

arrangement information of the gene sequencing read of the diseased cell at each site within the preset site interval.

In one possible implementation, the identifying module includes:

a generating sub-module, configured to obtain a feature matrix of the genetic variation candidate site according to the base arrangement features and the non-base arrangement features of the genetic variation candidate site, where first dimension features of the feature matrix correspond to the base arrangement features and the non-base arrangement features of the genetic variation candidate site, and second dimension features of the feature matrix correspond to sites within the preset site interval; and an identifying sub-module, configured to identify genetic variation of the genetic variation candidate site according to the feature matrix of the genetic variation candidate site.

In one possible implementation, the identifying sub-module is specifically configured to:

obtain a variation value for occurrence of genetic variation at the genetic variation candidate site according to the feature matrix of the genetic variation candidate site; and

in the case that the variation value is greater than or equal to a preset threshold, determine that genetic variation exists at the genetic variation candidate site.

In one possible implementation, the generating sub-module is specifically configured to:

generate a feature vector of each first dimension feature of the preset site interval according to the base arrangement features and the non-base arrangement features of the genetic variation candidate site; determine base arrangement feature vectors formed by the base arrangement features in the feature vectors; and randomly rank the base arrangement feature vectors to obtain the feature matrix of the genetic variation candidate site.

In one possible implementation, the first obtaining module includes:

a second obtaining sub-module, configured to obtain gene sequencing reads obtained by performing gene sequencing on a somatic gene; a comparing sub-module, configured to compare base sequences of the gene sequencing reads with a base sequence of the reference genome to obtain a comparison result; a fourth determining sub-module, configured to determine the genetic variation candidate site, at which a gene abnormality exists, of the somatic gene according to the comparison result; and a third obtaining sub-module, configured to obtain at least one gene sequencing read corresponding to the genetic variation candidate site.

A genetic variation identification apparatus is provided according to another aspect of the present disclosure, including: a processor; and a memory configured to store processor-executable instructions, where the processor is configured to execute the foregoing method.

A non-transitory computer-readable storage medium is provided according to another aspect of the present disclosure, having computer program instructions stored thereon, where when the computer program instructions are executed by a processor, the foregoing method is implemented.

According to the genetic variation identification solution provided in embodiments of the present disclosure, at least one gene sequencing read corresponding to a genetic variation candidate site is obtained; base arrangement features of the genetic variation candidate site are obtained; non-base arrangement features of the genetic variation candidate site are determined based on base arrangement information of the at least one gene sequencing read within a preset site interval; and thus genetic variation of the genetic variation candidate site is identified based on the base arrangement features and

obtain a variation value for occurrence of genetic variation at the genetic variation candidate site according to the feature matrix of the genetic variation candidate site; and

in the case that the variation value is greater than or equal to a preset threshold, determine that genetic variation exists at the genetic variation candidate site.

In one possible implementation, the generating sub-module is specifically configured to:

generate a feature vector of each first dimension feature of the preset site interval according to the base arrangement features and the non-base arrangement features of the genetic variation candidate site; determine base arrangement feature vectors formed by the base arrangement features in the feature vectors; and randomly rank the base arrangement feature vectors to obtain the feature matrix of the genetic variation candidate site.

In one possible implementation, the first obtaining module includes:

a second obtaining sub-module, configured to obtain gene sequencing reads obtained by performing gene sequencing on a somatic gene; a comparing sub-module, configured to compare base sequences of the gene sequencing reads with a base sequence of the reference genome to obtain a comparison result; a fourth determining sub-module, configured to determine the genetic variation candidate site, at which a gene abnormality exists, of the somatic gene according to the comparison result; and a third obtaining sub-module, configured to obtain at least one gene sequencing read corresponding to the genetic variation candidate site.

In some embodiments, functions provided by or modules included in the apparatus provided in embodiments of the present disclosure can be used for implementing the method described in the foregoing method embodiments. For specific implementation, refer to the description of the foregoing method embodiments. For the purpose of concision, details are not described here again.

FIG. 8 is a block diagram of a genetic variation identification apparatus 1900 according to one exemplary embodiment. For example, the apparatus 1900 is provided as a server. Referring to FIG. 8, the apparatus 1900 includes a processing component 1922, and further includes one or more processors, and a memory resource as represented by a memory 1932, configured to store instructions executable for the processing component 1922, such as an application program. The application program stored in the memory 1932 includes one or more modules corresponding to a set of instructions. In addition, the processing component 1922 is configured to execute the instructions so as to execute the foregoing method.

The apparatus 1900 further includes a power supply component 1926 configured to execute power supply management of the apparatus 1900, a wired or wireless network interface 1950 configured to connect the apparatus 1900 to a network, and a Input/Output (I/O) interface 1958. The apparatus 1900 can operate an operating system stored in the memory 1932, such as Windows Server™, Mac OS X™, Unix™, Linux™, FreeBSD™ or the like.

