(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 April 2017 (13.04.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/062970 A1

(51) International Patent Classification:
*C12Q 1/68* (2006.01) *G06F 19/20* (2011.01)
*C12N 15/11* (2006.01)

(21) International Application Number:
PCT/US2016/056314

(22) International Filing Date:
10 October 2016 (10.10.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/239,879 10 October 2015 (10.10.2015) US

(71) Applicant: **GUARDANT HEALTH, INC.** [US/US]; 505 Penobscot Drive, Redwood City, California 94063 (US).

(72) Inventors: **MOKHTARI, Reza**; 649 Catamaran Street, Apt. 3, Foster City, California 94404 (US). **KERMANI, Bahram Ghaffarzadeh**; 73 Third Street, Apt. 12, Los Altos, California 94022 (US).

(74) Agent: **AMODEO, Gabriele A.**; Wilson Sonsini Goodrich & Rosati, 650 Page Mill Road, Palo Alto, California 94304 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

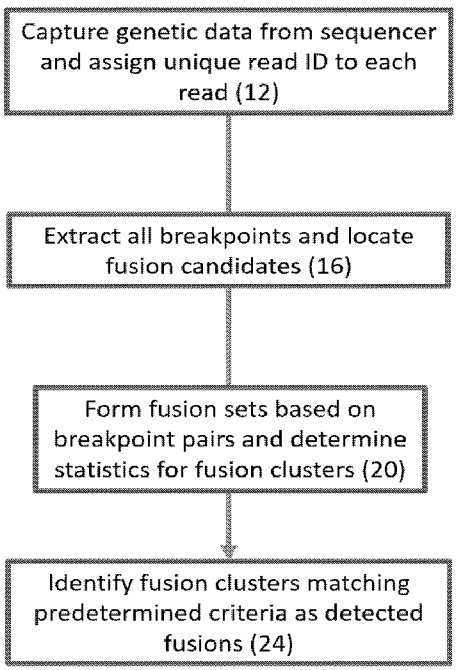
**Declarations under Rule 4.17**:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

*[Continued on next page]*

(54) Title: METHODS AND APPLICATIONS OF GENE FUSION DETECTION IN CELL-FREE DNA ANALYSIS

FIG. 1




(57) Abstract: Systems and methods are disclosed for determining gene fusion by determining a fused read containing sequencing data of a portion of a fused chromosome DNA molecule; determining a predetermined point on the genome with least one mapped portion of the fused read clipped at the predetermined point (a breakpoint); identifying two mapped read portions from two breakpoints (breakpoint pair) as a potential fusion candidate; creating one or more fusion sets based on breakpoint pairs and clustering the fusion sets into one or more fusion clusters; and identifying each fusion cluster meeting a predetermined criterion as a gene fusion.

WO 2017/062970 A1

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

**Published**:

— *with international search report (Art. 21(3))*

# METHODS AND APPLICATIONS OF GENE FUSION DETECTION IN CELL-FREE DNA ANALYSIS

## CROSS-REFERENCE

**[0001]** This application claims the benefit of U.S. Provisional Application No. 62/239,879, filed October 10, 2015, which application is incorporated herein by reference in its entirety.

## BACKGROUND OF THE INVENTION

**[0002]** Cancerous cells may have chromosomes that are fused together. If such a chromosome is sequenced, it will generate reads that can be mapped into two different zones (on the same or different chromosomes) of the genome. Gene fusion plays a role in the evolution of gene architecture. Duplication, sequence divergence, and recombination are the major contributors at work in gene evolution. When gene fusion happens in non-coding sequence region, it can lead to the misregulation of the expression of a gene now under the control of the cis-regulatory sequence of another gene. If it happens in coding sequences, gene fusion can cause the assembly of a new gene, allowing the appearance of new functions by adding peptide modules into a multi domain protein.

**[0003]** Chromosome banding analysis, fluorescence in situ hybridization (FISH), and reverse transcription polymerase chain reaction (RT-PCR) are common methods employed at diagnostic laboratories. These methods all have their distinct shortcomings due to the very complex nature of cancer genomes. Recent developments such as high-throughput sequencing and custom DNA microarrays bear promise of introduction of more efficient methods, but are still inadequate. High-throughput genome sequencing technologies have been used as a research tool and are currently being introduced in the clinics, and in the future of personalized medicine, whole genome sequence data may be an important tool to guide therapeutic intervention.

## SUMMARY OF THE INVENTION

**[0004]** In one aspect, systems and methods are disclosed for determining gene fusion by determining a fused read containing sequencing data of at least a portion of a fused chromosome DNA molecule; determining a predetermined point on the genome with least one mapped portion of the fused read clipped at the predetermined point (a breakpoint); identifying two mapped read portions from two breakpoints (breakpoint pair) as a potential fusion candidate; creating one or more fusion sets based on breakpoint pairs and clustering the fusion sets into one or more fusion clusters; and identifying each fusion cluster meeting a predetermined criterion as a gene fusion.

**[0005]** In one aspect, the present disclosure provides a method for processing genetic sequence read data from a sample, the method comprising: determining a fused read containing

sequencing data of at least a portion of a fused chromosome DNA molecule; determining a predetermined point on the genome with least one mapped portion of the fused read clipped at the predetermined point (a breakpoint); identifying two mapped read portions from two breakpoints (breakpoint pair) as a potential fusion candidate; creating one or more fusion sets based on breakpoint pairs and clustering the fusion sets into one or more fusion clusters; and identifying each fusion cluster meeting a predetermined criterion as a gene fusion.

**[0006]** In some embodiments, the method comprises assigning a unique molecule or read identifier (read ID) to each read. In some embodiments, the method comprises clipping each mapped portion of the reads from one or both sides. In some embodiments, the breakpoints are independent of the reads in identity and are identified by a sign, a chromosome and a position. In some embodiments, the breakpoints keep statistics including a number of reads and molecules that are clipped or split at the breakpoint, and a number of wild-type reads and molecules that pass over the breakpoint. In some embodiments, the method comprises selecting every two mapped read portions with common read IDs that belong to two breakpoints with appropriate signs as a potential fusion candidate. In some embodiments, the potential fusion candidate location in the original read before mapping shows the read portion as originally located next to each other. In some embodiments, the method comprises checking if read portions are mapped on one strand for differences in the breakpoints' signs. In some embodiments, the method comprises tracking fusion set statistics.

**[0007]** In some embodiments, the fusion set statistics are breakpoint IDs, number of molecules, or reads that are contained in the set. In some embodiments, the method comprises grouping fusion sets with similar breakpoints in a fusion cluster. In some embodiments, the similar breakpoints are breakpoints no more than 5 nucleotides, no more than 10 nucleotides or no more than 25 nucleotides apart. In some embodiments, the method comprises defining a fusion cluster between two regions in in the genome. In some embodiments, the method comprises determining for the fusion cluster a number of fused molecules for each partner. In some embodiments, the method comprises determining for the fusion cluster the number of fused reads for each partner. In some embodiments, the method comprises determining for the fusion cluster a number of wild-type molecules for each partner. In some embodiments, the method comprises determining for the fusion cluster a number of wild-type reads or molecules for each partner. In some embodiments, the method comprises determining for the fusion cluster a fusion percentage for each partner as a ratio of fused molecules over total molecule for each partner. In some embodiments, the total molecule comprises wild-type and clipped components. In some embodiments, the method comprises determining for the fusion cluster gene information for each partner. In some embodiments, the method comprises determining a downstream gene of

the fusion cluster. In some embodiments, the criteria comprises having more than one molecule in the cluster or having at least a molecule with both Watson and Crick strands.

**[0008]** In one aspect, the present disclosure provides a system to analyze genetic information, comprising a DNA sequencer; a processor coupled to the DNA sequencer, the processor running computer code to process genetic sequence read data from a sample, the computer code comprising instructions for: determining a fused read containing sequencing data of a portion of a fused chromosome DNA molecule; determining at least a predetermined point on the genome with least one mapped portion of the fused read clipped at the predetermined point (a breakpoint); identifying two mapped read portions from two breakpoints (breakpoint pair) as a potential fusion candidate; creating one or more fusion sets based on breakpoint pairs and clustering the fusion sets into one or more fusion clusters; and identifying each fusion cluster meeting a predetermined criterion as a gene fusion.

**[0009]** In one aspect, the present disclosure provides a method comprising: sequencing DNA molecules with a DNA sequencer to generate a collection of sequences; mapping the collection of sequences to a reference genome; identifying fused reads from the mapped collection, wherein a fused read contains sub-sequences, wherein a first sub-sequence maps to a first genetic locus and a second sub-sequence maps to a second, distinct genetic locus; for each fused read, identifying a first breakpoint at the first genetic locus and a second breakpoint at the second genetic locus, wherein a breakpoint is a point on the reference genome where a sequence of a fused read is clipped, and wherein the first and second breakpoints form a breakpoint pair; generating sets of fused reads, each set comprising fused reads having the same breakpoint pair; clustering sets of fused reads, wherein each cluster is formed from sets of fused reads having first breakpoints within a first predetermined nucleotide distance and second breakpoints within a second predetermined nucleotide distance; and determining a gene fusion for one or more clusters, wherein a gene fusion for a cluster has, as a first fusion gene breakpoint, a breakpoint selected from the first breakpoints in the cluster and, as a second fusion gene breakpoint, a breakpoint selected from the second breakpoints in the cluster, and wherein the first and second fusion gene breakpoints are each selected based on selection criteria.

**[0010]** In some embodiments, the distinct genetic loci are located on different chromosomes or on different genes of the same chromosome. In some embodiments, the first and second predetermined distances are each no more than 5 nucleotides, no more than 10 nucleotides or no more than 25 nucleotides. In some embodiments, the selection criteria include the breakpoint having the most fused reads in the cluster. In some embodiments, the method comprises determining a gene fusion for a plurality of gene clusters.

**[0011]** In one aspect, the present disclosure provides a method comprising: sequencing a plurality of DNA molecules with a DNA sequencer; tagging each of the plurality of sequences molecules with an identifier; mapping each tagged sequence to a reference genome; identifying clipped reads from the mapped tagged sequences, wherein a clipped read is a tagged sequence containing a mapped portion and a clipped portion, wherein the mapped portion maps to a genetic locus and the clipped portion does not map to the genetic locus; determining a breakpoint of each clipped read, wherein a breakpoint is a point on the reference genome where a sequence of a clipped read is clipped; creating breakpoint sets, each breakpoint set comprising identifiers of clipped reads having the same breakpoint; creating sets of breakpoint pairs by comparing pairs of breakpoint sets, each set of breakpoint pairs including identifiers present in both members of a compared pair of breakpoint sets; clustering sets of breakpoint pairs, wherein each cluster includes sets of breakpoint pairs having a first breakpoint of the pair within a first predetermined genetic distance and a second breakpoint of the pair within a second predetermined genetic distance; and determining a gene fusion for one or more of the clusters, wherein a gene fusion for a cluster has, as a first fusion gene breakpoint, a breakpoint selected from the first breakpoints in the cluster and, as a second fusion gene breakpoint, a breakpoint selected from the second breakpoints in the cluster, and wherein the first and second fusion gene breakpoints are each selected based on a selection criteria. In some embodiments, the selection criteria include the breakpoint having the most fused reads in the cluster.

**[0012]** In one aspect, the present disclosure provides a method for identifying a fusion gene breakpoint, the method comprising: determining a fused read containing sequencing data of at least a portion of a fused chromosome DNA molecule; determining a predetermined point on the genome with least one mapped portion of the fused read clipped at the predetermined point (a breakpoint); identifying two mapped read portions from two breakpoints (breakpoint pair) as a potential fusion candidate; creating one or more fusion sets based on breakpoint pairs and clustering the fusion sets into one or more fusion clusters; identifying each fusion cluster meeting a predetermined criterion as a gene fusion, and identifying a breakpoint of the gene fusion as the fusion gene breakpoint.

**[0013]** In one aspect, the present disclosure provides a method for diagnosing a condition in a subject, the method comprising: determining a fused read containing sequencing data of at least a portion of a fused chromosome DNA molecule; determining a predetermined point on the genome with least one mapped portion of the fused read clipped at the predetermined point (a breakpoint); identifying two mapped read portions from two breakpoints (breakpoint pair) as a potential fusion candidate; creating one or more fusion sets based on breakpoint pairs and clustering the fusion sets into one or more fusion clusters; and identifying each fusion cluster

meeting a predetermined criterion as a gene fusion, wherein said gene fusion is indicative of the condition.

**[0014]** In some embodiments, the condition is a cancer. In some embodiments, the cancer is selected from the group consisting of: a hematological cancer, a sarcoma, and a prostate cancer. In some embodiments, the method further comprises administering a treatment to the subject.

## INCORPORATION BY REFERENCE

**[0015]** All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference.

## BRIEF DESCRIPTION OF THE DRAWINGS

**[0016]** The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings of which:

**[0017]** FIG. 1 diagrams an exemplary process for detecting gene fusion.

**[0018]** FIG. 2 depicts possible different scenarios to create a fused chromosome from two other chromosomes.

**[0019]** FIG. 3A shows an exemplary +/- breakpoint with read portions that are clipped from left/right, respectively.

**[0020]** FIG. 3B shows exemplary merging process used with gene fusion detection.

**[0021]** FIG. 4 diagrams a system of the invention.

**[0022]** FIG. 5 shows an exemplary gene fusion between Chromosome A and Chromosome B, and DNA fused reads mapping across the breakpoint.

**[0023]** FIG. 6 shows the exemplary mapping of a DNA fragment to two locations in a reference genome.

**[0024]** FIG. 7 shows exemplary fused reads in which mapped breakpoints are differently located.

**[0025]** FIG. 8 shows exemplary grouping of mapped fused reads into sets and sets into clusters for calling a fusion gene breakpoint.

**[0026]** FIGS. 9A-9C show exemplary illustrations of the gene fusion detection process.

## DETAILED DESCRIPTION OF THE INVENTION

**[0027]** The present invention relates to systems and methods for detecting gene fusions.

**[0028]** Precise mapping of the breakpoint of a fusion gene is challenging. Errors in sequencing and difficulty aligning a fusion gene are just two of the difficulties encountered when trying to map a breakpoint. The systems and methods described herein may provide one or more of the following advantages. The system can identify fusion genes that can contribute to tumor formation because fusion genes can produce much more active abnormal protein than non-fusion genes. The system accurately determines cancer presence, as fusion genes are oncogenes that cause cancer; these include BCR-ABL, TEL-AML1 (ALL with t(12 ; 21)), AML1-ETO (M2 AML with t(8 ; 21)), and TMPRSS2-ERG with an interstitial deletion on chromosome 21, often occurring in prostate cancer. In the case of TMPRSS2-ERG, by disrupting androgen receptor (AR) signaling and inhibiting AR expression by oncogenic ETS transcription factor, the fusion product regulate the prostate cancer. Most fusion genes are found from hematological cancers, sarcomas, and prostate cancer. Oncogenic fusion genes may lead to a gene product with a new or different function from the two fusion partners. Alternatively, a proto-oncogene is fused to a strong promoter, and thereby the oncogenic function is set to function by an upregulation caused by the strong promoter of the upstream fusion partner. The latter is common in lymphomas, where oncogenes are juxtaposed to the promoters of the immunoglobulin genes. Oncogenic fusion transcripts may also be caused by trans-splicing or read-through event. Analysis of these gene fusions from a perspective of genome sequence and structure could provide relevant data to guide development of improved cancer diagnostics and targeted therapies.

**[0029]** FIG. 1 shows an exemplary process for determining gene fusion. In general, the process captures genetic data from sequencer and applies merging techniques that align and connect the sequenced Pair-End reads of which the insert size is smaller than the sum of the two read length and after merging assigns a unique read identifier (read ID) to each read (12). Next, the process extracts all breakpoints and locates fusion candidates (16). The process then forms fusion sets based on breakpoint pairs and determines statistics for fusion clusters (20). Fusion clusters are then identified by matching predetermined criteria as detected fusions (24).

**[0030]** Details on the process of FIG. 1 are discussed next. Cancerous cells may have chromosomes that are fused together. If such a chromosome is sequenced, it will generate reads that can be mapped into two different zones (on the same or different chromosomes) of the genome. This behavior is utilized to detect fusion.

**[0031]** Before mapping, a unique read identifier (read ID) is assigned to each read and would be encoded in the read header in one or more FASTQ file(s), as detailed below. Alternatively, a unique molecule, such as an oligonucleotide comprising a unique barcode, can be used instead of the read ID. Once the FASTQ file(s) is (are) mapped, this encoded read ID will be retrieved

and it can easily show what hits are coming from the same original read. Next, the process extracts all the breakpoints: a fused read (a read contains the sequencing data of a partial DNA molecule coming from a fused chromosome) cannot be mapped as a whole to the genome and the mapper maps different portions of them to different locations on the genome. This presents a challenge when using conventional techniques to attempt breakpoint mapping. Each mapped portion of such reads is clipped from one or both sides. A breakpoint is a point on the genome where at least one mapped portion of a fused read is clipped. Breakpoints are independent of the reads in their identity and are identified by their sign, chromosome and position. A +/- breakpoint has read portions that are clipped from left/right, respectively. All the reads that are clipped or split from the same side in a position are listed in the related breakpoint read list. Breakpoints also can keep other statistics, like number of reads and molecules that are clipped or split at the breakpoint; or the number of wild-type reads and molecules that pass over the breakpoint. The gene information at the breakpoint position also can be provided. By assigning breakpoints, with or without clustering, the methods and systems described herein can be used to accurately determine the locus where gene fusion has occurred.

**[0032]** The process then finds fusions: every two mapped read portions with common read IDs that belong to two breakpoints with appropriate signs is a potential fusion candidate. They also need to have the correct piece order (their location in original read before mapping), that shows the read portion were originally located next to each other, to be considered a true fusion candidate. In addition, the resulting fusion must be biologically possible in terms of sequence strands. This simply means that if the read portions are mapped on the same strand (both 5' strand or both 3' strand) then the breakpoints' signs must disagree and vice versa. An example of this is shown in FIG. 2.

**[0033]** All extracted fusion candidates are put in fusion sets based on the breakpoint pairs. Fusion sets also can keep statistics like breakpoint IDs and number of molecules and reads that are contained in the set. These statistics may be tracked.

**[0034]** Clustering is then performed. All fusion sets with close enough breakpoints will be grouped in a fusion cluster. As a result, a fusion cluster is defined between two regions in in the genome. The present disclosure also provides determining for a fusion cluster a number of fused molecules for each partner, the number of fused reads for each partner, a number of wild-type molecules for each partner, a number of wild-type reads or molecules for each partner, or a fusion percentage for each partner as a ratio of fused molecules over total molecule for each partner.

**[0035]** FIG. 5 shows a hypothetical gene fusion between Chromosome A and Chromosome B. As a result of crossing over, the gene fusion contains a portion of each chromosome. The point

of cross over is referred to as the breakpoint. In cell free DNA, DNA fragments may map across the breakpoint, such as Fused Read 1, Fused Read 2 and Fused Read 3.

**[0036]** Sequencing produces sequences of the DNA fragments. Software marks each sequence with an identifying tag. Software also maps these resultant sequences onto a reference genome. FIG. 6 shows hypothetical mapping of Fused Read 1 to a reference genome. Mapping software will map a sequence of a fused read to wherever in the reference genome sufficient homology is found. Ambiguous sequences can be mapped to multiple locations in the reference genome.

**[0037]** In the case of a fused read that maps across a breakpoint of a fusion gene, the software typically will map the sequence of a fused read twice, once to each chromosome. However, in each case, the mapping software cannot property map a portion of the sequence (a subsequence) to the reference genome. Accordingly, the mapped sequence will include both a sub-sequence that maps to the reference genome and a sub-sequence that, as a result of poor homology, does not map to the same locus. Such a sub-sequence is referred to as a "clipped" sequence. The point on the reference genome at which the read is clipped is the breakpoint.

**[0038]** Because each sequence bears an identifying tag, a sequence mapped to two different locations can be identified as originating from the same original sequence due to identical tags. So, for example, a sub-sequence of the sequence having sufficient homology is mapped to Chromosome A and a sub-sequence of the sequence having insufficient homologogy is clipped. Similarly, the mapping software will map the sequence to Chromosome B, clipping the sequence where there is insufficient homology.

**[0039]** However, as a result of several factors, including errors in sequencing and characteristics of the mapping algorithm, a DNA fragment including a breakpoint of a fusion gene may not map precisely to the breakpoint locus on each of the reference chromosomes. For example, the mapping software may identify the breakpoint of a sequence somewhat upstream or somewhat downstream of the actual breakpoint.

**[0040]** Because each sequence bears an identifying tag, a sequence mapped to two different locations can be identified as originating from the same original sequence due to identical tags. So, for example, a sub-sequence of the sequence having sufficient homology is mapped to Chromosome A and a sub-sequence of the sequence having insufficient homologogy is clipped. Similarly, the mapping software will map the sequence to Chromosome B, clipping the sequence where there is insufficient homology.

**[0041]** However, as a result of several factors, including errors in sequencing and characteristics of the mapping algorithm, a DNA fragment including a breakpoint of a fusion gene may not map precisely to the breakpoint locus on each of the reference chromosomes. For example, the mapping software may identify the breakpoint of a sequence somewhat upstream or somewhat

downstream of the actual breakpoint. These errors can affect the accuracy of, for example, a cancer diagnosis that depends upon accurate gene fusion information. The present disclosure

**[0042]** Several hypothetical mapping errors are shown in FIG. 7. Fused Read 1 maps properly, with breakpoints indicated in the reference chromosomes as Breakpoint A1 and Breakpoint B1 (first and second breakpoints). This fused read has breakpoint pair A1-B1. Fused Read 2 maps improperly, with the breakpoint for Chromosome 1 determined to be upstream, at Breakpoint A2 (first breakpoint). However, the breakpoint in Chromosome B is mapped correctly at Breakpoint B1 (second breakpoint). This fused read has breakpoint pair A2-B1. Fused Read 3 also maps improperly, with the breakpoint for Chromosome A mapped correctly at Breakpoint A1 (first breakpoint), but the breakpoint for Chromosome B determined to be downstream, at Breakpoint B2 second breakpoint). This fused read has breakpoint pair A1-B2. In such a situation, software has identified several breakpoints for the fusion gene.

**[0043]** According to a method of this disclosure, in order to call a breakpoint in a fusion gene, mapped sequences are grouped into sets based on common breakpoint pairs and then into clusters based on breakpoints within a predetermined base distance in the reference genome.

**[0044]** Such a method is described in FIG. 8. Sequences for Fused Reads 1, 2, 3, 4, 5 and 6 are mapped to Chromosomes A and B of the reference genome. Sequences in which the breakpoints on both sides of the fusion are grouped into a set. In the example, Fused Reads 1 and 4 share breakpoint pair A1 and B1, and are grouped into Set I. Fused Reads 2 and 5 share breakpoint pair A2 and B1, and are grouped into Set II. Fused Reads 3 and 6 share breakpoint pair A1 and B2, and are grouped into Set III.

**[0045]** In this example, breakpoints A1 and A2 are within a predetermined genetic distance A (e.g., 10 bases), and breakpoints B1 and B2 are within a predetermined genetic distance B. Accordingly, Sets I, II and III and grouped into a cluster.

**[0046]** The fusion gene breakpoint is called using selection criteria selected by the user. In some embodiments, the criteria comprise having more than one molecule in the cluster and/or having at least a molecule with both Watson and Crick strands. In one method, the breakpoint in each chromosome is determined by a voting method, in which the breakpoint among all the breakpoints having the most associated fused reads is called as the fusion gene breakpoint. In other methods, breakpoints of different sequences may be weighted using a quality algorithm. In the example of FIG. 8, in Chromosome A, Breakpoint A1 is associated with four fused reads, while Breakpoint A2 is associated with two fused reads. Therefore, the first gene fusion breakpoint is called at A1. In Chromosome B, Breakpoint B1 is associated with four fused reads, while Breakpoint B2 is associated with two fused reads. Therefore, the second gene fusion breakpoint is called at B1.

**[0047]** Another exemplary method is shown in FIGs. 9A-9C. DNA molecules are sequenced using a DNA sequencing system, such as a high throughput sequencer. Sequences can be analyzed to generate consensus sequences of original molecules in the collection. The collection of sequences generated are tagged with unique identifiers (in this case, 1-7). The sequences are mapped to a reference genome. In this example, the sequences each map to two different locations in the reference genome. Mapped portions are depicted as bars, while clipped portions are depicted as dashed lines. Breakpoints of all the sequences are identified. In this example, the breakpoints on chromosome A are A1, A2 and A3. The breakpoints on chromosome B are B1, B2 and B3. Mapped reads are organized into sets based on common breakpoints. Breakpoint pairs are determined as pairs of sequences having the same identifier and the same breakpoints on each chromosome. A predetermined distance on the chromosome to cluster breakpoints is determined. In this example, the cluster includes breakpoints A1 and A2, and B1 and B2. Breakpoints A3 and B3 are outside of the predetermined distance and, therefore, are not included in the cluster. A breakpoint pair in the original molecules is called based on selected criteria. In this example, the criteria are based on voting. Accordingly, breakpoints A1 and B1 are called as the breakpoint pair based on having a majority of the molecules within the cluster.

**[0048]** The system can handle multiple alignments of short sequences that is commonly seen in splicing. Longer sequences can be obtained around the fusion point to confirm their alignments. To reduce the false-positive rate, a list of filters, including read number, sequence similarity, read position distribution filter can be used to provide results with high specificity. In addition, expression estimation tool RSEM (RNA-Seq by Expectation Maximization) can be used to apply the Expectation Maximization (EM) algorithm with sparse optimization to estimate chimerical transcript abundance. Furthermore, this abundance quantification can be increase the identification accuracy. Taken together, these features allow the system to provide a more complete view of gene fusion events.

**[0049]** The system can encompass obtaining a plurality of sequence reads from one or more samples in any suitable file format, identifying sets of duplicative sequence reads, and storing only one read for each set of the duplicative sequence reads. Suitable file formats include the FASTA and FASTQ file formats. FASTA and FASTQ are common file formats used to store raw sequence reads from high throughput sequencing. FASTQ files store an identifier for each sequence read, the sequence, and the quality score string of each read. FASTA files store the identifier and sequence only. These two file formats are the inputs to many common sequencing alignment and assembly algorithms. The invention recognizes that the read sequence information for FASTQ and FASTA files within and across samples tends to be highly redundant or duplicative. This means that many of the sequence reads will consist of the same

sequence. Methods of the invention exploit this redundancy to achieve a many-fold reduction in file size, and there is no loss in the retrieval of the stored data. For example, the invention can be used to read the FASTA/FASTQ file associated with a sample and store only the unique read sequences in a master read sequence file.

**[0050]** The system further encompasses collecting meta information, such as a read identifier, for each read that has the same sequence as the identified unique sequence. This meta information can then be written to a file for that sample in which the meta information is correlated to the unique sequence reads identified in original FASTA/FASTQ file and now stored in a master read sequence file. Because this new file does not contain the duplicative information found in the original file, it is smaller and easier to transfer than the original file. Moreover, the compressed file need not contain any actual sequence data at all. In certain aspects, the compressed file may simply contain the identifier for the sequence read indexed to the unique sequences stored in the master file.

**[0051]** Sequence data can be compressed by obtaining—using a computer system comprising a processor coupled to a non-transitory memory—a plurality of sequence reads. Each sequence read can include a sequence string as well as meta information. The sequence reads may be provided in the format of one or more FASTA or FASTQ files, for example, with the meta information including the description line (preceded by the ">" character) and optionally, in the case of FASTQ, quality scores. The sequence string preferably represents nucleotide sequence data, e.g., using IUPAC nucleotide codes. A subset of the sequence strings that contains only unique entries is identified. Systems and methods of the invention may then be used to write output that includes the identified subset and—for each of the plurality of sequence reads—the meta information for that sequence read with an indicator of the unique entry in the subset that represents that sequence read.

**[0052]** In some embodiments, the subset (i.e., containing only unique sequence reads) is written to a master reads file, which may be a text file. Preferably, the unique sequence reads are represented in the master reads file using IUPAC nucleotide codes so that the files are human-readable and further processing (e.g., using a scripting language such as Perl or Python) can be easily performed. The meta information may be written to compressed output files corresponding to the input FASTA or FASTQ files.

**[0053]** Methods can include reconstituting the original input from only the output and in certain embodiments the retrieval is lossless, even perfectly lossless. That is, the output may be processed to create new FASTA or FASTQ files comprising the plurality of sequence reads. Where the retrieval is lossless, the new FASTA or FASTQ files contain the same information as the FASTA or FASTQ files.

**[0054]** The invention is amenable with any suitable type of data file. In addition to the aforementioned FASTA and FASTQ files, sequence reads can also be captured in Variant Call Format (VCF) files. With advances in high throughput sequencing, it is common for multiple sequencing centers to detect variants in the human genome and report them through these VCF files. The invention can facilitate the development of a unified database to store variant information in VCF files from different sources in a way that allows researchers to perform complex allele-, sample-, and population-level queries across centers. The unified database can consolidate variant information in the VCF files from different samples by storing every unique allele (e.g. unique sequence read) on one universal allele table and by storing references of these unique alleles to associated samples and sample-level meta-data.

