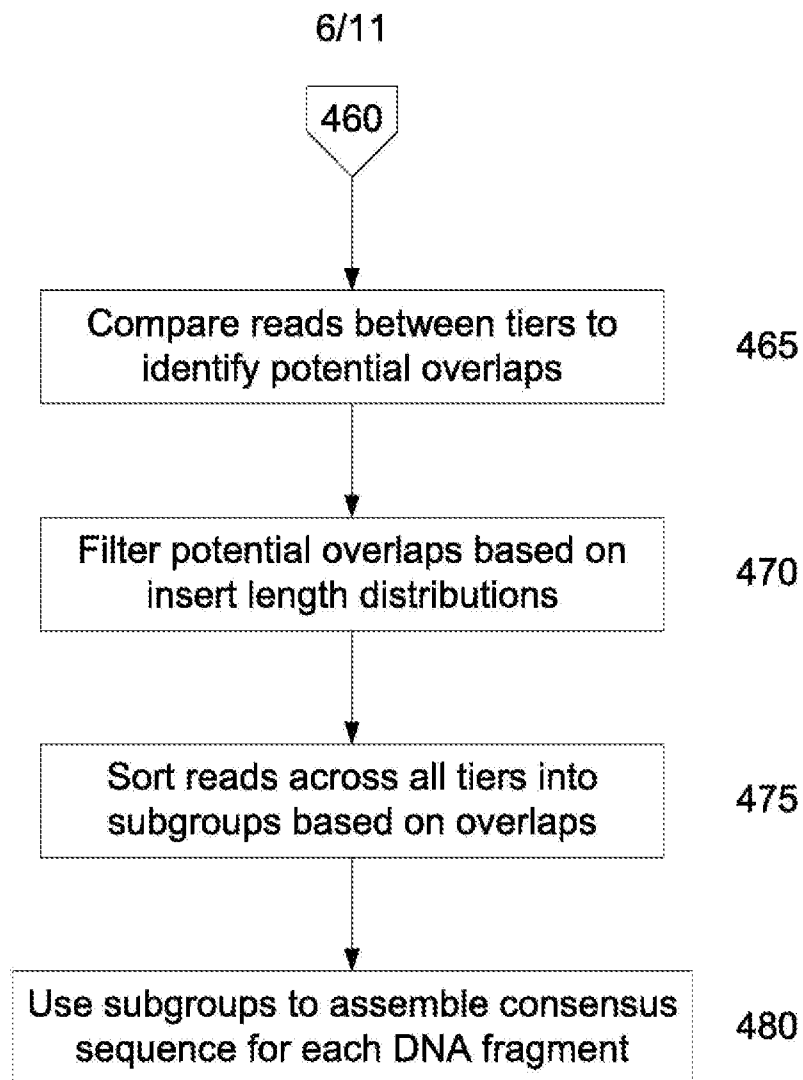
Fig. 2E

3/11

**Fig. 3**

**Fig. 4A**

