

METHOD AND DEVICE FOR ANALYZING GENES

TECHNICAL FIELD

The present invention relates to a method and device for analyzing genes, and more particularly, to a method and device for analyzing data for translocation genes.

BACKGROUND ART

A genome refers to the entirety of genetic information of an organism. For sequencing of the genome of an individual, various technologies such as a DNA chip and next generation sequencing technology, a next-next-generation sequencing technology, and the like are being developed. Analysis of genetic information, such as nucleic acid sequences, protein information, and the like is used to detect genes expressing diseases such as diabetes and cancer or to examine the relationship between genetic diversity and expression characteristics of individuals. In particular, genetic data collected from individuals are important to verify individuals' genetic characteristics related to different symptoms or progression of a disease. Thus, genetic data such as individual's nucleic acid sequences, protein, and the like are core data to prevent diseases by examining current and further disease-related information or to select an optimal treatment in an initial stage of a disease. Recently, due to advances in sequencing technologies, attempts on discovering various types of structure variations have been actively conducted. However, a considerable amount of false positive or false negative results is still produced, which reflects that there are still many challenges remaining in bioinformatics.

DETAILED DESCRIPTION OF THE INVENTION

TECHNICAL PROBLEM

One or more embodiments of the present invention provide a method and device for analyzing genes. The present invention is not limited to the aforementioned technical goals of these embodiments and other technical goals may be inferred from the following embodiments.

TECHNICAL SOLUTION

According to an aspect of an embodiment, a method of analyzing a gene includes acquiring data for split reads and discordantly aligned paired-end (PE) reads from

next-generation sequencing (NGS) data of a sample of a subject, extracting first candidate gene pairs having a possibility of translocation in a chromosome of the sample by using the split reads and the PE reads, and identifying translocation genes from among the first candidate gene pairs, based on breakpoints of the split reads and a fusion direction of the first candidate gene pairs.

In addition, the identifying includes extracting, from among the extracted first candidate gene pairs, second candidate gene pairs comprising genes in which a plurality of split reads having breakpoints within the same coverage is aligned, and the translocation genes are identified from the extracted second candidate gene pairs.

In addition, in the genes included in the extracted second candidate gene pairs, the number of the split reads having breakpoints within the same coverage may be a predetermined threshold value or greater.

In addition, the identifying includes extracting, from among the extracted second candidate gene pairs, third candidate gene pairs in which the direction of fusion between different genes goes from 5' end to 3' end or from 3' end to 5' end, and the translocation genes are identified from the extracted second candidate gene pairs.

In addition, the NGS data includes data in a binary version of sequence alignment/map (SAM) (BAM) format or in an SAM format.

In addition, the acquiring includes acquiring, from the data in a BAM format or in an SAM format, data of an FLAG and a compact idiosyncratic gapped alignment report (CIGAR) string for each of the split reads and the PE reads.

In addition, the NGS data are generated by targeted sequencing for identifying base sequences of target genes from the sample.

In addition, the sample is a biopsy sample or a formalin-fixed, paraffin-embedded (FFPE) sample.

According to an aspect of another embodiment, there is provided a computer-readable recording medium recording a program for executing the method in a computer.

According to an aspect of another embodiment, an apparatus for analyzing a gene includes a read analyzing unit to acquire, from next-generation sequencing (NGS) data of a sample of a subject, data for split reads and discordantly aligned paired-end (PE) reads and a translocation identifying unit to extract first candidate gene pairs having a possibility of translocation in a chromosome of the sample by using the split reads and the PE reads and

to identify translocation genes from among the first candidate gene pairs, based on breakpoints of the split reads and a fusion direction of the first candidate gene pairs.

In addition, the translocation identifying unit extracts, from among the extracted first candidate gene pairs, second candidate gene pairs comprising genes in which a plurality of split reads having breakpoints within the same coverage is aligned, and the translocation genes are identified from the extracted second candidate gene pairs.

In addition, in the genes included in the extracted second candidate gene pairs, the number of the split reads having breakpoints within the same coverage may be a predetermined threshold value or greater.

In addition, the translocating identifying unit extracts, from among the extracted second candidate gene pairs, third candidate gene pairs in which the direction of fusion between different genes goes from 5' end to 3' end or from 3' end to 5' end, and the translocation genes are identified from the extracted second candidate gene pairs.

In addition, the NGS data include data in a binary version of sequence alignment/map (SAM) (BAM) or in an SAM format.

