

CLAIMS

1. A method of sequencing a first polynucleotide strand having a first polynucleotide sequence, the first polynucleotide strand resembling a second polynucleotide strand having a known second polynucleotide sequence, the method employing a data set which, for one or more fragment(s) of the second polynucleotide sequence, contains:

for each position along each said fragment:

(i) first probe data describing the hybridization intensity of the first polynucleotide strand with a respective first probe designed to bind to a portion of the second polynucleotide strand centered at said position; and

(ii) second probe data describing the respective hybridization intensities of the first polynucleotide strand with each of a set of second probes, each said second probe being designed to bind with a respective mutation of the corresponding portion of the second polynucleotide sequence which is formed by mutating the corresponding portion of the second polynucleotide sequence at said position, the data set including said second probe data for every possible said mutation;

the method comprising:

for each said position;

(a) obtaining from the dataset a first numerical parameter characterizing the hybridization intensity of the first polynucleotide strand with a corresponding first probe in comparison to the hybridization intensities of the first polynucleotide strand with the corresponding second probes;

said first numerical parameter being indicative of whether a nucleic acid of the first polynucleotide sequence is equal to a nucleic acid the second polynucleotide sequence at said position;

2. A method according to claim 1 further comprising, at each said position,

(b) obtaining at least one corresponding second numerical parameter indicative of data abnormalities in the first probe data and second probe data relating to said position; and

(c) determining whether:

(i) said first numerical parameter indicates that the nucleic acid of the first polynucleotide sequence is equal to the nucleic acid of the second polynucleotide sequence at said position; and

(ii) said at least one second numerical parameter does not indicate abnormalities in the first probe data and the second probe data; and

if said determinations are both positive, determining that the nucleic acid of the first polynucleotide sequence is equal to the nucleic acid of the second polynucleotide sequence at said position.

23. A method according to claim 12 in which said at least one second numerical parameter for each said position includes a parameter comparing the mean and the standard deviation of the corresponding first probe data and second probe data.

34. A method according to claim 12 or claim 23 including identifying for each said position the perfect match probe which is the one of the corresponding first probe and second probes having the highest hybridization intensities, and, if either of said determinations is negative, performing a verification algorithm using perfect match data describing the hybridization intensities with the first polynucleotide strand of the respective perfect match probes for the neighbouring positions.

45. A method according to claim 34 in which the verification algorithm comprises a first determination of whether the perfect match data for the neighbouring positions is indicative of a divergence between the nucleic acid of the first and second polynucleotide sequences at said position.

56. A method according to claim 45 in which said first determination is positive if the average of the perfect match data for one or more nearest neighbouring positions is lower than the perfect match data for neighbouring positions further from said position than said nearest neighboring positions.

67. A method according to any of claims 34 to 56 in which the verification algorithm comprises a second determination of whether there is a likelihood of a substitution bias at said position.

78. A method according to claim 67 in which the second determination is calculated as a ratio of:

$$f_{\text{obs}} = \frac{\#(b_1 b_2 b_3 b_4)}{\#(b_1 b_2 b_3 b_4) + \#(b_1 b_2 b_4 b_3) + \#(b_1 b_3 b_2 b_4) + \#(b_1 b_3 b_4 b_2) + \#(b_1 b_4 b_2 b_3) + \#(b_1 b_4 b_3 b_2)}, \text{ and}$$

$$f_{\text{rand}} = \frac{\#(b_1 b_2)}{t} \times \frac{\#(b_2 b_3)}{t} \times \frac{\#(b_3 b_4)}{t},$$

wherein b_1 denotes the base encoded by the perfect match probe, b_2 , b_3 and b_4 denote the bases encoded by the other of the first and second probes, $\{b_1, b_2, b_3, b_4\} = \{A, C, G, T\}$, the hybridization intensity reduction order in the position is $b_1 b_2 b_3 b_4$, and for any order of the bases denoted by $wxyz$, the function $\#(wxyz)$ denotes the number of positions, out of a number t of other positions at which the first polynucleotide sequence was determined to be b_1 , that the hybridization intensity reduction order was $wxyz$, and $\#(wx)$ denotes $\#(wxyz) + \#(wxzy)$.

89. A method according to claim 67 or 78 when dependent on claim 6, in which, upon said first determination being positive and said second determination being negative, it is determined that the nucleic acid at the first polynucleotide sequence differs from the second polynucleotide sequence at said position.

940. A method according to any preceding claim in which the fragments overlap in more than one part of the second polynucleotide strand.

