

GENE SEQUENCING ANALYSIS METHOD AND APPARATUS, STORAGE MEDIUM, AND COMPUTER DEVICE

PRIORITY INFORMATION

[0001] This application claims priority to and the benefit of Patent Application No. 202110698855.6, filed to the China National Intellectual Property Administration on June 23, 2021, which is incorporated herein by reference in its entirety.

FIELD

[0002] The present disclosure relates to the technical field of biological information, and particularly, to a gene sequencing analysis method and apparatus, a storage medium, and computer device.

BACKGROUND

[0003] DNA sequencing technology has been playing a vital role in driving the development of molecular biology since its emergence. This technical field has undergone tremendous changes from early manual sequencing developed by Frederick Sanger and the first-generation automated sequencers based on the Sanger's sequencing, to the current new-generation sequencing platforms.

[0004] The sequencing process of the new generation sequencing platform mainly includes the following steps: during each cycle of sequencing, adding four nucleotides labeled with different fluorophores and DNA polymerases simultaneously into flow-through channels, to extend the DNA chain in accordance with the principle of complementary pairing; acquiring fluorescence images, with base-specific fluorescence labeling revealing which nucleotides are newly added to the chain in this cycle, so as obtain the DNA sequences at this position in the template; and further performing the next cycle of reaction. Such a process is repeated multiple

times to obtain multiple DNA sequences. For example, a DNA sequence including 50 bases can be obtained by performing 50 cycles of sequencing. The single sequence obtained by this method is very short, and thus it is also referred to as a short sequence.

[0005] In this regard, for the new-generation sequencing platforms, it is necessary to complete the biochemical and imaging operations before base recognition, and the recognition results shall be converted into other forms for storage. Thereafter, the stored files are transmitted to a primary storage for performing data quality control, checking sample information, and sequencing information, and data is archived using the secondary storage. Such a sequencing is time-consuming and has a low transmission efficiency, resulting in a relatively slow analysis process.

SUMMARY

[0006] The present disclosure aims to solve at least one of the above-mentioned technical drawbacks, in particular the technical drawbacks of the related art such as time-consuming sequencing and low transmission efficiency, which further leads to the relatively slow analysis process.

[0007] In a first aspect, the present disclosure provides a gene sequencing analysis method. The gene sequencing analysis method includes the following steps of: acquiring slice data transmitted from a sequencing platform in real time, the slice data being a set of short sequences read after performing at least one cycle of sequencing on a sequencing library; inputting the slice data into a memory; and performing, by calling an encapsulation program pre-set in the memory, data processing and detection analysis on the slice data, to obtain a corresponding analysis result, wherein the encapsulation program is a custom program.

[0008] According to an embodiment of the present disclosure, the above-mentioned gene sequencing analysis method may further comprise at least one of the following additional technical features.

[0009] According to an embodiment of the present disclosure, the step of acquiring the slice data includes: determining a slice size; and slicing, based on the slice size, all unsliced short sequences currently transmitted from the sequencing platform and read after a single cycle of

sequencing, a set of the sliced short sequences being used as the slice data.

[0010] According to an embodiment of the present disclosure, the step of determining the slice size includes: acquiring a current central processing unit (CPU) idleness state; and adjusting the slice size based on the current CPU idleness state.

5 **[0011]** According to an embodiment of the present disclosure, the encapsulation program includes a data quality control program, a data processing program, and a detection analysis program. The step of performing, by calling the encapsulation program pre-set in the memory, the data processing and detection analysis on the slice data, to obtain the corresponding analysis result includes: performing, by calling the data quality control program pre-set in the memory,
10 quality control processing on the slice data, to obtain a quality control result; performing, by calling the data processing program pre-set in the memory, data processing on the quality control result, to obtain a processing result; and performing, by calling the detection analysis program pre-set in the memory, detection analysis on the processing result, to obtain a final analysis result.

15 **[0012]** According to an embodiment of the present disclosure, the step of performing, by calling the data quality control program pre-set in the memory, the quality control processing on the slice data includes: by calling the data quality control program pre-set in the memory, calculating a position error rate of short sequences in the slice data, statistically analyzing a base distribution in the slice data, and/or removing low-quality bases from the slice data,
20 wherein the low-quality bases are bases with a base quality below a quality threshold.

[0013] According to an embodiment of the present disclosure, the step of performing, by calling the data processing program pre-set in the memory, the data processing on the quality control result includes: performing, by calling the data processing program pre-set in the memory, at least one of alignment, ordering, deduplication, and re-alignment on a short
25 sequence in the quality control result.

[0014] According to an embodiment of the present disclosure, the step of performing, by calling the data processing program pre-set in the memory, at least one of alignment, ordering, deduplication, and re-alignment on the short sequence in the quality control result includes: performing, by calling the data processing program pre-set in the memory, an alignment
30 between the short sequence in the quality control result and a reference gene, to determine a

position of the short sequence on the reference gene; determining a position tag of the short sequence based on the position of the short sequence on the reference gene; determining, among pre-set threads in the memory, a target thread that is currently in an idle state; and performing, by calling the target thread and with reference to the position tag of the short sequence, at least one of ordering, deduplication, and re-alignment of the short sequence.

[0015] According to an embodiment of the present disclosure, the step of performing, by calling the detection analysis program pre-set in the memory, the detection analysis on the processing result includes: performing, by calling the detection analysis program pre-set in the memory, mutation detection and interpretation on the processing result.

[0016] In a second aspect, the present disclosure further provides a gene sequencing analysis apparatus. The gene sequencing analysis apparatus includes: a data acquisition module configured to acquire slice data transmitted from a sequencing platform in real time, the slice data being a set of short sequences read after performing at least one cycle of sequencing on a sequencing library; a slice transmission module configured to input the slice data into a memory; and a processing analysis module configured to perform, by calling an encapsulation program pre-set in the memory, data processing and detection analysis on the slice data, to obtain a corresponding analysis result, wherein the encapsulation program is a custom program.

[0017] In a third aspect, the present disclosure also provides a storage medium having computer readable instructions stored thereon. The computer readable instructions, when being executed by one or more processors, cause the one or more processors to perform steps of the gene sequencing analysis method as described in any of the above embodiments.