In addition, the read analyzing unit acquires, from the data in a BAM format or in an SAM format, data of an FLAG and a compact idiosyncratic gapped alignment report (CIGAR) string for each of the split reads and the PE reads.

In addition, the NGS data are generated by targeted sequencing for identifying base sequences of target genes from the sample.

In addition, the sample is a biopsy sample or a formalin-fixed, paraffin-embedded (FFPE) sample.

ADVANTAGEOUS EFFECTS OF THE INVENTION

According to one or more embodiments of the present invention, the presence of translocation genes may be more accurately analyzed from genes extracted from a sample of a subject.

DESCRIPTION OF THE DRAWINGS

FIG. 1 is a view for explaining a gene analysis apparatus according to an embodiment.

FIG 2 is a block diagram illustrating hardware elements of a gene analysis apparatus according to an embodiment.

FIG. 3 is a view for explaining paired-end (PE) reads, according to an embodiment.

FIG. 4 is a view for explaining discordantly aligned PE reads, according to an embodiment.

FIG. 5 is a view for explaining split reads, according to an embodiment.

5 FIG. 6 is a view for explaining an integrative genomics viewer (IGV) screenshot for the comparison between reads collected from a biopsy sample of a subject and reference genetic data, according to an embodiment.

FIG. 7 is a view for explaining an IGV screenshot for the comparison between reads collected from a formalin-fixed, paraffin-embedded (FFPE) sample of a subject and
10 reference genetic data, according to an embodiment;

FIG. 8 is a flowchart for a method of identifying translocation genes by extracting candidate gene pairs from a translocation identifying unit, according to an embodiment.

FIG. 9 is a view for explaining a process of extracting second candidate gene pairs by using breakpoints of split reads, according to an embodiment.

15 FIG. 10 is a view for explaining a process of extracting third candidate gene pairs by using adequacy of a fusion direction, according to an embodiment.

FIG. 11 is a view for explaining identification results of translocation genes of echinoderm microtubule-associated protein-like 4 (EML4) and anaplastic lymphoma kinase (ALK), according to an embodiment.

20 FIG. 12 is a flowchart for a gene analysis method according to an embodiment.

FIG. 13 is a block diagram illustrating hardware elements of a computing device according to an embodiment.

BEST MODE

25 According to an aspect, a method of analyzing a gene includes acquiring data for split reads and discordantly aligned paired-end (PE) reads from next-generation sequencing (NGS) data of a sample of a subject, extracting first candidate gene pairs having a possibility of translocation in a chromosome of the sample by using the split reads and the PE reads, and identifying translocation genes from among the first candidate gene pairs,
30 based on breakpoints of the split reads and a fusion direction of the first candidate gene pairs.

According to another aspect, there is provided a computer-readable recording medium recording a program for executing the method in a computer.

According to another aspect, an apparatus for analyzing a gene includes a read analyzing unit to acquire, from next-generation sequencing (NGS) data of a sample of a subject, data for split reads and discordantly aligned paired-end (PE) reads and a translocation identifying unit to extract first candidate gene pairs having a possibility of translocation in a chromosome of the sample by using the split reads and the PE reads and to identify translocation genes from among the first candidate gene pairs, based on breakpoints of the split reads and a fusion direction of the first candidate gene pairs.

MODE OF THE INVENTION

General terminologies used currently and widely are selected as terminologies used in the present embodiments in consideration of functions of the present embodiments. However, the terminologies may vary according to intents or precedents of one of ordinary skill in the art, emergence of novel technologies, and the like. In addition, there are terminologies arbitrarily selected for particular cases, for which detailed meanings are explained in the description of the corresponding embodiment. Hence, the terminologies used in the present embodiments have to be defined based on not the names of the terminologies, but the meanings of the terminologies and contents throughout the present embodiments.

Throughout the description of embodiments, when it is described that a certain part is connected to another part, it should be understood that the certain part may be directly connected to the other part or electrically connected to the other part via another element in the middle. In addition, when a certain part includes another part, this does not exclude any other parts but means that any other parts can be further included unless otherwise stated herein. In addition, the terms "unit" and "module" described in embodiments refer to a unit processing at least one function or operation and can be realized as hardware or software or a combination of hardware and software.

The terms "comprising", "including", and the like as used in the present embodiments should not be interpreted as necessarily including all the elements or all the processes described in the specification, but should be interpreted as not including some of the elements or the processes or further including additional elements or processes.