**[0055]** Implementations of the system can include a method for compressing sequence data. The method includes: obtaining—using a computer system comprising a processor coupled to a non-transitory memory—a plurality of sequence reads, each sequence read comprising a sequence string and meta information; identifying a subset of the sequence strings that contains only unique entries; writing output comprising the subset and—for each of the plurality of sequence reads—the meta information for that sequence read with an indicator of the unique entry in the subset that represents that sequence read. Preferably the output comprises one or more text files that store the subset using IUPAC nucleotide codes. Preferably, the output is stored as plain text (e.g., and may be opened using a text editor program and read on-screen by a person). In a preferred embodiment, the sequence read data is stored without loss. The method may include processing the output to create new FASTA or FASTQ files comprising the plurality of sequence reads. The plurality of sequence reads may be obtained as FASTA or FASTQ files and the new FASTA or FASTQ files may contain the same information as the FASTA or FASTQ files. In some embodiments, the output occupies less than % of the disk space required to store the obtained plurality of sequence reads.

**[0056]** General methods for obtaining samples, generating sequencing reads, and various types of sequencing useful for practicing the invention will now be described. It is to be understood that these exemplary methods are not limiting and may be modified as necessary by those skilled in the art.

**[0057]** Obtaining a plurality of sequence reads can include sequencing a nucleic acid from a sample to generate the sequence reads. As explained in detail below, obtaining a plurality of sequence reads can also include receiving sequencing data from a sequencer. Nucleic acid in a sample can be any nucleic acid, including for example, genomic DNA in a tissue sample, cDNA amplified from a particular target in a laboratory sample, or mixed DNA from multiple organisms. In some embodiments, the sample includes homozygous DNA from a haploid or

diploid organism. For example, a sample can include genomic DNA from a patient who is homozygous for a rare recessive allele. In other embodiments, the sample includes heterozygous genetic material from a diploid or polyploidy organism with a somatic mutation such that two related nucleic acids are present in allele frequencies other than 50 or 100%, i.e., 20%, 5%, 1%, 0.1%, or any other allele frequency.

**[0058]** In one embodiment, nucleic acid template molecules (e.g., DNA or RNA) are isolated from a biological sample containing a variety of other components, such as proteins, lipids, and non-template nucleic acids. Nucleic acid template molecules can be obtained from any cellular material, obtained from animal, plant, bacterium, fungus, or any other cellular organism. Biological samples for use in the present invention also include viral particles or preparations. Nucleic acid template molecules can be obtained directly from an organism or from a biological sample obtained from an organism, e.g., from blood, serum, plasma, urine, cerebrospinal fluid, saliva, stool, lymph fluid, synovial fluid, cystic fluid, ascites, pleural effusion, amniotic fluid, chorionic villus sample, fluid from a preimplantation embryo, a placental sample, lavage and cervical vaginal fluid, interstitial fluid, a buccal swab sample, sputum, bronchial lavage, a Pap smear sample, or ocular fluid. Any tissue or body fluid specimen (e.g., a human tissue of bodily fluid specimen) may be used as a source for nucleic acid to use in the invention. Nucleic acid template molecules can also be isolated from cultured cells, such as a primary cell culture or cell line. The cells or tissues from which template nucleic acids are obtained can be infected with a virus or other intracellular pathogen. A sample can also be total RNA extracted from a biological specimen, a cDNA library, viral, or genomic DNA. A sample may also be isolated DNA from a non-cellular origin.

**[0059]** Nucleic acid obtained from biological samples may be fragmented to produce suitable fragments for analysis. Template nucleic acids may be fragmented or sheared to a desired length, using a variety of mechanical, chemical, and/or enzymatic methods. DNA may be randomly sheared via sonication using, for example, an ultrasonicator sold by Covaris (Woburn, Mass.), brief exposure to a DNase, or using a mixture of one or more restriction enzymes, or a transposase or nicking enzyme. RNA may be fragmented by brief exposure to an RNase, heat plus magnesium, or by shearing. The RNA may be converted to cDNA. If fragmentation is employed, the RNA may be converted to cDNA before or after fragmentation. In one embodiment, nucleic acid is fragmented by sonication. In another embodiment, nucleic acid is fragmented by a hydroshear instrument. Generally, individual nucleic acid template molecules can be from about 2 kb bases to about 40 kb. In a particular embodiment, nucleic acids are about 6 kb-10 kb fragments. Nucleic acid molecules may be single-stranded, double-stranded, or double stranded with single-stranded regions (for example, stem- and loop-structures).

**[0060]** A biological sample may be lysed, homogenized, or fractionated in the presence of a detergent or surfactant as needed. Suitable detergents may include an ionic detergent (e.g., sodium dodecyl sulfate or N-lauroylsarcosine) or a nonionic detergent. Once a nucleic acid is extracted or isolated from the sample it may be amplified.

**[0061]** Amplification refers to production of additional copies of a nucleic acid sequence and is generally carried out using polymerase chain reaction (PCR) or other technologies known in the art. The amplification reaction may be any amplification reaction known in the art that amplifies nucleic acid molecules such as PCR. Other amplification reactions include nested PCR, PCR-single strand conformation polymorphism, ligase chain reaction, strand displacement amplification and restriction fragments length polymorphism, transcription based amplification system, rolling circle amplification, and hyper-branched rolling circle amplification, quantitative PCR, quantitative fluorescent PCR (QF-PCR), multiplex fluorescent PCR (MF-PCR), real time PCR (RTPCR), restriction fragment length polymorphism PCR (PCR-RFLP), in situ rolling circle amplification (RCA), bridge PCR, picotiter PCR, emulsion PCR, transcription amplification, self-sustained sequence replication, consensus sequence primed PCR, arbitrarily primed PCR, degenerate oligonucleotide-primed PCR, and nucleic acid based sequence amplification (NABSA). Amplification methods that can be used include those described in U.S. Pat. Nos. 5,242,794; 5,494,810; 4,988,617; and 6,582,938. In certain embodiments, the amplification reaction is PCR as described, for example, U.S. Pat. No. 4,683,195; and U.S. Pat. No. 4,683,202, hereby incorporated by reference. Primers for PCR, sequencing, and other methods can be prepared by cloning, direct chemical synthesis, and other methods known in the art. Primers can also be obtained from commercial sources such as Eurofins MWG Operon (Huntsville, Ala.) or Life Technologies (Carlsbad, Calif.).

**[0062]** Amplification adapters may be attached to the fragmented nucleic acid. Adapters may be commercially obtained, such as from Integrated DNA Technologies (Coralville, Iowa). In certain embodiments, the adapter sequences are attached to the template nucleic acid molecule with an enzyme. The enzyme may be a ligase or a polymerase. The ligase may be any enzyme capable of ligating an oligonucleotide (RNA or DNA) to the template nucleic acid molecule. Suitable ligases include T4 DNA ligase and T4 RNA ligase, available commercially from New England Biolabs (Ipswich, Mass.). Methods for using ligases are well known in the art. The polymerase may be any enzyme capable of adding nucleotides to the 3′ and the 5′ terminus of template nucleic acid molecules.

**[0063]** The ligation may be blunt ended or utilize complementary overhanging ends. In certain embodiments, the ends of the fragments may be repaired, trimmed (e.g. using an exonuclease), or filled (e.g., using a polymerase and dNTPs) following fragmentation to form blunt ends. In

some embodiments, end repair is performed to generate blunt end 5′ phosphorylated nucleic acid ends using commercial kits, such as those available from Epicentre Biotechnologies (Madison, Wis.). Upon generating blunt ends, the ends may be treated with a polymerase and dATP to form a template independent addition to the 3′-end and the 5′-end of the fragments, thus producing a single A overhanging. This single A is used to guide ligation of fragments with a single T overhanging from the 5′-end in a method referred to as T-A cloning. Alternatively, because the possible combinations of overhangs left by the restriction enzymes are known after a restriction digestion, the ends may be left as-is, i.e., ragged ends. In certain embodiments, double stranded oligonucleotides with complementary overhanging ends are used.

**[0064]** Embodiments of the invention involve attaching the barcode sequences to the template nucleic acids. In certain embodiments, a barcode is attached to each fragment. In other embodiments, a plurality of barcodes, e.g., two barcodes, are attached to each fragment. A barcode sequence generally includes certain features that make the sequence useful in sequencing reactions. For example the barcode sequences are designed to have minimal or no homo-polymer regions, i.e., 2 or more of the same base in a row such as AA or CCC, within the barcode sequence. The barcode sequences are also designed so that they are at least one edit distance away from the base addition order when performing base-by-base sequencing, ensuring that the first and last base do not match the expected bases of the sequence.

**[0065]** The barcode sequences are designed such that each sequence is correlated to a particular portion of nucleic acid, allowing sequence reads to be correlated back to the portion from which they came. In certain embodiments, the barcode sequences range from about 5 nucleotides to about 15 nucleotides. In a particular embodiment, the barcode sequences range from about 4 nucleotides to about 7 nucleotides. Since the barcode sequence is sequenced along with the template nucleic acid, the oligonucleotide length should be of minimal length so as to permit the longest read from the template nucleic acid attached. For example, a plurality of DNA barcodes can comprise various numbers of sequences of nucleotides. In certain embodiments, the barcode sequences comprise 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or more nucleotides. When attached to only one end of a polynucleotide, the plurality of DNA barcodes can produce 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or more different identifiers. Alternatively, when attached to both ends of a polynucleotide, the plurality DNA barcodes can produce 4, 9, 16, 25, 36, 49, 64, 81, 100, 121, 144, 169, 196, 225, 256, 289, 324, 361, 400 or more different identifiers (which is the ^2 of when the DNA barcode is attached to only 1 end of a polynucleotide).

**[0066]** Generally, the barcode sequences are spaced from the template nucleic acid molecule by at least one base (minimizes homo-polymeric combinations). In certain embodiments, the barcode sequences are attached to the template nucleic acid molecule, e.g., with an enzyme. The enzyme may be a ligase or a polymerase, as discussed below.

**[0067]** Amplification or sequencing adapters or barcodes, or a combination thereof, may be attached to the fragmented nucleic acid. Such molecules may be commercially obtained, such as from Integrated DNA Technologies (Coralville, Iowa). In certain embodiments, such sequences are attached to the template nucleic acid molecule with an enzyme such as a ligase. Suitable ligases include T4 DNA ligase and T4 RNA ligase, available commercially from New England Biolabs (Ipswich, Mass.). The ligation may be blunt ended or via use of complementary overhanging ends. In certain embodiments, following fragmentation, the ends of the fragments may be repaired, trimmed (e.g. using an exonuclease), or filled (e.g., using a polymerase and dNTPs) to form blunt ends. In some embodiments, end repair is performed to generate blunt end 5′ phosphorylated nucleic acid ends using commercial kits, such as those available from Epicentre Biotechnologies (Madison, Wis.). Upon generating blunt ends, the ends may be treated with a polymerase and dATP to form a template independent addition to the 3′-end and the 5′-end of the fragments, thus producing a single A overhanging. This single A can guide ligation of fragments with a single T overhanging from the 5′-end in a method referred to as T-A cloning. Alternatively, because the possible combinations of overhangs left by the restriction enzymes are known after a restriction digestion, the ends may be left as-is, i.e., ragged ends. In certain embodiments double stranded oligonucleotides with complementary overhanging ends are used.

**[0068]** After any processing steps (e.g., obtaining, isolating, fragmenting, amplification, or barcoding), nucleic acid can be sequenced.

**[0069]** Sequencing may be by any method known in the art. DNA sequencing techniques include classic dideoxy sequencing reactions (Sanger method) using labeled terminators or primers and gel separation in slab or capillary, sequencing by synthesis using reversibly terminated labeled nucleotides, pyrosequencing, 454 sequencing, Illumina/Solexa sequencing, allele specific hybridization to a library of labeled oligonucleotide probes, sequencing by synthesis using allele specific hybridization to a library of labeled clones that is followed by ligation, real time monitoring of the incorporation of labeled nucleotides during a polymerization step, polony sequencing, SOLiD sequencing targeted sequencing, single molecule real-time sequencing, exon sequencing, electron microscopy-based sequencing, panel sequencing, transistor-mediated sequencing, direct sequencing, random shotgun sequencing, whole-genome sequencing, sequencing by hybridization, , capillary electrophoresis, gel

electrophoresis, duplex sequencing, cycle sequencing, single-base extension sequencing, solid-phase sequencing, high-throughput sequencing, massively parallel signature sequencing, emulsion PCR, co-amplification at lower denaturation temperature-PCR (COLD-PCR), multiplex PCR, sequencing by reversible dye terminator, paired-end sequencing, near-term sequencing, exonuclease sequencing, sequencing by ligation, short-read sequencing, single-molecule sequencing, real-time sequencing, reverse-terminator sequencing, nanopore sequencing, MS-PET sequencing, and a combination thereof. In some embodiments, the sequencing method is massively parallel sequencing, that is, simultaneously (or in rapid succession) sequencing any of at least 100, 1000, 10,000, 100,000, 1 million, 10 million, 100 million, or 1 billion polynucleotide molecules. In some embodiments, sequencing can be performed by a gene analyzer such as, for example, gene analyzers commercially available from Illumina or Applied Biosystems. Sequencing of separated molecules has more recently been demonstrated by sequential or single extension reactions using polymerases or ligases as well as by single or sequential differential hybridizations with libraries of probes. Sequencing may be performed by a DNA sequencer (e.g., a machine designed to perform sequencing reactions).

**[0070]** A sequencing technique that can be used includes, for example, use of sequencing-by-synthesis systems. In the first step, DNA is sheared into fragments of approximately 300-800 base pairs, and the fragments are blunt ended. Oligonucleotide adaptors are then ligated to the ends of the fragments. The adaptors serve as primers for amplification and sequencing of the fragments. The fragments can be attached to DNA capture beads, e.g., streptavidin-coated beads using, e.g., Adaptor B, which contains 5′-biotin tag. The fragments attached to the beads are PCR amplified within droplets of an oil-water emulsion. The result is multiple copies of clonally amplified DNA fragments on each bead. In the second step, the beads are captured in wells (pico-liter sized). Pyrosequencing is performed on each DNA fragment in parallel. Addition of one or more nucleotides generates a light signal that is recorded by a CCD camera in a sequencing instrument. The signal strength is proportional to the number of nucleotides incorporated. Pyrosequencing makes use of pyrophosphate (PPi) which is released upon nucleotide addition. PPi is converted to ATP by ATP sulfurylase in the presence of adenosine 5′ phosphosulfate. Luciferase uses ATP to convert luciferin to oxyluciferin, and this reaction generates light that is detected and analyzed.

**[0071]** Another example of a DNA sequencing technique that can be used is SOLiD technology by Applied Biosystems from Life Technologies Corporation (Carlsbad, Calif.). In SOLiD sequencing, genomic DNA is sheared into fragments, and adaptors are attached to the 5′ and 3′ ends of the fragments to generate a fragment library. Alternatively, internal adaptors can be introduced by ligating adaptors to the 5′ and 3′ ends of the fragments, circularizing the

fragments, digesting the circularized fragment to generate an internal adaptor, and attaching adaptors to the 5′ and 3′ ends of the resulting fragments to generate a mate-paired library. Next, clonal bead populations are prepared in microreactors containing beads, primers, template, and PCR components. Following PCR, the templates are denatured and beads are enriched to separate the beads with extended templates. Templates on the selected beads are subjected to a 3′ modification that permits bonding to a glass slide. The sequence can be determined by sequential hybridization and ligation of partially random oligonucleotides with a central determined base (or pair of bases) that is identified by a specific fluorophore. After a color is recorded, the ligated oligonucleotide is removed and the process is then repeated.

**[0072]** Another example of a DNA sequencing technique that can be used is ion semiconductor sequencing using, for example, a system sold under the trademark ION TORRENT by Ion Torrent by Life Technologies (South San Francisco, Calif.). Ion semiconductor sequencing is described, for example, in Rothberg, et al., An integrated semiconductor device enabling non-optical genome sequencing, Nature 475:348-352 (2011); U.S. Pub. 2010/0304982; U.S. Pub. 2010/0301398; U.S. Pub. 2010/0300895; U.S. Pub. 2010/0300559; and U.S. Pub. 2009/0026082, the contents of each of which are incorporated by reference in their entirety.

**[0073]** Another example of a sequencing technology that can be used is Illumina sequencing. Illumina sequencing is based on the amplification of DNA on a solid surface using fold-back PCR and anchored primers. Genomic DNA is fragmented, and adapters are added to the 5′ and 3′ ends of the fragments. DNA fragments that are attached to the surface of flow cell channels are extended and bridge amplified. The fragments become double stranded, and the double stranded molecules are denatured. Multiple cycles of the solid-phase amplification followed by denaturation can create several million clusters of approximately 1,000 copies of single-stranded DNA molecules of the same template in each channel of the flow cell. Primers, DNA polymerase and four fluorophore-labeled, reversibly terminating nucleotides are used to perform sequential sequencing. After nucleotide incorporation, a laser is used to excite the fluorophores, and an image is captured and the identity of the first base is recorded. The 3′ terminators and fluorophores from each incorporated base are removed and the incorporation, detection and identification steps are repeated. Sequencing according to this technology is described in U.S. Pat. No. 7,960,120; U.S. Pat. No. 7,835,871; U.S. Pat. No. 7,232,656; U.S. Pat. No. 7,598,035; U.S. Pat. No. 6,911,345; U.S. Pat. No. 6,833,246; U.S. Pat. No. 6,828,100; U.S. Pat. No. 6,306,597; U.S. Pat. No. 6,210,891; U.S. Pub. 2011/0009278; U.S. Pub. 2007/0114362; U.S. Pub. 2006/0292611; and U.S. Pub. 2006/0024681, each of which are incorporated by reference in their entirety.

**[0074]** Another example of a sequencing technology that can be used includes the single molecule, real-time (SMRT) technology of Pacific Biosciences (Menlo Park, Calif.). In SMRT, each of the four DNA bases is attached to one of four different fluorescent dyes. These dyes are phospholinked. A single DNA polymerase is immobilized with a single molecule of template single stranded DNA at the bottom of a zero-mode waveguide (ZMW). It takes several milliseconds to incorporate a nucleotide into a growing strand. During this time, the fluorescent label is excited and produces a fluorescent signal, and the fluorescent tag is cleaved off. Detection of the corresponding fluorescence of the dye indicates which base was incorporated. The process is repeated.

**[0075]** Another example of a sequencing technique that can be used is nanopore sequencing (Soni & Meller, 2007, Progress toward ultrafast DNA sequence using solid-state nanopores, Clin Chem 53(11):1996-2001). A nanopore is a small hole, of the order of 1 nanometer in diameter. Immersion of a nanopore in a conducting fluid and application of a potential across it results in a slight electrical current due to conduction of ions through the nanopore. The amount of current which flows is sensitive to the size of the nanopore. As a DNA molecule passes through a nanopore, each nucleotide on the DNA molecule obstructs the nanopore to a different degree. Thus, the change in the current passing through the nanopore as the DNA molecule passes through the nanopore represents a reading of the DNA sequence.

**[0076]** Another example of a sequencing technique that can be used involves using a chemical-sensitive field effect transistor (chemFET) array to sequence DNA (for example, as described in U.S. Pub. 2009/0026082). In one example of the technique, DNA molecules can be placed into reaction chambers, and the template molecules can be hybridized to a sequencing primer bound to a polymerase. Incorporation of one or more triphosphates into a new nucleic acid strand at the 3′ end of the sequencing primer can be detected by a change in current by a chemFET. An array can have multiple chemFET sensors. In another example, single nucleic acids can be attached to beads, and the nucleic acids can be amplified on the bead, and the individual beads can be transferred to individual reaction chambers on a chemFET array, with each chamber having a chemFET sensor, and the nucleic acids can be sequenced.

**[0077]** Another example of a sequencing technique that can be used involves using an electron microscope as described, for example, by Moudrianakis, E. N. and Beer M., in Base sequence determination in nucleic acids with the electron microscope, III. Chemistry and microscopy of guanine-labeled DNA, PNAS 53:564-71 (1965). In one example of the technique, individual DNA molecules are labeled using metallic labels that are distinguishable using an electron microscope. These molecules are then stretched on a flat surface and imaged using an electron microscope to measure sequences.

**[0078]** Sequencing according to embodiments of the invention generates a plurality of reads. Reads according to the invention generally include sequences of nucleotide data less than about 150 bases in length, or less than about 90 bases in length. In certain embodiments, reads are between about 80 and about 90 bases, e.g., about 85 bases in length. In some embodiments, methods of the invention are applied to very short reads, i.e., less than about 50 or about 30 bases in length. Sequence read data can include the sequence data as well as meta information. Sequence read data can be stored in any suitable file format including, for example, VCF files, FASTA files or FASTQ files, as are known to those of skill in the art.

**[0079]** FASTA is originally a computer program for searching sequence databases and the name FASTA has come to also refer to a standard file format. See Pearson & Lipman, 1988, Improved tools for biological sequence comparison, PNAS 85:2444-2448. A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column. The word following the ">" symbol is the identifier of the sequence, and the rest of the line is the description (both are optional). There should be no space between the ">" and the first letter of the identifier. It is recommended that all lines of text be shorter than 80 characters. The sequence ends if another line starting with a ">" appears; this indicates the start of another sequence.

**[0080]** The FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. It is similar to the FASTA format but with quality scores following the sequence data. Both the sequence letter and quality score are encoded with a single ASCII character for brevity. The FASTQ format is a de facto standard for storing the output of high throughput sequencing instruments such as the Illumina Genome Analyzer. Cock et al., 2009, The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants, Nucleic Acids Res 38(6):1767-1771.

**[0081]** For FASTA and FASTQ files, meta information includes the description line and not the lines of sequence data. In some embodiments, for FASTQ files, the meta information includes the quality scores. For FASTA and FASTQ files, the sequence data begins after the description line and is present typically using some subset of IUPAC ambiguity codes optionally with "–". In a preferred embodiment, the sequence data will use the A, T, C, G, and N characters, optionally including "–" or U as-needed (e.g., to represent gaps or uracil).

**[0082]** In some embodiments, the at least one master sequence read file and the output file are stored as plain text files (e.g., using encoding such as ASCII; ISO/IEC 646; EBCDIC; UTF-8; or UTF-16). A computer system provided by the invention may include a text editor program capable of opening the plain text files. A text editor program may refer to a computer program

capable of presenting contents of a text file (such as a plain text file) on a computer screen, allowing a human to edit the text (e.g., using a monitor, keyboard, and mouse). Exemplary text editors include, without limit, Microsoft Word, emacs, pico, vi, BBEdit, and TextWrangler. Preferably, the text editor program is capable of displaying the plain text files on a computer screen, showing the meta information and the sequence reads in a human-readable format (e.g., not binary encoded but instead using alphanumeric characters as they would be used in print human writing).

**[0083]** While methods have been discussed with reference to FASTA or FASTQ files, methods and systems of the invention may be used to compress any suitable sequence file format including, for example, files in the Variant Call Format (VCF) format. A typical VCF file will include a header section and a data section. The header contains an arbitrary number of meta-information lines, each starting with characters '##', and a TAB delimited field definition line starting with a single '#' character. The field definition line names eight mandatory columns and the body section contains lines of data populating the columns defined by the field definition line. The VCF format is described in Danecek et al., 2011, The variant call format and VCFtools, Bioinformatics 27(15):2156-2158. The header section may be treated as the meta information to write to the compressed files and the data section may be treated as the lines, each of which will be stored in a master file only if unique.

**[0084]** Certain embodiments of the invention provide for the assembly of sequence reads. In assembly by alignment, for example, the reads are aligned to each other or to a reference. By aligning each read, in turn to a reference genome, all of the reads are positioned in relationship to each other to create the assembly. In addition, aligning or mapping the sequence read to a reference sequence can also be used to identify variant sequences within the sequence read. Identifying variant sequences can be used in combination with the methods and systems described herein to further aid in the diagnosis or prognosis of a disease or condition, or for guiding treatment decisions.

**[0085]** In some embodiments, any or all of the steps of the invention are automated. Alternatively, methods of the invention may be embodied wholly or partially in one or more dedicated programs, for example, each optionally written in a compiled language such as C++ then compiled and distributed as a binary. Methods of the invention may be implemented wholly or in part as modules within, or by invoking functionality within, existing sequence analysis platforms. In certain embodiments, methods of the invention include a number of steps that are all invoked automatically responsive to a single starting queue (e.g., one or a combination of triggering events sourced from human activity, another computer program, or a machine). Thus, the invention provides methods in which any or the steps or any combination

of the steps can occur automatically responsive to a queue. Automatically generally means without intervening human input, influence, or interaction (i.e., responsive only to original or pre-queue human activity).

**[0086]** The system also encompasses various forms of output, which includes an accurate and sensitive interpretation of the subject nucleic acid. The output of retrieval can be provided in the format of a computer file. In certain embodiments, the output is a FASTA file, FASTQ file, or VCF file. Output may be processed to produce a text file, or an XML file containing sequence data such as a sequence of the nucleic acid aligned to a sequence of the reference genome. In other embodiments, processing yields output containing coordinates or a string describing one or more mutations in the subject nucleic acid relative to the reference genome. Alignment strings known in the art include Simple UnGapped Alignment Report (SUGAR), Verbose Useful Labeled Gapped Alignment Report (VULGAR), and Compact Idiosyncratic Gapped Alignment Report (CIGAR) (Ning, Z., et al., Genome Research 11(10):1725-9 (2001)). These strings are implemented, for example, in the Exonerate sequence alignment software from the European Bioinformatics Institute (Hinxton, UK).

**[0087]** In some embodiments, a sequence alignment is produced—such as, for example, a sequence alignment map (SAM) or binary alignment map (BAM) file—comprising a CIGAR string (the SAM format is described, e.g., in Li, et al., The Sequence Alignment/Map format and SAMtools, Bioinformatics, 2009, 25(16):2078-9). In some embodiments, CIGAR displays or includes gapped alignments one-per-line. CIGAR is a compressed pairwise alignment format reported as a CIGAR string. A CIGAR string is useful for representing long (e.g. genomic) pairwise alignments. A CIGAR string is used in SAM format to represent alignments of reads to a reference genome sequence.

**[0088]** A CIGAR string follows an established motif. Each character is preceded by a number, giving the base counts of the event. Characters used can include M, I, D, N, and S (M=match; I=insertion; D=deletion; N=gap; S=substitution). The CIGAR string defines the sequence of matches/mismatches and deletions (or gaps). For example, the CIGAR string 2MD3M2D2M will mean that the alignment contains 2 matches, 1 deletion (number 1 is omitted in order to save some space), 3 matches, 2 deletions and 2 matches.

**[0089]** As contemplated by the invention, the functions described above can be implemented using a system of the invention that includes software, hardware, firmware, hardwiring, or any combinations of these. Features implementing functions can also be physically located at various positions, including being distributed such that portions of functions are implemented at different physical locations.

**[0090]** As one skilled in the art would recognize as necessary or best-suited for performance of the methods of the invention, a computer system or machines of the invention include one or more processors (e.g., a central processing unit (CPU) a graphics processing unit (GPU) or both), a main memory and a static memory, which communicate with each other via a bus.

**[0091]** FIG. 4 diagrams a system 701 suitable for performing methods of the invention. As shown in FIG. 7, system 701 may include one or more of a server computer 705, a terminal 715, a sequencer 715, a sequencer computer 721, a computer 749, or any combination thereof. Each such computer device may communicate via network 709. Sequencer 725 may optionally include or be operably coupled to its own, e.g., dedicated, sequencer computer 721 (including any input/output mechanisms (I/O), processor, and memory). Additionally or alternatively, sequencer 725 may be operably coupled to a server 705 or computer 749 (e.g., laptop, desktop, or tablet) via network 709. Computer 749 includes one or more processor, memory, and I/O. Where methods of the invention employ a client/server architecture, any steps of methods of the invention may be performed using server 705, which includes one or more of processor, memory, and I/O, capable of obtaining data, instructions, etc., or providing results via an interface module or providing results as a file. Server 705 may be engaged over network 709 through computer 749 or terminal 715, or server 705 may be directly connected to terminal 715. Terminal 515 is preferably a computer device. A computer according to the invention preferably includes one or more processor coupled to an I/O mechanism and memory.

**[0092]** A processor may be provided by one or more processors including, for example, one or more of a single core or multi-core processor. An I/O mechanism may include a video display unit (e.g., a liquid crystal display (LCD) or a cathode ray tube (CRT)), an alphanumeric input device (e.g., a keyboard), a cursor control device (e.g., a mouse), a disk drive unit, a signal generation device (e.g., a speaker), an accelerometer, a microphone, a cellular radio frequency antenna, and a network interface device (e.g., a network interface card (NIC), Wi-Fi card, cellular modem, data jack, Ethernet port, modem jack, HDMI port, mini-HDMI port, USB port), touchscreen (e.g., CRT, LCD, LED, AMOLED, Super AMOLED), pointing device, trackpad, light (e.g., LED), light/image projection device, or a combination thereof. Memory according to the invention refers to a non-transitory memory which is provided by one or more tangible devices which preferably include one or more machine-readable medium on which is stored one or more sets of instructions (e.g., software) embodying any one or more of the methodologies or functions described herein. The software may also reside, completely or at least partially, within the main memory, processor, or both during execution thereof by a computer within system 701, the main memory and the processor also constituting machine-readable media. The software may further be transmitted or received over a network via the network interface device.