[0018] In a fourth aspect, the present disclosure also provides a computer device having computer readable instructions stored thereon. The computer readable instructions, when being executed by one or more processors, cause the one or more processors to perform steps of the gene sequencing analysis method as described in any of the above embodiments.

[0019] It can be seen from the above technical solutions that the embodiments of the present disclosure have the following advantages.

[0020] In the gene sequencing analysis method and apparatus, the storage medium, and the computer device according to the present disclosure, short sequences transmitted from a sequencing platform in real time are first sliced to perform corresponding slicing processing,

and then slice data is input into a memory. Thus, a pre-loaded encapsulation program in the memory can be called to perform bioinformatic analysis on the slice data, thereby obtaining a corresponding analysis result. Such a process can acquire the slice data of the sequencing of the sequencing platform in real time and perform sequencing and analysis on the slice data, rather than performing the processing analysis only after the sequencing platform has finished the sequencing and the complete sequencing results are transmitted to the corresponding platform. Therefore, the present disclosure enables the whole of sequencing and analysis to be accelerated. In addition, the data for analysis is the slice data, which can be transmitted faster and takes less time compared with the whole sequencing result.

[0021] In addition, the analysis process according to the present disclosure is performed by means of a pre-set encapsulation program in memory. Thus, the sequencing data is unnecessarily transmitted to another platform for processing, which effectively reduces I/O congestion and memory consumption, thereby further improving the operating efficiency of a machine and prolonging the service life of the machine.

DESCRIPTION OF THE DRAWINGS

[0022] To clearly explain the technical solutions in the embodiments of the present disclosure or the related art, the accompanying drawings required for describing the embodiments or the related art are briefly introduced below. The accompanying drawings in the following description merely illustrate some embodiments of the present disclosure. Those skilled in the art can derive other drawings from these accompanying drawings without paying creative efforts.

[0023] FIG. 1 is a schematic flowchart of a gene sequencing analysis method according to an embodiment of the present disclosure;

[0024] FIG. 2 is a schematic diagram of a flow analysis process according to an embodiment of the present disclosure;

[0025] FIG. 3 is a schematic diagram of a gene sequencing analysis process according to an embodiment of the present disclosure;

[0026] FIG. 4 is a schematic diagram illustrating a sequencing process in correlation with

time in a gene sequencing analysis process according to an embodiment of the present disclosure;

[0027] FIG. 5 is a schematic structural diagram of a gene sequencing analysis apparatus according to an embodiment of the present disclosure; and

5 **[0028]** FIG. 6 is a schematic diagram of an internal structure of a computer device according to an embodiment of the present disclosure.

DETAILED DESCRIPTION

[0029] Technical solutions in the embodiments of the present disclosure are clearly described in detail below with reference to the accompanying drawings of the embodiments of
10 the present disclosure. The embodiments described below are merely part of rather than all the embodiments of the present disclosure. Other embodiments, which are obtained by those skilled in the art based on the embodiments in the present disclosure without paying creative efforts, shall fall within the protection scope of the present disclosure.

[0030] In short, sequencing is the conversion from DNA chemical signals into computer-
15 processable digital signals. From early manual sequencing developed by Frederick Sanger and the first-generation automated sequencers based on the Sanger sequencing, to the current new-generation sequencing platforms, the field of sequencing has undergone tremendous changes. The changes and breakthrough in sequencing technology play a great promotion role in the genomics research, disease medical research, drug research, breeding, and other fields.

20 **[0031]** At present, the sequencing process of the new generation sequencing platform mainly includes the following steps: during each cycle of sequencing, adding four nucleotides labeled with different fluorophores and DNA polymerases simultaneously into flow-through channels, to extend the DNA chain in accordance with the principle of complementary pairing; acquiring fluorescence images, with base-specific fluorescence labeling revealing which
25 nucleotides are newly added to the chain in this cycle, so as obtain the DNA sequences at this position in the template; and further performing the next cycle of reaction. Such a process is repeated multiple times to obtain multiple DNA sequences. For example, a DNA sequence including 50 bases can be obtained by performing 50 cycles of sequencing. The single sequence

obtained by this method is very short, and thus it is also referred to as a short sequence.

[0032] In view of the above, for the new-generation sequencing platforms, it is necessary to complete the biochemical and imaging operations before base recognition, and the recognition results shall be converted into other forms for storage. Thereafter, the stored files
5 are transmitted to a primary storage for performing data quality control, checking sample information, and sequencing information, and data is archived using the secondary storage. Such a sequencing is time-consuming and has a low transmission efficiency, resulting in a relatively slow analysis process.

[0033] Therefore, the present disclosure aims to solve the technical problems of the slow
10 analysis process caused by long time consumption and low transmission efficiency of the sequencing in the related art. The present disclosure provides the following technical solutions.

[0034] FIG. 1 is a schematic flowchart of a gene sequencing analysis method according to an embodiment of the present disclosure. As illustrated in FIG. 1, the present disclosure provides a gene sequencing analysis method, which specifically includes the following steps.

[0035] S110: slice data is acquired.
15

[0036] In this step, before performing bioinformatic analysis, the slice data transmitted from the sequencing platform is required to be acquired, in order to perform the corresponding bioinformatic analysis based on the slice data.

[0037] It can be understood that, in the present disclosure, since the analysis efficiency of
20 the subsequent platforms and the occupation situation of the memory, etc. are taken into consideration, the real-time incoming data from the sequencing platform is required to be sliced, and the slice size can be adaptively adjusted based on the idle degree of the server. Therefore, in the present disclosure, the acquired slice data is a set of short sequences read after performing at least one cycle of sequencing on a sequencing library and transmitted from the sequencing
25 platform in real time.

[0038] It should be noted that, the adaptive adjustment of the slice size based on the idle degree of the server may mainly include the following situations. If the server is not idle, the slice size can be adjusted to be able to slice short sequences that can be read after one cycle of sequencing. If the server is idle, the slice size can be adjusted to be able to slice short sequences
30 that can be read after multiple cycles of sequencing. Further, the first slicing may occur after

multiple cycles of sequencing have been performed. For example, after 10 cycles of sequencing, one slice is made by slicing.