The description of the following embodiments should not be interpreted as limiting the scope of the present invention, and embodiments that may be inferred by one of ordinary skill in the art without undue difficulty should be interpreted as pertaining to the

scope of embodiments. Hereinafter, embodiments will be described in detail with reference to the accompanying drawings and are provided only for illustrative purposes.

FIG. 1 is a view for explaining a gene analysis apparatus 10 according to an embodiment.

Referring to FIG. 1, the gene analysis apparatus 10 may identify, by using reference genetic data 20 of a normal group and genetic data to be analyzed 30 acquired from a biological sample of a subject, the presence of a translocation gene in genes to be analyzed of the sample of a subject.

The genetic data 30 received in the gene analysis apparatus 10 may be next generation sequencing (NGS) data acquired by NGS, and the NGS data may include genetic data in a BAM (binary version of sequence alignment/map (SAM)) format or an SAM format. The BAM format or the SAM format may be generally used as a format describing data for short reads. A file in a BAM format or a SAM format may include text data for starting points of reads, read directions, mapping quality, FLAG representing an order of alignment, a compact idiosyncratic gapped alignment report (CIGAR) string, and the like. In this regard, the FLAG may be an identifier for identifying an alignment pair from among a primary alignment-primary alignment pair, a primary alignment-secondary alignment pair, a secondary alignment-primary alignment pair, and a secondary alignment-secondary alignment pair. By generating various alignment pairs, various supporting reads may be secured.

The reference genetic data 20 may be acquired from database (DB) known in the art, such as National Center for Biotechnology Information (NCBI), Gene Expression Omnibus (GEO), or the like, or may be acquired from biological samples of humans gathered to analyze genes of a subject.

Meanwhile, reference genes included in the reference genetic data 20 or genes to be analyzed that are included in the genetic data 30 may be obtained from biopsy tissues, or formalin-fixed, paraffin-embedded tissues.

Translocation refers to a phenomenon in which a chromosome is partially cleaved and the cleaved fragment thereof is bound to another portion of the same chromosome or to another chromosome and ultimately means structure variations of chromosomes.

The gene analysis apparatus 10 may determine the presence of translocation genes in the genetic data 30 acquired from a sample of a subject, with respect to the reference genetic data 20 acquired from a normal group. In this regard, genes to be analyzed by the

gene analysis apparatus 10 may mean nucleic acids such as deoxyribonucleic acids (DNAs), ribonucleic acids (RNAs), and the like.

In the present embodiments, the normal group means a group of normal humans having no particular disease, for example, cancer, tumor, or the like, and the subject may mean a patient with a particular disease such as cancer, tumor, or the like. In addition, in the present embodiments, the normal group and the subject may include other animals in addition to humans.

The gene analysis apparatus 10 may be embodied as at least one processor for data processing to execute various commanders and various algorithms for identifying translocation genes by analyzing the genetic data 20 and 30.

FIG 2 is a block diagram illustrating hardware elements of a gene analysis apparatus 10 according to an embodiment.

Referring to FIG. 2, the gene analysis apparatus 10 may include a read analyzing unit 110 and a translocation identifying unit 120. Meanwhile, the gene analysis apparatus 10 of FIG. 2 is illustrated as including only elements related to the present embodiment in order to prevent ambiguity of the characteristics of the present embodiment, and thus the gene analysis apparatus 10 may further include other general-purpose elements in addition to the elements illustrated in FIG. 2.

The read analyzing unit 110 acquires data for split reads and discordantly aligned paired-end (PE) reads from NGS data included in the genetic data 30 described above with reference to FIG. 1.

The NGS data included in the genetic data 30 are data in a BAM format or an SAM format, and the read analyzing unit 110 may acquire, from the data in a BAM format or an SAM format, text data for a starting point of each of the split reads and the PE reads, read directions, mapping quality, an FLAG representing an order of alignment, a CIGAR string, and the like.

In general, NGS technologies such as whole genome sequencing (WGS), whole exome sequencing, and the like are known as a sequencing technology for analyzing base sequences of genes from a sample to be analyzed. In the present embodiment, however, the NGS data may be produced by targeted sequencing for identifying base sequences of some target genes, not the entire genome, from the sample.

The sample to be analyzed may be a biopsy sample obtained from a subject or a formalin-fixed, paraffin-embedded (FFPE) sample.

The translocation identifying unit 120 extracts first candidate gene pairs having the possibility of translocation in a chromosome of the sample of a subject by using the split reads and the discordantly aligned PE reads. Genes sequenced with split reads or genes sequenced with discordantly aligned PE reads may be regarded as a candidate having a high possibility in which base sequences of a part of the genes differ from reference genes (genes of a normal person).

