Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2852665 A1 2013/04/25

(21) 2 852 665

(12) DEMANDE DE BREVET CANADIEN
CANADIAN PATENT APPLICATION

(13) A1

(86) Date de dépôt PCT/PCT Filing Date: 2012/09/14
(87) Date publication PCT/PCT Publication Date: 2013/04/25
(85) Entrée phase nationale/National Entry: 2014/04/16
(86) N° demande PCT/PCT Application No.: US 2012/055362
(87) N° publication PCT/PCT Publication No.: 2013/058907
(30) Priorité/Priority: 2011/10/17 (US61/548,073)

(51) Cl.Int./Int.Cl. *C12Q 1/68* (2006.01), *G06F 19/18* (2011.01), *G06F 19/22* (2011.01)

(71) Demandeur/Applicant:
GOOD START GENETICS, INC., US

(72) Inventeurs/Inventors:
KENNEDY, CALEB J., US;
UMBARGER, MARK, US;
PORRECA, GREGORY, US

(74) Agent: SMART & BIGGAR

(54) Titre : METHODES D'IDENTIFICATION DE MUTATIONS ASSOCIEES A DES MALADIES
(54) Title: ANALYSIS METHODS

(57) Abrégé/Abstract:
The disclosure relates to methods for analyzing nucleic acids to identify mutations associated with diseases. More specifically, the methods of the invention involve obtaining nucleic acid from a subject having a disease, identifying at least one mutation in the nucleic acid, and comparing the mutation to a database of mutations known to be associated with the disease, wherein mutations that do not match to the database are identified as newly identified mutations.

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
25 April 2013 (25.04.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/058907 A1

(51) **International Patent Classification**:
C12Q 1/68 (2006.01)

(21) **International Application Number**:
PCT/US2012/055362

(22) **International Filing Date**:
14 September 2012 (14.09.2012)

(25) **Filing Language**: English

(26) **Publication Language**: English

(30) **Priority Data**:
61/548,073 17 October 2011 (17.10.2011) US

(71) **Applicant** *(for all designated States except US)*: **GOOD START GENETICS, INC.** [US/US]; 237 Putnam Avenue, Cambridge, MA 02139 (US).

(72) **Inventors; and**
(75) **Inventors/Applicants** *(for US only)*: **KENNEDY, Caleb, J.** [US/US]; 92 Oxford Street, Arlington, MA 02474 (US). **UMBARGER, Mark** [US/US]; 5 Winchester Street, Unit 101, Brookline, MA 02446 (US). **PORRECA, Gregory** [US/US]; 57 Reservoir Street, Cambridge, MA 02138 (US).

(74) **Agents**: **MEYERS, Thomas C.** et al.; Brown Rudnick LLP, One Financial Center, Boston, MA 02111 (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, 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, 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, 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, ML, MR, NE, SN, TD, TG).

Published:

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

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

WO 2013/058907 A1

(54) Title: ANALYSIS METHODS

(57) Abstract: The disclosure relates to methods for analyzing nucleic acids to identify mutations associated with diseases. More specifically, the methods of the invention involve obtaining nucleic acid from a subject having a disease, identifying at least one mutation in the nucleic acid, and comparing the mutation to a database of mutations known to be associated with the disease, wherein mutations that do not match to the database are identified as newly identified mutations.

ANALYSIS METHODS

Related Application

The present application claims the benefit of and priority to U.S. provisional application serial number 61/548,073, filed October 17, 2011, the content of which is incorporated by reference herein its entirety.

Field of the Invention

The invention generally relates to methods for analyzing nucleic acids to identify novel mutations associated with diseases.

Background

All genetic diseases are associated with some form of genomic instability. Abnormalities can range from a discrete mutation in a single base in the DNA of a single gene to a gross chromosome abnormality involving the addition or subtraction of an entire chromosome or set of chromosomes. Being able to identify the genetic abnormalities associated with a particular disease provides a mechanism by which one can diagnosis a subject as having the disease.