[0039] As an example, in the present disclosure, a next-generation sequencing platform may be employed to acquire the slice data. Next-Generation Sequencing (NGS) and Sanger sequencing are the same in that a computer is used to detect the fluorescent signals generated at the time of binding the fluorophore-labelled dNTPs to a DNA template with catalyst of DNA polymerase in each cycle of sequencing. However, unlike Sanger sequencing in which a single fragment is detected per unit time, NGS can detect signals from thousands of channels simultaneously, thereby greatly improving efficiency.

[0040] It can be understood that the next-generation sequencing platforms discussed herein include, but are not limited to, Illumina (Solexa) sequencing, Roche 454 sequencing, Ion torrent: Proton/PGM sequencing, and SOLiD sequencing.

[0041] As an example, the following explains how the next-generation sequencing platform obtains the slice data. Illustratively, the next-generation sequencing platforms in the present disclosure require a construction of a sequencing library before sequencing. The sequencing library refers to a DNA mixture of DNA fragments having specific DNA adapter sequences connected at both ends thereof. For example, a mixture of DNA fragments is obtained by disrupting genomic DNA with ultrasonics, end-filling the disrupted DNA fragments with enzymes, and ligating the adapter sequences by means of ligases, and such a mixture of DNA fragments is referred to as a "library".

[0042] Before the sequencing library is sequenced using the next-generation sequencing platform, PCR amplification may be performed on the sequencing library, in order to amplify the small amount of DNA fragments in the solution to be sequenced by several times or even dozens of times. In this way, a distribution density of the DNA fragments to be sequenced in the solution can be increased, enabling the small amount of DNA fragments to be acquired during sampling.

[0043] During the sequencing, a neutral solution may be added, and sequencing dNTP primers may be added to the neutral solution. Since 3'-end of dNTP is blocked by an azide group, the template chain can only be extended by one base during one cycle of sequencing. After one cycle of sequencing is completed, the azide group and fluorophore are cleaved by

adding a specific chemical reagent, to expose the hydroxyl group at the 3'-end of dNTP, and thus the base sequence can be read.

[0044] In the present disclosure, the biochemical and imaging parts of each cycle of sequencing in the sequencing process are transmitted to the memory in real time for base calling (Basecall) to obtain the corresponding short sequence, and the slice data, when satisfying the slice size, is output and stored in the memory.

[0045] For example, a fastq file is output after sequencing with the existing next-generation sequencing platform. The fastq file is a sequence file containing a quality value, where q indicates quality, for storing the original sequencing data, with an extension name of fastq or fq. The following is a common format of sequence in fastq files:

```
@DJB775P1:248:D0MDGACXX:7:1202:12362:49613
TGCTTACTCTGCGTTGATACCACTGCTTAGATCGGAAGAGCACACGTCTGAA
+
JJJJJJJJJJJHHHHGHFFFFFFFCEEEEEEDBD?DDDDDDDBDDABDDCA
@DJB775P1:248:D0MDGACXX:7:1202:12782:49716
CTCTGCGTTGATACCACTGCTTACTCTGCGTTGATACCACTGCTTAGATCGG
+
IIIIIIIIIIHHHHHHFFFFFFFECCCCBCECCCCCCCCCCCCCCCC
```

[0046] In view of the above format of sequence, every four rows in the fastq file are regarded as an independent unit, i.e., referring to as a short sequence. The first row begins with “@”, followed by DJB775P1 indicating the name of the short sequence, and the character string of the first row without space therebetween is converted according to the state information during sequencing and is a unique identifier of the short sequence. The second row indicates a sequence of sequencing read, i.e., the DNA sequence of real interest. The sequence is composed of five letters of A, C, G, T, and N, where N represents a bases that cannot be identified during the sequencing. The third row begins with “+”, and may repeat the information of the first row in the previous version of the FASTQ file. Currently, no information is added to the third row for saving storage space. The fourth row indicates the quality value of the sequencing read, which is as important as the base information of the second row, and the quality value of the sequencing read represents the reliability of each sequenced base, expressed by the ASCII code.

[0047] In the present disclosure, in order to improve the analysis efficiency, the sequencing and analysis processes are subjected to slicing, and a stream structure is used for sequential processing to simultaneously perform sequencing, encoding, transmission, and analysis. The result obtained in the biochemistry and imaging parts of each cycle of sequencing in the sequencing process is first transmitted to the memory for image recognition processing of the basecall (base calling) process, to obtain a plurality of short sequences. Thereafter, the slice size is determined based on the idle degree of the server. That is, the bases in the above-mentioned second row are sliced based on the idle degree of the server. If the server is idle, the bases are sliced as one piece of slice data. If the server is not idle, a single base is sliced as one piece of slice data. The slice data obtained after the slicing is saved in the memory.

[0048] When reading the slice data in the memory for analysis, the adapter can be first removed. The adapter is located at the beginning of the first slice, generally 6 to 8 bases, and the adapter data can be used to distinguish different samples, which is convenient for later output. Then, the remaining slice data is aligned on the human genome. Each base of each new piece of slice data enables a range of reads to be located on the human genome to continuously narrowed, and thus the alignment position is more accurate.

[0049] In addition, in the present disclosure, during the sequencing and analysis, the used platform may be a machine integrating sequencing and analysis, or it may be a machine connecting an existing sequencer and an analyzer through a network connection for data transmission and operation analysis, which is not limited herein.

[0050] S120: the slice data is input into memory.

[0051] In this step, after the slice data is acquired in step S110, the slice data can be input into a memory of the integration machine or a memory of an analyzer in connection with a sequencer. Thus, the data processing and detection analysis can be performed on the slice data by means of an encapsulation program in the memory.

[0052] It can be understood that, after the sequencing library is sequenced with an existing sequencer, the obtained sequencing result is generally output as a fastq file and stored in a hard disk. When performing alignment analysis, the fastq file is read from the hard disk, and aligned to the human genome (reference. fa). In contrast, in the present disclosure, the acquired slice data is directly transmitted to the memory for performing the alignment processing, without

being output and transmitted to a hard disk, thereby reducing the processes of writing to and reading from the hard disk and reducing I/O consumption.