The translocation identifying unit 120 identifies translocation genes from among the first candidate gene pairs, based on breakpoints of split reads and a fusion direction of the first candidate gene pairs.

More particularly, the translocation identifying unit 120 may extract, from the first candidate gene pairs, second candidate gene pairs including genes in which a plurality of split reads having breakpoints within the same coverage is aligned. That is, translocation genes may be identified from the second candidate gene pairs listed in a narrower range than the first candidate gene pairs. In the genes included in the second candidate gene pairs, the number of the split reads having breakpoints within the same coverage may be a predetermined threshold value or greater. The coverage refers to an error range of breakpoints that may be regarded as the same breakpoint, in consideration of sequencing errors. For example, when the predetermined threshold value is 3, genes in which three or more split reads having break points within the same coverage are aligned may be included in the second candidate gene pairs. The predetermined threshold value may be variously changed.

In addition, the translocation identifying unit 120 may extract, from the second candidate gene pairs, third candidate gene pairs in which fusion between different genes goes in the 5' to 3' end direction or in the 3' to 5' end direction. That is, translocation genes may be identified from the third candidate gene pairs listed in a narrower range than the second candidate gene pairs. For example, a fusion gene in which the 3' end of gene A and the 3' end of gene B are fused may be an insignificant fusion gene having no biological expression. Thus, the translocation identifying unit 120 extracts the third candidate gene pairs by filtering gene pairs having an inadequate fusion direction from among the second candidate gene pairs, in consideration of a fusion direction.

Lastly, the translocation identifying unit 120 may determine a gene pair included in the third candidate gene pairs as translocation genes.

FIG. 3 is a view for explaining PE reads, according to an embodiment.

Referring to FIG. 3, PE sequencing means sequencing of each of opposite ends of a gene to be analyzed of a sample of a subject. In the present embodiments, it has been described that target sequencing is performed, and thus it may be assumed that a nucleic acid 300, e.g., DNA, RNA, or the like of the sample to be sequenced has a size of 500 bp (base pair). In a case in which the size of reads is set to be 100 bp, PE reads 310 and 320 may be produced by sequencing from opposite ends of the nucleic acid 300. Since the size of reads is smaller than that of the nucleic acid 300, separate reads may not be produced for the remaining parts of the nucleic acid 300. Meanwhile, in the present embodiment, the PE reads 310 and 320 may be obtained through the PE sequencing together with sequencing of an EXON 305 and an intron. A reason for the use of the PE sequencing will now be described in more detail with reference to FIG. 4.

FIG. 4 is a view for explaining discordantly aligned PE reads, according to an embodiment.

A translocation gene may be obtained by combination of different genes in the same chromosome or combination (fusion) of different genes in different chromosomes. That is, base sequences of the translocation gene on the 5' and 3' ends are derived from different genes. Thus, it will be obvious that base sequences of PE reads 410 and 420 differ from base sequences of reference genes of a normal person.

Referring to FIG. 4, when the PE read 410 is mapped to a gene 401 existing in chromosome 2 and the PE read 420 is mapped to another gene 402 existing in chromosome 3, based on the base sequences of reference genes of a normal person, it may be inferred that each of the PE reads 410 and 420 may be derived from a fusion gene (translocation gene) existing in the sample of a subject. Such reads may be defined as discordantly aligned PE reads 410 and 420. The gene pair 401 and 402 to which the discordantly aligned PE reads 410 and 420 are mapped may be included in the first candidate gene pairs described above.

FIG. 5 is a view for explaining a split read 510, according to an embodiment.

Referring to FIG. 5, the split read 510 refers to a read having a base sequence that is identical to a part of a base sequence of a reference gene and inconsistent with the remaining part thereof. When a read corresponding to a gene of a sample of a subject has a base sequence identical to that of the corresponding reference gene, it may be regarded that there is no structural variation in the base sequence of the gene of a sample of a subject. However, as in the present embodiment, when the split read 510 has a base sequence that is

identical to only a part of a base sequence of a reference gene and inconsistent with the remaining part thereof, it may be inferred that the base sequence of the gene of a sample of a subject differs from the base sequence of the reference gene. Thus, genes 501 and 502 mapped in the split read 510 may be included in the first candidate gene pairs described above.

