CLAIMS

What is claimed is:

1. A base mutation detection method based on sequencing data, the base mutation detection method comprising:

determining an initial frequency of sequencing data of a plurality of samples to be detected being a specific base at an interested locus;

calculating, based on the initial frequency, an expected value of each of the plurality of samples to be detected being the specific base at the interested locus;

updating, by using each expected value, the initial frequency of the sequencing data of the plurality of samples to be detected being the specific base at the interested locus;

further calculating, by using the updated initial frequency, the expected value of each of the plurality of samples to be detected being the specific base at the interested locus, further updating, by using each new expected value, the initial frequency of the sequencing data of the plurality of samples to be detected being the specific base at the interested locus, and repeating the foregoing iteration until the expected value of each of the plurality of samples to be detected being the specific base at the interested locus converges; and

determining, based on each converging expected value, a base mutation type and a mutation confidence at the interested locus of each of the plurality of samples to be detected,

wherein the specific base comprises an adenine A base, a thymine T base, a cytosine C base, or a guanine G base;

wherein said determining, based on each converging expected value, the base mutation type and the mutation confidence at the interested locus of each of the plurality of samples to be detected comprises:

calculating, based on each converging expected value, a maximum likelihood estimate of the plurality of samples to be detected belonging to each one of four specific base mutation types at the interested locus;

calculating a ratio of the maximum likelihood estimates of two adjacent specific base mutation types;

processing the ratio according to a preset rule to obtain a probability corresponding to the ratio; and

when the probability is smaller than a set threshold, determining the specific base mutation type at the interested locus of each sample to be detected as the specific base mutation type corresponding

to a current denominator,

wherein the four specific base mutation types comprises mono-allelic mutation, bi-allelic mutation, tri-allelic mutation, and tetra-allelic mutation.

2. The method according to claim 1, wherein said determining the initial frequency of the sequencing data of a plurality of samples to be detected being the specific base at the interested locus comprises:

statistically calculating the number of the specific base in the sequencing data of the plurality of samples to be detected and the total number of bases of four types in the sequencing data of the plurality of samples to be detected, wherein the bases of four types comprises the adenine A bases, the thymine T bases, the cytosine C bases, and the guanine G bases; and

determining a quotient of dividing the number of the specific base by the total number of the bases of four types as the initial frequency of the sequencing data of the plurality of samples to be detected being the specific base at the interested locus.

3. The method according to claim 1, wherein said calculating, based on the initial frequency, the expected value of each of the plurality of samples to be detected being the specific base at the interested locus comprises:

calculating the expected value of each of the plurality of samples to be detected being the specific base at the interested locus in accordance with the following equation:

$$p(\mathrm{b}_{i,j} \left| p_j, \mathrm{d}_i \right.) = \frac{p(\mathrm{b}_{i,j} \left| p_j \right.) p(\mathrm{d}_i \left| \mathrm{b}_{i,j} \right.)}{\sum_{\mathrm{b}_{i,j=0}}^{3} p(\mathrm{b}_{i,j} \left| p_j \right.) p(\mathrm{d}_i \left| \mathrm{b}_{i,j} \right.)}$$

where $\mathrm{b}_{i,j}$ represents that a sample i to be detected is a specific base j at an interested locus; p_j represents an initial frequency of the sequencing data of the plurality of samples to be detected being the specific base j at the interested locus; d_i represents a set of bases in base sequences at the interested locus covered by the sample i to be detected; $p(\mathrm{b}_{i,j} \left| p_j, \mathrm{d}_i \right.)$ represents an expected value of the sample i to be detected being the specific base j at the interested locus; $p(\mathrm{b}_{i,j} \left| p_j \right.)$ represents a prior probability of the sample i to be detected being the specific base j at the interested locus under a given p_j; and $p(\mathrm{d}_i \left| \mathrm{b}_{i,j} \right.)$ represents a base quality score of the base sequences at the interested locus covered by the sample i to be detected.