In exemplary embodiments, a non-transitory computer-readable storage medium is further provided, for example, the memory 1932 including computer program instructions, which can be executed by the processing component 1922 of the apparatus 1900 to implement the foregoing method.

The present disclosure may be a system, a method, and/or a computer program product. The computer program product may include a computer-readable storage medium having computer-readable program instructions loaded thereon for causing a processor to carry out different aspects of the present disclosure.

The computer-readable storage medium may be a tangible device that can retain and store instructions for use by an instruction execution device. The computer-readable storage medium may be, for example, but is not limited to, an electronic storage device, a magnetic storage device, an optical storage device, an electromagnetic storage device, a semiconductor storage device, or any suitable combination of the foregoing.

CLAIMS

1. A genetic variation identification method, comprising:

obtaining at least one gene sequencing read corresponding to a genetic variation candidate site;

obtaining base arrangement features of the genetic variation candidate site;

determining non-base arrangement features of the genetic variation candidate site based on non-base arrangement information of the at least one gene sequencing read within a preset site interval, wherein the non-base arrangement features remain unchanged after a base arrangement order changes; and

identifying genetic variation of the genetic variation candidate site based on the base arrangement features and the non-base arrangement features of the genetic variation candidate site.

2. The method according to claim 1, wherein obtaining the base arrangement features of the genetic variation candidate site comprises:

determining the preset site interval where the genetic variation candidate site is located; and

obtaining the base arrangement features of the genetic variation candidate site according to base arrangement information of a reference genome within the preset site interval, wherein the base arrangement features are used for representing the base arrangement order.

3. The method according to claim 1 or 2, wherein determining the non-base arrangement features of the genetic variation candidate site based on non-base arrangement information of the at least one gene sequencing read within the preset site interval comprises:

obtaining the non-base arrangement information of the at least one gene sequencing read at each site within the preset site interval; and

determining the non-base arrangement features of the genetic variation candidate site based on the non-base arrangement information at each site within the preset site interval.

4. The method according to claim 3, wherein determining the non-base arrangement features of the genetic variation candidate site based on the non-base arrangement information at each site within the preset site interval comprises:

determining, from the gene sequencing reads, a first gene sequencing read having a consistent base type with the reference genome at the genetic variation candidate site; and

determining the non-base arrangement features of the genetic variation candidate site according to the number of first gene sequencing reads corresponding to each site within the preset site interval.

5. The method according to claim 3, wherein determining the non-base arrangement features of the genetic variation candidate site based on the non-base arrangement information at each site within the preset site interval comprises:

determining, from the gene sequencing reads, a first gene sequencing read having a consistent base type with the reference genome at the genetic variation candidate site;

determining the number of first gene sequencing reads having inconsistent base types with the reference genome at each site within the preset site interval, as the number of first gene sequencing reads to which variation occurs; and

determining the non-base arrangement features of the genetic variation candidate site according to

the number of first gene sequencing reads to which variation occurs.

6. The method according to claim 3, wherein determining the non-base arrangement features of the genetic variation candidate site based on the non-base arrangement information at each site within the preset site interval comprises:

determining, from the gene sequencing reads, a second gene sequencing read having a base type at the genetic variation candidate site consistent with a variant base type of the genetic variation candidate site; and

determining the non-base arrangement features of the genetic variation candidate site according to the number of second gene sequencing reads corresponding to each site within the preset site interval.

7. The method according to claim 3, wherein determining the non-base arrangement features of the genetic variation candidate site based on the non-base arrangement information at each site within the preset site interval comprises:

determining, from the gene sequencing reads, a second gene sequencing read having a base type at the genetic variation candidate site consistent with a variant base type of the genetic variation candidate site;

determining the number of second gene sequencing reads having inconsistent base types with the reference genome at each site within the preset site interval, as the number of second gene sequencing reads to which variation occurs; and

determining the non-base arrangement features of the genetic variation candidate site according to the number of second gene sequencing reads to which variation occurs.

8. The method according to claim 3, wherein determining the non-base arrangement features of the genetic variation candidate site based on the non-base arrangement information at each site within the preset site interval comprises:

determining a third gene sequencing read from the gene sequencing reads, wherein the base type of the third gene sequencing read at the genetic variation candidate site is inconsistent with the base type of the reference genome, and the base type of the third gene sequencing read at the genetic variation candidate site is inconsistent with the variant base type of the genetic variation candidate site; and

determining the non-base arrangement features of the genetic variation candidate site according to the number of third gene sequencing reads corresponding to each site within the preset site interval.