**[0093]** While the machine-readable medium can in an exemplary embodiment be a single medium, the term "machine-readable medium" should be taken to include a single medium or multiple media (e.g., a centralized or distributed database, and/or associated caches and servers) that store the one or more sets of instructions. The term "machine-readable medium" shall also be taken to include any medium that is capable of storing, encoding or carrying a set of instructions for execution by the machine and that cause the machine to perform any one or more of the methodologies of the present invention. Memory may be, for example, one or more of a hard disk drive, solid state drive (SSD), an optical disc, flash memory, zip disk, tape drive, "cloud" storage location, or a combination thereof. In certain embodiments, a device of the invention includes a tangible, non-transitory computer readable medium for memory. Exemplary devices for use as memory include semiconductor memory devices, (e.g., EPROM, EEPROM, solid state drive (SSD), and flash memory devices e.g., SD, micro SD, SDXC, SDIO, SDHC cards); magnetic disks, (e.g., internal hard disks or removable disks); and optical disks (e.g., CD and DVD disks).

**[0094]** In some embodiments, the methods and systems of the present disclosure can be used to diagnose a disease or condition, e.g., cancer. The term "diagnosis" as used herein refers to methods by which the skilled worker can estimate and/or determine whether or not a patient is suffering from a given disease or condition. In some embodiments, the methods of the invention can be used in the prognosis of a disease of a disease or condition, e.g., cancer. The term "prognosis" as used herein refers to the likelihood of a disease or condition progression, including recurrence of a disease or condition. In some embodiments, the methods of the invention can be used to assess the risk of developing a disease or condition, e.g., cancer. For example, the methods and systems described herein can be used to identify a breakpoint or gene fusion associated with a particular diagnosis, prognosis, or risk of developing a disease or condition. Moreover, the methods and systems described herein can be used to identify a breakpoint or gene fusion associated with a predicted therapeutic outcome. Thus, the methods and systems can be used to guide the treatment of the disease or condition (e.g., by administering a compound or agent to a subject), or to guide the preparation of a medicament for treatment of the disease or condition.

**[0095]** As used herein, "treating" a disease or condition refers to taking steps to obtain beneficial or desired results, including clinical results. Beneficial or desired clinical results include, but are not limited to, alleviation or amelioration of one or more symptoms associated with diseases or conditions. As used herein, "administering" or "administration of" a compound or an agent to a subject can be carried out using one of a variety of methods known to those skilled in the art. For example, a compound or an agent can be administered, intravenously,

arterially, intradermally, intramuscularly, intraperitonealy, intravenously, subcutaneously, ocularly, sublingually, orally (by ingestion), intranasally (by inhalation), intraspinally, intracerebrally, and transdermally (by absorbtion, e.g., through a skin duct). A compound or agent can also appropriately be introduced by rechargeable or biodegradable polymeric devices or other devices, e.g., patches and pumps, or formulations, which provide for the extended, slow, or controlled release of the compound or agent. Administering can also be performed, for example, once, a plurality of times, and/or over one or more extended periods. In some aspects, the administration includes both direct administration, including self-administration, and indirect administration, including the act of prescribing a drug. For example, as used herein, a physician who instructs a patient to self-administer a drug, or to have the drug administered by another and/or who provides a patient with a prescription for a drug is administering the drug to the patient. In some embodiments, a compound or an agent is administered orally, e.g., to a subject by ingestion, or intravenously, e.g., to a subject by injection. In some embodiments, the orally administered compound or agent is in an extended release or slow release formulation, or administered using a device for such slow or extended release.

**[0096]** As used herein, the term "cancer" includes, but is not limited to, various types of malignant neoplasms, most of which can invade surrounding tissues, and may metastasize to different sites (see, for example, PDR Medical Dictionary, 1st edition (1995), incorporated herein by reference in its entirety for all purposes). The terms "neoplasm" and "tumor" refer to an abnormal tissue that grows by cellular proliferation more rapidly than normal and continues to grow after the stimuli that initiated proliferation is removed. Such abnormal tissue shows partial or complete lack of structural organization and functional coordination with the normal tissue which may be either benign (such as a benign tumor) or malignant (such as a malignant tumor). Examples of general categories of cancer include, but are not limited to, carcinomas (malignant tumors derived from epithelial cells such as, for example, common forms of breast, prostate, lung and colon cancer), sarcomas (malignant tumors derived from connective tissue or mesenchymal cells), lymphomas (malignancies derived from hematopoietic cells), leukemias (malignancies derived from hematopoietic cells), and germ cell tumors (tumors derived from totipotent cells, in adults most often found in the testicle or ovary; in fetuses, babies and young children, most often found on the body midline, particularly at the tip of the tailbone), blastic tumors (a typically malignant tumor which resembles an immature or embryonic tissue) and the like. Examples of the types of neoplasms intended to be encompassed by the present invention include but are not limited to those neoplasms associated with cancers of neural tissue, blood forming tissue, breast, skin, bone, prostate, ovaries, uterus, cervix, liver, lung, brain, larynx, gallbladder, pancreas, rectum, parathyroid, thyroid, adrenal gland, immune system, head and

neck, colon, stomach, bronchi, and/or kidneys. In particular embodiments, types and number of cancers that may be detected include, but are not limited to, blood cancers, brain cancers, lung cancers, skin cancers, nose cancers, throat cancers, liver cancers, bone cancers, lymphomas, pancreatic cancers, skin cancers, bowel cancers, rectal cancers, thyroid cancers, bladder cancers, kidney cancers, mouth cancers, stomach cancers, solid state tumors, heterogeneous tumors, homogenous tumors and the like. In particular embodiments, the cancer is a hematological cancer, a sarcoma, or a prostate cancer.

**[0097]** While preferred embodiments of the present invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the invention. It should be understood that various alternatives to the embodiments of the invention described herein may be employed in practicing the invention. It is intended that the following claims define the scope of the invention and that methods and structures within the scope of these claims and their equivalents be covered thereby.

## CLAIMS

WHAT IS CLAIMED IS:

1. A method for processing genetic sequence read data from a sample, the method comprising:

determining a fused read containing sequencing data of at least a portion of a fused chromosome DNA molecule;

determining a predetermined point on the genome with least one mapped portion of the fused read clipped at the predetermined point (a breakpoint);

identifying two mapped read portions from two breakpoints (breakpoint pair) as a potential fusion candidate;

creating one or more fusion sets based on breakpoint pairs and clustering the fusion sets into one or more fusion clusters; and

identifying each fusion cluster meeting a predetermined criterion as a gene fusion.

2. The method of claim 1, comprising assigning a unique molecule or read identifier (read ID) to each read.

3. The method of claim 1, comprising clipping each mapped portion of the reads from one or both sides.

4. The method of claim 1, wherein the breakpoints are independent of the reads in identity and are identified by a sign, a chromosome and a position.

5. The method of claim 4, wherein the breakpoints keep statistics including a number of reads and molecules that are clipped or split at the breakpoint, and a number of wild-type reads and molecules that pass over the breakpoint.

6. The method of claim 2, comprising selecting every two mapped read portions with common read IDs that belong to two breakpoints with appropriate signs as a potential fusion candidate.

7. The method of claim 6, wherein the potential fusion candidate location in the original read before mapping shows the read portion as originally located next to each other.

8. The method of claim 6, comprising checking if read portions are mapped on one strand for differences in the breakpoints' signs.

9. The method of claim 1, comprising tracking fusion set statistics

10. The method of claim 9, wherein the fusion set statistics are breakpoint IDs, number of molecules, or reads that are contained in the set.

11. The method of claim 1, comprising grouping fusion sets with similar breakpoints in a fusion cluster.

12. The method of claim 11, wherein the similar breakpoints are breakpoints no more than 5 nucleotides, no more than 10 nucleotides or no more than 25 nucleotides apart.

13. The method of claim 1, comprising defining a fusion cluster between two regions in in the genome.

14. The method of claim 1, comprising determining for the fusion cluster a number of fused molecules for each partner.

15. The method of claim 1, comprising determining for the fusion cluster the number of fused reads for each partner.

16. The method of claim 1, comprising determining for the fusion cluster a number of wild-type molecules for each partner.

17. The method of claim 1, comprising determining for the fusion cluster a number of wild-type reads or molecules for each partner.

18. The method of claim 1, comprising determining for the fusion cluster a fusion percentage for each partner as a ratio of fused molecules over total molecule for each partner.

19. The method of claim 1, wherein the total molecule comprises wild-type and clipped components.

20. The method of claim 18, comprising determining for the fusion cluster gene information for each partner.

21. The method of claim 1, comprising determining a downstream gene of the fusion cluster.

22. The method of claim 1, wherein the criteria comprises having more than one molecule in the cluster or having at least a molecule with both Watson and Crick strands.

23. A system to analyze genetic information, comprising

a DNA sequencer;

a processor coupled to the DNA sequencer, the processor running computer code to process genetic sequence read data from a sample, the computer code comprising instructions for:

determining a fused read containing sequencing data of a portion of a fused chromosome DNA molecule;

determining at least a predetermined point on the genome with least one mapped portion of the fused read clipped at the predetermined point (a breakpoint);

identifying two mapped read portions from two breakpoints (breakpoint pair) as a potential fusion candidate;

creating one or more fusion sets based on breakpoint pairs and clustering the fusion sets into one or more fusion clusters; and

identifying each fusion cluster meeting a predetermined criterion as a gene fusion.

24. A method comprising:

(a) sequencing DNA molecules with a DNA sequencer to generate a collection of sequences;

(b) mapping the collection of sequences to a reference genome;

(c) identifying fused reads from the mapped collection, wherein a fused read contains sub-sequences, wherein a first sub-sequence maps to a first genetic locus and a second sub-sequence maps to a second, distinct genetic locus;

(d) for each fused read, identifying a first breakpoint at the first genetic locus and a second breakpoint at the second genetic locus, wherein a breakpoint is a point on the reference genome where a sequence of a fused read is clipped, and wherein the first and second breakpoints form a breakpoint pair;

(e) generating sets of fused reads, each set comprising fused reads having the same breakpoint pair;

(f) clustering sets of fused reads, wherein each cluster is formed from sets of fused reads having first breakpoints within a first predetermined nucleotide distance and second breakpoints within a second predetermined nucleotide distance; and

(g) determining a gene fusion for one or more clusters, wherein a gene fusion for a cluster has, as a first fusion gene breakpoint, a breakpoint selected from the first breakpoints in the cluster and, as a second fusion gene breakpoint, a breakpoint selected from the second breakpoints in the cluster, and wherein the first and second fusion gene breakpoints are each selected based on selection criteria.

25. The method of claim 24, wherein the distinct genetic loci are located on different chromosomes or on different genes of the same chromosome.

26. The method of claim 24 wherein the first and second predetermined distances are each no more than 5 nucleotides, no more than 10 nucleotides or no more than 25 nucleotides.

27. The method of claim 24, wherein the selection criteria include the breakpoint having the most fused reads in the cluster.

28. The method of claim 24, comprising determining a gene fusion for a plurality of gene clusters.

29. A method comprising:

(a) sequencing a plurality of DNA molecules with a DNA sequencer;

(b) tagging each of the plurality of sequences molecules with an identifier;

(c) mapping each tagged sequence to a reference genome;

(d) identifying clipped reads from the mapped tagged sequences, wherein a clipped read is a tagged sequence containing a mapped portion and a clipped portion, wherein the mapped portion maps to a genetic locus and the clipped portion does not map to the genetic locus;

(e) determining a breakpoint of each clipped read, wherein a breakpoint is a point on the reference genome where a sequence of a clipped read is clipped;

(f) creating breakpoint sets, each breakpoint set comprising identifiers of clipped reads having the same breakpoint;

(g) creating sets of breakpoint pairs by comparing pairs of breakpoint sets, each set of breakpoint pairs including identifiers present in both members of a compared pair of breakpoint sets;

(h) clustering sets of breakpoint pairs, wherein each cluster includes sets of breakpoint pairs having a first breakpoint of the pair within a first predetermined genetic distance and a second breakpoint of the pair within a second predetermined genetic distance; and

(i) determining a gene fusion for one or more of the clusters, wherein a gene fusion for a cluster has, as a first fusion gene breakpoint, a breakpoint selected from the first breakpoints in the cluster and, as a second fusion gene breakpoint, a breakpoint selected from the second breakpoints in the cluster, and wherein the first and second fusion gene breakpoints are each selected based on a selection criteria.

30. The method of claim 29, wherein the selection criteria include the breakpoint having the most fused reads in the cluster.

31. A method for identifying a fusion gene breakpoint, the method comprising:

(a) determining a fused read containing sequencing data of at least a portion of a fused chromosome DNA molecule;

(b) determining a predetermined point on the genome with least one mapped portion of the fused read clipped at the predetermined point (a breakpoint);

(c) identifying two mapped read portions from two breakpoints (breakpoint pair) as a potential fusion candidate;

(d) creating one or more fusion sets based on breakpoint pairs and clustering the fusion sets into one or more fusion clusters;

(e) identifying each fusion cluster meeting a predetermined criterion as a gene fusion, and

(f) identifying a breakpoint of the gene fusion as the fusion gene breakpoint.

32. A method for diagnosing a condition in a subject, the method comprising:

(a) determining a fused read containing sequencing data of at least a portion of a fused chromosome DNA molecule;

(b) determining a predetermined point on the genome with least one mapped portion of the fused read clipped at the predetermined point (a breakpoint);

(c) identifying two mapped read portions from two breakpoints (breakpoint pair) as a potential fusion candidate;

(d) creating one or more fusion sets based on breakpoint pairs and clustering the fusion sets into one or more fusion clusters; and

(e) identifying each fusion cluster meeting a predetermined criterion as a gene fusion,

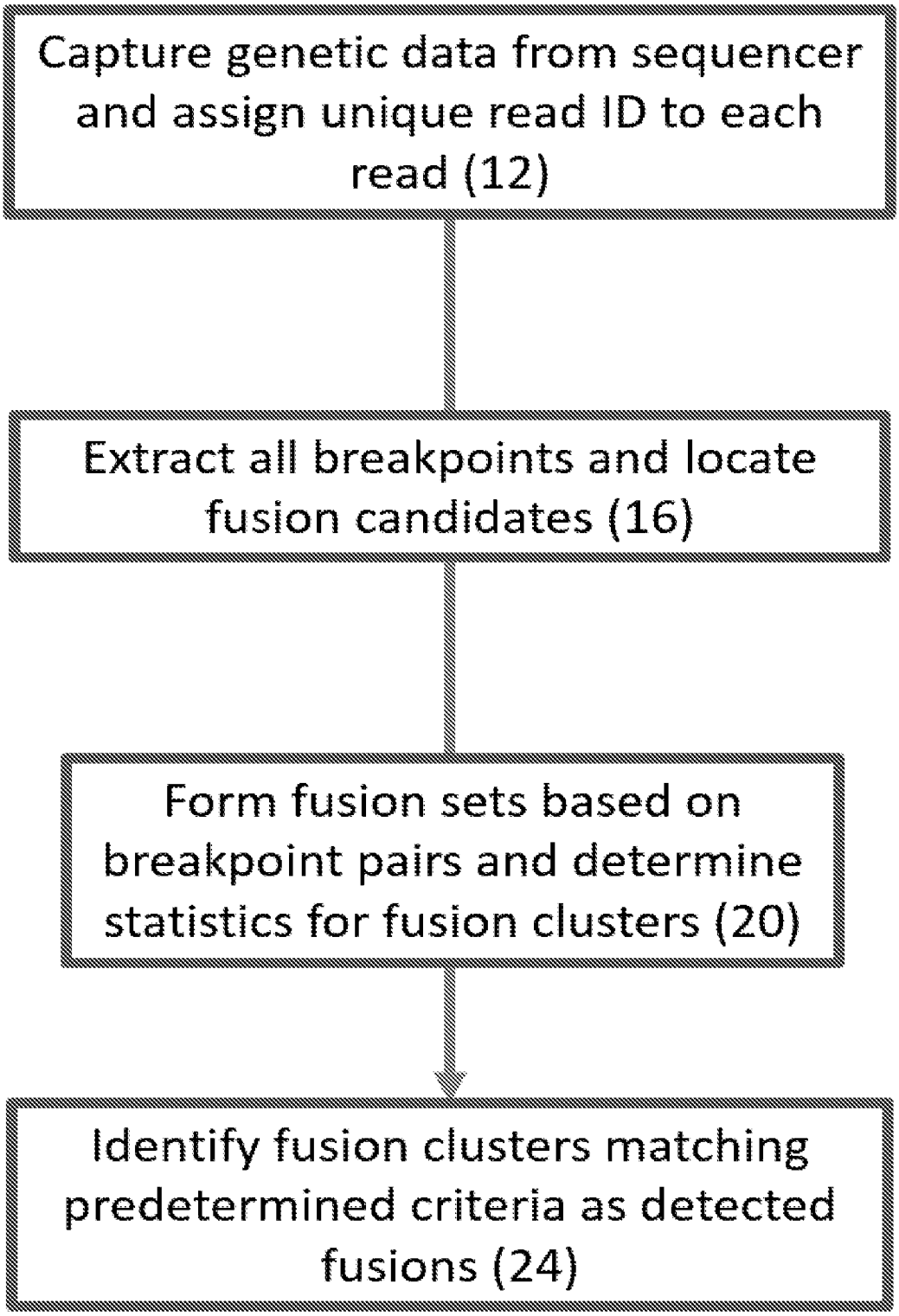
wherein said gene fusion is indicative of the condition.

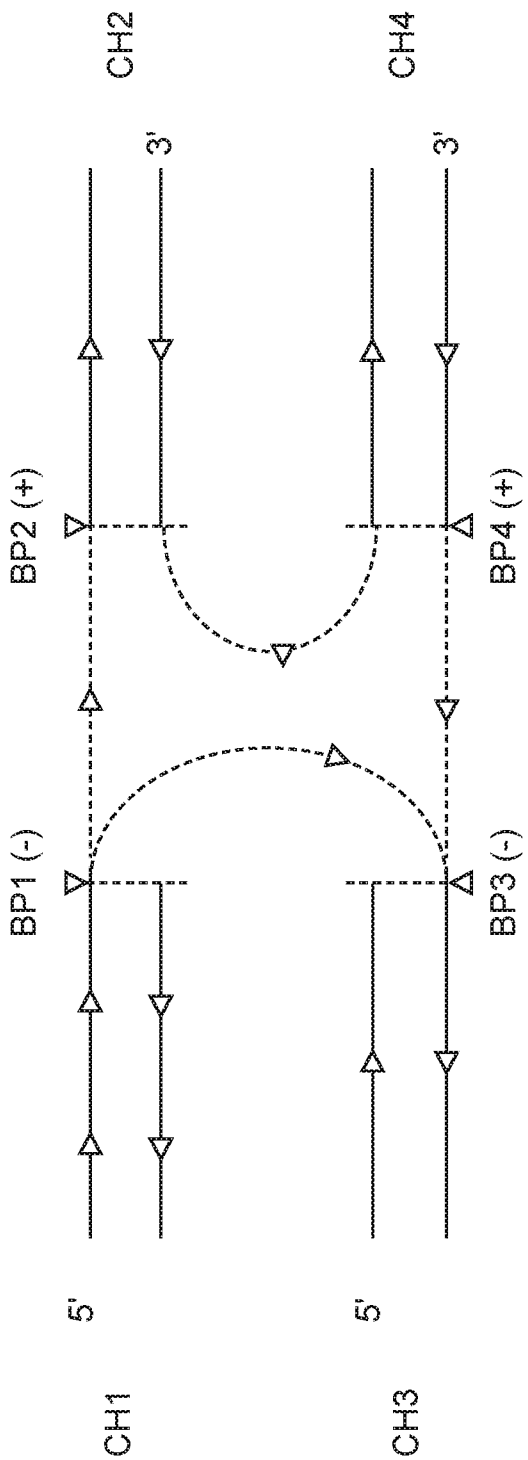
33. The method of claim 32, wherein the condition is a cancer.

34. The method of claim 33, wherein the cancer is selected from the group consisting of: a hematological cancer, a sarcoma, and a prostate cancer.

35. The method of claim 34, further comprising administering a treatment to the subject.

FIG. 1


Capture genetic data from sequencer and assign unique read ID to each read (12)
Extract all breakpoints and locate fusion candidates (16)
Form fusion sets based on breakpoint pairs and determine statistics for fusion clusters (20)
Identify fusion clusters matching predetermined criteria as detected fusions (24)




| Possible fusions | | | | | |
|---|---|---|---|---|---|
| | | | (upstream) | (downstream) | |
| fusion | breakpoint 1 | breakpoint 2 | Strand 1 | Strand 2 | |
| CH1-CH2 | BP1 (-) | BP2 (+) | CH1:5' | CH2:5' | reverse complement of each other |
| CH1-CH2 | BP2 (+) | BP1 (-) | CH2:3' | CH1:3' | |
| | | | | (The same fusion) | |
| CH1-CH3 | BP1 (-) | BP3 (-) | CH1:5' | CH3:3' | reverse complement of each other |
| CH3-CH1 | BP3 (-) | BP1 (-) | CH3:5' | CH1:3' | |
| CH2-CH4 | BP2 (+) | BP4 (+) | CH2:3' | CH4:5' | reverse complement of each other |
| CH4-CH2 | BP4 (+) | BP2 (+) | CH4:3' | CH2:5' | |

FIG. 2

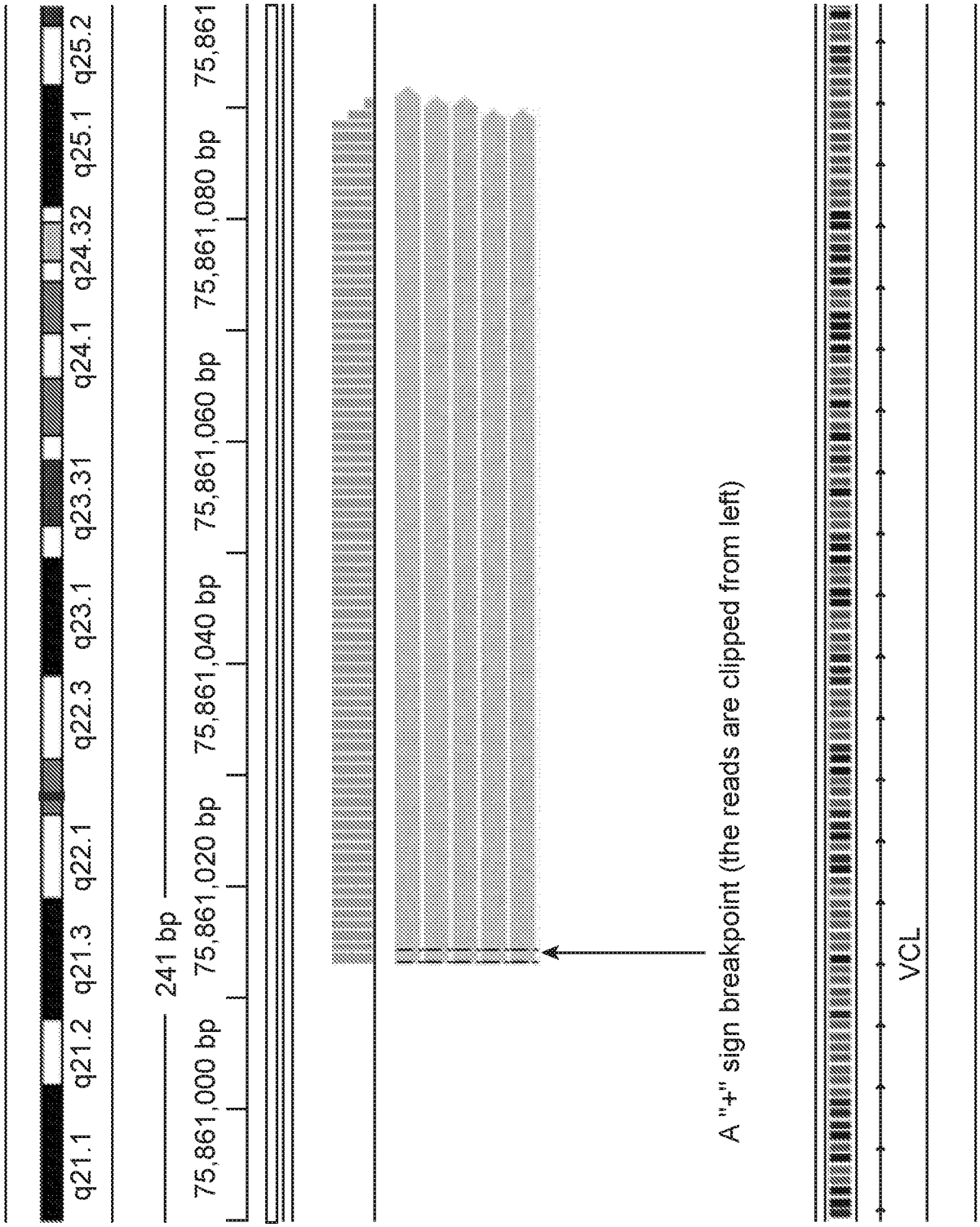

q21.1
q21.2
q21.3
q22.1
q22.3
q23.1
q23.31
q24.1
q24.32
q25.1
q25.2
241 bp
75,861,000 bp
75,861,020 bp
75,861,040 bp
75,861,060 bp
75,861,080 bp
75,861
A "+" sign breakpoint (the reads are clipped from left)
VCL

FIG. 3A

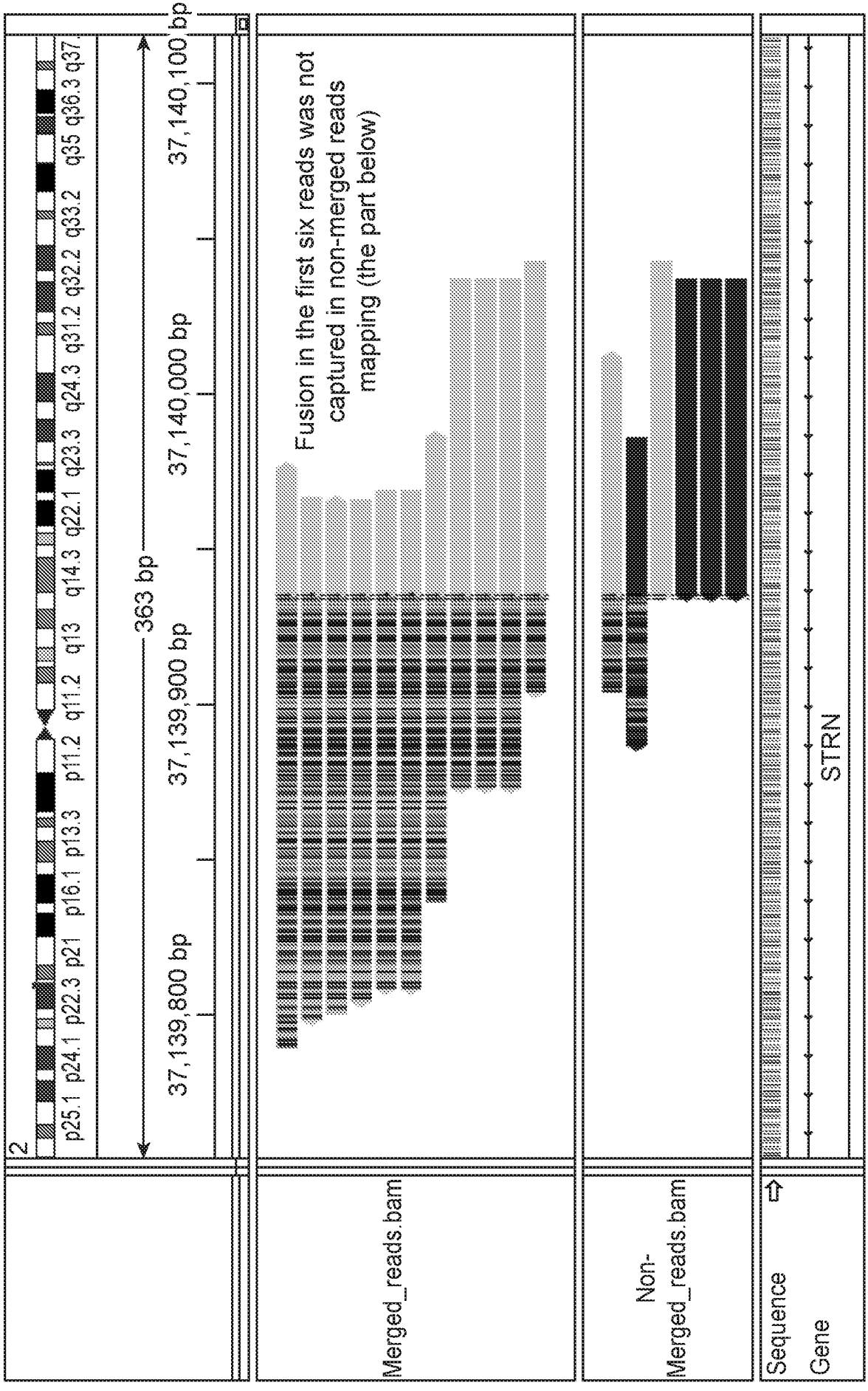

2
p25.1 p24.1 p22.3 p21 p16.1 p13.3 p11.2 q11.2 q13 q14.3 q22.1 q23.3 q24.3 q31.2 q32.2 q33.2 q35 q36.3 q37.
363 bp
37,139,800 bp
37,139,900 bp
37,140,000 bp
37,140,100 bp
Fusion in the first six reads was not captured in non-merged reads mapping (the part below)
Merged_reads.bam
Non-Merged_reads.bam
Sequence
Gene
STRN