[0053] Additionally, it should be noted that the memory may be a Double Data Rate (DDR) memory. Due to the large gene data and the limited server memory, it is often extended to the DDR memory. The data is respectively transmitted at a rising edge and a falling edge of clock signal in the DDR memory, leading to that a data transmission speed of the DDR memory is twice that of a conventional SDRAM. Moreover, the energy consumption is not increased, as only the falling edge signal is additionally adopted. The addressing and control signals are identical to those of the conventional SDRAM and are transmitted only on the rising edge of clock signal.

[0054] S130: data processing and detection analysis is performed on the slice data by calling an encapsulation program in memory, to obtain an analysis result.

[0055] In this step, when processing the slice data, the pre-set encapsulation program in memory needs to be called to the perform data processing and detection analysis on the slice data, and a corresponding analysis result is obtained.

[0056] It should be noted that the encapsulation program herein refers to as a software encapsulating a plurality of programs in one program, and the encapsulation program is a custom program, which may encapsulate a plurality of different programs in one program according to user's requirements, for performing the data processing and detection analysis on the slice data.

[0057] In addition, the encapsulation program herein is optimized based on the existing internationally recognized program, thereby increasing memory cohesion, increasing multithreading, and adding the processing functions of slice data. During the data processing by using the encapsulation program, the corresponding encapsulation program in a disk can be enabled and loaded into the memory as a task process, thereby effectively reducing I/O consumption, reducing reading and writing time, and improving analysis efficiency.

[0058] For example, it is necessary to perform the data processing on the slice data before the slice data is subjected to the detection analysis, and the data processing process can be implemented by one or more programs, such as a data quality control program, a data processing program, etc. Thereafter, a detection analysis program is connected to perform the

detection analysis on a result obtained after the data processing. In this way, the final analysis result can be more accurate.

[0059] FIG. 2 is a schematic diagram of a flow analysis process according to an embodiment of the present disclosure. As illustrated in FIG. 2,

5 **[0060]** a sequencer can use the biochemical and imaging system to perform biochemical imaging on the sequencing library, and transmit the images obtained after the biochemical imaging to a central control server in a slice transmission manner. The central control server can transmit the images to an analysis instrument for data processing and detection analysis. The data quality control process includes that, a server memory in a data quality control server
10 obtains the sample and sequencing information in a production information system and the slice data transmitted by the central control server; the data quality control is performed on the slice data, followed by data archiving and delivery; and finally, the analysis result is stored, thereby realizing the flow analysis process.

[0061] It can be seen from FIG. 2 that, the set of short sequences transmitted from the
15 sequencing platform is sliced, a part of the analysis time is hided within the original sequencing time, and the data processing process and the detection analysis process are linked to each other via the memory. In this way, the flow analysis process used in the present disclosure can effectively reduce I/O consumption, shorten reading and writing time, and improve analysis efficiency.

20 **[0062]** With the gene sequencing analysis method provided in the above-mentioned embodiment, short sequences transmitted from a sequencing platform in real time are first sliced to perform corresponding slicing processing, and then the slice data is input into a memory. Thus, a pre-loaded encapsulation program in the memory can be called to perform bioinformatic analysis on the slice data, thereby obtaining a corresponding analysis result. Such a process can
25 acquire the slice data of the sequencing of the sequencing platform in real time and perform sequencing and analysis on the slice data, rather than performing the processing analysis only after the sequencing platform has finished the sequencing and the complete sequencing results are transmitted to the corresponding platform. Therefore, the present disclosure enables the whole of sequencing and analysis to be accelerated. In addition, the data for analysis is the slice
30 data, which can be transmitted faster and takes less time compared with the whole sequencing

result.

[0063] In addition, the analysis process according to the present disclosure is performed by means of a pre-set encapsulation program in memory. Thus, the sequencing data is unnecessarily transmitted to another platform for processing, which effectively reduces I/O congestion and memory consumption, thereby further improving the operating efficiency of a machine and prolonging the service life of the machine.

[0064] The gene sequencing analysis method of the present disclosure will be further explained by means of the following embodiments. As described below, the process of acquiring slice data is described in detail in the following embodiments.

[0065] In an embodiment, the step of acquiring slice data in step S110 may include the following steps:

[0066] S111: a slice size is determined.

[0067] S112: based on the slice size, all unsliced short sequences, which are currently transmitted from the sequencing platform and read after a single cycle of sequencing, are sliced, and the sliced short sequences are used as the slice data.

[0068] In this embodiment, the biochemical and imaging parts in the sequencing process of each cycle are transmitted to the memory in real time for base recognition. After the corresponding short sequences are obtained, it is required to determine the slice size first, and the short sequences satisfying the slice size, as a combination, is output as the slice data, and the slice data is stored in the memory.

[0069] For example, every $m + n$ cycles (cycles of sequencing) are taken as one piece of slice data, where m is the first slice, $m \geq 10$, n refers to all the subsequent slices, $n \geq 1$ and $n \leq 20$, and $m + n$ can adaptively adjust the slice size according to the analysis speed.

[0070] FIG. 3 is a schematic diagram of a gene sequencing analysis process according to an embodiment of the present disclosure, and FIG. 4 is a schematic diagram of a sequencing process in correlation with time in a gene sequencing analysis process according to an embodiment of the present disclosure. As illustrated in FIG. 3 and FIG. 4, in the present disclosure, by acquiring the slice data from the sequencing on a sequencing platform in real time and performing the flow sequencing and analysis on the slice data, the process of sequencing and analysis can be accelerated. Thereafter, a mutation detection or an alignment to

an RNA pathogen library, and then subsequent annotation are performed, thereby ending the analysis process, and completing the delivery.

[0071] The process of acquiring the slice data is described in detail in the above embodiment. The step of determining the slice size in the above embodiment will be further described below.

[0072] In an embodiment, the step of determining the slice size in step S111 may include the following steps.

[0073] S1111: a current central processing unit (CPU) idleness state is acquired.

[0074] S1112: the slice size is adjusted based on the current CPU idleness state.