**Fig. 4B**

**Fig. 4C**

7/11

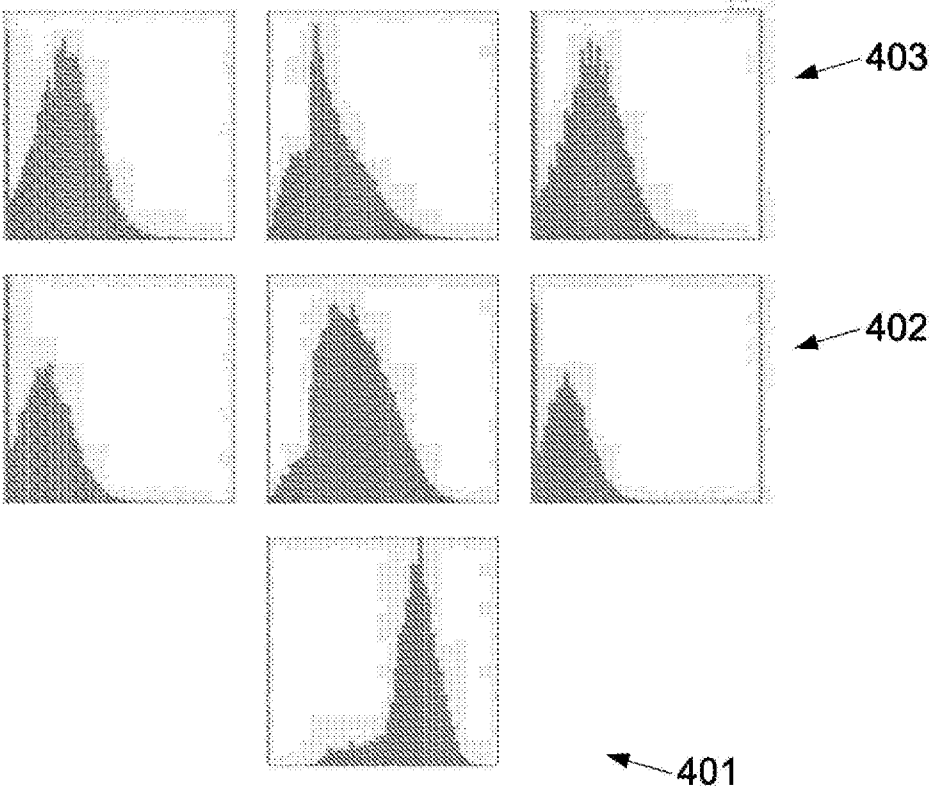