Summary

The invention generally relates to methods for analyzing nucleic acids to identify novel mutations associated with diseases. Methods of the invention involve obtaining nucleic acid from a subject having a disease, identifying at least one mutation in the nucleic acid, and comparing the mutation to a database of mutations known to be associated with the disease, wherein mutations that do not match to the database are identified as novel mutations.

Numerous methods of identifying mutations in nucleic acids are known by those of skill in the art and any of those methods may be used with methods of the invention. In certain embodiments, identifying a mutation in a nucleic acid from a sample involves sequencing the nucleic acid, and comparing the sequence of the nucleic acid from the sample to a reference sequence. Any sequencing technique known in the art may be used, such as sequencing-by-synthesis and more particularly single molecule sequencing-by-synthesis. The reference sequence may be a consensus human sequence or a sequence from a non-diseased sample.

Certain aspects of the invention are especially amenable for implementation using a computer. Such systems generally include a central processing unit (CPU) and storage coupled to the CPU. The storage stores instructions that when executed by the CPU, cause the CPU execute the method steps described above and throughout the present application.

Brief Description of the Drawings

Figure 1 is an illustration of EIR for a simple homopolymeric sequence.

Figure 2 is an illustration of the CFTR exon 10 5' boundary (hg18).

Figure 3 illustrates a system for performing methods of the invention.

Detailed Description

The invention generally relates to methods for analyzing nucleic acids to identify novel mutations associated with diseases. Methods of the invention involve obtaining nucleic acid from a subject having a disease, identifying at least one mutation in the nucleic acid, and comparing the mutation to a database of mutations known to be associated with the disease, wherein mutations that do not match to the database are identified as novel mutations.

Obtaining a Tissue Sample and Extraction of nucleic acid

Methods of the invention involve obtaining a sample, e.g., tissue, blood, bone, that is suspected to be associated with a disease. Such samples may include tissue from brain, kidney, liver, pancreas, bone, skin, eye, muscle, intestine, ovary, prostate, vagina, cervix, uterus, esophagus, stomach, bone marrow, lymph node, and blood. Once the sample is obtained, nucleic acids are extracted.

Nucleic acids may be obtained by methods known in the art. Generally, nucleic acids can be extracted from a biological sample by a variety of techniques such as those described by Maniatis, et al., Molecular Cloning: A Laboratory Manual, Cold Spring Harbor, N.Y., pp. 280-281, (1982), the contents of which is incorporated by reference herein in its entirety.

It may be necessary to first prepare an extract of the cell and then perform further steps--i.e., differential precipitation, column chromatography, extraction with organic solvents and the like--in order to obtain a sufficiently pure preparation of nucleic acid. Extracts may be prepared using standard techniques in the art, for example, by chemical or mechanical lysis of the cell. Extracts

then may be further treated, for example, by filtration and/or centrifugation and/or with chaotropic salts such as guanidinium isothiocyanate or urea or with organic solvents such as phenol and/or $HCCl_3$ to denature any contaminating and potentially interfering proteins.

Capture of target sequences

Any method known in the art for capturing target sequences may be used with methods of the invention. In certain embodiments, an oligonucleotide-driven annealing reaction is performed between genomic DNA and target-specific probes to form open loop complexes, where the target sequence is flanked by the ends of each oligo. Then, polymerase and ligase enzymes are added to fill and seal the gap between the two oligonucleotide probe ends, forming a covalently-closed circular molecule that contains the target sequence. Finally, an exonuclease mix is added to degrade any non-circular DNA (un-reacted probe, genomic DNA). What remains is circular DNA containing the set of targets captured by the reaction. Further details are provided for example in the following US patents: 5,866,337; 7,790,388; 6,858,412; 7,993,880; 7,700,323; 6,558,928; 6,235,472; 7,320,860; 7,351,528; 7,074,564; 5,871,921; 7,510,829; 7,862,999; and 7,883,849, the content of each of which is incorporated by reference herein in its entirety.