FIG. 3B

FIG. 4

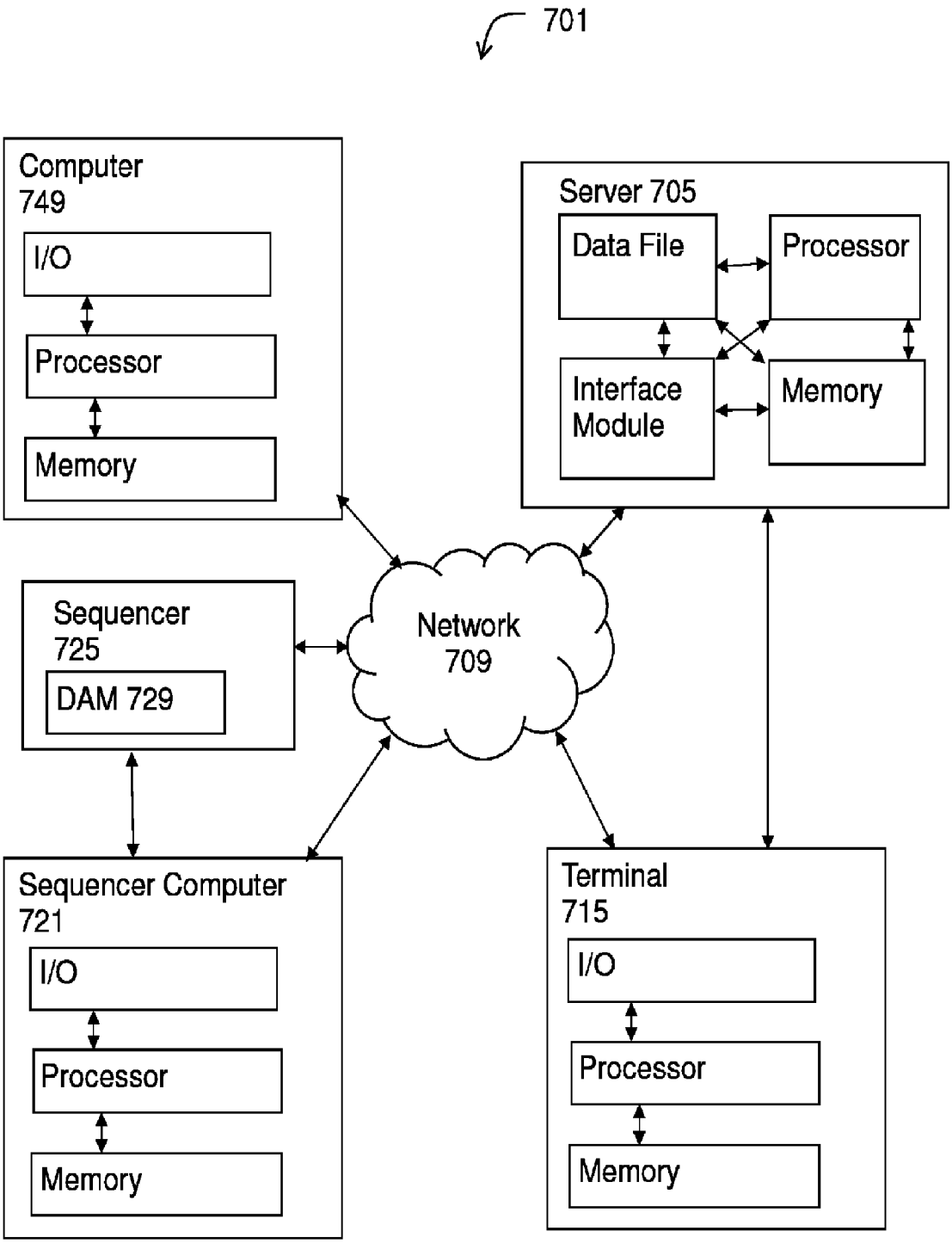

701
Computer
749
I/O
Processor
Memory
Server 705
Data File
Processor
Interface
Module
Memory
Network
709
Sequencer
725
DAM 729
Sequencer Computer
721
I/O
Processor
Memory
Terminal
715
I/O
Processor
Memory

FIG. 5

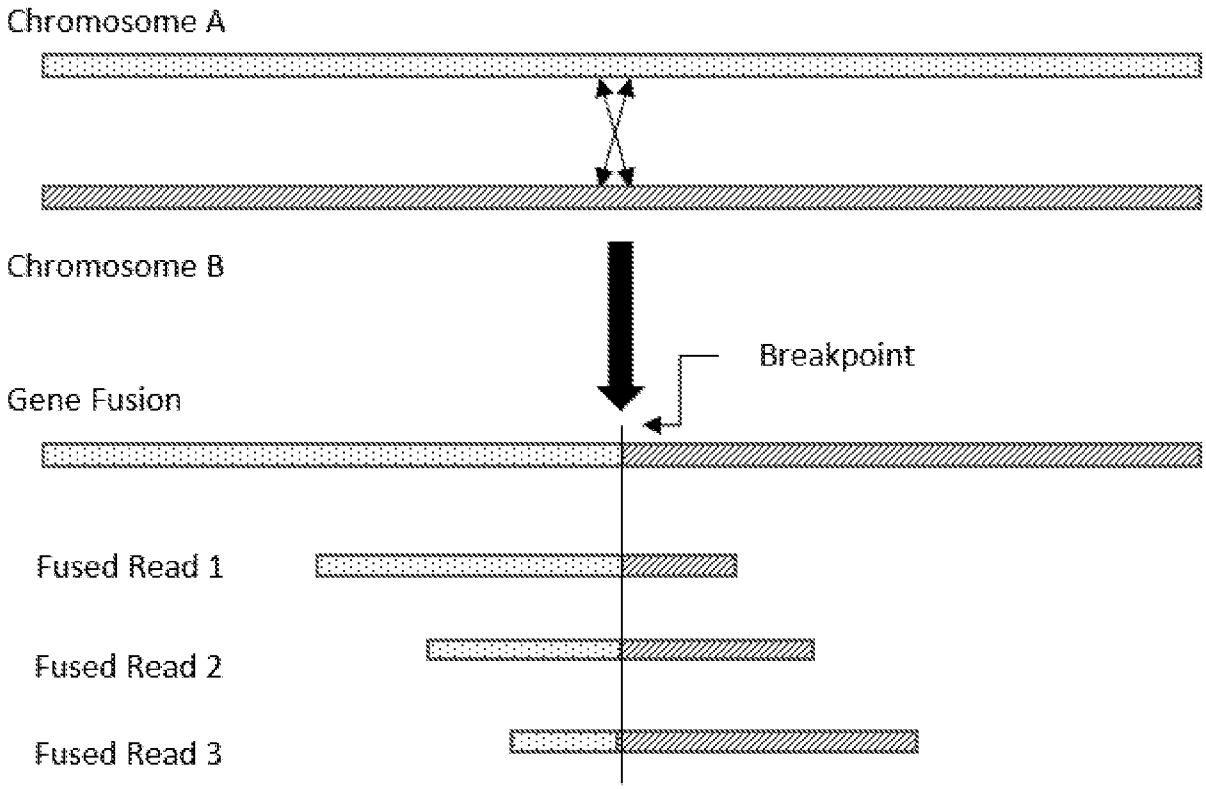

Chromosome A
Chromosome B
Breakpoint
Gene Fusion
Fused Read 1
Fused Read 2
Fused Read 3

FIG. 6

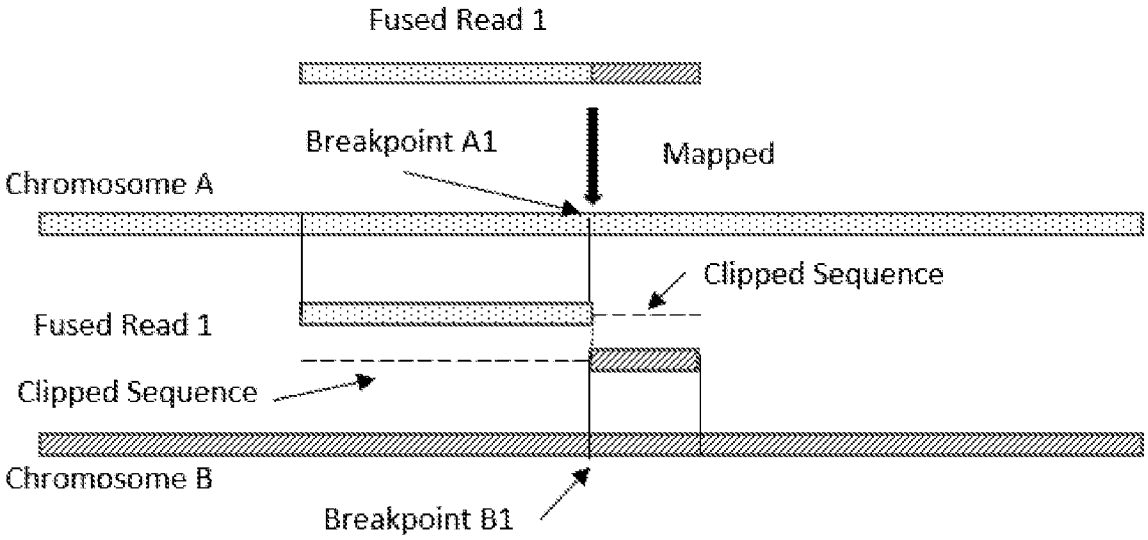

Fused Read 1
Breakpoint A1
Mapped
Chromosome A
Clipped Sequence
Fused Read 1
Clipped Sequence
Chromosome B
Breakpoint B1

FIG. 7

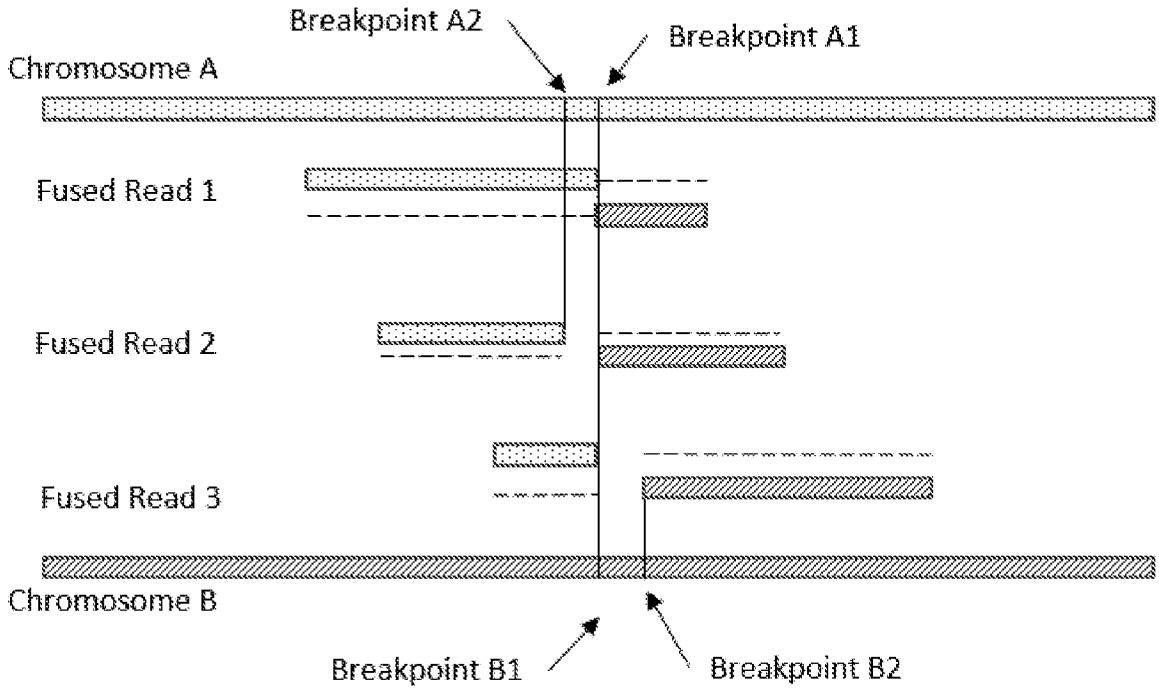
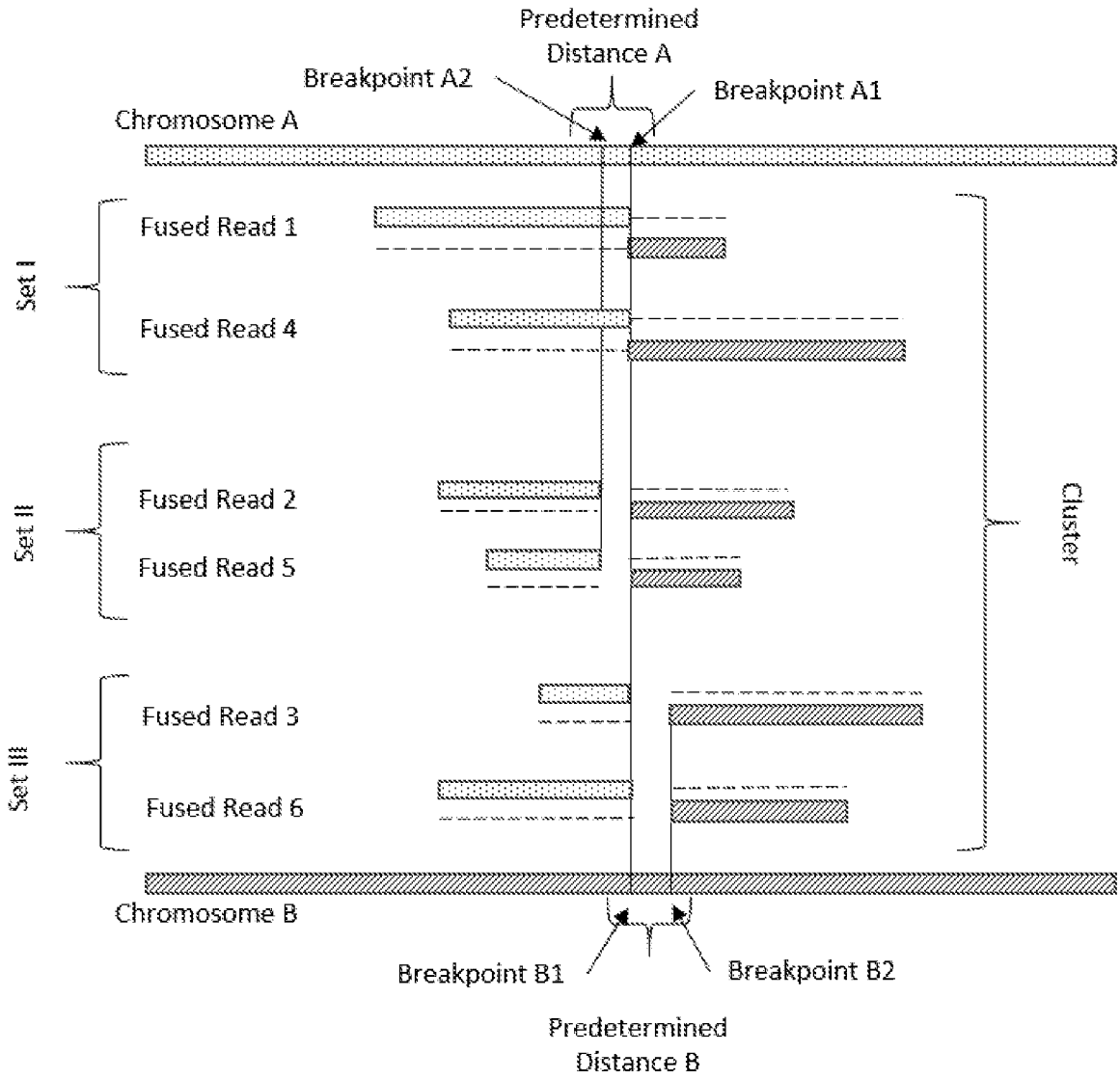

Breakpoint A2
Breakpoint A1
Chromosome A
Fused Read 1
Fused Read 2
Fused Read 3
Chromosome B
Breakpoint B1
Breakpoint B2

FIG. 8




Chromosome A:
Breakpoint A1: Fused Reads 1, 4, 3, 6 = four
Breakpoint A2: Fused Reads 2, 5 = two
First Gene Fusion Breakpoint Call: A1

Chromosome B:
Breakpoint B1: Fused Reads 1, 4, 2, 5 = four
Breakpoint B2: Fused Reads 3, 6 = two
Second Gene Fusion Breakpoint Call: B1

Gene Fusion Breakpoint pair: A1-B1

FIG. 9A

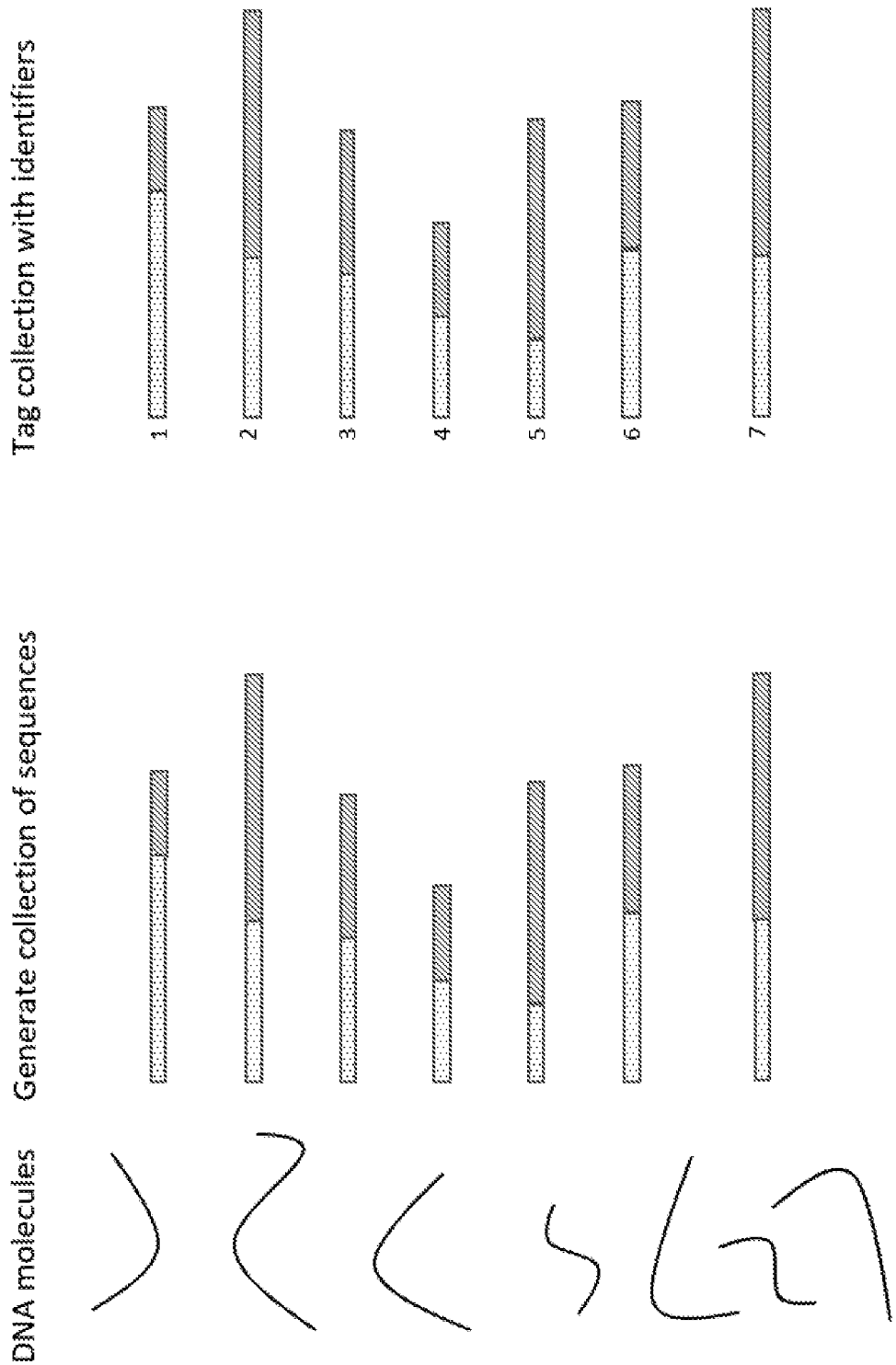

DNA molecules
Generate collection of sequences
Tag collection with identifiers
1
2
3
4
5
6
7

FIG. 9B

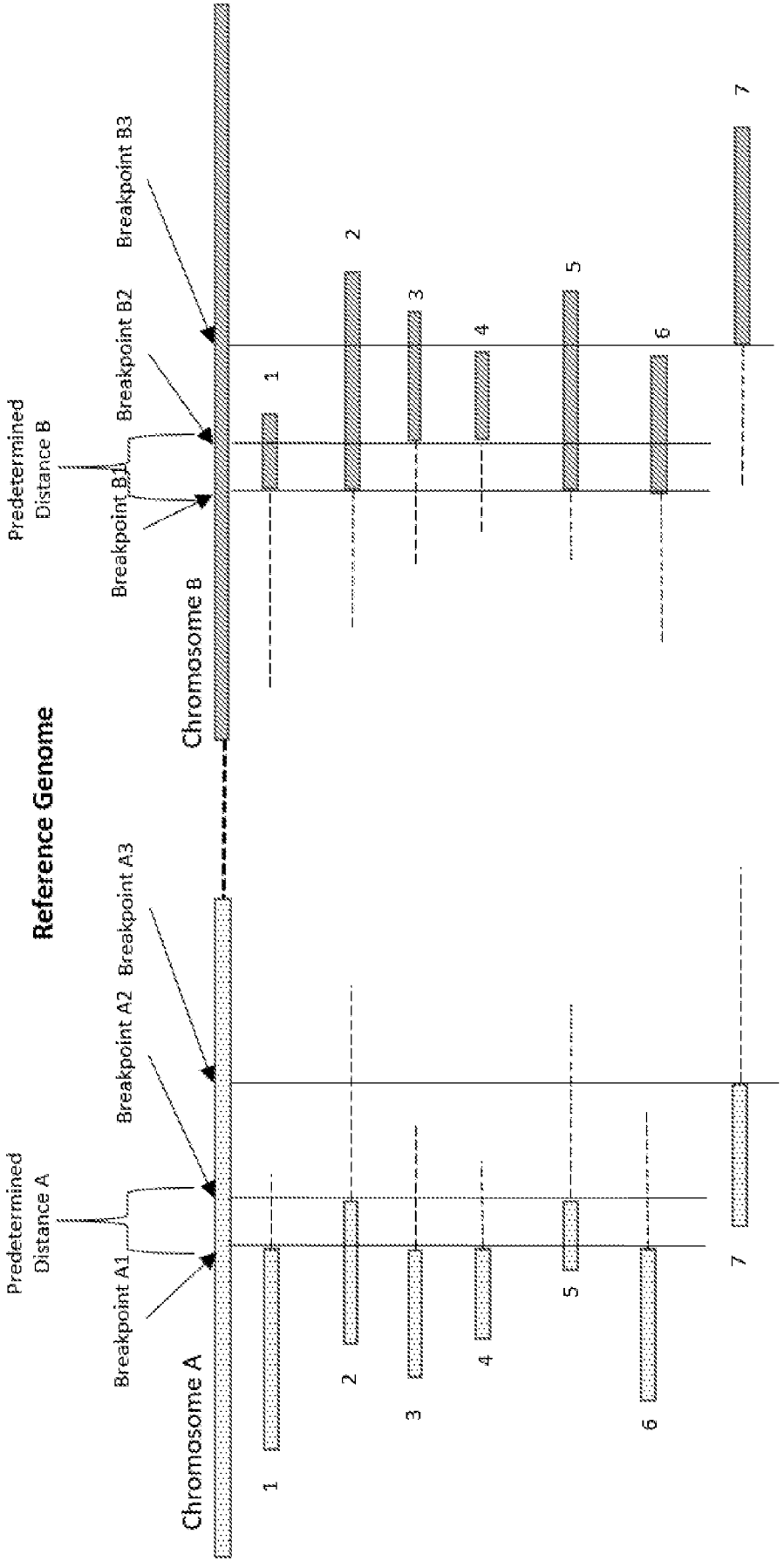
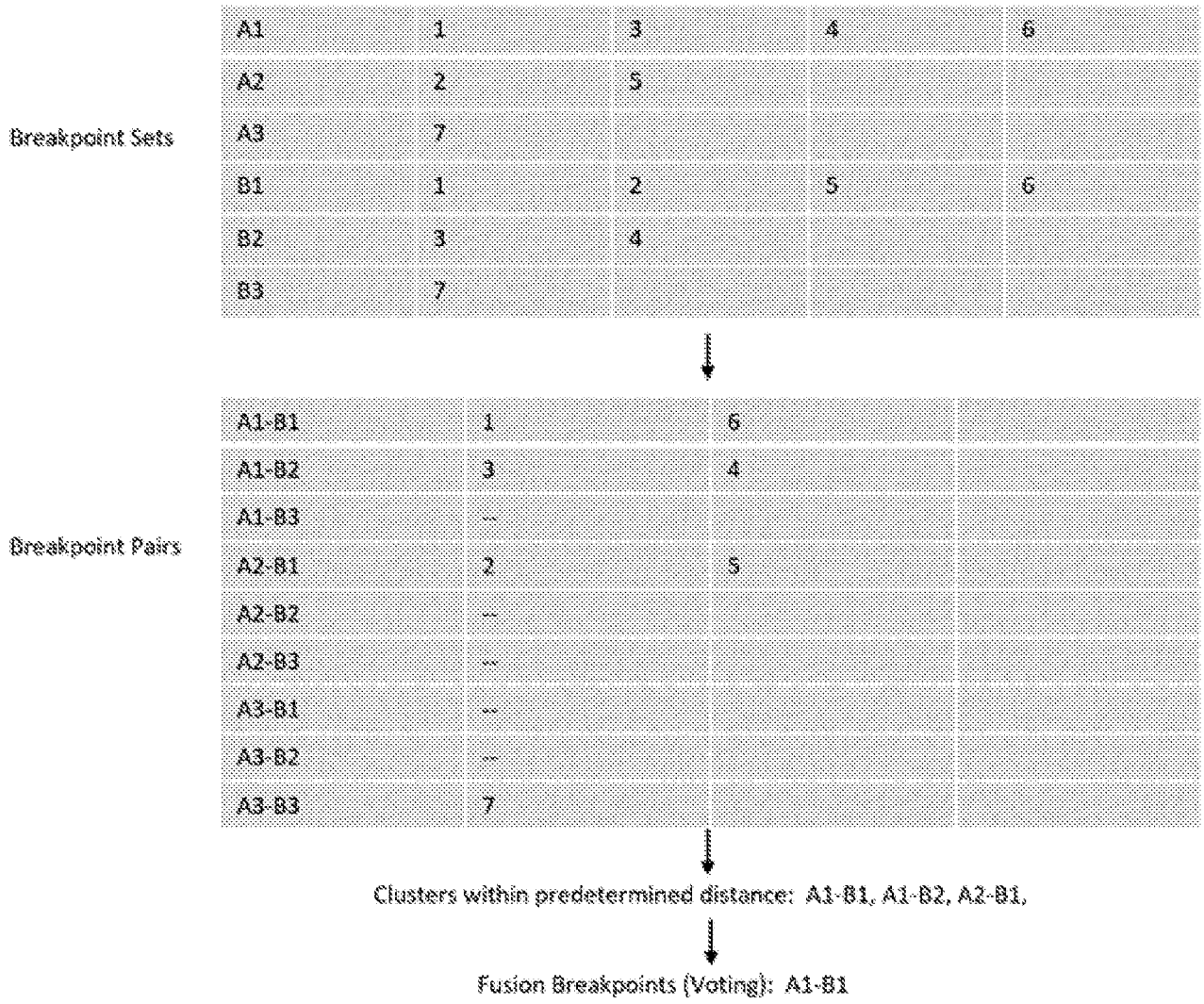