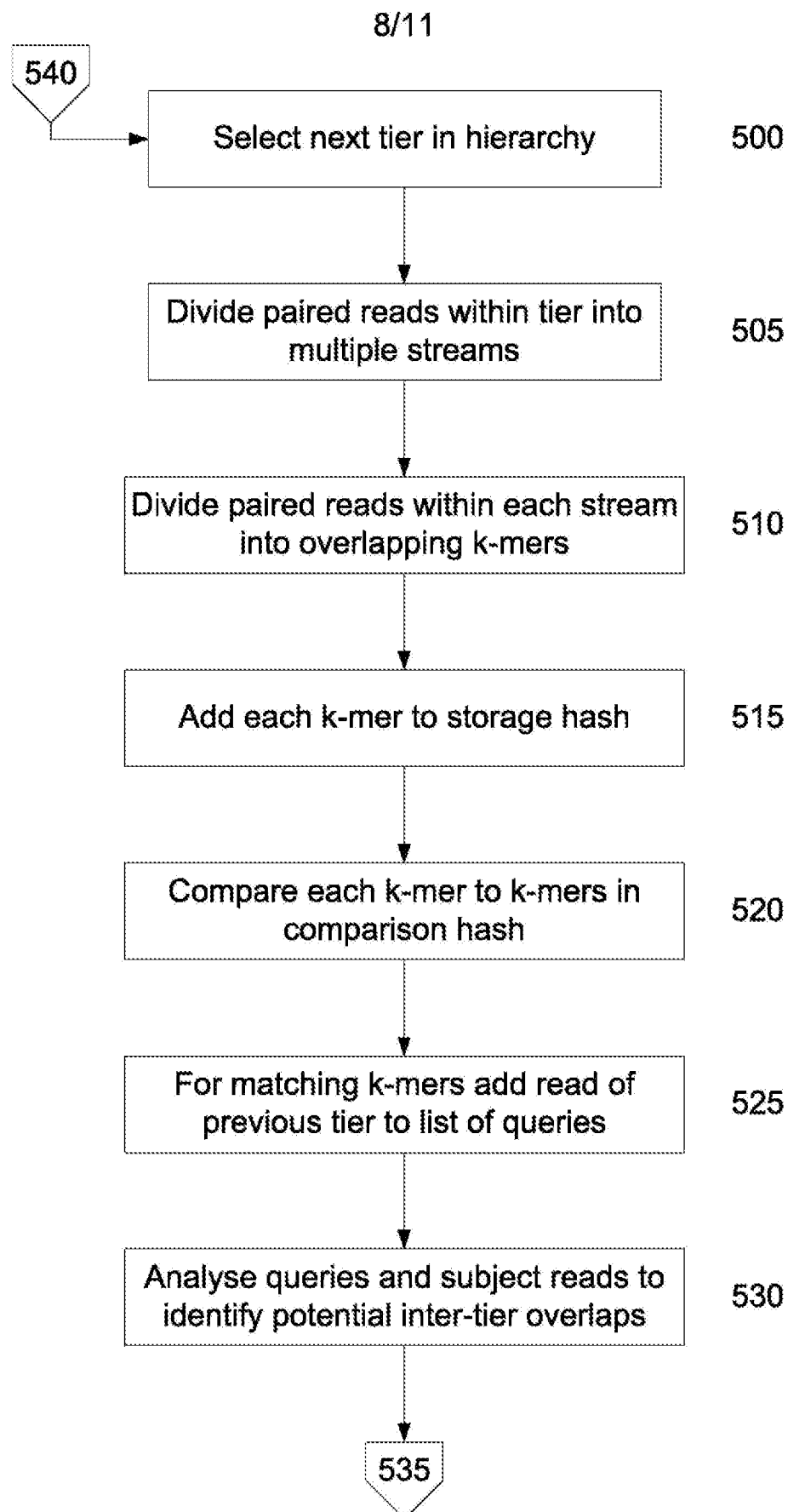
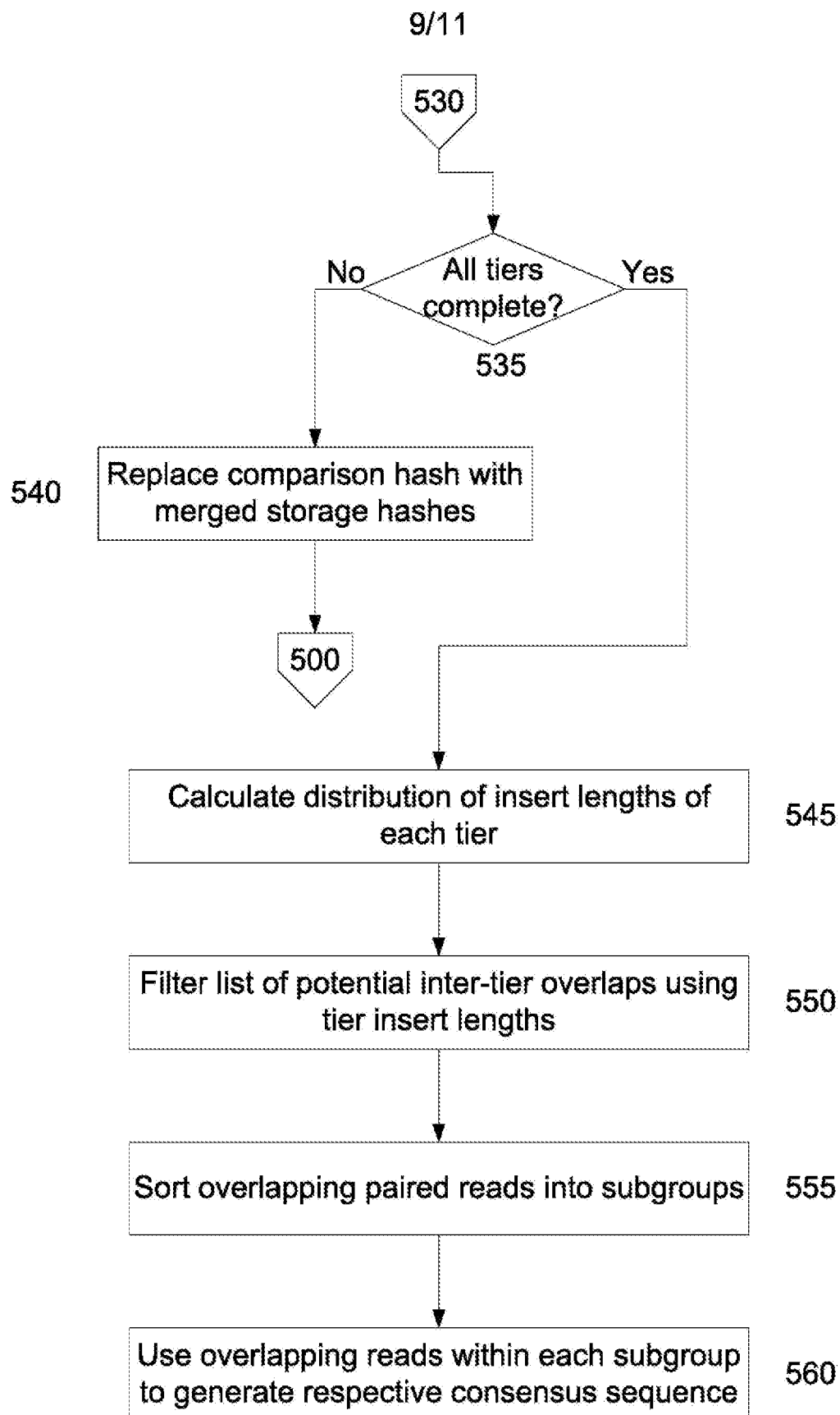