Barcode Sequences

In certain embodiments, at least one barcode sequence is attached to or incorporated into a nucleic acid template prior to sequencing. Strategies for barcoding nucleic acid templates are described for example in Porreca et al. (U.S. patent application serial number 13/081,660) and Umbarger et al. (U.S. patent application serial number 13/081,660), the content of each of which is incorporated by reference herein in its entirety. In embodiments that use more than one barcode, the barcode sequences may be attached to the template such that a first barcode sequence is attached to a 5' end of the template and a second barcode sequence is attached to a 3' end of the template. The first and second barcode sequences may be the same, or they may be different. Barcode sequence may be incorporated into a contiguous region of a template that includes the target to be sequenced.

Exemplary methods for designing sets of barcode sequences and other methods for attaching barcode sequences are shown in U.S. paten numbers 6,138,077; 6,352,828; 5,636,400;

6,172,214; 6235,475; 7,393,665; 7,544,473; 5,846,719; 5,695,934; 5,604,097; 6,150,516; RE39,793; 7,537,897; 6172,218; and 5,863,722, the content of each of which is incorporated by reference herein in its entirety.

The barcode sequence generally includes certain features that make the sequence useful in sequencing reactions. For example the barcode sequences can be designed to have minimal or no homopolymer 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 can also be designed so that they do not overlap the target region to be sequence or contain a sequence that is identical to the target.

The first and second barcode sequences are designed such that each pair of sequences is correlated to a particular sample, allowing samples to be distinguished and validated. Methods of designing sets of barcode sequences is shown for example in Brenner et al. (U.S. patent number 6,235,475), the contents of which are incorporated by reference herein in their entirety. In certain embodiments, the barcode sequences range from about 2 nucleotides to about 50; and preferably from about 4 to about 20 nucleotides. Since the barcode sequence is sequenced along with the template nucleic acid or may be sequenced in a separate read, the oligonucleotide length should be of minimal length so as to permit the longest read from the template nucleic acid attached. Generally, the barcode sequences are spaced from the template nucleic acid molecule by at least one base.

Methods of the invention involve attaching the barcode sequences to the template nucleic acids. Template nucleic acids are able to be fragmented or sheared to desired length, e.g. generally from 100 to 500 bases or longer, using a variety of mechanical, chemical and/or enzymatic methods. DNA may be randomly sheared via sonication, exposed to a DNase or one or more restriction enzymes, 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 before or after fragmentation.

Barcode sequence is integrated with template using methods known in the art. Barcode sequence is integrated with template using, for example, a ligase, a polymerase, Topo cloning (e.g., Invitrogen's topoisomerase vector cloning system using a topoisomerase enzyme), or chemical ligation or conjugation. 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 (such ligases are available commercially, from New England

Biolabs). 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. Barcode sequence can be incorporated via a PCR reaction as part of the PCR primer.

The ligation may be blunt ended or via use of over hanging 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. 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 combination 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 over hanging ends are used.

Sequencing

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, 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, and SOLiD sequencing. 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.

A sequencing technique that can be used in the methods of the provided invention includes, for example, Helicos True Single Molecule Sequencing (tSMS) (Harris T. D. et al. (2008) Science 320:106-109). In the tSMS technique, a DNA sample is cleaved into strands of approximately 100 to 200 nucleotides, and a polyA sequence is added to the 3' end of each DNA strand. Each strand is labeled by the addition of a fluorescently labeled adenosine nucleotide.

The DNA strands are then hybridized to a flow cell, which contains millions of oligo-T capture sites that are immobilized to the flow cell surface. The templates can be at a density of about 100 million templates/cm^2. The flow cell is then loaded into an instrument, e.g., HeliScope.TM. sequencer, and a laser illuminates the surface of the flow cell, revealing the position of each template. A CCD camera can map the position of the templates on the flow cell surface. The template fluorescent label is then cleaved and washed away. The sequencing reaction begins by introducing a DNA polymerase and a fluorescently labeled nucleotide. The oligo-T nucleic acid serves as a primer. The polymerase incorporates the labeled nucleotides to the primer in a template directed manner. The polymerase and unincorporated nucleotides are removed. The templates that have directed incorporation of the fluorescently labeled nucleotide are detected by imaging the flow cell surface. After imaging, a cleavage step removes the fluorescent label, and the process is repeated with other fluorescently labeled nucleotides until the desired read length is achieved. Sequence information is collected with each nucleotide addition step. Further description of tSMS is shown for example in Lapidus et al. (U.S. patent number 7,169,560), Lapidus et al. (U.S. patent application number 2009/0191565), Quake et al. (U.S. patent number 6,818,395), Harris (U.S. patent number 7,282,337), Quake et al. (U.S. patent application number 2002/0164629), and Braslavsky, et al., PNAS (USA), 100: 3960-3964 (2003), the contents of each of these references is incorporated by reference herein in its entirety.