104. A method according to any preceding claim in which the dataset further comprises further data describing the hybridization intensity of the first polynucleotide with one or more sets of plurality of additional mismatch probes,

each set of additional mismatch probes being designed to bind with mutations of a respective hotspot portion of the second polynucleotide strand known to contain a plurality of hotspots, and comprising an additional mismatch probe for every possible mutation of the corresponding hotspot portion of the second nucleotide portion in at least one of the hotspot positions.

112. A method of sequencing a pair of first polynucleotide strands which are complementary strands having complementary first polynucleotide sequences,

comprising performing a method according to claim 1 for each first polynucleotide strand using a respective second polynucleotide strand, the second polynucleotide strand having complementary respective second polynucleotide sequences, for each corresponding position in the second polynucleotide sequences, said verification algorithm being performed upon a determination that said first numerical parameters are indicative of the two first polynucleotide sequences not being complementary in that position.

13. ~~A method of producing an array for sequencing a first polynucleotide strand having a first polynucleotide sequence, the method employing data encoding a second polynucleotide sequence of a second polynucleotide strand resembling the first polynucleotide strand, the method comprising:~~

~~(a) defining one or more fragment(s) of the second polynucleotide sequence;~~

~~(b) constructing the array, the array comprising:~~

~~(i) for each position along each said fragment of the second polynucleotide sequence, a first probe designed to bind to a portion of the second polynucleotide sequence centred at said position; and~~

~~(ii) for each first probe, a plurality of second probes, each said second probe being designed to bind with a respective mutation of the corresponding portion of the second polynucleotide sequence which is formed by mutating a nucleic acid of the second polynucleotide sequence at said position, there being a respective said second probe for every possible said mutation.~~

14. ~~A method according to claim 13 in which said step (a) of defining the one or more fragments includes:~~

~~—identifying one or more critical regions of said second polynucleotide sequence; and~~

~~—defining at least one of said fragments to include at least one of said critical regions;~~

~~said critical regions being any one or more of:~~

~~——(a) drug-binding sites;~~

~~——(b) structural components; and~~

~~——(c) mutation hotspots.~~

15. ~~An array for sequencing a first polynucleotide strand having a first polynucleotide sequence and resembling a second polynucleotide strand having a second, known~~

polynucleotide sequence, the array comprising, for each of one or more fragment(s) of the second polynucleotide sequence:

- (i) for each position along each said fragment of the second polynucleotide sequence, a first probe designed to bind to a portion of the second polynucleotide sequence centred at said position; and
- (ii) for each first probe, a plurality of second probes, each said second probe being designed to bind with a respective mutation of the corresponding portion of the second polynucleotide sequence which is formed by mutating a nucleic acid of the second polynucleotide sequence at said position, there being a respective said second probe for every possible said mutation.

16. An array according to claim 15, wherein the second polynucleotide sequence comprises at least one sequence selected from the group consisting of SEQ ID NOs:1-8.

17. An array according to either claim 15 or 16, wherein the mismatch probes are fragments at least one sequence selected from the group consisting of SEQ ID NOs:1-8 comprising at least one mutation.

18. A method according to any of claims 1 to 12 in which the dataset is derived using an array which is produced by a method according to claim 13 or 14, or an array according to any one of claims 15 to 17.

129. A method according to any of claims 1 to 114 or 16, or an array according to any one of claims 15 to 17, in which the second polynucleotide strand is RNA or DNA of a virus.

1320. A method according to any of claims 1 to 114 or 16, or an array according to any one of claims 15 to 17, in which the second polynucleotide strand is of an influenza A virus.

1424. A method according to any of claims 1 to 114 or 16, or an array according to any one of claims 15 to 17, in which the second polynucleotide strand is of an H1N1 influenza A virus.

1522. A system comprising a processor and a data storage device, the data storage device storing program instructions readable by the processor to cause the processor to perform a method according to any one of claims 1 to 114.

1623. A computer program product, such as a tangible data storage device, encoding program instructions readable by a computer processor to cause the processor to perform a method according to any of claims 1 to 114.

1724. A kit comprising:

(a) RT-PCR primers used for amplification,

(b) an array for sequencing a first polynucleotide strand having a first polynucleotide sequence and resembling a second polynucleotide strand having a second, known polynucleotide sequence, the array comprising, for each of one or more fragment(s) of the second polynucleotide sequence:

(b) (i) for each position along each said fragment of the second polynucleotide sequence, a first probe designed to bind to a portion of the second polynucleotide sequence centred at said position; and

(ii) for each first probe, a plurality of second probes, each said second probe being designed to bind with a respective mutation of the corresponding portion of the second polynucleotide sequence which is formed by mutating a nucleic acid of the second polynucleotide sequence at said position, there being a respective said second probe for every possible said mutation according to claim 15, and

(c) a computer readable medium encoding program instructions readable by a computer processor to cause the processor to enable of carrying out the method according to any one of claims 1 to 112.