[0075] In the present embodiment, when the slice data is acquired for detection analysis, the slice data is a set of short sequences transmitted from a sequencing platform in real time and read after performing at least one cycle of sequencing on gene fragments in a sequencing library.

[0076] The slice size can be adaptively adjusted according to the analysis speed, and the analysis speed can be obtained by acquiring and analyzing the current CPU idleness state. The current CPU idleness state is related to a current occupation of the memory. More current occupation of the memory indicates that the memory is not idle, and thus the analysis speed of the memory is relatively slow. Therefore, the slice size can be adjusted to be able to slice a set of short sequences that are read after multiple cycles of sequencing. Less current occupation of the memory indicates that the memory is idle, and thus the analysis speed of the memory is relative fast. Therefore, the slice size can be adjusted to be able to slice a set of short sequences that are read after one cycle of sequencing.

[0077] For example, every $m + n$ cycles (cycles of sequencing) are taken as one piece of slice data, and the slice after m slices is automatically controlled. That is, a size of the slice n is determined based on the idleness state of the machine from the beginning of the second slice to the end of the last slice. In the machine is relatively idle, the transmission is performed once after each cycle, i.e., $n = 1$; and if the machine is relatively busy, the transmission is performed after 20 cycle, i.e., $n = 20$.

[0078] It can be understood that, according to the alignment principle in the present disclosure, the slice data is transmitted to the alignment process and aligned with the human

genome, and the alignment range of the slice data is determined. If the slice data is larger, the alignment range is reduced more accurately.

[0079] The steps of determining the slice size are further explained in the above embodiments, and the encapsulation program and the process of calling the encapsulation program for data processing and detection analysis will be described in detail in the following embodiments.

[0080] In an embodiment, the encapsulation program may include a data quality control program, a data processing program, and a detection analysis program.

[0081] In step S130, the step of performing, by calling the encapsulation program pre-set in the memory, the data processing and detection analysis on the slice data includes the following steps.

[0082] S131: quality control processing is performed on the slice data by calling the data quality control program pre-set in the memory, to obtain a quality control result.

[0083] S132: data processing is performed on the quality control result by calling the data processing program pre-set in the memory, to obtain a processing result.

[0084] S133: detection analysis is performed on the processing result by calling the detection analysis program pre-set in the memory, to obtain a final analysis result.

[0085] In this embodiment, when processing the slice data, the pre-set encapsulation program in memory may be called to perform the data processing and detection analysis on the slice data, to obtain the corresponding analysis result.

[0086] The encapsulation program pre-set in the memory may be a custom program. The custom program may be one program encapsulated by multiple programs. For example, the encapsulation program may include a data quality control program, a data processing program, and a detection analysis program.

[0087] The data quality control program is mainly used to perform the quality control processing on the slice data, to obtain the quality control result. The data processing program is mainly used to perform the data processing on the quality control result, to obtain the processing result. The detection analysis program is used to perform the detection analysis on the processing result, to obtain the final analysis result.

[0088] In the above-mentioned embodiments, the encapsulation program, and the process

of performing the data processing and detection analysis by calling the encapsulation program are described in detail. The process of performing the quality control processing by calling the data quality control program is described in detail below.

[0089] In an embodiment, the step S131 of performing, by calling the data quality control program pre-set in the memory, the quality control processing on the slice data may include the following steps: by calling the data quality control program pre-set in the memory, calculating a position error rate of short sequence in the slice data, statistically analyzing a base distribution in the slice data, and/or removing low-quality bases from the slice data. The low-quality bases are bases with a base quality below a quality threshold.

[0090] In this embodiment, when performing the quality control processing on the slice data, the data quality control program in an encapsulation program is called for performing the quality control processing.

[0091] Specifically, the data quality control program is mainly used to perform the quality control processing on the slice data, and the quality control processing may include: calculating the position error rate of the short sequence in the slice data, statistically analyzing the base distribution in the slice data, removing the low-quality bases from the slice data, and other quality control processing applicable to the short sequences. One or more quality control processing and tools to be used in the quality control processing can be specifically selected according to actual conditions.

[0092] For example, when the low-quality bases need to be removed from the slice data, the average quality of the bases in a window with a certain length can be calculated by sliding the window. If the average quality is too low, all bases are directly removed.

[0093] In the above-mentioned embodiment, the process of performing the quality control processing by calling the data quality control program is described in detail. The process of performing the data processing by calling the data processing program is described in detail by way of embodiments.

[0094] In an embodiment, the step S132 of performing, by calling the data processing program pre-set in the memory, the data processing on the quality control result may include: performing, by calling the data processing program pre-set in the memory, at least one of alignment, ordering, deduplication, and re-alignment on the short sequence in the quality

control result.

[0095] In this embodiment, after the quality control processing is performed on the slice data, the data processing program pre-set in the memory may be called to perform the alignment, ordering, deduplication, and/or re-alignment on the short sequences in the quality control result
5 obtained after the quality control processing.

[0096] Although the slice data are all from a genome having a sequence, different short sequences in the slice data has completely lost their sequential relation after DNA library construction and sequencing. Therefore, two adjacent short sequences in the slice data are random short sequences at certain positions in the original genome, and the positional relation
10 between two adjacent short sequences in the slice data is unknown.

[0097] Therefore, in the present disclosure, after performing the quality control processing on the slice data, the data processing program pre-set in memory may be called, to align the short sequences in the slice data corresponding to the quality control result with the reference gene of the corresponding species one by one, thereby finding the position of each short
15 sequence on the reference gene. Thereafter, the short sequences are arranged in an order. Such a process is referred to as an alignment process of the slice data.

[0098] It should be noted that the reference gene described herein refers to the genomic sequence of the species, i.e., the assembled complete genomic sequence, which is often used as a standard reference for the species.

[0099] After the alignment of the short sequences is completed, the short sequences are subjected to a sorting procedure. The sorting procedure is mainly to sort the previously aligned short sequences according to the order of their positions, to form a continuous position relation
20 of the short sequences.

[00100] After the ordering procedure, a deduplication operation may be performed on the
25 short sequences in the slice data, in order to remove the repeated short sequences in the slice data, thereby facilitating the subsequent detection analysis.