4. The method according to claim 3, wherein said updating, by using each expected value, the

initial frequency of the sequencing data of the plurality of samples to be detected being the specific base at the interested locus comprises:

updating, by using each expected value, the initial frequency of the sequencing data of the plurality of samples to be detected being the specific base at the interested locus in accordance with the following equation:

$$\hat{p}_j = \frac{\sum_{i=1}^{N} p(\mathrm{b}_{i,j} \mid p_j, \mathrm{d}_i)}{n}$$

where $\hat{p}_j$ represents an updated initial frequency of the sequencing data of the plurality of samples to be detected being the specific base at the interested locus; $p(\mathrm{b}_{i,j} \mid p_j, \mathrm{d}_i)$ represents an expected value of the sample i to be detected being the specific base j at the interested locus; N represents the number of the plurality of samples to be detected; and n represents the total number of bases of four types in the sequencing data of the plurality of samples to be detected.

~~5. The method according to claim 1 or 4, wherein said determining, based on each converging expected value, the base mutation type and the mutation confidence at the interested locus of each of the plurality of samples to be detected comprises:~~

~~calculating, based on each converging expected value, a maximum likelihood estimate of the plurality of samples to be detected belonging to each one of four specific base mutation types at the interested locus;~~

~~calculating a ratio of the maximum likelihood estimates of two adjacent specific base mutation types;~~

~~processing the ratio according to a preset rule to obtain a probability corresponding to the ratio; and~~

~~when the probability is smaller than a set threshold, determining the specific base mutation type at the interested locus of each sample to be detected as the specific base mutation type corresponding to a current denominator,~~

~~wherein the four specific base mutation types comprises mono-allelic mutation, bi-allelic mutation, tri-allelic mutation, and tetra-allelic mutation.~~

~~6~~5. The method according to claim ~~5~~4, wherein said calculating, based on each converging expected value, the maximum likelihood estimate of the plurality of samples to be detected belonging to each one of four specific base mutation types at the interested locus comprises:

calculating the maximum likelihood estimate of the plurality of samples to be detected belonging to each one of four specific base mutation types at the interested locus in accordance with

the following equation:

$$p(\mathrm{D}\left|p_j\right.)=\prod_{i=1}^{N}\sum_{j=0}^{3}\left[p(\mathrm{b}_{i,j}\left|p_j\right.)\mathrm{p}(\mathrm{d}_i\left|b_{i,j}\right.)\right]$$

where D represents observation data composed of sets of bases in base sequences at the interested locus covered by all of the plurality of samples to be detected; p_j represents a frequency of the sequencing data of the plurality of samples to be detected being a specific base j at the interested locus, which is obtained based on each converging expected value; $p(\mathrm{D}\left|p_j\right.)$ represents a maximum likelihood estimate of the plurality of samples to be detected being a base mutation type corresponding to j at the interested locus; $p(\mathrm{b}_{i,j}\left|p_j\right.)$ represents a prior probability of the sample i to be detected being the specific base j at the interested locus under a given p_j; and $p(\mathrm{d}_i\left|\mathrm{b}_{i,j}\right.)$ represents a base quality score of a base sequence at the interested locus covered by the sample i; when j=0, the specific base mutation type corresponding to j is the mono-allelic mutation; when j=1, the specific base mutation type corresponding to j is the bi-allelic mutation; when j=2, the specific base mutation type corresponding to j is the tri-allelic mutation; and when j=3 , the specific base mutation type corresponding to j is the tetra-allelic mutation.

~~7~~6. The method according to claim ~~6~~5, wherein said calculating the ratio of the maximum likelihood estimates of the two adjacent specific base mutation types comprises:

determining $\frac{f_3}{f_4}$ as the ratio of the maximum likelihood estimates of the two adjacent specific base mutation types, provided that a maximum likelihood estimate of the interested locus being the tetra-allelic mutation is f_4, and a maximum likelihood estimate of the interested locus being the tri-allelic mutation is f_3.