Fig. 4D

**Fig. 5A**

**Fig. 5B**

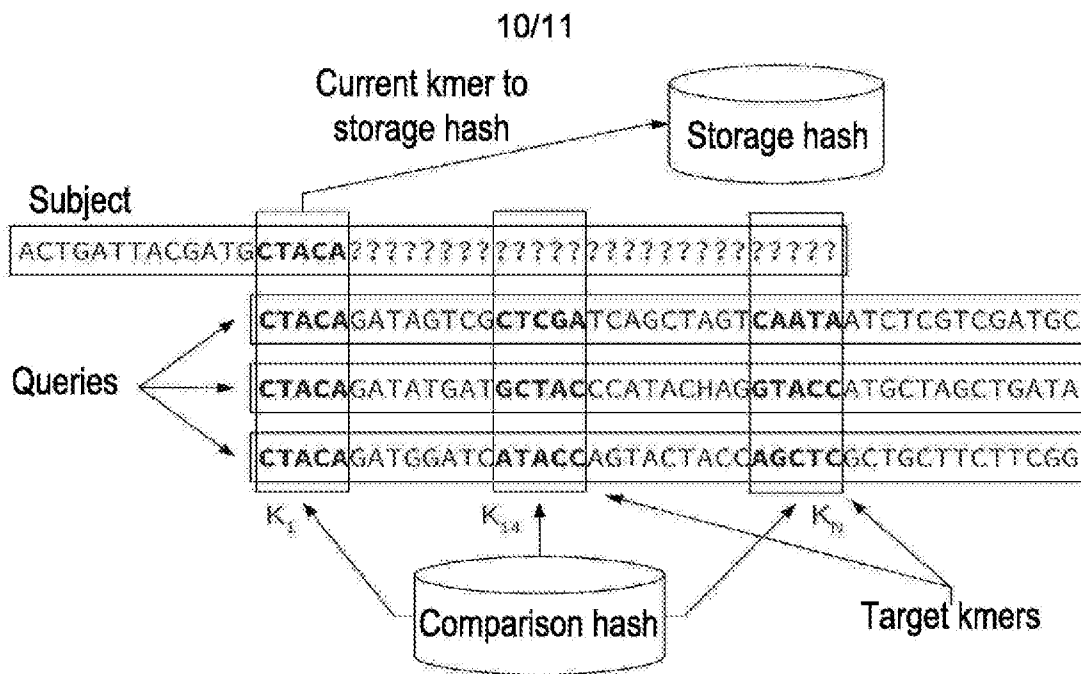


Fig. 5C

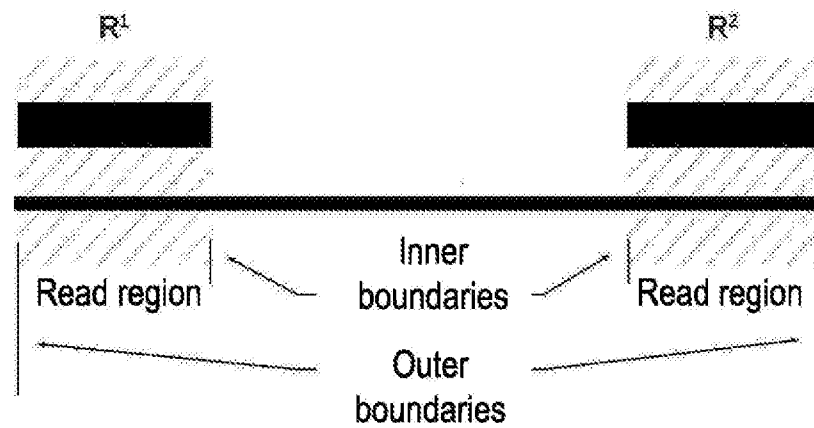
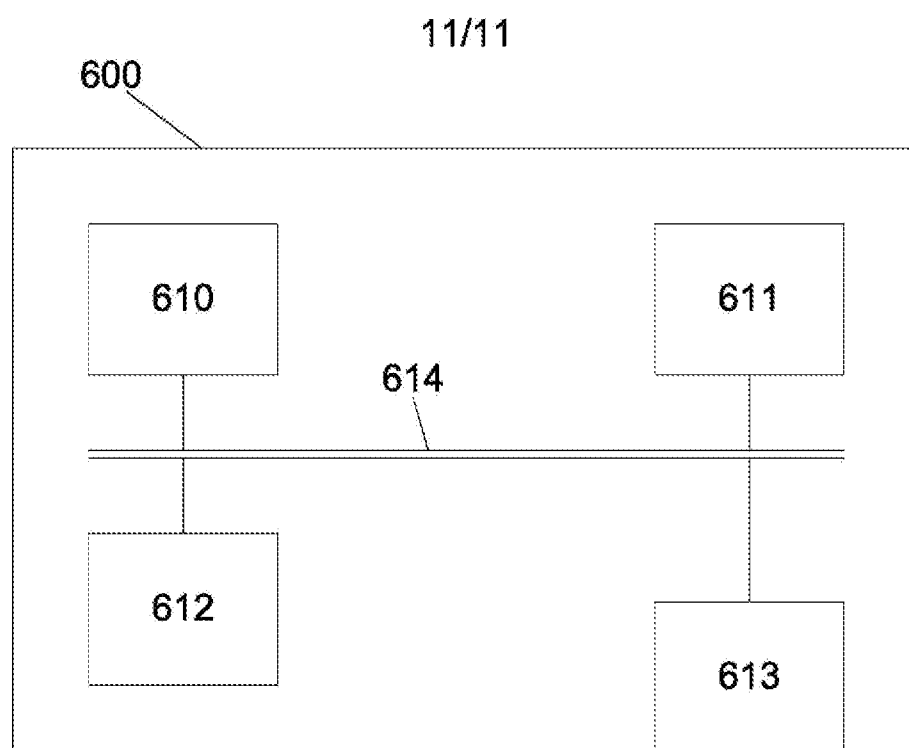


Fig. 5D



Fig. 5E

**Fig. 6**

INTERNATIONAL SEARCH REPORT

International application No.
PCT/AU2014/050446

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68 (2006.01)

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

MEDLINE, HCAPLUS, BIOSIS, EMBASE, WPIAP, EPODOC, EPOQUE XFULL, TXTE (all English full text databases): nested, ordered, sequential, differential, successive, deletion, sequencing, exonuclease III, mung bean nuclease, S1 nuclease, Bal31, sequence, read, identity, consensus, similar, tier, inter-tier?, over-lap, align, assemble, and like terms and synonyms. GOOGLE SCHOLAR, PUBMED, WIPO, AUSPAT: Applicant/inventor name search.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	Documents are listed in the continuation of Box C	



