

**CLAIMS**

1. A method for determining a frequency distribution of a plurality of nucleotide sequence variants (501, 502, 503) of a nucleotide sequence sample (101, 410, 500),

5 wherein the plurality of nucleotide sequence variants comprises a nucleotide reference sequence (501), and

wherein each nucleotide sequence variant (502, 503), which is not identical to the nucleotide reference sequence, differs from the nucleotide reference sequence by at least one intermediate part (504, 505) and coincides with the nucleotide reference sequence by at least two parts (506) of the nucleotide reference sequence,

10 the method comprising:

- generating (110) a reference variant list (430, 510) of sequence reference variants comprising unique nucleotide sequence variants (511, 512, 513) of the plurality of nucleotide sequence variants;
- generating (120) a reference set (440, 520) of sequence read data by performing a  
15 plurality of reference sequence reads (432) for each sequence reference variant of the reference variant list, wherein each reference sequence read has a read start site, RSS, (521);
- for each RSS and for the intermediate part of each sequence reference variant, determining (130, 442) whether or not the intermediate part is detectable by  
20 corresponding reference sequence read based on alignment of the corresponding reference sequence read to the reference variant list;
- generating (140, 414, 464) a sample set (470, 630) of sequence read data by performing one or more sample sequence reads (464) for each of a plurality of nucleotide sequence variant fragments of the nucleotide sequence sample, wherein  
25 each sample sequence read corresponds to a reference sequence read and has a corresponding RSS (621, 622);
- for each RSS and for the intermediate part of each sequence reference variant, determining (150, 435) a total number of occurrences (480, 640) in the sample set of

sequence read data based on alignment of corresponding sample sequence read to the reference variant list; and

- determining (180, 497) the frequency distribution (499) of the plurality of nucleotide sequence variants based on the total number of occurrences and the determination of whether or not the intermediate part is detectable for each RSS.

5

2. The method of claim 1, further comprising defining the nucleotide reference sequence.

3. The method of any of claims 1 through 2, wherein the intermediate part is one or more of: an insertion, a deletion, an alteration, and a translocation.

4. The method of any of claims 1 through 3, wherein determining, for an RSS and for the intermediate part of a sequence reference variant, whether or not the intermediate part is detectable by corresponding reference sequence read comprises:

10

determining, via alignment, a detection correlation value for the RSS between the corresponding reference sequence read and the sequence reference variant;

15

determining the intermediate part to be detectable at the RSS by the corresponding reference sequence read when the detection correlation value is above a detection threshold value; and

determining the intermediate part to be non-detectable at the RSS by the corresponding reference sequence read when the detection correlation value is not above the detection threshold value.

5. The method of any of claims 1 through 4, wherein determining, for an RSS and for the intermediate part of a sequence reference variant, a total number of occurrences in the sample set comprises:

20

determining, via alignment, a sample correlation value for the RSS between the corresponding sample sequence read and the sequence reference variant;

25

counting an occurrence when the sample correlation value is above an occurrence threshold value; and

refraining from counting an occurrence when the sample correlation value is not above the occurrence threshold value.

6. The method of any of claims 1 through 5, further comprising receiving an input signal indicative of one or more of the plurality of nucleotide sequence variants and the nucleotide reference sequence.
7. The method of any of claims 1 through 6, further comprising transferring a reporting signal indicative of the frequency distribution to a user interface.
8. A computer program product comprising a non-transitory computer readable medium (300), having thereon a computer program comprising program instructions, the computer program being loadable into a data processing unit and configured to cause execution of the method according to any of claims 1 through 7 when the computer program is run by the data processing unit.
9. An apparatus for determining a frequency distribution of a plurality of nucleotide sequence variants (501, 502, 503) of a nucleotide sequence sample (101, 410, 500), wherein the plurality of nucleotide sequence variants comprises a nucleotide reference sequence (501), and wherein each nucleotide sequence variant (502, 503), which is not identical to the nucleotide reference sequence, differs from the nucleotide reference sequence by at least one intermediate part (504, 505) and coincides with the nucleotide reference sequence by at least two parts (506) of the nucleotide reference sequence, the apparatus comprising controlling circuitry (200) configured to cause:
- generation of a reference variant list (430, 510) of sequence reference variants comprising unique nucleotide sequence variants (511, 512, 513) of the plurality of nucleotide sequence variants;
  - generation of a reference set (440, 520) of sequence read data by performing a plurality of reference sequence reads (432) for each sequence reference variant of the reference variant list, wherein each reference sequence read has a read start site, RSS, (521);

- for each RSS and for the intermediate part of each sequence reference variant, determination of whether or not the intermediate part is detectable by corresponding reference sequence read based on alignment of the corresponding reference sequence read to the reference variant list;
- 5       - generation of a sample set (470, 630) of sequence read data by performing one or more sample sequence reads (464) for each of a plurality of nucleotide sequence variant fragments of the nucleotide sequence sample, wherein each sample sequence read corresponds to a reference sequence read and has a corresponding RSS (621, 622);
- 10       - for each RSS and for the intermediate part of each sequence reference variant, determination of a total number of occurrences (480, 640) in the sample set of sequence read data based on alignment of corresponding sample sequence read to the reference variant list; and
- 15       - determination of the frequency distribution (499) of the plurality of nucleotide sequence variants based on the total number of occurrences and the determination of whether or not the intermediate part is detectable for each RSS.

10. A server node comprising the apparatus of claim 9.