[00101] Finally, the re-alignment process of short sequences is mainly to re-correct the regions of potential sequence insertions or deletions, which have been found in the above-mentioned alignment process, thereby allowing the result of subsequent detection analysis to
30 be more accurate.

[00102] It can be understood that the above-mentioned data processing processes of alignment, ordering, deduplication, and re-alignment may be selectively set according to actual situations, and the sequential order may also be manually set, which are not specifically limited herein.

5 **[00103]** For example, the normal analysis procedures are alignment, ordering, deduplication, re-alignment, mutation detection, and annotation. During the alignment, the input file is a fastq file, and the output file is a bam file. The output is continuously a bam file until the re-alignment. The output file of the mutation detection is a vcf file, and the annotation is based on the result of the mutation detection. When the gene sequencing analysis method of the present disclosure and the next-generation sequencing platform for sequencing are used, the alignment is required
10 for the analysis processes of all raw data of the next-generation fastq, followed by subsequent processes or not. The bam file at any stage can also be input for subsequent analysis processing when not starting from the raw data.

[00104] The alignment process of the present disclosure will be further described below. For
15 example, in the alignment of the slice data according to the present disclosure, the bases of the short sequences in the slice data can be aligned with the whole genome to find the position of each short sequence on the reference gene. After all the slice data are positioned at specific positions on the whole genome, the matching alignment information is integrated to perform screening and extension on adjacent alignment positions. In addition, gaps (i.e., the positions
20 where the alignment fails) are allowable, and thus a longer sequence chain aligned to the whole genome can be obtained.

[00105] Furthermore, in the above-mentioned alignment process, a dynamic programming algorithm integrating global and local alignment situations can be used to select the optimal alignment result information, and then the alignment result of different samples is respectively
25 output to the memory according to the adapter information.

[00106] For example, the common BWA-MEM algorithm mainly uses a seed-and-extend strategy. In a seed stage, BWA takes the base fragment of reads to perform an accurate mapping on reference and selects the read fragment satisfying certain requirements on matching times and length as seed. This stage algorithm focuses on the accurate matching of FM-index. In an
30 extend stage, BWA uses the Smith-Waterman algorithm to extend the seed on reads and

reference toward both sides for the alignment (gap-tolerant), thereby finding the eligible global match of the entire read on reference.

[00107] The above embodiments mainly describe how the data processing program is called to perform the data processing procedure, and specific embodiments of the data processing will be described further below.

[00108] In an embodiment, the step of performing, by calling the data processing program pre-set in the memory, at least one of alignment, ordering, deduplication, and re-alignment on the short sequence in the quality control result may include the following steps.

[00109] A11: an alignment between the short sequence in the quality control result and a reference gene is performed by calling the data processing program pre-set in the memory, to determine a position of the short sequence on the reference gene

[00110] A12: a position tag of the short sequence is determined based on the position of the short sequence on the reference gene.

[00111] A13: among pre-set threads in the memory, a target thread that is currently in an idle state is determined.

[00112] A14: at least one of ordering, deduplication, and re-alignment of the short sequence is performed by calling the target thread and with reference to the position tag of the short sequence.

[00113] In this embodiment, when performing the data processing on the short sequence in a quality control result, the position tag can be added to the aligned short sequence; and then it is determined whether there is a thread in an idle state among the respective pre-set threads in the current memory. If there is a thread in the idle state, this thread serves as the target thread, and with reference to the position tag of the short sequence, the target thread is called to perform the ordering, deduplication, and/or re-alignment on the short sequence. In this way, the ordering and deduplication are more efficient.

[00114] For example, after the short sequence in the quality control result is aligned with the reference gene by calling the data processing program and the position label corresponding to the short sequence is determined based on the position of the short sequence on the reference gene, the target thread can be called to perform operations such as ordering and deduplication on the short sequence. For example, in the ordering process, when multi-thread is used for

processing, the ordering can be performed with reference to the position tag of the short sequence, in order to improve the processing efficiency of multi-thread. In this way, the ordering result obtained after multi-thread ordering is accurate and the ordering efficiency is higher.

[00115] The process of the data processing is further explained in the above embodiments, and the process of performing the detection analysis on the processing result by calling the detection analysis program is described in detail below.

[00116] In an embodiment of the present disclosure, the step S133 of performing, by calling the detection analysis program pre-set in the memory, the detection analysis on the processing result may include: performing, by calling the detection analysis program pre-set in the memory, mutation detection and interpretation on the processing result.

[00117] In this embodiment, one goal of performing the gene sequencing analysis is to obtain an accurate mutation set of the sample, which needs to be determined by mutation detection, interpretation, etc. Currently, the results of mutation detection include SNP, Indel, CNV, and SV. The process of mutation detection may use algorithms applying Bayesian inference or HaplotypeCaller algorithm.

[00118] For example, when the HaplotypeCaller algorithm is employed for the mutation detection, the haploid combination of the population will be inferred, a probability of each combination is calculated, and then the genotype combination of each sample will be deduced according to this information.

[00119] Furthermore, when performing the detection analysis such as the mutation detection on the processing result, the mutation detection can be performed without waiting for the completion of alignment. Different chromosomes in the whole genome can be divided into several regions. When a part of the alignment results is accumulated for each region, the high-mutation region can be found for detection. The continuously verified mutation situation of the existing high-mutation region, including mismatch/insertion/deletion and the like is added into the subsequent alignment result. After an assembly of the data of the high-mutation region, the simplified haplotype data can be obtained. The maximum likelihood estimation of haplotype is estimated by using Hidden Markov Model (HMM) to obtain the typing result of each site in this region, and the mutation information is output.

[00120] It can be understood that, in order to improve the analysis efficiency of the whole

bioinformatic analysis, when performing the mutation detection on the processing result in the present disclosure, the mutation detection is performed on the high-mutation region rather than on all the sites in the whole genome. The high-mutation region can be selected based on the probability of mutation occurrence at each site. For example, after calculating the probability of mutation occurrence at each site, the high-mutation region can be determined based on a predetermined probability threshold.

[00121] Furthermore, during the assembly of the data of the high-mutation region, if there is a duplication in the set of fragments corresponding to the reference genome, the length of the short fragment will be increased progressively until no duplication exists or reaching the maximum length limit, and the edge obtained by the assembly can be assigned with a weight according to the number of short sequences obtained by alignment.