Further documents are listed in the continuation of Box C



See patent family annex

* "A"	Special categories of cited documents: document defining the general state of the art which is not considered to be of particular relevance	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"E"	earlier application or patent but published on or after the international filing date	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means	"&"	document member of the same patent family
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search
23 March 2015

Date of mailing of the international search report
23 March 2015

Name and mailing address of the ISA/AU

AUSTRALIAN PATENT OFFICE
PO BOX 200, WODEN ACT 2606, AUSTRALIA
Email address: pct@ipaustalia.gov.au

Authorised officer

Susan MacKenzie
AUSTRALIAN PATENT OFFICE
(ISO 9001 Quality Certified Service)
Telephone No. 0262832690

INTERNATIONAL SEARCH REPORT		International application No.
C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		PCT/AU2014/050446
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2012/050920 A1 (ILLUMINA, INC.) 19 April 2012 paragraphs [0076], [0077], [0083], [0098], [0153]-[0156], claim 24, [0091]	1-30
X	WO 2009/052214 A2 (COMPLETE GENOMICS, INC.) 23 April 2009 [0128],[0132], claims	1-30
X	US 6248569 B1 (DUNN, J.J. et al.) 19 June 2001 Lines 60-67, examples	1-30
X	US 6258571 B1 (CHUMAKOV, I. et al.) 10 July 2001 Examples 3 and 4, claim 24	1-30
X	SLATKO, B. et al., 'Constructing nested deletions for use in DNA sequencing', Current Protocols in Molecular Biology, 1991, pages 7.2.1-7.2.20 whole document	1-30
X	MCCOMBIE, W. R. et al., 'The Use of Exonuclease III Deletions in Automated DNA sequencing', Methods: A companion to Methods in Enzymology, Vol. 3, No. 1, 1991 Pages 33-40 whole document	1-30
X	REN, Z. et al., 'A powerful approach for generating and sequencing DNA deletions: sequencing from outside in', Analytical Biochemistry, 1997, Vol. 245, pages 112-114 whole document	1-30
X	HENIKOFF, S. 'Ordered deletions for DNA sequencing and in vitro mutagenesis by polymerase extension and exonuclease III gapping of circular templates', Nucleic Acids Research, 1990, Vol. 18, pages 2961-2966 whole document	1-30

Form PCT/ISA/210 (fifth sheet) (July 2009)

INTERNATIONAL SEARCH REPORT		International application No.	
Information on patent family members		PCT/AU2014/050446	
This Annex lists known patent family members relating to the patent documents cited in the above-mentioned international search report. The Australian Patent Office is in no way liable for these particulars which are merely given for the purpose of information.			
Patent Document/s Cited in Search Report		Patent Family Member/s	
Publication Number	Publication Date	Publication Number	Publication Date
WO 2012/050920 A1	19 April 2012	US 2013244886 A1	19 Sep 2013
		US 8759038 B2	24 Jun 2014
		US 2014287935 A1	25 Sep 2014
WO 2009/052214 A2	23 April 2009	US 2009111115 A1	30 Apr 2009
		US 8951731 B2	10 Feb 2015
US 6248569 B1	19 June 2001	US 5928908 A	27 Jul 1999
		US 5968786 A	19 Oct 1999
US 6258571 B1	10 July 2001	AU 754202 B2	07 Nov 2002
		AU 3048499 A	01 Nov 1999
		CA 2321592 A1	21 Oct 1999
		EP 1068305 A2	17 Jan 2001
		EP 1068305 B1	21 Jul 2004
		JP 2002511255 A	16 Apr 2002
		US 6022716 A	08 Feb 2000
		US 2002068333 A1	06 Jun 2002
		US 6955902 B2	18 Oct 2005
		WO 9953044 A2	21 Oct 1999
End of Annex			
Due to data integration issues this family listing may not include 10 digit Australian applications filed since May 2001.			