For example, when a CIGAR string of a split read 511 is 75M25S, the split read 511 has 75 base sequences identical to gene A 501 and the remaining 25 base sequences inconsistent with gene A 501. In addition, when a CIGAR string of a split read 512 is 80M20S, the split read 512 has 80 base sequences consistent with the gene A 501 and the remaining 20 base sequences inconsistent with the gene A 501. The read analyzing unit 110 (see FIG. 1) may acquire data of such split reads from data in a BAM or SAM format.

FIG. 6 is a view for explaining an integrative genomics viewer (IGV) screenshot 600 for the comparison between reads acquired from a biopsy sample of a subject and reference genetic data, according to an embodiment.

Referring to FIG. 6, in the IGV screenshot 600, reads 610 identical to base sequences of a reference gene (e.g., anaplastic lymphoma kinase (ALK) gene) are represented by a gray color. Meanwhile, reads 620 inconsistent with the base sequences of the reference gene are represented by various colors instead of the gray color. That is, the reads 620 represented by various colors have a high possibility of corresponding to, for example, discordantly aligned PE reads or split reads. Thus, the read analyzing unit 110 (see FIG. 1) acquires data for the reads 620 having different base sequences from those of the reference gene.

FIG. 7 is a view for explaining an IGV screenshot 700 for the comparison between reads acquired from a formalin-fixed, paraffin-embedded (FFPE) sample of a subject and reference genetic data, according to an embodiment.

In the IGV screenshot 700 illustrated in FIG. 7, the position of a chromosome of the reference gene (e.g., ALK gene) is similar to that of a chromosome of the IGV screenshot 600 illustrated in FIG. 6.

However, unlike what has been described above with reference to FIG. 6, the IGV screenshot 700 is more colorful than the IGV screenshot 600 (see FIG. 6). This means that reads (e.g., PE reads and split reads) having base sequences inconsistent with a base sequence of the reference gene exists in a greater amount than that in FIG. 6. This is because, in the case of FIG. 7, the reference gene is obtained from an FFPE sample. A

biopsy sample has a short lifespan and thus FFPE is an essential treatment to maintain biochemical properties of the biopsy sample for a long period of time. Unlike a biopsy sample, an FFPE sample undergoes chemical variation and structural variation by FFPE treatment, and thus inconsistent reads exist more in FIG. 7 than in FIG. 6. As a result, a plurality of false-positive or false-negative determinations may be included in gene analysis results. According to the present embodiments, however, even though the reference gene and the gene of the sample of a subject is obtained from a biopsy sample or an FFPE sample, false-positive or false-negative determinations may be eliminated or reduced. This will be described below in further detail.

FIG. 8 is a flowchart for a method of identifying translocation genes by extracting candidate gene pairs from a translocation identifying unit, according to an embodiment.

In operation 801, the translocation identifying unit 120 extracts first candidate gene pairs having a possibility of translocation by using data for split reads and discordantly aligned PE reads acquired by the read analyzing unit 110. For example, the translocation identifying unit 120 may extract first candidate gene pairs from various combinations of genes mapped to the split reads and genes mapped to the discordantly aligned PE reads by using data for reads having a base sequence identical to that of the reference gene in the IGV screenshot 600 or 700 illustrated in FIG. 6 or 7.

In operation 802, the translocation identifying unit 120 extracts second candidate gene pairs including genes in which a plurality of split reads having breakpoints within the same coverage is aligned. As described above, when the number of the split reads having breakpoints within the same coverage is a predetermined threshold value or more, genes mapped to the split reads may be regarded as having a higher possibility of having an actual breakpoint of a translocation gene. Thus, the translocation identifying unit 120 selects as second candidate gene pairs, from the first candidate gene pairs, the genes in which a plurality of split reads having breakpoints within the same coverage is aligned. That is, the genes included in the second candidate gene pairs may have a higher possibility of being a translocation gene than that of the genes included in the first candidate gene pairs.

In operation 803, the translocation identifying unit 120 extracts third candidate gene pairs in which fusion between different genes goes in the 5' to 3' end direction or in the 3' to 5' end direction. Even though there are gene pairs expected to be a fusion gene (translocation gene), included in the second candidate gene pairs, the gene pairs are not

regarded as a translocation gene if the fusion direction is inadequate. Thus, the translocation identifying unit 120 determines whether the direction of fusion between different genes is adequate, i.e., whether fusion between gene pairs included in the second candidate gene pairs adequately goes in the 5' to 3' end direction or in the 3' to 5' end direction, and filters the gene pairs, a fusion direction of which is adequate, as third candidate gene pairs,. That is, genes included in the third candidate gene pairs may have a higher possibility of being a translocation gene than that of the genes included in the second candidate gene pairs.