[00122] Still further, after obtaining the mutation detection result, the mutation detection result may be subjected to a quality control and filtering, in order to further judge the quality of the mutation detection result.

[00123] The gene sequencing analysis apparatus provided in the embodiments of the present disclosure is described below. The gene sequencing analysis apparatus described below can be correspondingly referred to the gene sequencing analysis method described above

[00124] FIG. 5 is a schematic structural diagram of a gene sequencing analysis apparatus according to an embodiment of the present disclosure. In an embodiment, as illustrated in FIG. 5, the present disclosure also provides a gene sequencing analysis apparatus, which includes a data acquisition module 210, a slice transmission module 220, and a processing analysis module 230.

[00125] The data acquisition module 210 is configured to acquire slice data transmitted from a sequencing platform in real time, the slice data being a set of short sequences read after performing at least one cycle of sequencing on a sequencing library.

[00126] The slice transmission module 220 is configured to input the slice data into a memory.

[00127] The processing analysis module 230 is configured to perform, by calling an encapsulation program pre-set in the memory, data processing and detection analysis on the slice data, to obtain a corresponding analysis result, wherein the encapsulation program is a

custom program.

[00128] With the gene sequencing analysis apparatus provided in the above-mentioned embodiment, short sequences transmitted from a sequencing platform in real time are first sliced to perform corresponding slicing processing, and then the slice data is input into a memory.

5 Thus, a pre-loaded encapsulation program in the memory can be called to perform bioinformatic analysis on the slice data, thereby obtaining a corresponding analysis result. Such a process can acquire the slice data of the sequencing of the sequencing platform in real time and perform sequencing and analysis on the slice data, rather than performing the processing analysis only after the sequencing platform has finished the sequencing and the complete sequencing results
10 are transmitted to the corresponding platform. Therefore, the present disclosure enables the whole of sequencing and analysis to be accelerated. In addition, the data for analysis is the slice data, which can be transmitted faster and takes less time compared with the whole sequencing result.

[00129] In addition, the analysis process according to the present disclosure is performed by
15 means of a pre-set encapsulation program in memory. Thus, the sequencing data is unnecessarily transmitted to another platform for processing, which effectively reduces I/O congestion and memory consumption, thereby further improving the operating efficiency of a machine and prolonging the service life of the machine.

[00130] In an embodiment, the data acquisition module 210 may include a determination
20 module 211 configured to determine a slice size, and a slicing module 212 configured to slice, based on the slice size, all unsliced short sequences currently transmitted from the sequencing platform and read after a single cycle of sequencing, a set of the sliced short sequences being used as the slice data.

[00131] In an embodiment, the determination module 211 may include a memory condition
25 acquisition module 2111 configured to acquire a current idleness state of the memory, and a slice adjustment module 2112 configured to adjust the slice size based on the current idleness state of the memory.

[00132] In an embodiment, the encapsulation program in the processing analysis module 230
30 may include a data quality control program, a data processing program, and a detection analysis program.

[00133] The processing analysis module 230, which is configured to perform, by calling an encapsulation program pre-set in the memory, data processing and detection analysis on the slice data, to obtain a corresponding analysis result, may include: a data quality control module 231 configured to perform, by calling the data quality control program pre-set in the memory, quality control processing on the slice data, to obtain a quality control result; a data processing module 232 configured to perform, by calling the data processing program pre-set in the memory, data processing on the quality control result, to obtain a processing result; and a detection analysis module 233 configured to perform, by calling the detection analysis program pre-set in the memory, detection analysis on the processing result, to obtain a final analysis result.

[00134] In an embodiment, the data quality control module 231 may include a first processing module 2311. The first processing module 2311 is configured to, by calling the data quality control program pre-set in the memory, calculate a position error rate of short sequence in the slice data, statistically analyze a base distribution in the slice data, and/or remove low-quality bases from the slice data. The low-quality bases are bases with a base quality below a quality threshold.

[00135] In an embodiment, the data processing module 232 may include a second processing module 2321. The second processing module 2321 is configured to perform, by calling the data processing program pre-set in the memory, at least one of alignment, ordering, deduplication, and re-alignment on the short sequence in the quality control result.

[00136] In an embodiment, the second processing module 2321 may include an alignment module 310, a determination tag module 311, a determining thread module 312, and a multi-thread operation module 313.

[00137] The alignment module 310 is configured to perform, by calling the data processing program pre-set in the memory, an alignment between the short sequence in the quality control result and a reference gene, to determine a position of the short sequence on the reference gene.

[00138] The determination tag module 311 is configured to determine a position tag of the short sequence based on the position of the short sequence on the reference gene.

[00139] The determining thread module 312 is configured to determine, among pre-set threads in the memory, a target thread that is currently in an idle state.

[00140] The multi-thread operation module 313 is configured to perform, by calling the target thread and with reference to the position tag of the short sequence, at least one of ordering, deduplication, and re-alignment of the short sequence.

[00141] In an embodiment, the detection analysis module 233 may include a third processing module 2331. The third processing module 2331 is configured to perform, by calling the detection analysis program pre-set in the memory, mutation detection and interpretation on the processing result.

[00142] In an embodiment, the present disclosure also provides a storage medium having computer readable instructions stored thereon. The computer readable instructions, when being executed by one or more processors, cause the one or more processors to perform steps of the gene sequencing analysis method as described in any of the above embodiments.

[00143] In an embodiment, the present disclosure also provides computer device having computer readable instructions stored thereon. The computer readable instructions, when being executed by one or more processors, cause the one or more processors to perform steps of the gene sequencing analysis method as described in any of the above embodiments.

[00144] FIG. 6 is a schematic diagram of an internal structure of a computer device according to an embodiment of the present disclosure. The computer device 300 may be provided as a server. Referring to FIG. 6, the computer device 300 includes: a processing component 302, which further includes one or more processors; and a memory resource represented by memory 301 and configured to store an instruction such as an application program, which is executable by the processing component 302. The application program stored in the memory 301 may include one or more modules, each of which corresponds to a set of instructions. Further, the processing component 302 is configured to execute an instruction to perform the gene sequencing analysis method as described in any of the embodiments described above.