~~8~~7. The method according to claim ~~6~~5, wherein said calculating the ratio of the maximum likelihood estimates of the two adjacent specific base mutation types comprises:

determining $\frac{f_2}{f_3 \min}$ as the ratio of the maximum likelihood estimates of the two adjacent specific base mutation types, provided that a maximum likelihood estimate of the interested locus being the bi-allelic mutation is f_2, and a minimum value of maximum likelihood estimates of four mutation combinations in the tri-allelic mutations is $f_3 min$.

~~9~~8. The method according to claim ~~6~~5, wherein said calculating the ratio of the maximum likelihood estimates of the two adjacent specific base mutation types comprises:

determining $\frac{f_1}{f_2 \min}$ as the ratio of the maximum likelihood estimates of the two adjacent specific base mutation types, provided that a maximum likelihood estimate of the interested locus being the mono-allelic mutation is f_1, and a minimum value of maximum likelihood estimates of 16 mutation combinations in the bi-allelic mutations is $f_2 min$.

~~10~~9. The method according to any one of claims ~~7~~ 6 to ~~9~~8, wherein said processing the ratio according to the preset rule to obtain the probability corresponding to the ratio comprises:

calculating a natural logarithm of the ratio to obtain a first result;

multiplying the obtained first result by -2 to obtain a second result; and

obtaining a probability corresponding to the second result by looking up a chi-square distribution table.

~~11~~10. The method according to claim ~~9~~8, wherein said determining, based on each converging expected value, the mutation confidence at the interested locus of each sample to be detected comprises:

performing a regular Phred-scale transformation on a probability corresponding to $\frac{f_1}{f_2 \min}$ to obtain a Phred quality score; and

determining the Phred quality score as the mutation confidence at the interested locus of each sample to be detected.

~~12~~11. A base mutation detection apparatus based on sequencing data, the base mutation detection apparatus comprising:

an initial frequency determination module configured to determine an initial frequency of an initial frequency of sequencing data of a plurality of samples to be detected being a specific base at an interested locus;

an expected value calculation module configured to calculate, based on the initial frequency, an expected value of each of the plurality of samples to be detected being the specific base at the interested locus;

an update module configured to update, by using each expected value, the initial frequency of the sequencing data of the plurality of samples to be detected being the specific base at the interested locus;

an iteration module configured to: further calculate, by using the updated initial frequency, the

expected value of each of the plurality of samples to be detected being the specific base at the interested locus, further update, by using each new expected value, the initial frequency of the sequencing data of the plurality of samples to be detected being the specific base at the interested locus, and repeat the foregoing iteration until the expected value of each of the plurality of samples to be detected being the specific base at the interested locus converges; and

a mutation type determination module configured to determine, based on each converging expected value, a base mutation type and a mutation confidence at the interested locus of each of the plurality of samples to be detected~~,~~; wherein the mutation type determination module comprisies:

a maximum likelihood estimate calculation module, configured to calculate, based on each converging expected value, a maximum likelihood estimate of the plurality of samples to be detected belonging to each one of four specific base mutation types at the interested locus;

a ratio calculation module, configured to calculate a ratio of the maximum likelihood estimates of two adjacent specific base mutation types;

a probability calculation module, configured to process the ratio according to a preset rule to obtain a probability corresponding to the ratio; and when the probability is smaller than a set threshold, determining the specific base mutation type at the interested locus of each sample to be detected as the specific base mutation type corresponding to a current denominator, wherein the four specific base mutation types comprises mono-allelic mutation, bi-allelic mutation, tri-allelic mutation, and tetra-allelic mutation;

wherein the specific base comprises an adenine A base, a thymine T base, a cytosine C base, or a guanine G base.

~~13~~12. A storage medium, comprising a computer-executable instruction, wherein the computer-executable instruction, when executed by a computer processor, implements the base mutation detection method based on sequencing data according to any one of claims 1 to ~~11~~10.