Reference Genome
Chromosome A
Chromosome B
Predetermined Distance A
Predetermined Distance B
Breakpoint A1
Breakpoint A2
Breakpoint A3
Breakpoint B1
Breakpoint B2
Breakpoint B3
1
2
3
4
5
6
7

FIG. 9C


Breakpoint Sets
A1 1 3 4 6
A2 2 5
A3 7
B1 1 2 5 6
B2 3 4
B3 7
Breakpoint Pairs
A1-B1 1 6
A1-B2 3 4
A1-B3 --
A2-B1 2 5
A2-B2 --
A2-B3 --
A3-B1 --
A3-B2 --
A3-B3 7
Clusters within predetermined distance: A1-B1, A1-B2, A2-B1,
Fusion Breakpoints (Voting): A1-B1

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/US16/56314

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - C12Q 1/68; C12N 15/11; G06F 19/20 (2016.01)
CPC - C12Q 1/6886, 1/6883, 1/6869; G06F 19/20
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC(8): C12Q 1/68; C12N 15/11, 1/21, 15/00, 15/52, 9/00, 15/09; G06F 19/20; C10G 32/00 (2016.01)
CPC: C12Q 1/6886, 1/6883, 1/6869, 1/6858, 2525/155, 2537/143, 2549/119, 2600/106, 2600/156; G06F 19/20; C10G 32/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PatSeer (US, EP, WO, JP, DE, GB, CN, FR, KR, ES, AU, IN, CA, INPADOC Data); Google Scholar; Pubmed; EBSCO
Keywords: gene, fusion, DNA, chromosome, breakpoint, distance, cluster, statistics, nucleotide, cancer

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| A | (THIEME, S et al.) Genome Fusion Detection: a Novel Method to Detect Fusion Genes from SNP-Array Data. Bioinformatics. 22 January 2013; Vol. 29, No. 6; pages 671-677; page 672, column 2, paragraphs 1-2; page 673, column 2, paragraph 1; page 673, column 1, paragraph 6 – column 2, paragraph 1; page 674, column 1, paragraph 6; page 674, column 2, paragraph 2; page 675, column 1, paragraph 2; column 2, paragraph 2; page 676, column 1, paragraph 1; figure 1B; DOI: 10.1093/bioinformatics/btt028. | 1-35 |
| A | (SCHRODER, J et al.) Socrates: Identification of Genomic Rearrangements in Tumour Genomes by Re-aligning Soft Clipped Reads. Bioinformatics. 02 January 2014; Vol. 30, No. 8; pages 1064-1072; abstract; page 1065, column 2, paragraph 2; page 1066, column 1, paragraph 2; page 1066, column 2, paragraphs 2, 5; page 1068, column 1, paragraph 1; DOI: 10.1093/bioinformatics/btt767 | 1-35 |
| A | (WANG, J et al.) CREST Maps Somatic Structural Variation in Cancer Genomes with Base-Pair Resolution. Nature Methods. 12 June 2011; Vol. 8, No. 8; pages 1-10; page 2, paragraph 3; page 5, paragraph 1; DOI: 10.1038/nmeth.1628. | 1-35 |
| A | US 2009/0202999 A1 (MORLEY, AA.) 13 August 2009; abstract; paragraph [0039]. | 1-35 |
| A | (MCPHERSON, A et al.) DeFuse: An Algorithm for Gene Fusion Discovery in Tumor RNA-Seq Data. PLoS Computational Biology. 19 May 2011; Vol. 7, No. 5; pages 1-16; page 2, column 2, paragraph 3; DOI: 10.1371/journal.pcbi.1001138. | 1-23 |
| A | US 2014/0222443 A1 (DANENBERG, K et al.) 07 August 2014; paragraphs [0031], [0227], [0265], [0333] | 23 |

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 08 December 2016 (08.12.2016) | 12 JAN 2017 |
| Name and mailing address of the ISA/US<br>Mail Stop PCT, Attn: ISA/US, Commissioner for Patents<br>P.O. Box 1450, Alexandria, Virginia 22313-1450<br>Facsimile No. 571-273-8300 | Authorized officer<br>Shane Thomas<br>PCT Helpdesk: 571-272-4300<br>PCT OSP: 571-272-7774 |

Form PCT/ISA/210 (second sheet) (January 2015)

(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 108368546 A
(43)申请公布日 2018.08.03

(21)申请号 201680071449.4

(22)申请日 2016.10.10

(30)优先权数据
62/239,879 2015.10.10 US

(85)PCT国际申请进入国家阶段日
2018.06.06

(86)PCT国际申请的申请数据
PCT/US2016/056314 2016.10.10

(87)PCT国际申请的公布数据
WO2017/062970 EN 2017.04.13

(71)申请人 夸登特健康公司
地址 美国加利福尼亚州

(72)发明人 雷扎·莫克塔里
巴赫拉姆·加法尔扎德·克曼尼

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
代理人 刘晓杰

(51)Int.Cl.
*C12Q 1/686*(2018.01)
*C12N 15/11*(2006.01)
*G06F 19/22*(2011.01)

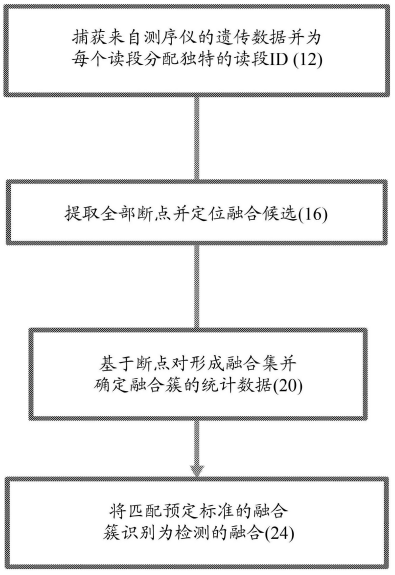
权利要求书3页 说明书16页 附图11页

(54)发明名称

无细胞DNA分析中基因融合检测的方法和应用

(57)摘要

公开了用于通过以下来确定基因融合的系统和方法:确定含有融合的染色体DNA分子的一部分的测序数据的融合读段;确定基因组上的预定点,其中融合读段的至少一个映射部分在所述预定点(断点)处剪切;将来自两个断点(断点对)的两个映射读段部分识别为潜在的融合候选;基于断点对创建一个或更多个融合集并将融合集聚类到一个或更多个融合簇中;和将符合预定标准的每个融合簇识别为基因融合。




CN 108368546 A

1.一种用于处理来自样品的遗传序列读段数据的方法,所述方法包括:

确定包含融合染色体DNA分子的至少一部分的测序数据的融合读段;

确定基因组上的预定点,其中所述融合读段的至少一个映射部分在所述预定点(断点)处剪切;

将来自两个断点(断点对)的两个映射读段部分识别为潜在融合候选;

基于断点对创建一个或更多个融合集并将所述融合集聚类到一个或更多个融合簇中;和

将符合预定标准的每个融合簇识别为基因融合。

2.如权利要求1所述的方法,包括向每个读段分配独特的分子或读段标识符(读段ID)。

3.如权利要求1所述的方法,包括从一侧或两侧剪切所述读段的每个映射部分。

4.如权利要求1所述的方法,其中所述断点的身份独立于所述读段并且由符号、染色体和位置来标识。

5.如权利要求4所述的方法,其中所述断点记录统计数据,所述统计数据包括在所述断点处被剪切或分裂的读段和分子的数目,以及跳过所述断点的野生型读段和分子的数目。

6.如权利要求2所述的方法,包括选择具有公共读段ID的每两个映射读段部分作为潜在融合候选,所述公共读段ID属于具有适当符号的两个断点。

7.如权利要求6所述的方法,其中映射之前的原始读段中的潜在融合候选位置将所述读段部分显示为原先彼此相邻地定位。

8.如权利要求6所述的方法,包括检查读段部分是否映射在一条链上以获得断点符号的差异。

9.如权利要求1所述的方法,包括追踪融合集统计数据。

10.如权利要求9所述的方法,其中所述融合集合统计数据是断点ID、包含在所述集中的分子或读段的数目。

11.如权利要求1所述的方法,包括将具有相似断点的融合集分组为融合簇。

12.如权利要求11所述的方法,其中所述相似断点是相距不超过5个核苷酸、不超过10个核苷酸或不超过25个核苷酸的断点。

13.如权利要求1所述的方法,包括定义所述基因组中的两个区域之间的融合簇。

14.如权利要求1所述的方法,包括为所述融合簇确定每个配偶体的融合分子的数目。

15.如权利要求1所述的方法,包括为所述融合簇确定每个配偶体的融合读段的数目。

16.如权利要求1所述的方法,包括为所述融合簇确定每个配偶体的野生型分子的数目。

17.如权利要求1所述的方法,包括为所述融合簇确定每个配偶体的野生型读段或分子的数目。

18.如权利要求1所述的方法,包括为所述融合簇确定每个配偶体的融合百分比为融合分子与每个配偶体的总分子的比率。

19.如权利要求1所述的方法,其中所述总分子包含野生型组分和剪切组分。

20.如权利要求18所述的方法,包括为所述融合簇确定每个配偶体的基因信息。

21.如权利要求1所述的方法,包括确定所述融合簇的下游基因。

22.如权利要求1所述的方法,其中所述标准包括在所述簇中具有多于一个分子或者具

有至少一个具有两条沃森和克里克链的分子。

23.一种分析遗传信息的系统,包括

DNA测序仪;

处理器,所述处理器偶联到所述DNA测序仪,所述处理器运行计算机代码以处理来自样品的遗传序列读段数据,所述计算机代码包括用于以下的指令:

确定包含融合染色体DNA分子的一部分的测序数据的融合读段;

确定基因组上的至少一个预定点,其中融合读段的至少一个映射部分在所述预定点(断点)处剪切;

将来自两个断点(断点对)的两个映射读段部分识别为潜在融合候选;

基于断点对创建一个或更多个融合集并将所述融合集聚类到一个或更多个融合簇中;和

将符合预定标准的每个融合簇识别为基因融合。

24.一种方法,包括:

(a)用DNA测序仪对DNA分子进行测序以生成序列的集合;

(b)将序列的集合映射到参考基因组;

(c)从映射集合识别融合读段,其中融合读段包含子序列,其中第一子序列映射到第一遗传基因座且第二子序列映射到第二、独特的遗传基因座;

(d)对于每个融合读段,识别所述第一遗传基因座处的第一断点和所述第二遗传基因座处的第二断点,其中断点是所述参考基因组上融合读段的序列被剪切的点,并且其中所述第一断点和所述第二断点形成断点对;

(e)生成融合读段集,每个集包括具有相同断点对的融合读段;

(f)聚类融合读段集,其中每个簇由具有第一预定核苷酸距离内的第一断点和第二预定核苷酸距离内的第二断点的融合读段集形成;和

(g)确定一个或更多个簇的基因融合,其中簇的基因融合具有选自所述簇中的第一断点的断点作为第一融合基因断点和选自所述簇中的第二断点的断点作为第二融合基因断点,并且其中所述第一融合基因断点和所述第二融合基因断点各自基于选择标准选择。

25.如权利要求24所述的方法,其中所述独特的遗传基因座位于不同染色体上或位于同一染色体的不同基因上。

26.如权利要求24所述的方法,其中所述第一预定距离和所述第二预定距离各自不超过5个核苷酸、不超过10个核苷酸或不超过25个核苷酸。

27.如权利要求24所述的方法,其中所述选择标准包括所述断点在所述簇中具有最多融合读段。

28.如权利要求24所述的方法,包括确定多个基因簇的基因融合。

29.一种方法,包括:

(a)用DNA测序仪对多个DNA分子进行测序;

(b)用标识符对所述多个序列分子的每一个加标签;

(c)将每个加标签的序列映射到参考基因组;

(d)从所映射的加标签的序列识别剪切读段,其中剪切读段是包含映射部分和剪切部分的加标签的序列,其中所述映射部分映射到遗传基因座且所述剪切部分不映射到所述遗

传基因座；

(e)确定每个剪切读段的断点，其中断点是所述参考基因组上剪切读段的序列被剪切的点；

(f)创建断点集，每个断点集包括具有相同断点的剪切读段的标识符；

(g)通过比较断点集的对创建断点对集，每个断点对集包括存在于被比较的断点集的对的两个成员中的标识符；

(h)聚类断点对集，其中每个簇包括具有第一预定遗传距离内的所述对的第一断点和第二预定遗传距离内的所述对的第二断点的断点对集；

(i)确定所述簇中的一个或更多个簇的基因融合，其中簇的基因融合具有选自所述簇中的第一断点的断点作为第一融合基因断点和选自所述簇中的第二断点的断点作为第二融合基因断点，并且其中所述第一融合基因断点和所述第二融合基因断点各自基于选择标准选择。

30.如权利要求29所述的方法，其中所述选择标准包括所述断点在所述簇中具有最多融合读段。

31.一种用于识别融合基因断点的方法，所述方法包括：

(a)确定包含融合染色体DNA分子的至少一部分的测序数据的融合读段；

(b)确定基因组上的预定点，其中融合读段的至少一个映射部分在所述预定点(断点)处剪切；

(c)将来自两个断点(断点对)的两个映射读段部分识别为潜在融合候选；

(d)基于断点对创建一个或更多个融合集并将所述融合集聚类到一个或更多个融合簇中；

(e)将符合预定标准的每个融合簇识别为基因融合，和

(f)将所述基因融合的断点识别为融合基因断点。

32.一种用于诊断受试者中的状况的方法，所述方法包括：

(a)确定包含融合染色体DNA分子的至少一部分的测序数据的融合读段；

(b)确定基因组上的预定点，其中融合读段的至少一个映射部分在所述预定点(断点)处剪切；

(c)将来自两个断点(断点对)的两个映射读段部分识别为潜在融合候选；

(d)基于断点对创建一个或更多个融合集并将所述融合集聚类到一个或更多个融合簇中；和

(e)将符合预定标准的每个融合簇识别为基因融合，

其中所述基因融合指示所述状况。

33.如权利要求32所述的方法，其中所述状况是癌症。

34.如权利要求33所述的方法，其中所述癌症选自由血液学癌症、肉瘤和前列腺癌组成的组。

35.如权利要求34所述的方法，所述方法还包括向所述受试者施用治疗。

# 无细胞DNA分析中基因融合检测的方法和应用

[0001] 交叉引用

[0002] 本申请要求于2015年10月10日提交的美国临时申请号62/239,879的权益,该申请通过引用以其整体并入本文。

[0003] 发明背景

[0004] 癌性细胞可以具有融合到一起的染色体。如果对这样的染色体进行测序,它将生成可以被映射到基因组的两个不同区域(在相同或不同染色体上)的读段。基因融合在基因结构的进化中起作用。重复、序列趋异和重组是基因进化工作的主要参与者。当基因融合在非编码序列区域中发生时,它可以导致目前在另一个基因的顺式调节序列的控制下的基因的表达的错误调节。如果它在编码序列中发生,基因融合可以引起新基因的装配,允许通过在多结构域蛋白中增加肽模块而出现新功能。

[0005] 染色体显带分析(chromosome banding analysis)、荧光原位杂交(FISH)和逆转录聚合酶链式反应(RT-PCR)是在诊断实验室使用的常见方法。由于癌症基因组的非常复杂的性质,这些方法都具有其独特的缺点。近期开发诸如高通量测序和定制DNA微阵列具有引进更高效的方法的希望,但仍然不足。高通量基因组测序技术已经被用作研究工具并且目前正在被引入临床,而在个性化医学的未来,全基因组测序数据可能是指导治疗干预的重要工具。

[0006] 发明概述

[0007] 在一个方面,公开了用于通过以下来确定基因融合的系统和方法:确定含有融合的染色体DNA分子的至少一部分的测序数据的融合读段;确定基因组上的预定点,其中融合读段的至少一个映射部分在预定点(断点)处剪切(clipped);将来自两个断点(断点对)的两个映射的读段部分识别为潜在的融合候选;基于断点对创建一个或更多个融合集并将融合集聚类到一个或更多个融合簇中;和将符合预定标准的每个融合簇识别为基因融合。

[0008] 在一个方面,本公开内容提供了用于处理来自样品的遗传序列读段数据的方法,所述方法包括:确定含有融合的染色体DNA分子的至少一部分的测序数据的融合读段;确定基因组上的预定点,其中融合读段的至少一个映射部分在所述预定点(断点)处剪切;将来自两个断点(断点对)的两个映射的读段部分识别为潜在的融合候选;基于断点对创建一个或更多个融合集并将融合集聚类到一个或更多个融合簇中;和将符合预定标准的每个融合簇识别为基因融合。

[0009] 在一些实施方案中,所述方法包括向每个读段分配独特的分子或读段标识符(读段ID)。在一些实施方案中,所述方法包括从一侧或两侧剪切所述读段的每个映射部分。在一些实施方案中,断点的身份独立于所述读段并且由符号、染色体和位置来标识。在一些实施方案中,断点记录统计数据,所述统计数据包括在所述断点处被剪切或分裂的读段和分子的数目,以及跳过所述断点的野生型读段和分子的数目。在一些实施方案中,所述方法包括选择具有公共读段ID的每两个映射读段部分作为潜在融合候选,所述公共读段ID属于具有适当符号的两个断点。在一些实施方案中,映射之前的原始读段中的潜在融合候选位置将所述读段部分显示为原先彼此相邻地定位。在一些实施方案中,所述方法包括检查读段

部分是否映射在一条链上以获得断点符号的差异。在一些实施方案中，所述方法包括追踪融合集统计数据。

[0010] 在一些实施方案中，所述融合集统计数据是断点ID、包含在所述集中的分子或读段的数目。在一些实施方案中，所述方法包括将具有相似断点的融合集分组为融合簇。在一些实施方案中，所述相似断点是相距不超过5个核苷酸、不超过10个核苷酸或不超过25个核苷酸的断点。在一些实施方案中，所述方法包括定义基因组中两个区域之间的融合簇。在一些实施方案中，所述方法包括为所述融合簇确定每个配偶体的融合分子的数目。在一些实施方案中，所述方法包括为所述融合簇确定每个配偶体的融合读段的数目。在一些实施方案中，所述方法包括为所述融合簇确定每个配偶体的野生型分子的数目。在一些实施方案中，所述方法包括为所述融合簇确定每个配偶体的野生型读段或分子的数目。在一些实施方案中，所述方法包括为所述融合簇确定每个配偶体的融合百分比为融合分子与每个配偶体的总分子的比率。在一些实施方案中，总分子包括野生型和剪切组分。在一些实施方案中，所述方法包括为所述融合簇确定每个配偶体的基因信息。在一些实施方案中，所述方法包括确定所述融合簇的下游基因。在一些实施方案中，所述标准包括在所述簇中具有多于一个分子或者具有至少一个具有两条沃森和克里克链的分子。

[0011] 在一个方面，本公开内容提供了分析遗传信息的系统，包括：DNA测序仪；处理器，所述处理器偶联至所述DNA测序仪，所述处理器运行处理来自样品的遗传序列读段数据的计算机代码，所述计算机代码包括用于以下的指令：确定含有融合的染色体DNA分子的一部分的测序数据的融合读段；确定基因组上的至少一个预定点，其中融合读段的至少一个映射部分在所述预定点(断点)处剪切；将来自两个断点(断点对)的两个映射的读段部分识别为潜在的融合候选；基于断点对创建一个或更多个融合集并将融合集聚类到一个或更多个融合簇中；和将符合预定标准的每个融合簇识别为基因融合。

[0012] 在一个方面，本公开内容提供了包括以下的方法：用DNA测序仪对DNA分子进行测序以生成序列的集合；将序列的集合映射到参考基因组；从映射的集合识别融合读段，其中融合读段包含子序列，其中第一子序列映射至第一遗传基因座且第二子序列映射到第二、独特的遗传基因座；对于每个融合读段，识别第一遗传基因座处的第一断点和第二遗传基因座处的第二断点，其中断点是融合读段的序列被剪切的参考基因组上的点，且其中第一断点和第二断点形成断点对；生成融合读段集，每个集包含具有相同断点对的融合读段；聚类融合读段集，其中每个簇由具有第一预定多核苷酸距离内的第一断点和第二预定核苷酸距离内的第二断点的融合读段集形成；和确定一个或更多个簇的基因融合，其中簇的基因融合具有选自簇中的第一断点的断点作为第一融合基因断点和选自簇中的第二断点的断点作为第二融合基因断点，并且其中第一融合基因断点和第二融合基因断点各自基于选择标准选择。

[0013] 在一些实施方案中，独特的遗传基因座位于不同染色体上或位于同一染色体的不同基因上。在一些实施方案中，第一预定距离和第二预定距离各自不超过5个核苷酸、不超过10个核苷酸或不超过25个核苷酸。在一些实施方案中，选择标准包括断点在簇中具有最多融合读段。在一些实施方案中，所述方法包括确定多个基因簇的基因融合。

[0014] 在一个方面，本公开内容提供了包括以下的方法：用DNA测序仪对多个DNA分子进行测序；用标识符为多个序列分子的每一个加标签；将每个加标签的序列映射到参考基因

组;从映射的加标签的序列识别剪切读段,其中剪切读段是包含映射部分和剪切部分的加标签的序列,其中映射部分映射到遗传基因座且剪切部分不映射到遗传基因座;确定每个剪切读段的断点,其中断点是剪切读段的序列被剪切的参考基因组上的点;生成断点集,每个断点集包括具有相同断点的剪切读段的标识符;通过比较断点集的对创建断点对的集,断点对的每个集包括存在于比较的断点集的对的两个成员中的标识符;聚类断点对的集,其中每个簇包括具有第一预定遗传距离内的对的第一断点和第二预定遗传距离内的对的第二断点的断点对的集;和确定一个或更多个簇的基因融合,其中簇的基因融合具有选自簇中的第一断点的断点作为第一融合基因断点和选自簇中的第二断点的断点作为第二融合基因断点,并且其中第一融合基因断点和第二融合基因断点各自基于选择标准选择。在一些实施方案中,选择标准包括断点在簇中具有最多融合读段。

[0015] 在一个方面,本公开内容提供了用于识别融合基因断点的方法,所述方法包括:确定含有融合的染色体DNA分子的至少一部分的测序数据的融合读段;确定基因组上的预定点,其中融合读段的至少一个映射部分在预定点(断点)处剪切;将来自两个断点(断点对)的两个映射的读段部分识别为潜在的融合候选;基于断点对创建一个或更多个融合集并将融合集聚类到一个或更多个融合簇中;将符合预定标准的每个融合簇识别为基因融合,并将基因融合的断点识别为融合基因断点。

[0016] 在一个方面,本公开内容提供了用于诊断受试者中的状况的方法,所述方法包括:确定含有融合的染色体DNA分子的至少一部分的测序数据的融合读段;确定基因组上的预定点,其中融合读段的至少一个映射部分在预定点(断点)处剪切;将来自两个断点(断点对)的两个映射的读段部分识别为潜在的融合候选;基于断点对创建一个或更多个融合集并将融合集聚类到一个或更多个融合簇中;和将符合预定标准的每个融合簇识别为基因融合,其中所述基因融合指示该状况。

[0017] 在一些实施方案中,状况是癌症。在一些实施方案中,癌症选自由血液学癌症、肉瘤和前列腺癌组成的组。在一些实施方案中,所述方法还包括向受试者施用治疗。

[0018] 通过引用并入

[0019] 本说明书中提及的所有出版物、专利和专利申请通过引用并入本文,其程度如同每一个单独的出版物、专利或专利申请被具体和单独地指明通过引用并入的相同程度。

[0020] 附图简述

[0021] 本发明的新颖特征在所附权利要求中具体地阐述。通过参考下文的详细描述和附图将获得对本发明的特征和优点的更好理解,详细描述阐述了利用本发明的原理的说明性实施例,在附图中:

[0022] 图1示意了用于检测基因融合的示例性过程。

[0023] 图2描绘了从两个其他染色体产生融合染色体的可能的不同情景。

[0024] 图3A示出了示例性+/-断点和分别从左侧/右侧被剪切的读段部分。

[0025] 图3B示出了基因融合检测使用的示例性合并过程。

[0026] 图4示意了本发明的系统。

[0027] 图5示出了染色体A和染色体B之间的示例性基因融合,以及跨断点映射的DNA融合读段。

[0028] 图6示出了DNA片段到参考基因组中的两个位置的示例性映射。

[0029] 图7示出了其中映射的断点被不同地定位的示例性融合读段。

[0030] 图8示出了映射的融合读段示例性分组到集中并将集分组成簇，用于判定融合基因断点。

[0031] 图9A-9C示出了基因融合检测过程的示例性图解。

[0032] 发明的详细描述

[0033] 本发明涉及用于检测基因融合的系统和方法。

[0034] 融合基因的断点的准确映射是具有挑战性的。当试图映射断点时，测序中的错误和比对融合基因的困难仅仅是遇到的两个困难。本文描述的系统和方法可以提供以下优势中的一个或更多个。系统可以识别可能导致肿瘤形成的融合基因，因为融合基因可以比非融合基因产生活跃得多的异常的蛋白。当融合基因是引起癌症的致癌基因时，系统准确地确定癌症存在；这些包括BCR-ABL、TEL-AML1(具有t(12;21)的ALL)、AML1-ETO(具有t(8;21)的M2AML)、和在染色体21上具有中间缺失的TMPRSS2-ERG，通常发生于前列腺癌中。在TMPRSS2-ERG的情况下，通过用致癌基因ETS转录因子破坏雄激素受体(AR)信号通路和抑制AR表达，融合产物调节前列腺癌。多数融合基因从血液学癌症、肉瘤和前列腺癌发现。致癌融合基因可以导致具有新功能或不同于两个融合配偶体的功能的基因产物。可选地，将原致癌基因融合至强启动子，并且从而致癌基因功能通过由上游融合配偶体的强启动子引起的上调发挥功能。后者常见于淋巴瘤，其中致癌基因与免疫球蛋白基因的启动子并列。致癌基因融合转录物还可以由反式剪接或连读事件引起。从基因组序列和结构的角度分析这些基因融合可以提供引导开发改进的癌症诊断和靶向疗法的相关数据。

[0035] 图1示出了用于确定基因融合的示例性过程。一般来说，该过程捕获来自测序仪的遗传数据并且应用比对和连接测序的对-末端读段的融合技术，该对-末端读段的插入尺寸小于两个读段长度之和，并且在融合后为每个读段分配独特的读段标识符(读段ID)(12)。接下来，该过程提取所有断点并定位融合候选(16)。随后，该过程基于断点对形成融合集并确定融合簇的统计数据(20)。随后，通过匹配预定标准将融合簇识别为检测的融合(24)。

[0036] 下面讨论关于图1的过程的细节。癌性细胞可以具有融合到一起的染色体。如果对这样的染色体进行测序，它将生成可以被映射到基因组的两个不同区域(在相同或不同染色体上)的读段。利用此行为来检测融合。

[0037] 映射前，独特的读段标识符(读段ID)被分配给每个读段并将在一个或更多个FASTQ文件中的读段标题(header)中编码，如以下详述的。可选地，独特分子，诸如包含独特条形码的寡核苷酸，可以代替读段ID使用。一旦FASTQ文件被映射，此编码的读段ID将被检索并且它可以容易地示出何种匹配来自相同原始读段。随后，该过程提取所有断点：融合读段(含有来自融合染色体的部分DNA分子的测序数据的读段)不能以整体映射到基因组并且映射器将其不同的部分映射到基因组上的不同位置。当使用传统技术来尝试断点映射时，这呈现了挑战。此类读段的每个映射部分从一侧或两侧被剪切。断点是融合读段的至少一个映射部分被剪切的基因组上的点。断点的身份独立于所述读段并且由其符号、染色体和位置来标识。+/-断点具有分别从左侧/右侧被剪切的读段部分。从一个位置的同一侧剪切或分裂的全部读段列在相关的断点读段列表中。断点还可以记录其他统计数据，如在所述断点处被剪切或分裂的读段和分子的数目；或者跳过断点的野生型读段和分子的数目。还可以提供断点位置处的基因信息。通过分配断点，使用或不使用聚类，本文描述的方法和系

统可被用于准确地确定发生基因融合的基因座。

[0038] 随后，此过程发现融合：具有公共读段ID的每两个映射读段部分是潜在的融合候选，所述公共读段ID属于具有适当符号的两个断点。它们还需要具有表明读段部分最初彼此相邻地定位的正确的片段顺序(它们在映射之前在原始读段中的位置)，以被认为是真正的融合候选。此外，所得的融合必须在序列链方面是生物学上可能的。这仅意味着，如果读段部分被映射到相同链上(都在5’链或都在3’链)，则断点的符号必须不一致，且反之亦然。在图2中示出这个的实例。