In operation 804, the translocation identifying unit 120 identifies gene pairs included in the third candidate gene pairs as a translocation gene, when the third candidate gene pairs are extracted.

FIG. 9 is a view for explaining a process of extracting second candidate gene pairs by using breakpoints of split reads 910, according to an embodiment.

Referring to FIG. 9, the split reads 910 may be mapped to a gene of a sample of a subject, i.e., gene X 900. The gene X 900 with the split reads 910 mapped therein may be included in the first candidate gene pairs. Data for a breakpoint 920 may be mapped for each of the split reads 910. The split reads 910 may exist due to various causes such as sequencing errors, gene insertion, gene deletion, and the like and thus, even though the split reads 910 are mapped to the gene X 900, the gene X 900 cannot be determined as corresponding to a part of a translocation gene.

However, in a case in which split reads 930 having breakpoints 940 within the same coverage among the split reads 910 mapped to the gene X 900 are mapped, it may be regarded that the gene X 900 has a high possibility of having a breakpoint of a translocation gene. Thus, when the number of the split reads 930 having breakpoints 940 within the same coverage is a predetermined threshold value or more, the gene X 900 is identified as being highly possible to correspond to a part of a translocation gene. That is, the gene X 900 may be extracted as genes included in the second candidate gene pairs.

Meanwhile, actual breakpoints of the split reads 930 may not completely match with each other due to various causes such as sequencing errors and the like. Thus, the translocation identifying unit 120 may determine whether breakpoints exist within a certain range (i.e., coverage), rather than determine whether breakpoints are the same value.

FIG. 10 is a view for explaining a process of extracting third candidate gene pairs by using adequacy of a fusion direction, according to an embodiment.

Referring to FIG. 10, when a gene X 1010 on a chromosome 2 1001 and a gene Y 1020 on a chromosome 3 1002 are included in the second candidate gene pairs, the translocation identifying unit 120 may determine the direction of fusion between the gene X 1010 and the gene Y 1020.

5 In the chromosome 2 1001 and the chromosome 3 1002, fusion genes 1030 and 1040 are formed such that the 3' end of the gene X 1010 is linked to the 5' end of the gene Y 1020, and thus the translocation identifying unit 120 may determine that the fusion direction of the fusion gene 1030 is adequate.

Among the second candidate gene pairs, the fusion genes 1030 and 1040, the fusion
10 direction of which is adequate, correspond to the third candidate gene pairs, and the translocation identifying unit 120 ultimately determines the gene pairs included in the third candidate gene pairs as a translocation gene.

In the FFPE sample illustrated in FIG. 7, even though false-positive reads are extracted, actual translocation genes may be identified by eliminating or reducing the
15 false-positive determination through the breakpoint and fusion direction determination described above.

FIG. 11 is a view for explaining identification results of translocation genes of echinoderm microtubule-associated protein-like 4 (EML4) and anaplastic lymphoma kinase (ALK), according to an embodiment.

20 In FIG. 11, data for the translocation genes of EML4-ALK, identified through the methods of analyzing translocation genes described above, are illustrated. An IGV screenshot 1101 on the right side shows reads mapped to EML4, and an IGV screenshot 1102 on the left side shows reads mapped to ALK. Therefrom, it is confirmed that the reads mapped to EML4 are split in a breakpoint coverage of 42536701 to 42559688, and
25 the reads mapped to ALK are split in a breakpoint coverage of 29415639 to 29446500. In addition, the number of supporting reads for identification of the translocation genes of EML4-ALK is 39. FIG. 11 illustrates only simulation results obtained by verifying identification results of translocation genes by applying the gene analysis methods described in the present embodiments to a sample of an actual patient, and thus the present
30 embodiments are not limited by the embodiment of FIG. 11.

FIG. 12 is a flowchart for a gene analysis method according to an embodiment. Referring to FIG. 12, the gene analysis method includes operations time-serially performed in the gene analysis apparatus 10 described in the foregoing drawings. Thus, even though

the operations are not entirely described below, the operations described above with reference to the drawings may also be applied to the gene analysis method of FIG. 12.

In operation 1201, the read analyzing unit 110 acquires data for split reads and discordantly aligned PE reads from NGS data.

5 In operation 1202, the translocation identifying unit 120 extracts first candidate gene pairs having a possibility of translocation in a chromosome of a sample of a subject by using the split reads and the discordantly aligned PE reads.

In operation 1203, the translocation identifying unit 120 identifies translocation genes from among the first candidate gene pairs, based on breakpoints of the split reads
10 and a fusion direction of the first candidate gene pairs.