[00145] The computer device 300 may further include a power supply component 303 configured to perform power management of the computer device 300, a wired or wireless network interface 304 configured to connect the computer device 300 to a network, and an input/output (I/O) interface 305. The computer device 300 may run an operating system stored in memory 301, such as Windows Server TM, Mac OS X TM, Unix TM, Linux TM, Free

BSDTM, or the like.

[00146] It will be appreciated by those skilled in the art that the structure illustrated in FIG. 6 is merely a block diagram of a part of the structure relevant to the solution of the present disclosure and does not constitute a limitation on the computer device to which the solution of the present disclosure is applied. The specific computer device may include components than those illustrated in the figures, or combine some components, or have a different arrangement of components.

[00147] Finally, it should be noted that in this specification, relational terms such as “first” and “second” are used only to differentiate an entity or operation from another entity or operation, and these terms do not necessarily require or imply that any actual relation or sequence exists between these entities or operations. Furthermore, the terms “include”, “comprise”, or any other mutation thereof, are intended to cover a non-exclusive inclusion, such that a process, method, article, or equipment that includes a list of elements may further include other elements that are not expressly listed or inherent to this process, method, article, or equipment. Without more limitations, the elements defined by the expression “including a...” do not exclude the existence of other identical elements in the process, method, article, or equipment including the elements.

[00148] Various embodiments are described in this specification in a progressive manner, with each embodiment focusing on differences from the other embodiments, and these embodiments may be combined as needed, with the same and similar parts referring to each other.

[00149] The above embodiments of the present disclosure are provided to enable those skilled in the art to implement or practice the present disclosure. Various modifications to these embodiments are readily apparent to those skilled in the art, and the generic principles defined herein can be implemented in other embodiments without departing from the spirit or scope of the present disclosure. Therefore, the present disclosure is not limited to the embodiments illustrated herein, but shall fall within the widest scope consistent with the principles and novel features disclosed herein.

CLAIMS

What is claimed is:

1. A gene sequencing analysis method, comprising:

acquiring slice data transmitted from a sequencing platform in real time, the slice data
5 being a set of short sequences read after performing at least one cycle of sequencing on a sequencing library;

inputting the slice data into a memory; and

performing, by calling an encapsulation program pre-set in the memory, data processing
and detection analysis on the slice data, to obtain a corresponding analysis result, wherein the
10 encapsulation program is a custom program.

2. The gene sequencing analysis method according to claim 1, wherein said acquiring the slice data comprises:

determining a slice size;

slicing, based on the slice size, all unsliced short sequences currently transmitted from the
15 sequencing platform and read after a single cycle of sequencing, a set of the sliced short sequences being used as the slice data.

3. The gene sequencing analysis method according to claim 2, wherein said determining the slice size comprises:

acquiring a current central processing unit (CPU) idleness state; and

20 adjusting the slice size based on the current CPU idleness state.

4. The gene sequencing analysis method according to claim 1, wherein the encapsulation program comprises a data quality control program, a data processing program, and a detection analysis program, and

wherein said performing, by calling the encapsulation program pre-set in the memory, the
25 data processing and detection analysis on the slice data, to obtain the corresponding analysis result comprises:

performing, by calling the data quality control program pre-set in the memory, quality control processing on the slice data, to obtain a quality control result;

performing, by calling the data processing program pre-set in the memory, data processing on the quality control result, to obtain a processing result; and

performing, by calling the detection analysis program pre-set in the memory, detection analysis on the processing result, to obtain a final analysis result.

5 5. The gene sequencing analysis method according to claim 4, wherein said performing, by calling the data quality control program pre-set in the memory, the quality control processing on the slice data comprises:

by calling the data quality control program pre-set in the memory, calculating a position error rate of short sequences in the slice data, statistically analyzing a base distribution in the
10 slice data, and/or removing low-quality bases from the slice data, wherein the low-quality bases are bases with a base quality below a quality threshold.

6. The gene sequencing analysis method according to claim 4, wherein said performing, by calling the data processing program pre-set in the memory, the data processing on the quality control result comprises:

15 performing, by calling the data processing program pre-set in the memory, at least one of alignment, ordering, deduplication, re-alignment, and base quality value correction on a short sequence in the quality control result.

7. The gene sequencing analysis method according to claim 6, wherein said performing, by calling the data processing program pre-set in the memory, at least one of alignment, ordering,
20 deduplication, and re-alignment on the short sequence in the quality control result comprises:

performing, by calling the data processing program pre-set in the memory, an alignment between the short sequence in the quality control result and a reference gene, to determine a position of the short sequence on the reference gene;

determining a position tag of the short sequence based on the position of the short sequence
25 on the reference gene;

determining, among pre-set threads in the memory, a target thread that is currently in an idle state; and

performing, by calling the target thread and with reference to the position tag of the short sequence, at least one of ordering, deduplication, and re-alignment of the short sequence.

30 8. The gene sequencing analysis method according to claim 4, wherein said performing,

by calling the detection analysis program pre-set in the memory, the detection analysis on the processing result comprises:

performing, by calling the detection analysis program pre-set in the memory, mutation detection and interpretation on the processing result.

5 9. A gene sequencing analysis apparatus, comprising:

a data acquisition module configured to acquire slice data transmitted from a sequencing platform in real time, the slice data being a set of short sequences read after performing at least one cycle of sequencing on a sequencing library;

a slice transmission module configured to input the slice data into a memory; and

10 a processing analysis module configured to perform, by calling an encapsulation program pre-set in the memory, data processing and detection analysis on the slice data, to obtain a corresponding analysis result, wherein the encapsulation program is a custom program.

15 10. A storage medium, having computer readable instructions stored thereon, wherein the computer readable instructions, when being executed by one or more processors, cause the one or more processors to perform steps of the gene sequencing analysis method according to any one of claims 1 to 8.

20 11. A computer device, having computer readable instructions stored thereon, wherein the computer readable instructions, when being executed by one or more processors, cause the one or more processors to perform steps of the gene sequencing analysis method according to any one of claims 1 to 8.