9. The method according to claim 3, wherein determining the non-base arrangement features of the genetic variation candidate site based on the non-base arrangement information at each site within the preset site interval comprises:

determining a third gene sequencing read from the gene sequencing reads, wherein the base type of the third gene sequencing read at the genetic variation candidate site is inconsistent with the base type of the reference genome, and the base type of the third gene sequencing read at the genetic variation candidate site is inconsistent with the variant base type of the genetic variation candidate site;

determining the number of third gene sequencing reads having inconsistent base types with the reference genome at each site within the preset site interval, as the number of third gene sequencing reads to which variation occurs; and

determining the non-base arrangement features of the genetic variation candidate site according to the number of third gene sequencing reads to which variation occurs.

10. The method according to any one of claims 3 to 9, wherein determining the non-base arrangement

features of the genetic variation candidate site based on the non-base arrangement information at each site within the preset site interval comprises:

determining, from the at least one gene sequencing read, a gene sequencing read derived from a normal cell; and

determining the non-base arrangement features of the genetic variation candidate site based on non-base arrangement information of the gene sequencing read of the normal cell at each site within the preset site interval.

11. The method according to any one of claims 3 to 9, wherein determining the non-base arrangement features of the genetic variation candidate site based on the non-base arrangement information at each site within the preset site interval comprises:

determining, from the at least one gene sequencing read, a gene sequencing read derived from a diseased cell; and

determining the non-base arrangement features of the genetic variation candidate site based on non-base arrangement information of the gene sequencing read of the diseased cell at each site within the preset site interval.

12. The method according to any one of claims 1 to 11, wherein identifying the genetic variation of the genetic variation candidate site based on the base arrangement features and the non-base arrangement features of the genetic variation candidate site comprises:

obtaining a feature matrix of the genetic variation candidate site according to the base arrangement features and the non-base arrangement features of the genetic variation candidate site, wherein first dimension features of the feature matrix correspond to the base arrangement features and the non-base arrangement features of the genetic variation candidate site, and second dimension features of the feature matrix correspond to sites within the preset site interval; and

identifying the genetic variation of the genetic variation candidate site according to the feature matrix of the genetic variation candidate site.

13. The method according to claim 12, wherein identifying the genetic variation of the genetic variation candidate site according to the feature matrix of the genetic variation candidate site comprises:

obtaining a variation value for occurrence of genetic variation at the genetic variation candidate site according to the feature matrix of the genetic variation candidate site; and

in the case that the variation value is greater than or equal to a preset threshold, determining that genetic variation exists at the genetic variation candidate site.

14. The method according to claim 12, wherein obtaining the feature matrix of the genetic variation candidate site according to the base arrangement features and the non-base arrangement features of the genetic variation candidate site comprises:

generating a feature vector of each first dimension feature of the preset site interval according to the base arrangement features and the non-base arrangement features of the genetic variation candidate site;

determining base arrangement feature vectors formed by the base arrangement features in the feature vectors; and

randomly ranking the base arrangement feature vectors to obtain the feature matrix of the genetic

variation candidate site.

15. The method according to any one of claims 1 to 14, wherein obtaining at least one gene sequencing read corresponding to the genetic variation candidate site comprises:

obtaining gene sequencing reads obtained by performing gene sequencing on a somatic gene;

comparing base sequences of the gene sequencing reads with a base sequence of the reference genome to obtain a comparison result;

determining the genetic variation candidate site, at which a gene abnormality exists, of the somatic gene according to the comparison result; and

obtaining at least one gene sequencing read corresponding to the genetic variation candidate site.

16. A genetic variation identification apparatus, comprising:

a first obtaining module, configured to obtain at least one gene sequencing read corresponding to a genetic variation candidate site;

a second obtaining module, configured to obtain base arrangement features of the genetic variation candidate site;

a determining module, configured to determine non-base arrangement features of the genetic variation candidate site based on non-base arrangement information of the at least one gene sequencing read within a preset site interval, wherein the non-base arrangement features remain unchanged after a base arrangement order changes; and

an identifying module, configured to identify genetic variation of the genetic variation candidate site based on the base arrangement features and the non-base arrangement features of the genetic variation candidate site.

17. The apparatus according to claim 16, wherein the second obtaining module comprises:

a first determining sub-module, configured to determine the preset site interval where the genetic variation candidate site is located; and

a second determining sub-module, configured to obtain the base arrangement features of the genetic variation candidate site according to base arrangement information of a reference genome within the preset site interval, wherein the base arrangement features are used for representing the base arrangement order.

18. The apparatus according to claim 16 or 17, wherein the determining module comprises:

a first obtaining sub-module, configured to obtain non-base arrangement information of the at least one gene sequencing read at each site within the preset site interval; and

a third determining sub-module, configured to determine the non-base arrangement features of the genetic variation candidate site based on the non-base arrangement information at each site within the preset site interval.

19. A genetic variation identification apparatus, comprising:

a processor; and

a memory configured to store processor-executable instructions, wherein

the processor is configured to execute the method according to any one of claims 1 to 15.

20. A non-transitory computer-readable storage medium, having computer program instructions stored

thereon, wherein when the computer program instructions are executed by a processor, the method according to any one of claims 1 to 15 is implemented.