Another example of a DNA sequencing technique that can be used in the methods of the provided invention is 454 sequencing (Roche) (Margulies, M et al. 2005, Nature, 437, 376-380). 454 sequencing involves two steps. 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.

Another example of a DNA sequencing technique that can be used in the methods of the provided invention is SOLiD technology (Applied Biosystems). 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 cleaved and removed and the process is then repeated.

Another example of a DNA sequencing technique that can be used in the methods of the provided invention is Ion Torrent sequencing (U.S. patent application numbers 2009/0026082, 2009/0127589, 2010/0035252, 2010/0137143, 2010/0188073, 2010/0197507, 2010/0282617, 2010/0300559), 2010/0300895, 2010/0301398, and 2010/0304982), the content of each of which is incorporated by reference herein in its entirety. In Ion Torrent sequencing, 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 a surface and is attached at a resolution such that the fragments are individually resolvable. Addition of one or more nucleotides releases a proton (H^+), which signal detected and recorded in a sequencing instrument. The signal strength is proportional to the number of nucleotides incorporated.

Another example of a sequencing technology that can be used in the methods of the provided invention 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.

Another example of a sequencing technology that can be used in the methods of the provided invention includes the single molecule, real-time (SMRT) technology of Pacific Biosciences. 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). A ZMW is a confinement structure which enables observation of incorporation of a single nucleotide by DNA polymerase against the background of fluorescent nucleotides that rapidly diffuse in an out of the ZMW (in microseconds). 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.

Another example of a sequencing technique that can be used in the methods of the provided invention is nanopore sequencing (Soni G V and Meller A. (2007) Clin Chem 53: 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.

Another example of a sequencing technique that can be used in the methods of the provided invention involves using a chemical-sensitive field effect transistor (chemFET) array to sequence DNA (for example, as described in US Patent Application Publication No. 20090026082). 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.

Another example of a sequencing technique that can be used in the methods of the provided invention involves using a electron microscope (Moudrianakis E. N. and Beer M. Proc Natl Acad Sci USA. 1965 March; 53:564-71). 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.

Analysis

Alignment and/or compilation of sequence results obtained from the image stacks produced as generally described above utilizes look-up tables that take into account possible sequences changes (due, e.g., to errors, mutations, etc.). Essentially, sequencing results obtained as described herein are compared to a look-up type table that contains all possible reference sequences plus 1 or 2 base errors. Sequence alignment algorithms and methods are described for example in U.S. patent number 8,209,130, the content of which is incorporated by reference herein in its entirety.

In some embodiments, de novo assembly proceeds according to so-called greedy algorithms. For assembly according to greedy algorithms, one of the reads of a group of reads is selected, and it is paired with another read with which it exhibits a substantial amount of overlap--generally it is paired with the read with which it exhibits the most overlap of all of the other

reads. Those two reads are merged to form a new read sequence, which is then put back in the group of reads and the process is repeated. Assembly according to a greedy algorithm is described, for example, in Schatz, et al., Genome Res., 20:1165-1173 (2010) and U.S. Pub. 2011/0257889, each of which is hereby incorporated by reference in its entirety.