[0039] 所有提取的融合候选基于断点对放置在融合集中。融合集还可以记录统计数据，如断点ID和分子的数目以及包含在所述集中的读段。这些统计数据可以被追踪。

[0040] 随后进行聚类。具有足够靠近的断点的所有融合集将分组在一个融合簇中。结果是，在基因组的两个区域之间定义融合簇。本公开内容还提供了对于融合簇，确定每个配偶体的融合分子的数目、每个配偶体的融合读段的数目、每个配偶体的多个野生型分子、每个配偶体的野生型读段或分子的数目、或者每个配偶体的融合百分比作为融合分子相对于每个配偶体的总分子的比率。

[0041] 图5示出了染色体A和染色体B之间的假设性基因融合。作为交换的结果，基因融合包含每个染色体的一部分。交换的点被称为断点。在无细胞DNA中，DNA片段可以跨断点映射，诸如融合读段1、融合读段2和融合读段3。

[0042] 测序产生DNA片段的序列。软件用识别标签标记每个序列。软件还将这些产生的序列映射到参考基因组上。图6示出了融合读段1向参考基因组的假设性映射。映射软件将融合读段的序列映射到发现足够同源性的参考基因组中的任何位置。有歧义的序列可以被映射到参考基因组中的多个位置。

[0043] 在融合读段跨融合基因的断点映射的情况中，软件通常将融合读段的序列映射两次，向每个染色体映射一次。但是，在每个情况中，映射软件不能适当地将序列的部分(子序列)映射到参考基因组。因此，映射序列将包括映射到参考基因组的子序列和由于同源性差而未映射到同一基因座的子序列二者。此类子序列被称为“剪切”序列。读段被剪切处的参考基因组上的点是断点。

[0044] 由于每个序列携带识别标签，映射到两个不同位置的序列可以由于相同标签被鉴定为源自相同原始序列。所以，例如，具有足够同源性的序列的子序列被映射到染色体A，且具有不足同源性的序列的子序列被剪切。类似地，映射软件将序列映射到染色体B，在存在不足同源性处剪切序列。

[0045] 然而，作为几个因素，包括测序中的错误和映射算法的特征的结果，包括融合基因的断点的DNA片段可能不准确地映射到每个参考染色体上的断点基因座。例如，映射软件可以识别实际断点稍微上游或者稍微下游的序列的断点。

[0046] 由于每个序列携带识别标签，映射到两个不同位置的序列可以由于相同标签被鉴定为源自相同原始序列。所以，例如，具有足够同源性的序列的子序列被映射到染色体A，且具有不足同源性的序列的子序列被剪切。类似地，映射软件将序列映射到染色体B，在存在不足同源性处剪切序列。

[0047] 然而，作为几个因素，包括测序中的错误和映射算法的特征的结果，包括融合基因的断点的DNA片段可能不准确地映射到每个参考染色体上的断点基因座。例如，映射软件可

以识别实际断点稍微上游或者稍微下游的序列的断点。这些错误可以影响例如取决于准确的基因融合信息的癌症诊断的准确性。本公开内容

[0048] 几个假设的映射错误在图7中示出。融合读段1适当地映射，断点在参考染色体中被指示为断点A1和断点B1(第一断点和第二断点)。此融合读段具有断点对A1-B1。融合读段2不适当地映射，染色体1的断点被确定为在上游，在断点A2(第一断点)处。然而，染色体B中的断点被正确映射在断点B1(第二断点)处。此融合读段具有断点对A2-B1。融合读段3也不适当地映射，染色体A的断点在断点A1(第一断点)处正确地映射，但染色体B的断点被确定为在下游，在断点B2(第二断点)处。此融合读段具有断点对A1-B2。在此情况下，软件已经识别了融合基因的几个断点。

[0049] 根据本公开内容的方法，为了判定融合基因中的断点，映射序列基于共有断点对被分组成集，并随后基于参考基因组中的预定碱基距离内的断点被分组成簇。

[0050] 这种方法在图8中描述。融合读段1、2、3、4、5和6的序列被映射到参考基因组的染色体A和B。其中断点在融合的两侧的序列被分组成集。在此实例中，融合读段1和4共有断点对A1和B1，并且被分组到集I。融合读段2和5共有断点对A2和B1，并且被分组到集II。融合读段3和6共有断点对A1和B2，并且被分组到集III。

[0051] 在此实例中，断点A1和A2在预定遗传距离A(例如10个碱基)内，并且断点B1和B2在预定遗传距离B内。因此，集I、II和III被分组到一个簇内。

[0052] 使用用户选择的选择标准判定融合基因断点。在一些实施方案中，标准包括在所述簇中具有多于一个分子和/或者具有至少一个具有两条沃森和克里克链的分子。在一个方法中，每个染色体中的断点由投票方法确定，其中所有断点中具有最多相关融合读段的断点被判定为融合基因断点。在其他方法中，不同序列的断点可以使用质量算法来加权。在图8的实例中，在染色体A中，断点A1与4个融合读段相关，而断点A2与2个融合读段相关。因此，第一基因融合断点被判定为在A1处。在染色体B中，断点B1与4个融合读段相关，而断点B2与2个融合读段相关。因此，第二基因融合断点被判定为在B1处。

[0053] 另一个示例性方法在图9A-9C中示出。DNA分子使用DNA测序系统测序，诸如高通量测序仪。序列可以被分析以生成集合中的原始分子的共有序列。生成的序列的集合用独特的标识符加标签(在此情况下，1-7)。序列被映射到参考基因组。在此实例中，序列各自映射到参考基因组中的两个不同位置。映射部分被描绘为短棒，而剪切部分被描绘为虚线。所有序列的断点被识别。在此实例中，染色体A上的断点是A1、A2和A3。染色体B上的断点是B1、B2和B3。映射读段基于共有断点被组织成集。断点对被确定为具有相同标识符和每个染色体上的相同断点的序列的对。确定染色体与簇断点上的预定距离。在此实例中，簇包括断点A1和A2以及B1和B2。断点A3和B3在预定距离以外，并且因此不被包括在簇中。基于选择标准判定原始分子中的断点对。在此实例中，标准是基于投票。因此，断点A1和B1基于具有在簇内的多数分子而被判定为断点对。

[0054] 系统可以处理剪接中常见的短序列的多个比对。更长的序列可以在融合点周围获得以确认其比对。为了降低假阳性率，可以使用一系列过滤器，包括读段数目、序列相似性、序列位置分布过滤器以提供具有高特异性的结果。另外，可以使用表达估计工具RSEM(通过期望最大化的RNA-Seq，RNA-Seq by Expectation Maximization)来应用具有稀疏优化的期望最大化(EM)算法来估计幻想的转录物丰度。此外，此丰度定量可以增加识别准确性。综

合考虑,这些特征允许系统提供基因融合事件的更完整的综述。

[0055] 系统可以包括以任何合适的文件格式从一个或更多个样品获得多个序列读段、识别重复序列读段集和对于每个重复序列读段集仅存储一个读段。合适的文件格式包括FASTA和FASTQ文件格式。FASTA和FASTQ是用于存储来自高通量测序的原始序列读段的常用文件格式。FASTQ文件存储每个序列读段的标识符、序列和每个读段的质量评分串。FASTA文件仅存储标识符和序列。这两种文件格式是许多常见测序比对和装配算法的输入。本发明认识到,样品内和跨样品的FASTQ和FASTA文件的读段序列信息倾向于高度冗余或重复。这意味着,许多序列读段将由相同序列组成。本发明的方法利用此冗余性来实现文件大小的多倍减小,并且在存储数据的检索方面没有损失。例如,本发明可以用于读取与样品相关的FASTA/FASTQ文件并且在主读段序列文件中仅存储独特的读段序列。

[0056] 系统还包括为与识别的独特序列具有相同序列的每个读段收集元信息(meta information),诸如读段标识符。然后可以将该元信息写入该样品的文件中,其中元信息与原始FASTA/FASTQ文件中识别的独特序列读段相关联,并且现在存储在主读段序列文件中。由于这个新文件不包含原始文件中发现的重复信息,它比原始文件更小并且更容易传输。而且,压缩文件根本不需要包含任何实际的序列数据。在某些方面,压缩文件可以简单地包含索引到存储在主文件中的独特序列的序列读段的标识符。

[0057] 序列数据可以通过获得——使用包括偶联到非瞬时性存储器的处理器的计算机系统——多个序列读段来压缩。每个序列读段可以包括序列串以及元信息。序列读段可以以一个或更多个FASTA或FASTQ文件的格式提供,例如,元信息包括描述行(前面带有“>”字符),并且任选地在FASTQ的情况下包括质量评分。序列串优选地表示核苷酸序列数据,例如使用IUPAC核苷酸编码。识别仅包含独特条目的序列串的子集。然后可以使用本发明的系统和方法来写入包括所识别的子集和——对于多个序列读段的每一个——用该子集中表示该序列读段的独特条目的指示符的序列读段的元信息的输出。

[0058] 在一些实施方案中,子集(即仅包含独特序列读段)被写入主读段文件,其可以是文本文件。优选地,使用IUPAC核苷酸编码在主读段文件中表示独特序列读段,使得文件是人类可读的并且进一步处理可以容易地执行(例如使用脚本语言诸如Perl或Python)。元信息可写入对应于输入FASTA或FASTQ文件的压缩输出文件。

[0059] 方法可以包括仅从输出重构原始输入,并且在某些实施方案中,检索是无损的,甚至是完全无损的。也就是说,输出可以被处理以创建包含多个序列读段的新FASTA或FASTQ文件。在检索是无损的情况下,新FASTA或FASTQ文件包含与FASTA或FASTQ文件相同的信息。

[0060] 本发明适用于任何合适类型的数据文件。除了前面提到的FASTA和FASTQ文件,序列读段还可以在变体判别格式(Variant Call Format,VCF)文件中捕获。随着高通量测序的进展,多个测序中心检测人类基因组中的变体并通过这些VCF文件报告它们是常见的。本发明可以促进统一数据库的开发,以便在不同来源的VCF文件中存储变体信息,从而使研究人员跨中心执行复杂的等位基因、样本和群体水平的查询。统一数据库可以通过在一个通用等位基因表上存储每个独特等位基因(例如,独特序列读段)并通过将这些独特等位基因的参考存储到关联样品和样品水平元数据来整合来自不同样品的VCF文件中的变体信息。

[0061] 系统的实现可以包括用于压缩序列数据的方法。该方法包括:使用包括偶联到非瞬时性存储器的处理器的计算机系统获得多个序列读段,每个序列读段包括序列串和元信

息;识别仅包含独特条目的序列串的子集;写入包括所述子集和——对于所述多个序列读段中的每一个——用该子集中代表该序列读段的独特条目的指示符的序列读段的元信息的输出。优选地,输出包括使用IUPAC核苷酸编码存储子集的一个或更多个文本文件。优选地,输出被存储为纯文本(例如,并且可以使用文本编辑器程序打开并由人在屏幕上阅读)。在优选实施方案中,序列读段无损地存储。该方法可以包括处理输出以创建包含多个序列读段的新FASTA或FASTQ文件。多个序列读段可以作为FASTA或FASTQ文件获得,且新FASTA或FASTQ文件可以包含与FASTA或FASTQ文件相同的信息。在一些实施方案中,输出占用小于存储获得的多个序列读段所需的磁盘空间的%。

[0062] 现在将描述用于获得样品、生成测序读段和用于实施本发明的各种类型的测序的一般方法。应该理解,这些示例性方法不是限制性的,并且可以根据需要由本领域技术人员进行修改。

[0063] 获得多个序列读段可以包括对来自样品的核酸进行测序以产生序列读段。如下文详细解释的,获得多个序列读段还可以包括接收来自测序仪的测序数据。样品中的核酸可以是任何核酸,包括例如组织样品中的基因组DNA、从实验室样品中特定靶扩增的cDNA或来自多个生物体的混合DNA。在一些实施方案中,样品包括来自单倍体或二倍体生物体的纯合DNA。例如,样品可以包括来自罕见隐性等位基因纯合的患者的基因组DNA。在其他实施方案中,样品包括来自具有体细胞突变使得两种相关核酸以除了50%或100%的等位基因频率,即20%、5%、1%、0.1%或任何其他等位基因频率存在的二倍体或多倍体生物体的杂合遗传物质。

[0064] 在一个实施方案中,从含有多种其他组分(诸如蛋白、脂质和非模板核酸)的生物样品分离核酸模板分子(例如DNA或RNA)。核酸模板分子可以从获自动物、植物、细菌、真菌或任何其他细胞生物体的任何细胞材料获得。用于在本发明中使用的生物样品还包括病毒颗粒或制备物。核酸模板分子可直接从生物体或从生物体获得的生物样品获得,例如从血液、血清、血浆、尿液、脑脊液、唾液、粪便、淋巴液、滑液、囊液、腹水、胸膜渗出物、羊水、绒毛膜绒毛样品、来自植入前胚胎的液体、胎盘样品、灌洗液和宫颈阴道液、组织间隙液、颊拭子样品、痰、支气管灌洗液、巴氏涂片样品(Pap smear sample)或眼液。任何组织或体液样本(例如体液样本的人类组织)可以用作在本发明中使用的核酸的来源。核酸模板分子也可以从培养的细胞诸如原代细胞培养物或细胞系分离。从其获得模板核酸的细胞或组织可以被病毒或其他细胞内病原体感染。样品也可以是从生物样本中提取的总RNA、cDNA文库、病毒或基因组DNA。样品也可以是来自非细胞来源的分离的DNA。

[0065] 从生物样品获得的核酸可以被片段化以产生合适的片段用于分析。模板核酸可以使用各种机械、化学和/或酶学方法片段化或修剪(sheared)成期望的长度。DNA可以经由使用例如由Covaris(Woburn,Mass)销售的超声波仪超声处理、短暂暴露于DNA酶、或使用一种或更多种限制酶的混合物、或转座酶或切口酶随机修剪。RNA可以通过短暂暴露于RNA酶、加热加上镁或通过修剪来片段化。RNA可以被转化成cDNA。如果使用片段化,则RNA可以在片段化之前或之后转化成cDNA。在一个实施方案中,核酸通过超声片段化。在另一个实施方案中,核酸通过液压剪床(hydroshear)仪器来片段化。通常,单个核酸模板分子可以是从约2kb碱基至约40kb。在特定的实施方案中,核酸是约6kb-10kb片段。核酸分子可以是单链、双链或具有单链区的双链(例如,茎和环结构)。

[0066] 生物样品可以根据需要在洗涤剂或表面活性剂的存在下裂解、均质化或分级。合适的洗涤剂可以包括离子型洗涤剂(例如十二烷基硫酸钠或N-月桂酰肌氨酸)或非离子型洗涤剂。一旦核酸从样品中提取或分离出来,就可以被扩增。

[0067] 扩增是指产生核酸序列的另外的拷贝,并且通常使用聚合酶链式反应(PCR)或本领域已知的其他技术进行。扩增反应可以是扩增核酸分子的本领域已知的任何扩增反应诸如PCR。其他扩增反应包括巢式PCR、PCR-单链构象多态性、连接酶链式反应、链置换扩增和限制性片段长度多态性、基于转录的扩增系统、滚环扩增、和超分枝滚环扩增、定量PCR、定量荧光PCR(QF-PCR)、多重荧光PCR(MF-PCR)、实时PCR(RTPCR)、限制性片段长度多态性PCR(PCR-RFLP)、原位滚环扩增(RCA)、桥式PCR、皮升PCR(picotiter PCR)、乳液PCR、转录扩增、自我维持序列复制、共有序列引物PCR、任意引物PCR、简并寡核苷酸引物PCR和基于核酸的序列扩增(NABSA)。可以使用的扩增方法包括在美国专利第5,242,794号;第5,494,810号;第4,988,617号;和第6,582,938号中描述的那些。在某些实施方案中,扩增反应为例如美国专利第4,683,195号;以及美国专利第4,683,202号中描述的PCR,特此通过引用并入。用于PCR、测序和其他方法的引物可以通过克隆、直接化学合成和本领域已知的其他方法来制备。引物也可以从商业来源获得,诸如Eurofins MWG Operon(Huntsville,Ala.)或Life Technologies(Carlsbad,Calif.)。

[0068] 扩增衔接子可以附接到片段化的核酸上。衔接子可以是商购获得的,例如来自Integrated DNA Technologies(Coralville,Iowa)。在某些实施方案中,衔接子序列用酶附接到模板核酸分子。酶可以是连接酶或聚合酶。连接酶可以是能够将寡核苷酸(RNA或DNA)连接至模板核酸分子的任何酶。合适的连接酶包括可从New England Biolabs(Ipswich,Mass.)商购可得的T4DNA连接酶和T4RNA连接酶。使用连接酶的方法是本领域熟知的。聚合酶可以是能够将核苷酸添加到模板核酸分子的3'和5'末端的任何酶。

[0069] 连接可以是平端的或利用互补的悬垂末端。在某些实施方案中,片段的末端可以在片段化之后被修复、修剪(trimmed)(例如使用核酸外切酶)、或填充(例如使用聚合酶和dNTP)以形成平端。在一些实施方案中,使用商业试剂盒(诸如从Epicenter Biotechnologies(Madison,Wis.)可获得的试剂盒)进行末端修复以生成平端5'磷酸化核酸末端。在生成平端时,末端可以用聚合酶和dATP处理以形成独立添加至片段的3'末端和5'末端的模板,从而产生单个A悬垂。这个单个A用于指导被称为T-A克隆的方法中从5'末端的单个T悬垂的片段的连接。可选地,因为限制性消化后限制酶留下的悬垂的可能的组合是已知的,末端可以保持原样,即不齐的末端。在某些实施方案中,使用具有互补悬垂末端的双链寡核苷酸。

[0070] 本发明的实施方案涉及将条形码序列附接至模板核酸。在某些实施方案中,条形码被附接到每个片段。在其他实施方案中,多个条形码(例如两个条形码)被附接到每个片段。条形码序列通常包括使序列在测序反应中有用的某些特征。例如,条形码序列被设计为在条形码序列内具有最小均聚物区域或者不具有均聚物区域,所述均聚物区域即在一行具有2个或更多个相同碱基,诸如AA或CCC。条形码序列还被设计为使得它们在执行逐个碱基测序时距离碱基添加顺序至少有一个编辑距离,确保第一个碱基和最后一个碱基与序列的预期碱基不匹配。

[0071] 条形码序列被设计为使得每个序列与核酸的特定部分相关,允许序列读段与它们

来自于其的部分相关联。在某些实施方案中，条形码序列的范围从约5个核苷酸至约15个核苷酸。在特定实施方案中，条形码序列的范围从约4个核苷酸至约7个核苷酸。由于条形码序列与模板核酸一起测序，寡核苷酸长度应该是最小的长度，以便允许附接来自模板核酸的最长读段。例如，多个DNA条形码可以包含各个数目的核苷酸序列。在某些实施方案中，条形码序列包含2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30或更多个核苷酸。当仅附接到多核苷酸的一个末端时，多个DNA条形码可以产生2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30或更多个不同的标识符。可选地，当附接到多核苷酸的两个末端时，多个DNA条形码可以产生4、9、16、25、36、49、64、81、100、121、144、169、196、225、256、289、324、361、400或更多个不同的标识符(当DNA条形码仅附接到多核苷酸的一个末端时为^2)。

[0072] 通常，条形码序列与模板核酸分子间隔至少一个碱基(使均聚物组合最小化)。在某些实施方案中，条形码序列用例如酶附接到模板核酸分子。如下所述，酶可以是连接酶或聚合酶。

[0073] 扩增或测序衔接子或条形码或其组合可以附接到片段化的核酸上。这样的分子可以商购获得，例如从Integrated DNA Technologies(Coralville,Iowa)。在某些实施方案中，这样的序列用酶诸如连接酶附接至模板核酸分子。合适的连接酶包括可从New England Biolabs(Ipswich,Mass.)商购可得的T4DNA连接酶和T4RNA连接酶。连接可以平端的或者通过使用互补的悬垂末端。在某些实施方案中，在片段化之后，片段的末端可以被修复、修整(例如使用核酸外切酶)或填充(例如使用聚合酶和dNTP)以形成平端。在一些实施方案中，使用商业试剂盒(诸如从Epicenter Biotechnologies(Madison,Wis.)可获得的试剂盒)进行末端修复以生成平端5'磷酸化核酸末端。在生成平端时，末端可以用聚合酶和dATP处理以形成独立添加至片段的3'末端和5'末端的模板，从而产生单个A悬垂。这个单个A可以指导被称为T-A克隆的方法中从5'末端的单个T悬垂的片段的连接。可选地，因为限制性消化后限制酶留下的悬垂的可能的组合是已知的，末端可以保持原样，即不齐的末端。在某些实施方案中，使用具有互补悬垂末端的双链寡核苷酸。

[0074] 在任何处理步骤(例如获得、分离、片段化、扩增或加条形码)后，可以对核酸进行测序。

[0075] 测序可以是通过本领域已知的任何方法。DNA测序技术包括使用标记的终止子或引物和在平板或毛细管中的凝胶分离的经典双脱氧测序反应(Sanger法)、使用可逆终止的标记的核苷酸的合成测序、焦磷酸测序、454测序、Illumina/Solexa测序、与标记的寡核苷酸探针的文库的等位基因特异性杂交、使用与标记的克隆的文库的等位基因特异性杂交然后连接的合成测序、实时监测聚合步骤期间标记的核苷酸的掺入、群落测序(polony sequencing)、SOLiD测序靶向测序、单分子实时测序、外显子测序、基于电子显微术的测序、小组测序(panel sequencing)、晶体管介导的测序、直接测序、随机鸟枪测序、全基因组测序、杂交测序、毛细管电泳、凝胶电泳、双链测序、循环测序、单碱基延伸测序、固相测序、高通量测序、大规模平行签名测序(massively parallel signature sequencing)、乳液PCR、低变性温度共扩增-PCR(COLD-PCR)、多重PCR、可逆染料终止子测序、配对的末端测序、近末端测序(near-term sequencing)、核酸外切酶测序、连接测序、短读段测序、单分子测序、实时测序、反向终止子测序、纳米孔测序、MS-PET测序及其组合。在一些实施方案中，测序方法

可以是大规模平行测序,即,同时(或以快速相继)测序至少100、1000、10,000、100,000、1百万、1千万、1亿、或10亿个多核苷酸分子的任一个。在一些实施方案中,测序可以通过基因分析仪进行,诸如例如可以从Illumina或Applied Biosystems商业获得的基因分析仪。最近已经通过使用聚合酶或连接酶的连续或单一延伸反应以及通过与探针文库的单一或顺序差异杂交证明了分离的分子的测序。测序可以通过DNA测序仪(例如设计用于执行测序反应的机器)来执行。

[0076] 可以使用的测序技术包括例如使用合成测序系统。在第一步中,将DNA修剪成大约300-800个碱基对的片段,并将片段平端化。然后将寡核苷酸衔接子连接到片段的末端。衔接子充当用于片段的扩增和测序的引物。片段可以使用例如含有5'-生物素标签的衔接子B附接至DNA捕获珠,例如链霉亲和素包被的珠。附接于珠的片段在油-水乳液的液滴内进行PCR扩增。结果是每个珠上克隆扩增的DNA片段的多个拷贝。在第二步中,珠被捕获在孔中(皮升大小)。焦磷酸测序在每个DNA片段上平行进行。添加一个或更多个核苷酸生成由测序仪器中的CCD照相机记录的光信号。信号强度与掺入的核苷酸数目成正比。焦磷酸测序利用了在添加核苷酸时释放的焦磷酸(PPi)。PPi在腺苷5'磷酰硫酸的存在下被ATP硫酸化酶转化为ATP。萤光素酶使用ATP将萤光素转化为氧化萤光素,并且该反应生成被检测和分析的光。

[0077] 可以使用的DNA测序技术的另一个实例是来自Life Technologies Corporation(Carlsbad,Calif.)的Applied Biosystems的SOLiD技术。在SOLiD测序中,将基因组DNA修剪成片段,并将衔接子附接到片段的5'和3'末端以生成片段文库。可选地,内部衔接子可以通过将衔接子连接到片段的5'和3'末端、使片段环化、消化环化的片段以生成内部衔接子而被引入,并且将衔接子附接到所得片段的5'和3'末端来生成匹配配对文库。接下来,在含有珠、引物、模板和PCR组分的微量反应器中制备克隆珠群体。PCR后,将模板变性并富集珠以分离具有延伸模板的珠。所选的珠上的模板经受3'修饰,这允许键合到玻璃载玻片。该序列可通过部分随机寡核苷酸与由特定荧光团识别的中心确定碱基(或碱基对)的连续杂交和连接来确定。记录颜色后,去除连接的寡核苷酸,且然后重复该过程。

[0078] 可以使用的DNA测序技术的另一个实例是离子半导体测序,使用例如Life Technologies(South San Francisco,Calif.)的Ion Torrent的以商标IONTORRENT出售的系统。离子半导体测序在例如Rothberg等人,Anintegrated semiconductor device enabling non-optical genome sequencing,Nature 475:348-352(2011);美国公布2010/0304982;美国公布2010/0301398;美国公布2010/0300895;美国公布2010/0300559;和美国公布2009/0026082中描述,其每一个的内容通过引用整体并入。

[0079] 可以使用的测序技术的另一个实例是Illumina测序。Illumina测序是基于使用折回PCR(fold-back PCR)和锚定引物在固体表面上扩增DNA。基因组DNA被片段化,且衔接子被添加到片段的5'和3'末端。附着于流通池通道的表面的DNA片段被延伸并被桥式扩增。片段变成双链的,且双链分子被变性。固相扩增随后变性的多个循环可在流通池的每个通道中产生相同模板的约1,000个拷贝的单链DNA分子的几百万个簇。使用引物、DNA聚合酶和四种荧光团标记的可逆终止核苷酸进行顺序测序。掺入核苷酸后,使用激光激发荧光团并捕获图像并记录第一个碱基的身份。来自每个掺入的碱基的3'终止子和荧光团被去除并且重复掺入、检测和鉴定步骤。根据此技术的测序在美国专利第7,960,120号;美国专利第7,

835,871号;美国专利第7,232,656号;美国专利第7,598,035号;美国专利第6,911,345号;美国专利第6,833,246号;美国专利第6,828,100号;美国专利第6,306,597号;美国专利第6,210,891号;美国公布2011/0009278;美国公布2007/0114362;美国公布2006/0292611;和美国公布2006/0024681中描述,其每一个通过引用整体并入。

[0080] 可以使用的测序技术的另一个实例包括Pacific Biosciences(Menlo Park, Calif.)的单分子实时(SMRT)技术。在SMRT中,四种DNA碱基中的每一种都与四种不同荧光染料中的一种附接。这些染料是磷酸连接的。单个DNA聚合酶与模板单链DNA的单个分子固定在零模波导(ZMW)的底部。将一个核苷酸掺入到增长链需要几毫秒。在此期间,荧光标签被激发并产生荧光信号,且荧光标签被裂解。染料的相应荧光的检测指示掺入了哪个碱基。该过程重复进行。

[0081] 可以使用的测序技术的另一个例子是纳米孔测序(Soni&Meller,2007,Progress toward ultrafast DNA sequence using solid-state nanopores,Clin Chem 53(11):1996-2001)。纳米孔是直径为1纳米的级别的小孔。由于离子通过纳米孔的传导,将纳米孔浸没在导电流体中并在其上施加电势产生轻微的电流。流动的电流的量对纳米孔的尺寸敏感。当DNA分子穿过纳米孔时,DNA分子上的每个核苷酸以不同程度阻碍纳米孔。因此,当DNA分子穿过纳米孔时流经纳米孔的电流的变化代表DNA序列的读取。

[0082] 可以使用的测序技术的另一个实例包括使用化学敏感的场效应晶体管(chemFET)阵列对DNA进行测序(例如,如美国公布2009/0026082中所述的)。在该技术的一个实例中,DNA分子可以被置于反应室中,并可将模板分子与结合聚合酶的测序引物杂交。在测序引物的3'末端将一个或更多个三磷酸掺入新核酸链可以通过chemFET的电流变化来检测。阵列可以具有多个chemFET传感器。在另一个实例中,单核酸可以附接到珠上,并且核酸可以在珠上扩增,并且单个珠可以被转移到chemFET阵列上的单个反应室中,每个室具有chemFET传感器,并且核酸可以被测序。

[0083] 可以使用的测序技术的另一个实例包括使用电子显微镜,如例如由Moudrianakis,E.N.和Beer M.在Base sequence determination in nucleic acids with the electron microscope,III.Chemistry and microscopy of guanine-labeled DNA,PNAS 53:564-71(1965)中描述的。在该技术的一个实例中,单个DNA分子使用可用电子显微镜区分的金属标记物来标记。然后将这些分子在平坦表面上拉伸并使用电子显微镜成像以测量序列。

[0084] 根据本发明的实施方案的测序生成多个读段。根据本发明的读段通常包括长度小于约150个碱基或长度小于约90个碱基的核苷酸数据的序列。在某些实施方案中,读段在约80和约90个碱基之间,例如约85个碱基长度。在一些实施方案中,本发明的方法应用于非常短即长度小于约50或约30个碱基的读段。序列读段数据可以包括序列数据以及元信息。如本领域技术人员已知的,序列读段数据可以以任何合适的文件格式存储,包括例如VCF文件、FASTA文件或FASTQ文件。