FIG. 13 is a block diagram illustrating hardware elements of a computing device 1 according to an embodiment.

Referring to FIG. 13, the computing device 1 includes a gene analysis apparatus (processor) 10, a data interface 11, and a memory 12. However, the computing device 1 of
15 FIG. 13 is illustrated as including only elements related to the present embodiment in order to prevent ambiguity of the characteristics of the present embodiment, and thus the computing device 1 may further include other general-purpose elements.

The data interface 11 receives the reference genetic data 20 of a normal group and the genetic data 30 of a sample of a subject, as described above in FIG. 1. That is, the data
20 interface 11 may be realized as a hardware of a wired/wireless network interface to allow the computing device 1 to communicate with external other devices. The data interface 11 transmits the received genetic data 20 and 30 to the gene analysis apparatus (processor) 10.

The data interface 11 may receive the genetic data 30 of a subject from an external NGS device for sequencing genes to be analyzed of a subject, a microarray, or the like.

25 The memory 12 is a hardware for storing data to be processed in the computing device 1 and processing-completed results, and non-limiting examples thereof include memory chips such as a random access memory (RAM) and a read-only memory (ROM), and storages such as a hard disk drive (HDD) and a solid state drive (SSD). That is, the memory 12 may store the genetic data 20 and 30 received by the data interface 11 and also
30 store data for the first to third candidate gene pairs processed by the gene analysis apparatus (processor) 10, data for the identified translocation genes, and the like.

The gene analysis apparatus (processor) 10 which is a module including one or more processing units may be embodied as a combination of a microprocessor having an

array of a plurality of logic gates and a memory module in which a program capable of being executed in the microprocessor is stored. The gene analysis apparatus (processor) 10 may be embodied as a module form of an application program. The gene analysis apparatus (processor) 10 is a hardware device to process the analysis of genes described
5 above with reference to FIGS. 1 to 12.

Information on the translocation genes identified by the gene analysis apparatus (processor) 10 may be transmitted, via the data interface 11, to external other devices, e.g., a display device, another computing device, or the like, or onto an external network, e.g., the Internet or an open database (DB) server.

10 According to the aforementioned embodiments, translocation genes may be detected from a cancer tissue of a subject (e.g., a patient with cancer). In addition, even though genes (genes to be analyzed) of a cancer tissue (a sample) obtained from a subject are chemically damaged slightly by FFPE treatment, translocation genes may be accurately determined.

15 The apparatus according to the present embodiments may include a processor, a memory for storing and executing program data, a permanent storage unit such as a disk drive, a communication port for handling communications with external devices, and a user interface device such as a touch panel, keys, and buttons. The methods may be implemented as software modules or algorithms, and may be stored as program
20 instructions or computer-readable codes executable on the processor on a computer-readable recording medium. Here, examples of the computer-readable recording medium include a magnetic storage medium (e.g., a read-only memory (ROM), a random-access memory (RAM), a floppy disk, or a hard disk), and optical reading medium (e.g., a compact disk (CD)-ROM or a digital versatile disk (DVD)). The
25 computer-readable recording medium may be distributed over network coupled computer systems so that the computer-readable code may be stored and executed in a distributed fashion. The computer-readable recording medium may be read by the computer, stored in the memory, and executed by the processor.

30 The present embodiments may be described in terms of functional blocks and various processing steps. Such functional blocks may be realized by any number of hardware or/and software components configured to perform specific functions. For example, the present embodiments may employ various integrated circuit components, e.g., memory elements, processing elements, logic elements, and look-up tables, and the like,

which may carry out a variety of functions under the control of one or more microprocessors or other control devices. Similarly, where the elements of the present embodiments are implemented using software programming or software elements, the present embodiments may be implemented with any programming or scripting language, such as C, C++, Java, or assembler, with various algorithms being implemented with any combination of data structures, objects, processes, routines, or other programming elements. Functional aspects may be implemented in algorithms that execute on one or more processors. Furthermore, the present embodiments could employ any number of conventional techniques for electronics configuration, signal processing and/or control, data processing, and the like. The terms “mechanism”, “element”, “means”, and “configuration” are broadly used, and are not limited to mechanical and physical embodiments, but can include software routines in connection with processors, etc.