In other embodiments, assembly proceeds by pairwise alignment, for example, exhaustive or heuristic (e.g., not exhaustive) pairwise alignment. Exhaustive pairwise alignment, sometimes called a "brute force" approach, calculates an alignment score for every possible alignment between every possible pair of sequences among a set. Assembly by heuristic multiple sequence alignment ignores certain mathematically unlikely combinations and can be computationally faster. One heuristic method of assembly by multiple sequence alignment is the so-called "divide-and-conquer" heuristic, which is described, for example, in U.S. Pub. 2003/0224384. Another heuristic method of assembly by multiple sequence alignment is progressive alignment, as implemented by the program ClustalW (see, e.g., Thompson, et al., Nucl. Acids. Res., 22:4673-80 (1994)). Assembly by multiple sequence alignment in general is discussed in Lecompte, O., et al., Gene 270:17-30 (2001); Mullan, L. J., Brief Bioinform., 3:303-5 (2002); Nicholas, H. B. Jr., et al., Biotechniques 32:572-91(2002); and Xiong, G., Essential Bioinformatics, 2006, Cambridge University Press, New York, N.Y.

An alignment according to the invention can be performed using any suitable computer program known in the art.

One exemplary alignment program, which implements a BWT approach, is Burrows-Wheeler Aligner (BWA) available from the SourceForge web site maintained by Geeknet (Fairfax, Va.). BWA can align reads, contigs, or consensus sequences to a reference. BWT occupies 2 bits of memory per nucleotide, making it possible to index nucleotide sequences as long as 4G base pairs with a typical desktop or laptop computer. The pre-processing includes the construction of BWT (i.e., indexing the reference) and the supporting auxiliary data structures.

BWA implements two different algorithms, both based on BWT. Alignment by BWA can proceed using the algorithm bwa-short, designed for short queries up to .about.200 bp with low error rate (<3%) (Li H. and Durbin R. Bioinformatics, 25:1754-60 (2009)). The second algorithm, BWA-SW, is designed for long reads with more errors (Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. Bioinformatics, Epub.). The BWA-SW component performs heuristic Smith-Waterman-like alignment to find

high-scoring local hits. One skilled in the art will recognize that bwa-sw is sometimes referred to as "bwa-long", "bwa long algorithm", or similar. Such usage generally refers to BWA-SW.

An alignment program that implements a version of the Smith-Waterman algorithm is MUMmer, available from the SourceForge web site maintained by Geeknet (Fairfax, Va.). MUMmer is a system for rapidly aligning entire genomes, whether in complete or draft form (Kurtz, S., et al., Genome Biology, 5:R12 (2004); Delcher, A. L., et al., Nucl. Acids Res., 27:11 (1999)). For example, MUMmer 3.0 can find all 20-basepair or longer exact matches between a pair of 5-megabase genomes in 13.7 seconds, using 78 MB of memory, on a 2.4 GHz Linux desktop computer. MUMmer can also align incomplete genomes; it can easily handle the 100s or 1000s of contigs from a shotgun sequencing project, and will align them to another set of contigs or a genome using the NUCmer program included with the system. If the species are too divergent for a DNA sequence alignment to detect similarity, then the PROmer program can generate alignments based upon the six-frame translations of both input sequences.

Another exemplary alignment program according to embodiments of the invention is BLAT from Kent Informatics (Santa Cruz, Calif.) (Kent, W. J., Genome Research 4: 656-664 (2002)). BLAT (which is not BLAST) keeps an index of the reference genome in memory such as RAM. The index includes of all non-overlapping k-mers (except optionally for those heavily involved in repeats), where k=11 by default. The genome itself is not kept in memory. The index is used to find areas of probable homology, which are then loaded into memory for a detailed alignment.

Another alignment program is SOAP2, from Beijing Genomics Institute (Beijing, CN) or BGI Americas Corporation (Cambridge, Mass.). SOAP2 implements a 2-way BWT (Li et al., Bioinformatics 25(15):1966-67 (2009); Li, et al., Bioinformatics 24(5):713-14 (2008)).

Another program for aligning sequences is Bowtie (Langmead, et al., Genome Biology, 10:R25 (2009)). Bowtie indexes reference genomes by making a BWT.