[0085] FASTA最初是用于搜索序列数据库的计算机程序,并且名称FASTA也指标准文件格式。参见Pearson&Lipman,1988,Improved tools for biological sequence comparison,PNAS 85:2444-2448。FASTA格式的序列以单行描述开始,接着是序列数据行。描述行通过第一列中的大于(">")符号与序列数据区分开。">"符号后面的单词是序列的标识符,且该行

的其余部分是描述(都是任选的)。“>”和标识符的第一个字母之间不应有空格。建议文本的所有行少于80个字符。如果出现以“>”开头的另一行,则序列结束;这表示另一个序列的开始。

[0086] FASTQ格式是用于存储生物序列(通常是核苷酸序列)及其相应的质量评分的基于文本的格式。它与FASTA格式相似,但是在序列数据之后具有质量评分。为简洁起见,序列字母和质量评分都使用单个ASCII字符编码。FASTQ格式是用于存储高通量测序仪器(诸如Illumina Genome Analyzer)的输出的实际标准。Cock等人,2009,The Sanger FASTQ file format for sequences with quality scores,and the Solexa/Illumina FASTQ variants,Nucleic Acids Res 38(6):1767-1771。

[0087] 对于FASTA和FASTQ文件,元信息包括描述行而不是序列数据行。在一些实施方案中,对于FASTQ文件,元信息包括质量评分。对于FASTA和FASTQ文件,序列数据在描述行之后开始,并且通常使用一些IUPAC模糊代码的子集(任选地带有“-”)呈现。在优选实施方案中,序列数据将使用A、T、C、G和N字符,任选地根据需要包括“-”或者包括U(例如,以表示空位或尿嘧啶)。

[0088] 在一些实施方案中,至少一个主序列读段文件和输出文件被存储为纯文本文件(例如使用诸如ASCII;ISO/IEC 646;EBCDIC;UTF-8或UTF-16的编码)。本发明提供的计算机系统可以包括能够打开纯文本文件的文本编辑器程序。文本编辑器程序可以指能够在计算机屏幕上呈现文本文件(诸如纯文本文件)的内容、允许人员编辑文本(诸如使用监视器、键盘和鼠标)的计算机程序。示例性文本编辑器包括但不限于Microsoft Word、emacs、pico、vi、BBEdit和TextWrangler。优选地,文本编辑器程序能够在计算机屏幕上显示纯文本文件,以人类可读格式显示元信息和序列读段(例如,不是二进制编码而是使用字母数字字符,因为它们将用于打印人类书写)。

[0089] 虽然已经参照FASTA或FASTQ文件讨论了方法,但是本发明的方法和系统可以用于压缩任何合适的序列文件格式,包括例如Variant Call Format(VCF)格式的文件。典型的VCF文件将包括标题部分和数据部分。标题包含任意数量的元信息行,每行都以字符‘##’开始,以及以单个‘#’字符开始的TAB分隔字段定义行。字段定义行命名八个必填列,而主体部分包含填充字段定义行定义的列的数据行。VCF格式在Danecek等人,2011,The variant call format and VCFtools,Bioinformatics 27(15):2156-2158中描述。标题部分可被视为要写入压缩文件的元信息,并且数据部分可被视为行,其中每一行只有在独特时才被存储在主文件中。

[0090] 本发明的某些实施方案提供了序列读段的装配。例如,在通过比对进行装配时,读段彼此比对或与参考比对。通过比对每个读段,继而与参考基因组比对,所有读段关于彼此定位以创建装配体。另外,将序列读段与参考序列比对或映射到参考序列也可以用于识别序列读段中的变体序列。识别变体序列可以与本文描述的方法和系统组合使用,以进一步帮助疾病或状况的诊断或预后或用于指导治疗决定。

[0091] 在一些实施方案中,本发明的任何或全部步骤是自动化的。可选地,本发明的方法可以全部或部分地在一个或更多个专用程序中实现,例如每一个任选地以编译语言诸如C++写入,然后以二进制编译和分发。本发明的方法可以全部或部分地作为现有序列分析平台内的模块或通过调用现有序列分析平台内的功能而实现。在某些实施方案中,本发明的方

法包括响应于单个启动队列(例如,源自人类活动、另一个计算机程序或机器的触发事件中的一个或组合)而自动调用的多个步骤。因此,本发明提供了其中任何步骤或步骤的任何组合可以响应于队列而自动发生的方法。自动地通常意味着不介入人类输入、影响或交互(即仅响应于原始或预先排队的人类活动)。

[0092] 该系统还包括各种形式的输出,其包括主题核酸的准确和灵敏的解释。检索的输出可以以计算机文件的格式提供。在某些实施方案中,输出是FASTA文件、FASTQ文件或VCF文件。输出可被处理以产生文本文件或含有序列数据诸如与参考基因组的序列比对的核酸的序列的XML文件。在其他实施方案中,处理产生包含坐标或描述主题核酸中相对于参考基因组的一个或更多个突变的串的输出。本领域中已知的比对串包括简单无空位比对报告(Simple UnGapped Alignment Report,SUGAR)、详细的有用的标记空位比对报告(Verbose Useful Labeled Gapped Alignment Report,VULGAR)和简要比对信息表达式(Compact Idiosyncratic Gapped Alignment Report,CIGAR)(Ning,Z.等人,Genome Research 11(10):1725-9(2001))。这些串例如在来自European Bioinformatics Institute(Hinxton,UK)的Exonerate序列比对软件中实现。

[0093] 在一些实施方案中,产生包含CIGAR串的序列比对——诸如,例如序列比对图(SAM)或二元比对图(BAM)文件(SAM格式在例如Li等人,The Sequence Alignment/Map format and SAMtools,Bioinformatics,2009,25(16):2078-9中描述)。在一些实施方案中,CIGAR显示或包括每行一个空位比对。CIGAR是一种报告为CIGAR串的压缩的成对比对格式。CIGAR串对于表示长(例如基因组)成对比对是有用的。在SAM格式中使用CIGAR串来表示读段与参考基因组序列的比对。

[0094] CIGAR串遵循建立的基序。每个字符前面是数字,给出事件的碱基计数。使用的字符可以包括M、I、D、N和S(M=匹配;I=插入;D=缺失;N=空位;S=置换)。CIGAR串定义匹配/不匹配和缺失(或空位)的序列。例如,CIGAR串2MD3M2D2M将意味着,比对包含2个匹配、1个缺失(为了节省一些空间省略数字1)、3个匹配、2个缺失和2个匹配。

[0095] 如本发明所预期的,上文描述的功能可以使用包括软件、硬件、固件、硬接线或者这些的任何组合的本发明的系统来实现。实现功能的特征还可以物理地位于各个位置,包括被分布为使得功能的各部分在不同的物理位置处实现。

[0096] 如本领域技术人员将认识到必要或最适合于执行本发明的方法,本发明的计算机系统或机器包括一个或更多个处理器(例如中央处理单元(CPU)、图形处理单元(GPU)或二者)、主存储器和静态存储器,其经由总线彼此通信。

[0097] 图4图示了适于执行本发明的方法的系统701。如图7中示出的,系统701可以包括服务器计算机705、终端715、测序仪715、测序仪计算机721、计算机749中的一个或更多个,或其任何组合。每个这样的计算机设备可以经由网络709通信。测序仪725可以任选地包括或可操作地偶联到其自己的、例如专用测序仪计算机721(包括任何输入/输出机构(I/O)、处理器和存储器)。另外地或可选地,测序仪725可以经由网络709可操作地偶联到服务器705或计算机749(例如,膝上型计算机、台式计算机或平板电脑)。计算机749包括一个或更多个处理器、存储器和I/O。在本发明的方法采用客户机/服务器结构的情况下,本发明的方法的任何步骤可以使用服务器705来执行,该服务器705包括能够获得数据、指令等的处理器、存储器和I/O中的一个或更多个,或者通过接口模块提供结果或将结果作为文件提供。

服务器705可以通过计算机749或终端715接合网络709，或者服务器705可以直接连接到终端715。终端515优选地是计算机设备。根据本发明的计算机优选地包括偶联到I/O机构和存储器的一个或更多个处理器。

[0098] 处理器可以由一个或更多个处理器提供，包括例如单核或多核处理器中的一个或更多个。I/O机构可以包括视频显示单元(例如，液晶显示器(LCD)或阴极射线管(CRT))、字母数字输入设备(例如键盘)、光标控制设备(例如鼠标)、磁盘驱动单元、信号发生设备(例如扬声器)、加速度计、麦克风、蜂窝射频天线以及网络接口设备(例如网络接口卡(NIC)、Wi-Fi卡、蜂窝调制解调器、数据插孔、以太网端口、调制解调器插孔、HDMI端口、迷你HDMI端口、USB端口)、触摸屏(例如CRT、LCD、LED、AMOLED、超级AMOLED)、指示设备、触控板、光(例如LED)、光/图像投影设备或其组合。根据本发明的存储器指的是由一个或更多个有形设备提供的非瞬时性存储器，所述有形设备优选地包括在其上存储一组或更多组指令的一个或更多个机器可读介质(例如软件)，所述指令实现本文描述的方法或功能中的任何一个或更多个。软件还可以在被系统701内的计算机执行期间完全或至少部分地驻留在主存储器、处理器或二者中，主存储器和处理器也构成机器可读介质。软件还可以经由网络接口设备通过网络发送或接收。

[0099] 尽管机器可读介质可以在示例性实施方案中被示为是单个介质，但是术语“机器可读介质”应该被认为包括存储一组或多组指令的单个介质或多个介质(例如，集中式或分布式数据库、和/或相关联的高速缓存和服务器)。术语“机器可读介质”还应当包括能够储存、编码或携带指令集的任何介质，其中这些指令集由机器执行或使机器执行本发明的任何一种或更多种方法。存储器可以是例如硬盘驱动器、固态驱动器(SSD)、光盘、闪存、压缩盘、磁带驱动器、“云”存储位置中的一个或更多个，或其组合。在某些实施方案中，本发明的设备包括用于存储器的有形的、非瞬时性计算机可读介质。用作存储器的示例性装置包括半导体存储器装置(例如EPROM、EEPROM、固态驱动器(SSD)和闪存装置例如SD、微SD、SDXC、SDIO、SDHC卡)；磁盘(例如内部硬盘或可移动磁盘)；和光盘(例如CD和DVD盘)。

[0100] 在一些实施方案中，本公开内容的方法和系统可以用于诊断疾病或状况，例如癌症。如本文使用的，术语“诊断”是指技术人员能够估计和/或确定患者是否患有给定疾病或状况的方法。在一些实施方案中，本发明的方法可以用于疾病或状况(例如癌症)的疾病的预后。如本文使用的，术语“预后”是指疾病或状况进展的可能性，包括疾病或状况的复发。在一些实施方案中，本发明的方法可用于评估发展疾病或状况(例如癌症)的风险。例如，本文描述的方法和系统可用于鉴定与特定诊断、预后或者发展疾病或状况的风险相关的断点或基因融合。此外，本文所述的方法和系统可用于鉴定与预测的治疗结果相关的断点或基因融合。因此，所述方法和系统可以用于指导疾病或状况的治疗(例如通过向受试者施用化合物或剂)或指导用于治疗疾病或状况的药物的制备。

[0101] 如本文使用的，“治疗”疾病或状况是指采取步骤以获得有益或期望的结果，包括临床结果。有益或期望的临床结果包括但不限于缓解或改善与疾病或状况相关的一种或更多种症状。如本文使用的，向受试者“施用(administering)”或“施用(administration of)”化合物或剂可以使用本领域技术人员已知的多种方法中的一种进行。例如，化合物或剂可以静脉内、动脉内、皮内、肌肉内、腹膜内、静脉内、皮下、眼部、舌下、经口(通过摄入)、鼻内(通过吸入)、脊椎内、脑内和透皮(通过吸收，例如通过皮肤导管)施用。化合物或剂也

可以通过可再充装或生物可降解的聚合物装置或其他装置(例如贴剂和泵)或制剂适当地引入,所述装置或制剂提供化合物或剂的延长、缓慢或控制释放。施用还可以进行例如一次、多次和/或在一个或更多个延长的时间段内进行。在某些方面,施用包括直接施用(包括自我施用)和间接施用(包括给药物开处方的行为)。例如,如本文使用的,指导患者自我施用药物或者由另一个人施用药物和/或向患者提供药物处方的医师正在向患者施用药物。在一些实施方案中,化合物或剂通过口服例如通过摄入施用至受试者,或者经静脉地例如通过注射施用至患者。在一些实施方案中,口服施用的化合物或剂处于延长释放或缓慢释放制剂中,或者使用用于这种缓慢或延长释放的装置施用。

[0102] 如本文使用的,术语“癌症”包括但不限于各种类型的恶性赘生物,其中大多数可侵入周围组织,并可以转移至不同部位(参见例如PDR Medical Dictionary,第1版(1995),其出于所有目的通过引用整体并入本文)。术语“赘生物(neoplasm)”和“肿瘤(tumor)”是指通过细胞增殖比正常组织更迅速地生长并且在去除引发增殖的刺激之后继续生长的异常组织。这种异常组织显示部分或完全缺乏结构组织和与可以是良性(例如良性肿瘤)或恶性(例如恶性肿瘤)的正常组织的功能协调。癌症的一般类别的实例包括但不限于上皮癌(carcinoma)(源自上皮细胞的恶性肿瘤诸如例如乳腺癌、前列腺癌、肺癌和结肠癌的常见形式)、肉瘤(源自结缔组织或间充质细胞的恶性肿瘤)、淋巴瘤(源自造血细胞的恶性肿瘤)、白血病(源自造血细胞的恶性肿瘤)和生殖细胞肿瘤(源自全能性细胞的肿瘤,在成人中最常见于睾丸或卵巢中;在胎儿、婴儿和年幼儿童中,最常见于身体中线,特别是在尾骨尖处)、急变性肿瘤(类似未成熟组织或胚胎组织的典型恶性肿瘤)等。意图被本发明涵盖的赘生物类型的实例包括但不限于与神经组织、血液形成组织、乳房、皮肤、骨骼、前列腺、卵巢、子宫、子宫颈、肝、肺、脑、喉、胆囊、胰腺、直肠、甲状旁腺、甲状腺、肾上腺、免疫系统、头和颈、结肠、胃、支气管和/或肾的癌症相关的那些赘生物。在特定实施方案中,可以被检测的癌症的类型和数目包括但不限于血癌、脑癌、肺癌、皮肤癌、鼻癌、喉癌、肝癌、骨癌、淋巴瘤、胰腺癌、皮肤癌、肠癌、直肠癌、甲状腺癌、膀胱癌、肾癌、口腔癌、胃癌、实体瘤(solid state tumors)、异质肿瘤、均质肿瘤等。在特定实施方案中,癌症是血液学癌症、肉瘤或前列腺癌。

[0103] 虽然本文已经显示和描述了本发明的优选实施方案,但对于本领域技术人员将明显的是,此类实施方案仅通过示例的方式提供。在不偏离本发明的情况下,本领域技术人员现在将想到许多变化、改变和替换。应当理解,在实践本发明时可以采用本文描述的本发明的实施方案的各种替代选择。以下权利要求意图界定本发明的范围,并且从而涵盖在这些权利要求范围内的方法和结构及其等同物。

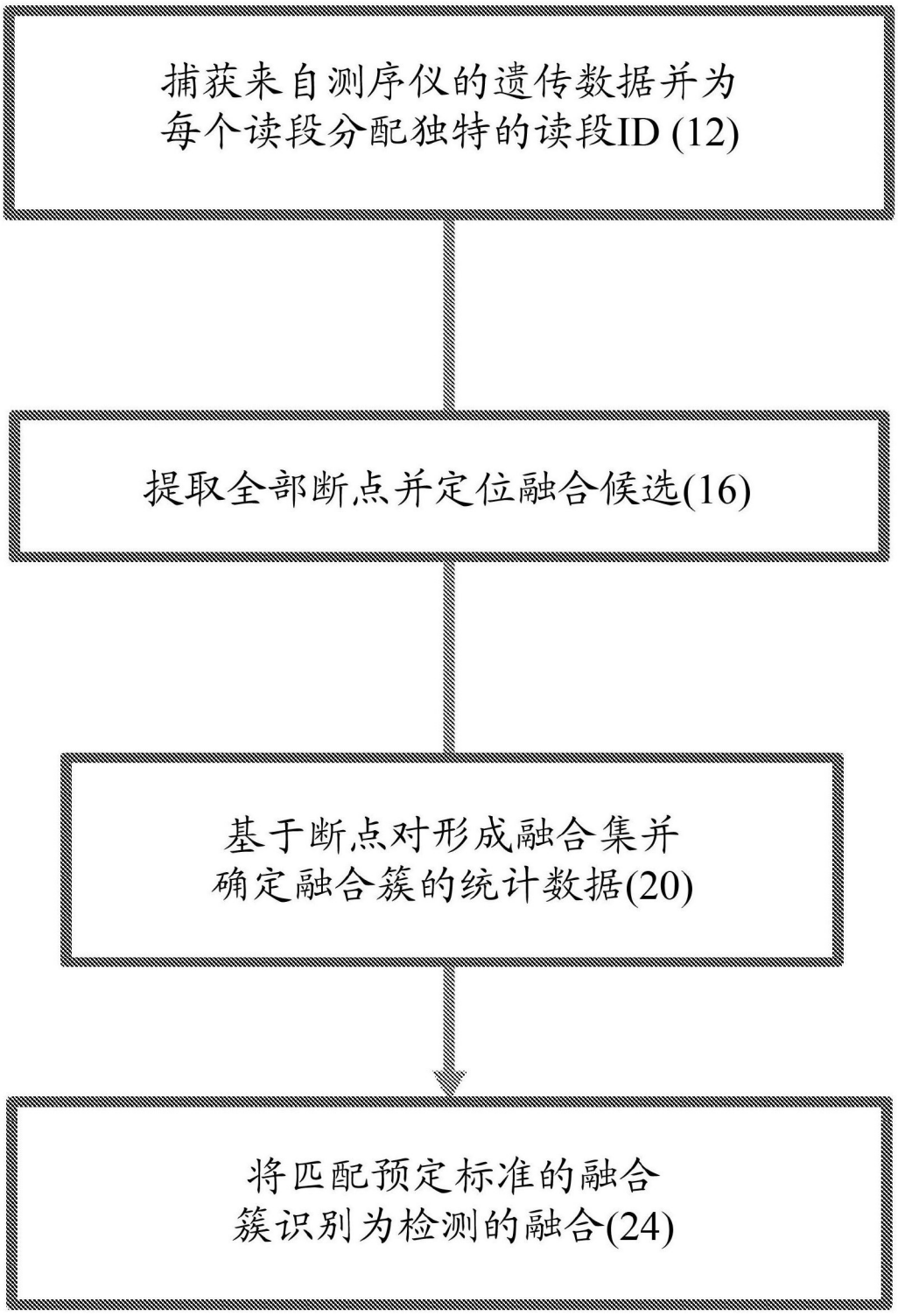
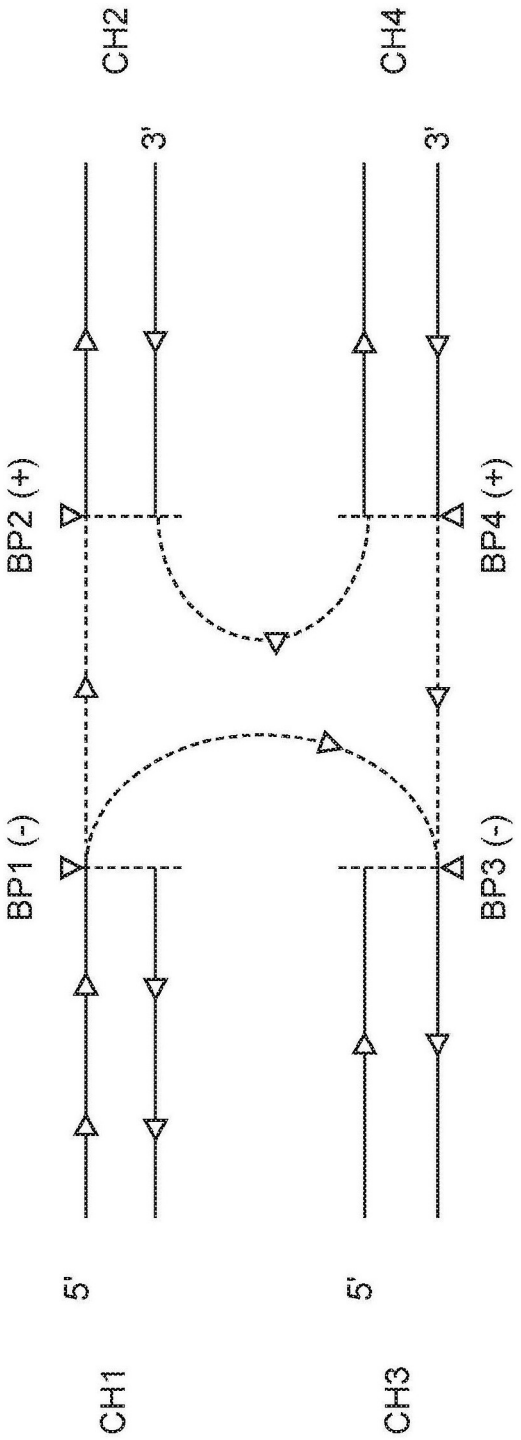

捕获来自测序仪的遗传数据并为
每个读段分配独特的读段ID (12)
提取全部断点并定位融合候选(16)
基于断点对形成融合集并
确定融合簇的统计数据(20)
将匹配预定标准的融合
簇识别为检测的融合(24)

图1




可能的融合

| 融合 | 断点1 | 断点2 | (上游)链1 | (下游)链2 | |
|---|---|---|---|---|---|
| CH1-CH2 | BP1 (-) | BP2 (+) | CH1:5' | CH2:5' | 彼此反向互补 |
| CH1-CH2 | BP2 (+) | BP1 (-) | CH2:3' | CH1:3' | |
| | | | (相同的融合) | | |
| CH1-CH3 | BP1 (-) | BP3 (-) | CH1:5' | CH3:3' | 彼此反向互补 |
| CH3-CH1 | BP3 (-) | BP1 (-) | CH3:5' | CH1:3' | |
| CH2-CH4 | BP2 (+) | BP4 (+) | CH2:3' | CH4:5' | 彼此反向互补 |
| CH4-CH2 | BP4 (+) | BP2 (+) | CH4:3' | CH2:5' | |

图2

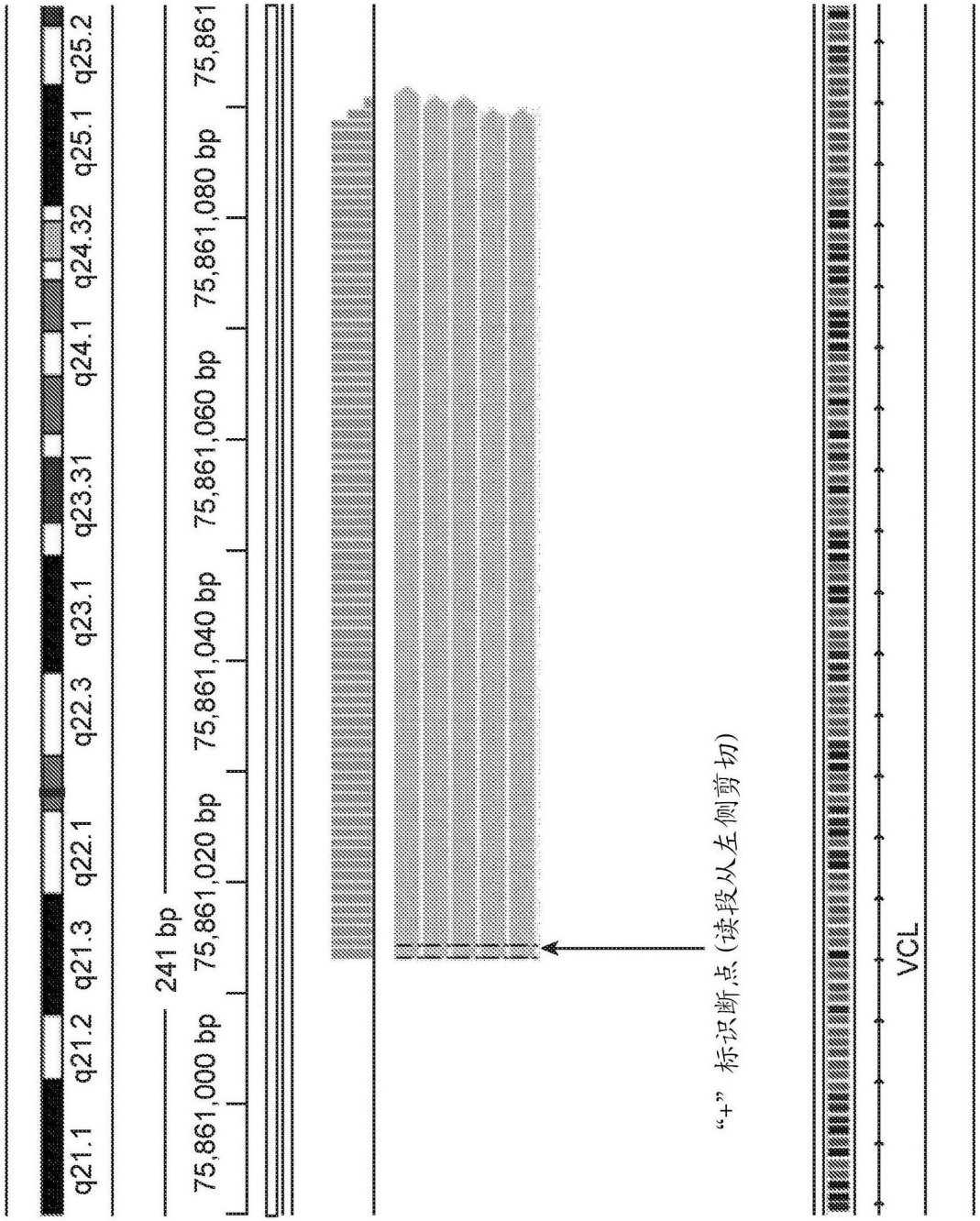

q21.1
q21.2
q21.3
q22.1
q22.3
q23.1
q23.31
q24.1
q24.32
q25.1
q25.2
241 bp
75,861,000 bp
75,861,020 bp
75,861,040 bp
75,861,060 bp
75,861,080 bp
75,861
“+”标识断点(读段从左侧剪切)
VCL

图3A

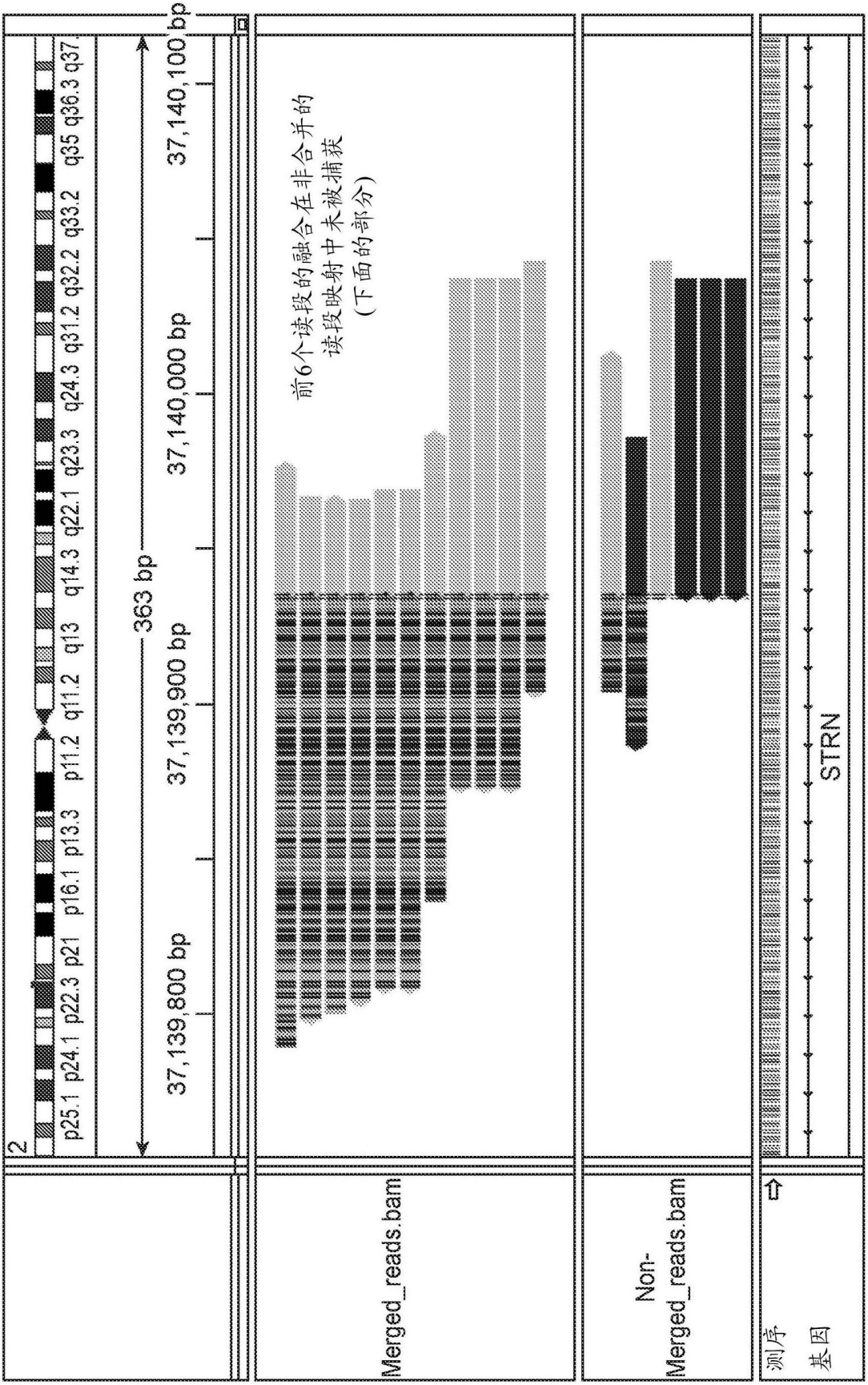

2
p25.1 p24.1 p22.3 p21 p16.1 p13.3 p11.2 q11.2 q13 q14.3 q22.1 q23.3 q24.3 q31.2 q32.2 q33.2 q35 q36.3 q37
37,139,800 bp
363 bp
37,139,900 bp
37,140,000 bp
37,140,100 bp
前6个读段的融合在非合并的
读段映射中未被捕获
(下面的部分)
Merged_reads.bam
Non-
Merged_reads.bam
测序
基因
STRN