The particular implementations shown and described herein are illustrative examples of the invention and are not intended to otherwise limit the scope of the invention in any way. For the sake of brevity, conventional electronics, control systems, software development and other functional aspects of the systems (and components of the individual operating components of the systems) may not be described in detail. Furthermore, the connecting lines, or connectors shown in the various figures presented are intended to represent exemplary functional relationships and/or physical or logical couplings between the various elements. It should be noted that many alternative or additional functional relationships, physical connections or logical connections, may be present in a practical device.

The use of the terms "a" and "an" and "the" and similar referents in the context of describing the invention (especially in the context of the following claims) are to be construed to cover both the singular and the plural. Furthermore, recitation of ranges of values herein are merely intended to serve as a shorthand method of referring individually to each separate value falling within the range, unless otherwise indicated herein, and each separate value is incorporated into the specification as if it were individually recited herein. Finally, the steps of all methods described herein can be performed in any suitable order unless otherwise indicated herein or otherwise clearly contradicted by context.

While one or more exemplary embodiments of the present invention have been described, it will be understood by those of ordinary skill in the art that various changes in form and details may be made herein without departing from the essential characteristics of

the present invention. Thus, it should be understood that the exemplary embodiments described herein should be considered in a descriptive sense only and not for purposes of limitation. The scope of the present invention is shown not in the above description but in the claims, and all differences within the range equivalent thereto will be interpreted as

5 being incorporated in the present invention.

CLAIMS

1. A method of analyzing a gene, the method comprising:

acquiring data for split reads and discordantly aligned paired-end (PE) reads from
5 next-generation sequencing (NGS) data of a sample of a subject acquired by
deep-sequencing;

extracting first candidate gene pairs having a possibility of translocation in a
chromosome of the sample by using break points that are indicated by the split reads and the
PE reads; and

10 extracting, from among the extracted first candidate gene pairs, second candidate
gene pairs comprising genes in which a plurality of split reads having breakpoints within
the same coverage are aligned;

extracting, from among the extracted second candidate gene pairs, third candidate
gene pairs in which the direction of fusion between different genes goes from 5' end to 3'
15 end or from 3' end to 5' end; and

identifying translocation genes from among the extracted third candidate gene pairs,
wherein, in the genes included in the extracted second candidate gene pairs, the
number of the split reads having the breakpoints within the same coverage is a
predetermined threshold value or greater.

20 2. The method of claim 1, wherein the NGS data comprises data in a binary
version of sequence alignment/map (SAM) (BAM) format or in an SAM format.

3. The method of claim 2, wherein the acquiring comprises acquiring, from the
25 data in a BAM format or in an SAM format, data of an FLAG and a compact idiosyncratic
gapped alignment report (CIGAR) string for each of the split reads and the PE reads.

4. The method of claim 1, wherein the NGS data is generated by targeted
sequencing for identifying base sequences of target genes from the sample.

30 5. The method of claim 1, wherein the sample is a biopsy sample or a
formalin-fixed, paraffin-embedded (FFPE) sample.

6. A computer-readable recording medium recording a program for executing the method according to any one of claims 1 to 5 in a computer.

7. An apparatus for analyzing a gene, the apparatus comprising:

5 a read analyzing unit to acquire, from next-generation sequencing (NGS) data of a sample of a subject acquired by deep-sequencing, data for split reads and discordantly aligned paired-end (PE) reads; and

a translocation identifying unit to extract first candidate gene pairs having a possibility of translocation in a chromosome of the sample by using break points that are indicated by the split reads and the PE reads, extract, from among the extracted first candidate gene pairs, second candidate gene pairs comprising genes in which a plurality of split reads having breakpoints within the same coverage are aligned, extract, from among the extracted second candidate gene pairs, third candidate gene pairs in which the direction of fusion between different genes goes from 5' end to 3' end or from 3' end to 5' end, and
10 identify translocation genes from among the extracted third candidate gene pairs,

wherein, in the genes included in the extracted second candidate gene pairs, the number of the split reads having the breakpoints within the same coverage is a predetermined threshold value or greater.

20 8. The apparatus of claim 7, wherein the NGS data comprises data in a binary version of sequence alignment/map (SAM) (BAM) or in an SAM format.

9. The apparatus of claim 8, wherein the read analyzing unit acquires, from the data in a BAM format or in an SAM format, data of an FLAG and a compact idiosyncratic
25 gapped alignment report (CIGAR) string for each of the split reads and the PE reads.

10. The apparatus of claim 7, wherein the NGS data is generated by targeted sequencing for identifying base sequences of target genes from the sample.

30 11. The apparatus of claim 7, wherein the sample is a biopsy sample or a formalin-fixed, paraffin-embedded (FFPE) sample.