图3B

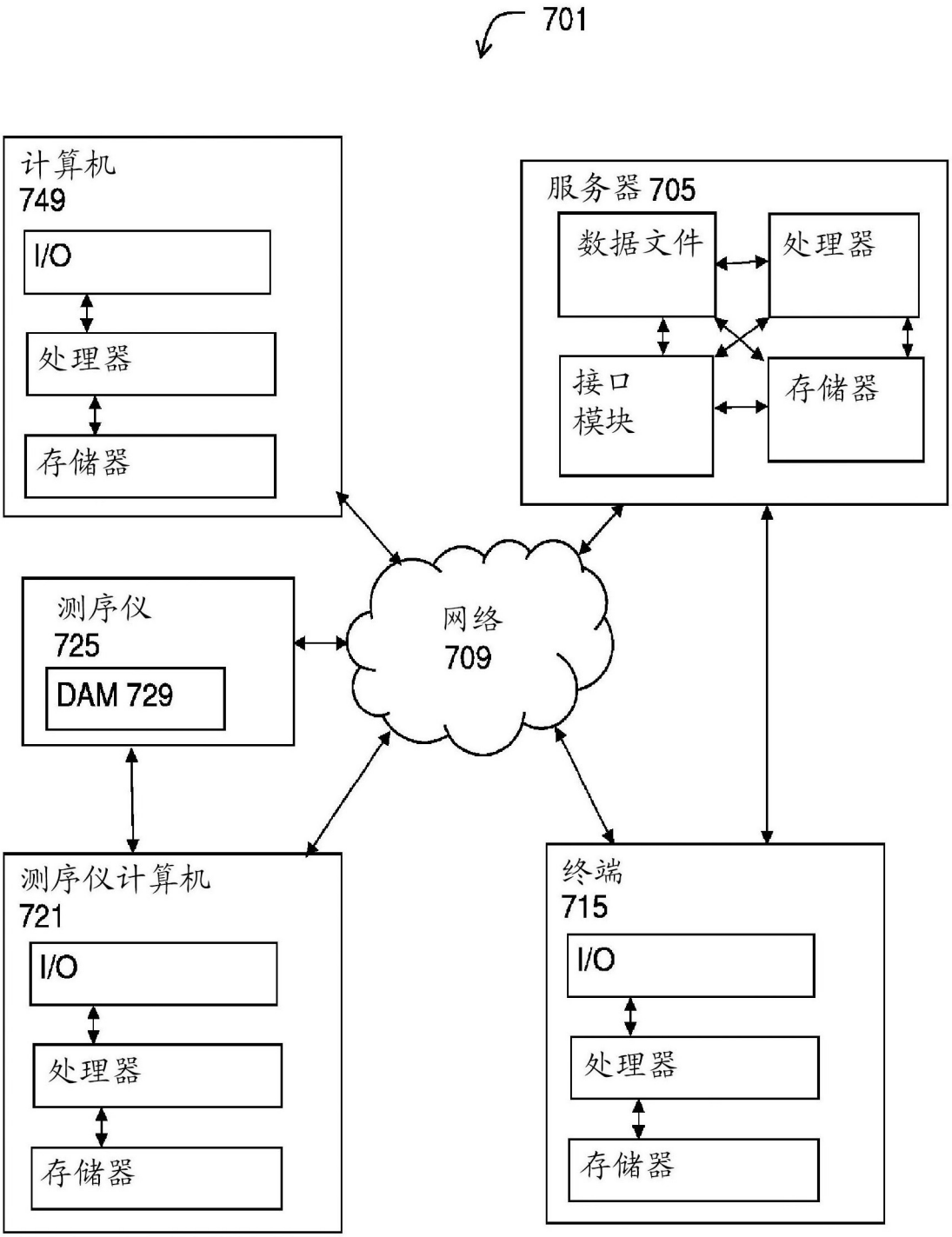

701
计算机
749
I/O
处理器
存储器
服务器 705
数据文件
处理器
接口
模块
存储器
网络
709
测序仪
725
DAM 729
测序仪计算机
721
I/O
处理器
存储器
终端
715
I/O
处理器
存储器

图4

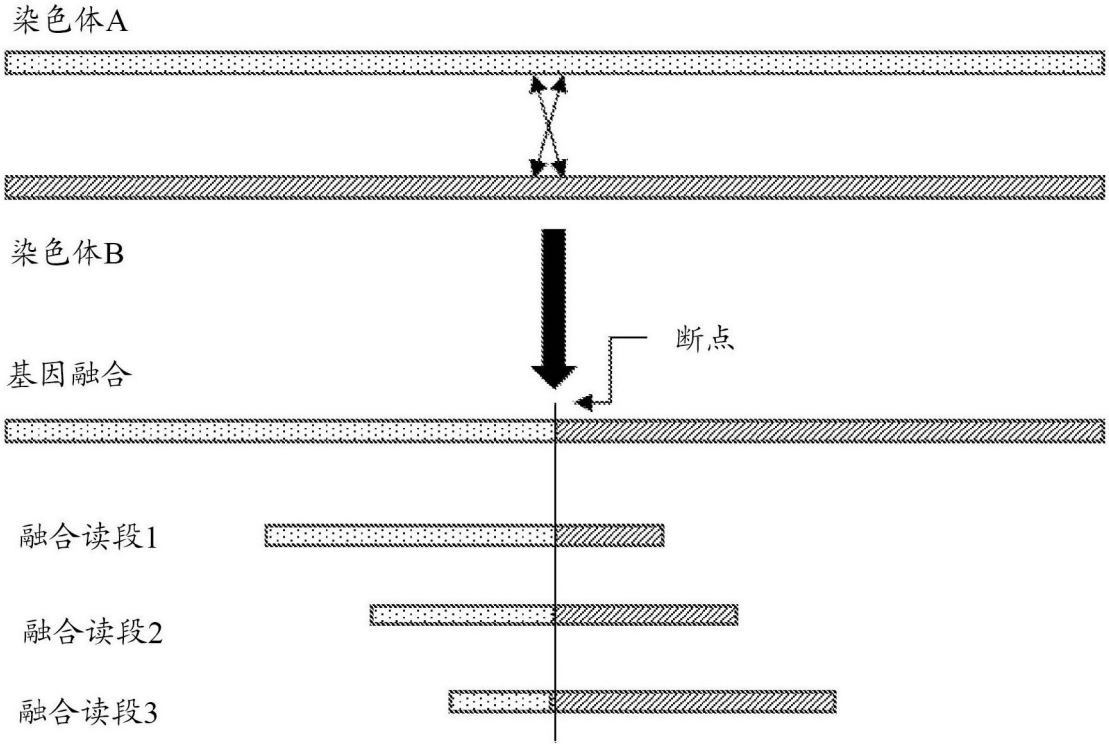

染色体A
染色体B
断点
基因融合
融合读段1
融合读段2
融合读段3

图5

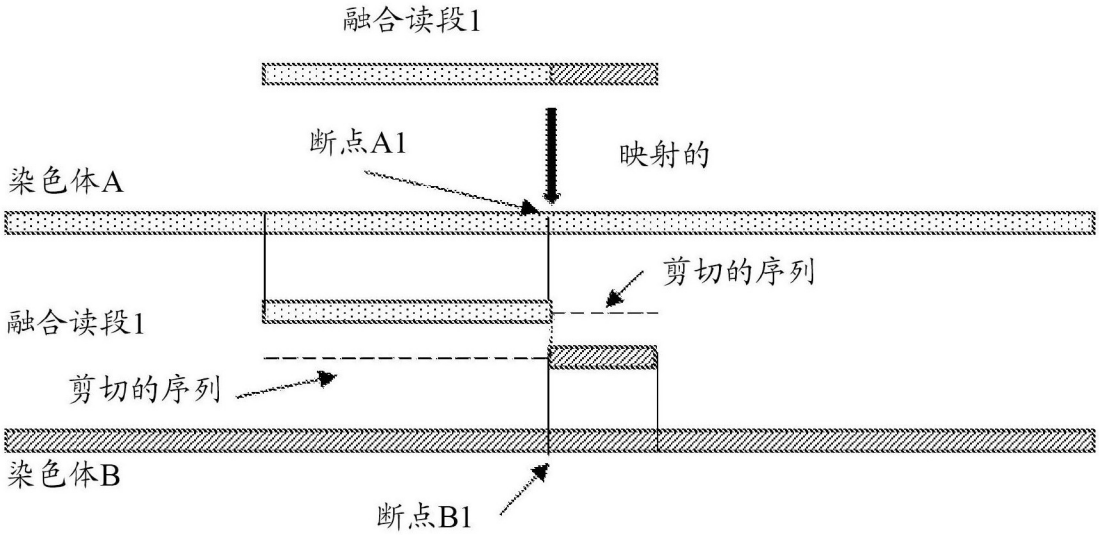

融合读段1
断点A1
映射的
染色体A
剪切的序列
融合读段1
剪切的序列
染色体B
断点B1

图6

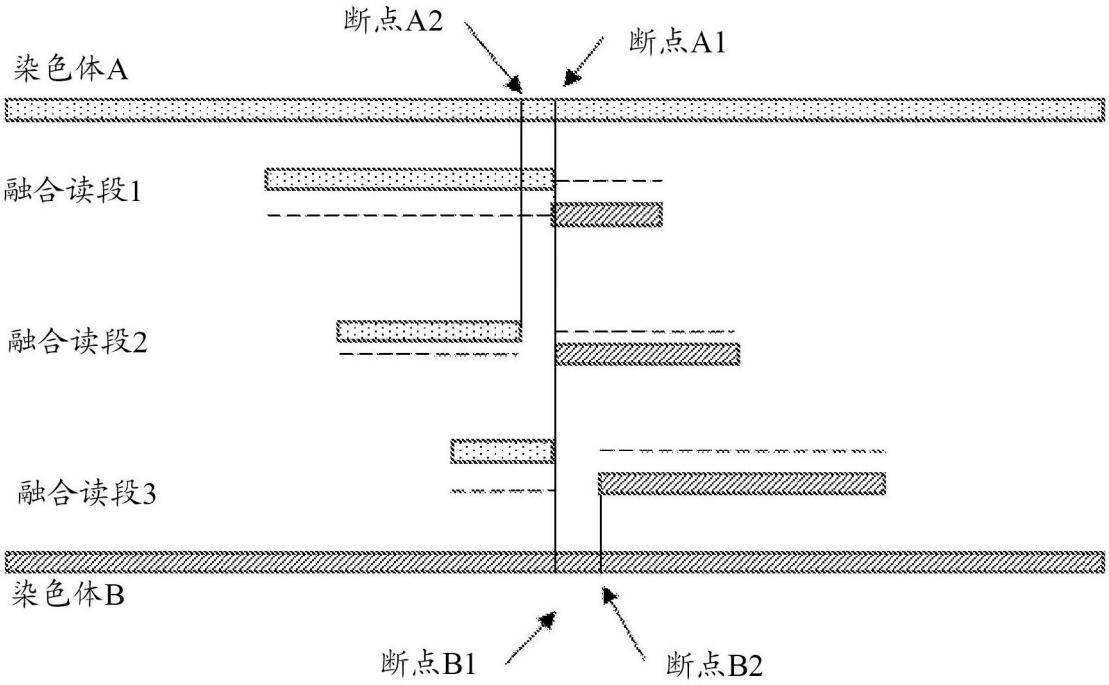
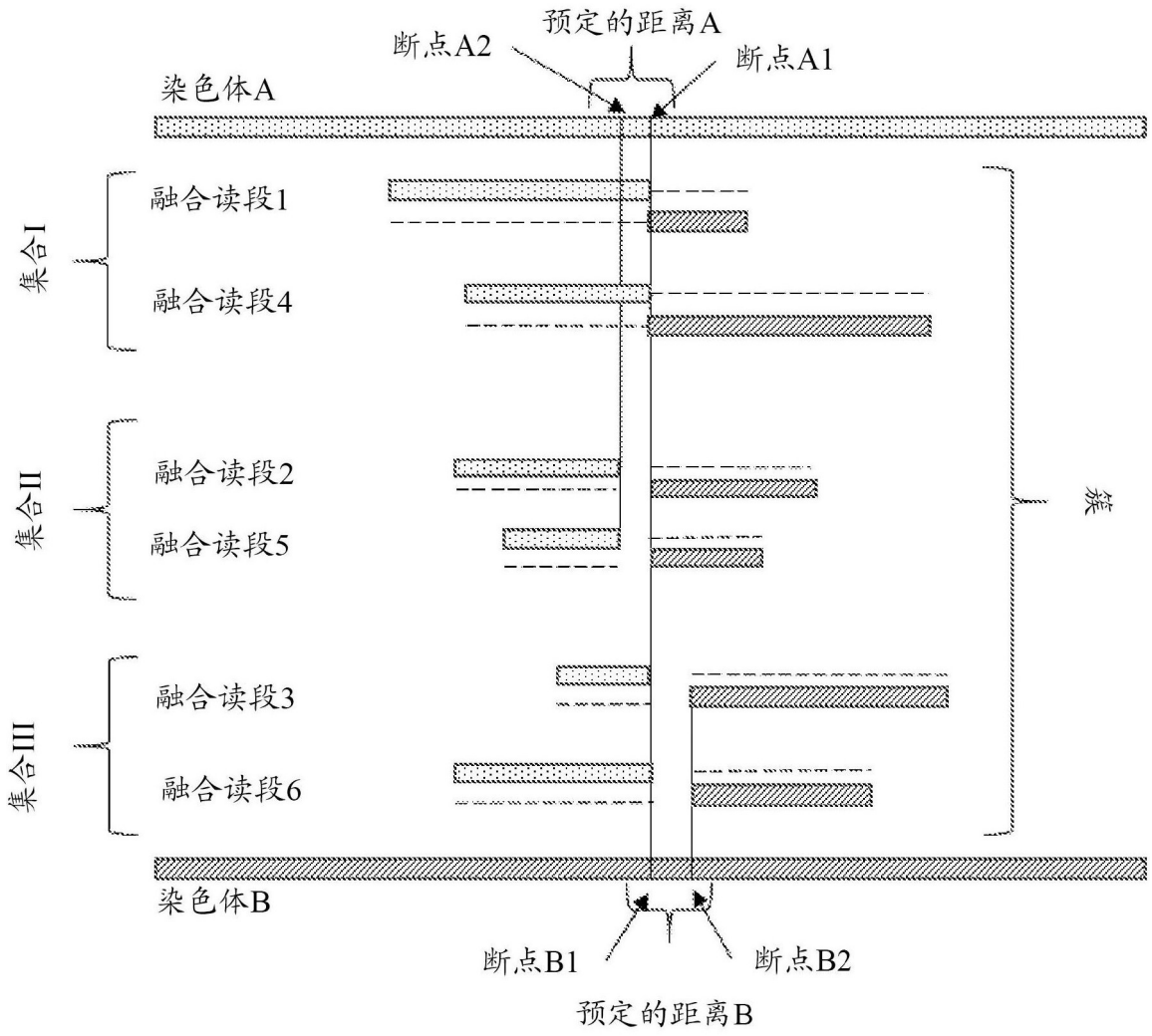



图7




染色体A:
断点A1：融合读段1、4、3、6＝4个
断点A2：融合读段2、5＝2个
第一基因融合断点判定：A1

染色体B:
断点B1：融合读段1、4、2、5＝4个
断点B2：融合读段3、6＝2个
第二基因融合断点判定：B1

基因融合断点对：A1-B1

图8

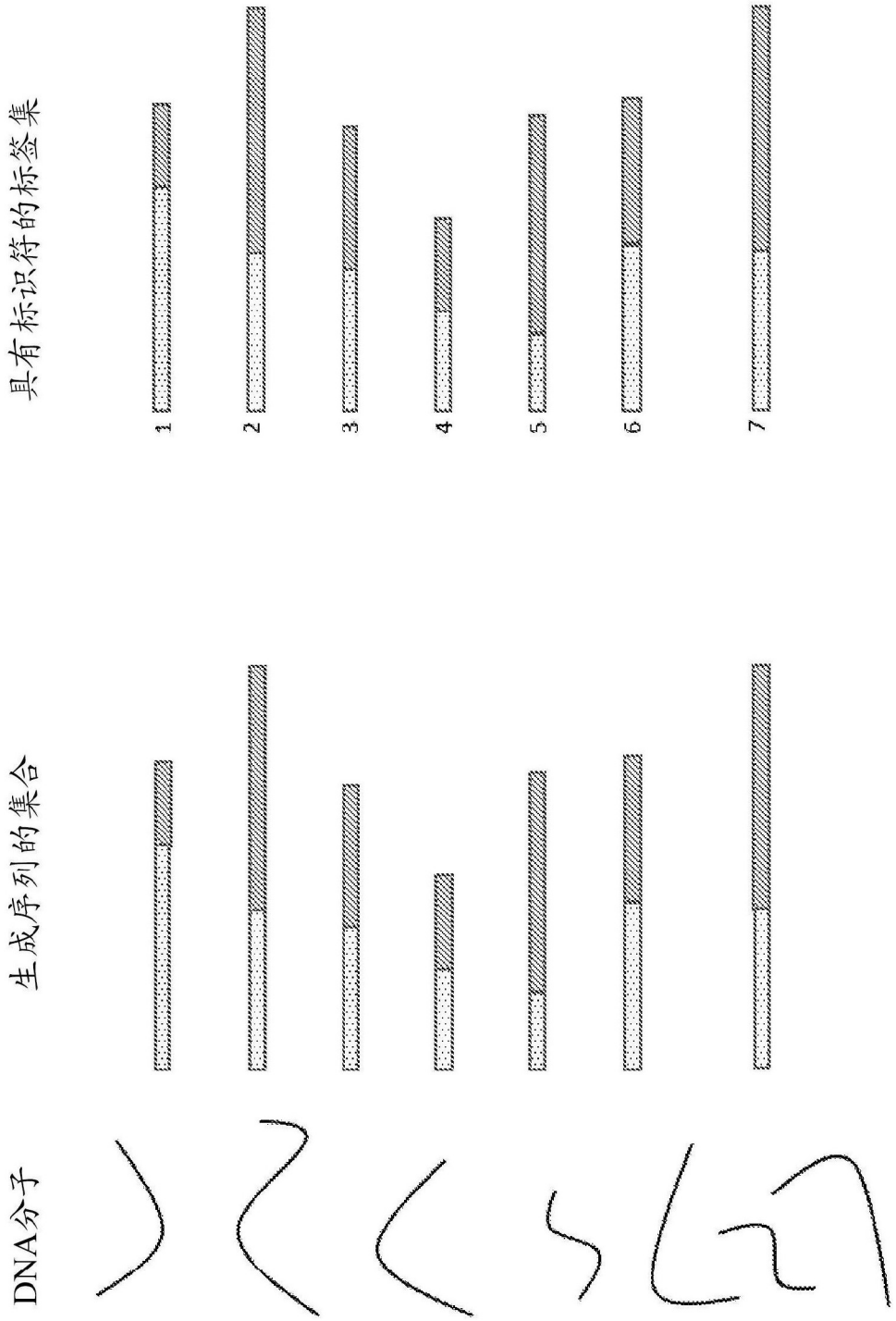



图9A

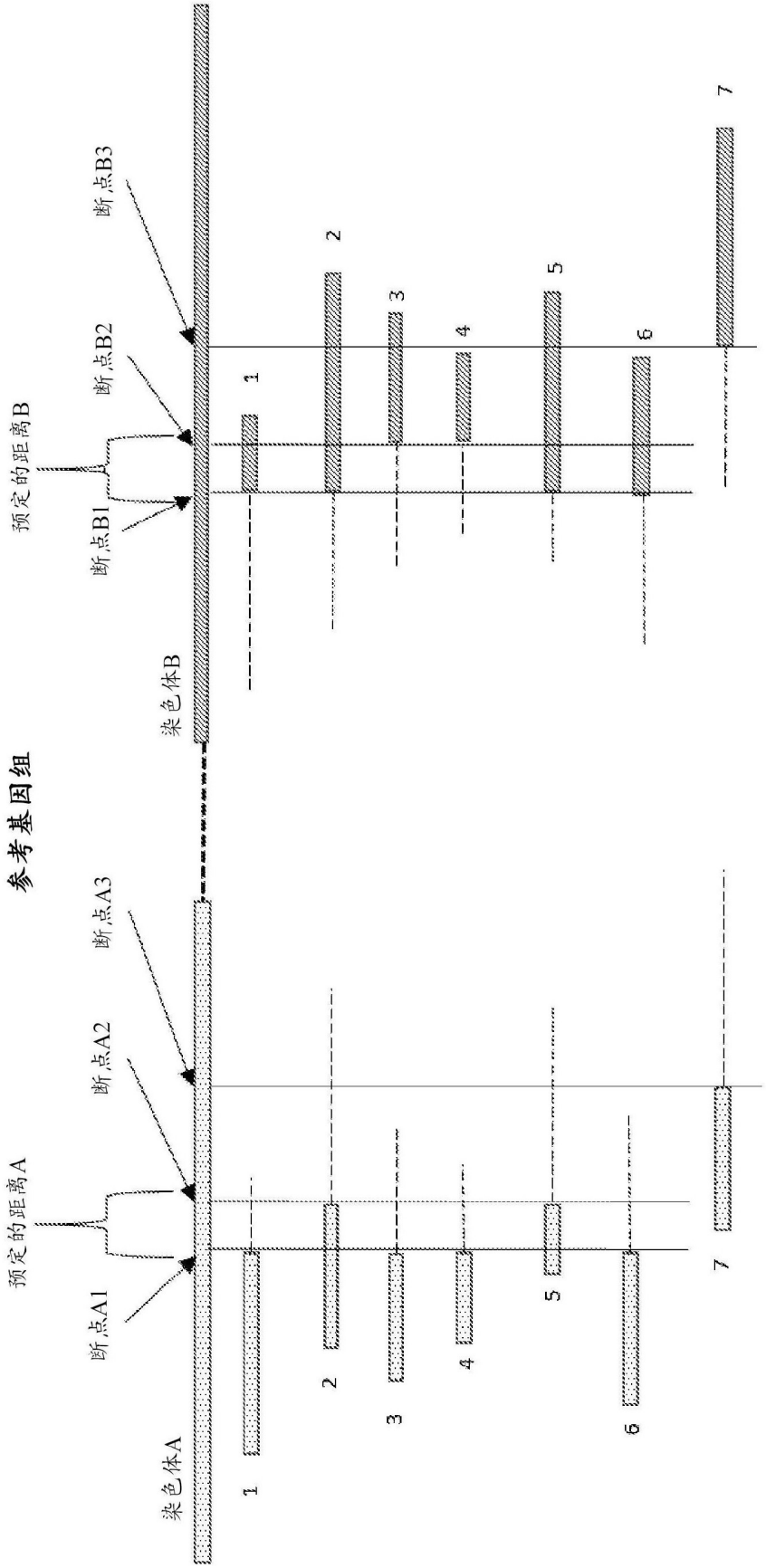

参考基因组
染色体A
染色体B
预定的距离A
预定的距离B
断点A1
断点A2
断点A3
断点B1
断点B2
断点B3
1
2
3
4
5
6
7

图9B

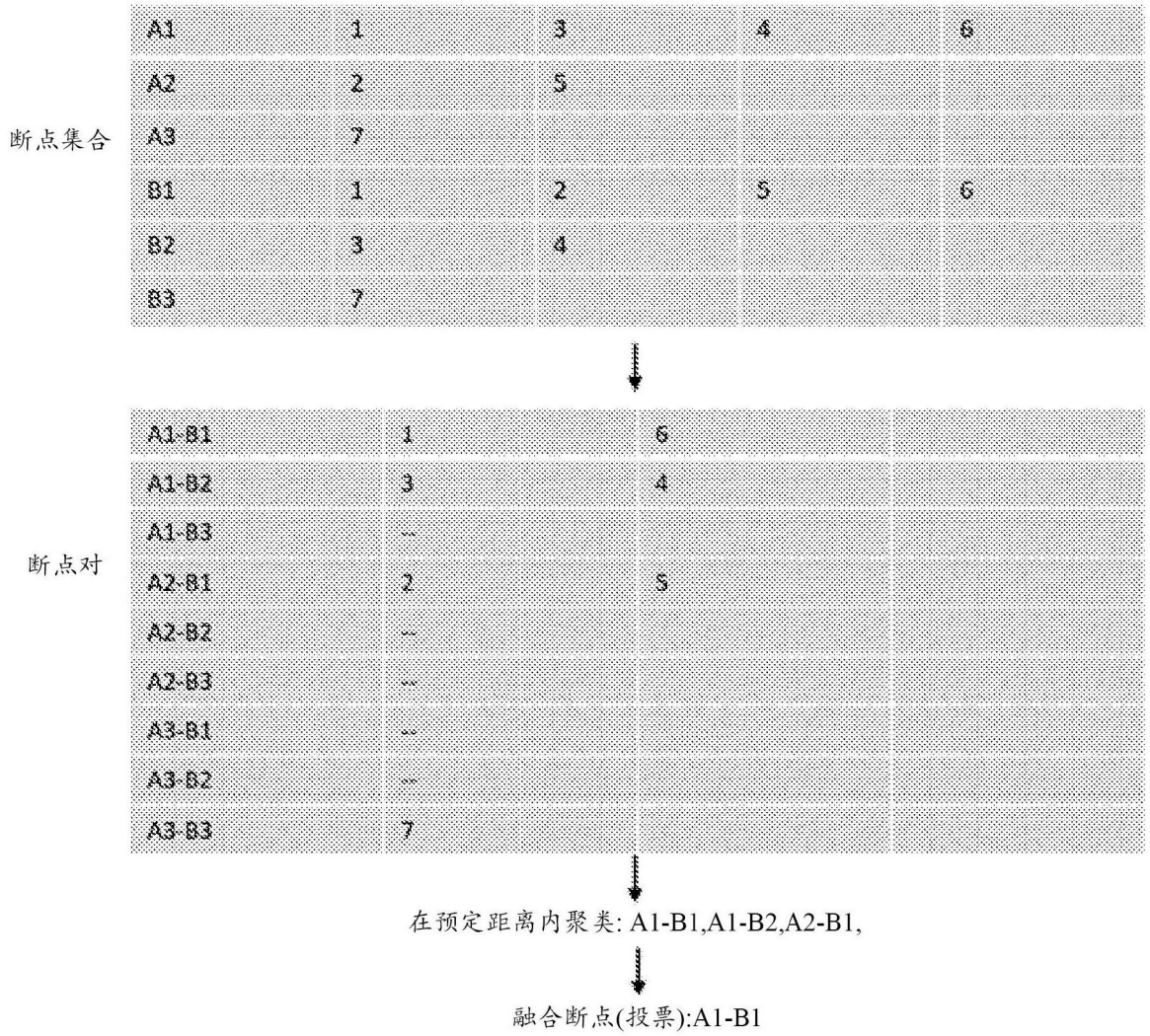

断点集合
A1 1 3 4 6
A2 2 5
A3 7
B1 1 2 5 6
B2 3 4
B3 7
断点对
A1-B1 1 6
A1-B2 3 4
A1-B3 --
A2-B1 2 5
A2-B2 --
A2-B3 --
A3-B1 --
A3-B2 --
A3-B3 7
在预定距离内聚类: A1-B1,A1-B2,A2-B1,
融合断点(投票):A1-B1

图9C