Other exemplary alignment programs include: Efficient Large-Scale Alignment of Nucleotide Databases (ELAND) or the ELANDv2 component of the Consensus Assessment of Sequence and Variation (CASAVA) software (Illumina, San Diego, Calif.); RTG Investigator from Real Time Genomics, Inc. (San Francisco, Calif.); Novoalign from Novocraft (Selangor, Malaysia); Exonerate, European Bioinformatics Institute (Hinxton, UK) (Slater, G., and Birney, E., BMC Bioinformatics 6:31(2005)), Clustal Omega, from University College Dublin (Dublin,

Ireland) (Sievers F., et al., Mol Syst Biol 7, article 539 (2011)); ClustalW or ClustalX from University College Dublin (Dublin, Ireland) (Larkin M. A., et al., Bioinformatics, 23, 2947-2948 (2007)); and FASTA, European Bioinformatics Institute (Hinxton, UK) (Pearson W. R., et al., PNAS 85(8):2444-8 (1988); Lipman, D. J., Science 227(4693):1435-41 (1985)).

Once the mutations in the nucleic acid sequence from the sample are determined, those mutations are compared to a database(s) of known mutations associated with the particular disease. Such databases are publically available and known to those of skill in the art. Mutations that do not match to the database are identified as novel mutations.

Novel insertions and deletion variants present a particular challenge for high-throughput sequencing technologies. Aligned reads with coordinate-altering variants require the use of penalized gaps in either the query or reference sequence to maintain global coordinate order. Extended gaps tend to reduce overall mappability leading to false negative insertions and deletions. Gaps are often inserted at the ends of reads to artificially maintain optimality leading to false positive insertion, deletion, and substitution variants. Realignment improves sensitivity (of insertions/deletions) and specificity (of substitutions); however, these techniques often use Smith-Waterman alignment algorithms without gaps. Without penalizing gaps FP insertions and deletions often result.

An additional complication results from the sequence context where the majority of insertions and deletion variants are found. Small insertions and deletions (less than 100 bp) commonly occur within tandem repeats where polymerase slippage or intra-chromosomal recombination leads to nucleotide expansion or contraction. Relative to the original (or reference) genome, the consequence of these processes appear as insertions or deletions, respectively. Insertions and deletions within tandem repeats are spatially ambiguous, that is, they may not be faithfully represented using a single genomic coordinate (Figure 1). It is necessary to calculate the variant's equivalent insertion/deletion region (EIR) which is essentially the contiguous block of DNA representing its associated tandem repeat. It is important to note that alignment algorithms arbitrarily assign variant positions within EIRs.

Due to the biological mechanisms mentioned above, naturally occurring insertion and deletion mutations tend to occur as tandem repeats (i.e., within EIRs) much more often than would be expected by chance. This fact can be exploited to distinguish true variants from false positions. For example, within capture regions of capture probes, 13 (21%) and 53 (100%) of

dbSNP insertion and deletion variants, respectively, have EIRs within lengths greater than one. Thus, known insertions and deletions are strongly associated with tandem repeats. Appropriate probability-based scores can be used to measure the mutual dependence between these two variables and reduce uncertainty about whether a caller variant represents a true position or a false positive. For example:

$$p(\text{deletion} \mid \text{repeat}) = \frac{p(\text{repeat} \mid \text{deletion})p(\text{deletion})}{p(\text{repeat})}$$

where $p(\text{repeat} \mid \text{deletion})$ is the likelihood of a repeat given a deletion (in the example above, this value equals 1.0), $p(\text{deletion})$ is the prior probability of a deletion in the absence of additional evidence, and $p(\text{repeat})$ is a normalization factor that accounts for local variability in sequence repetitiveness (the latter two values depend on the specific genomic regions under consideration). It is likely that probabilities would be calculated separately for different sized variants. In combination with other pieces of evidence, such as genotype qualities, a sample lookup table would provide additional confidence in any particular variant call given its presence in a repetitive region.

Once a particular insertion/deletion variant is determined to be real, the EIR required further to determine its precise functional or clinical significance. This is illustrated with reference to Figure 2. Consider a scenario of a three base pair homopolymeric repeat (GGG), that partially overlaps the exon boundary and its associate splice site (chr7:116975929-116975930). Depending on its size, a deletion of one or more nucleotides from within this repeat may be reported by detection algorithms at any of three equivalent positions (chr7:116975929-11697931) within the EIR chr7:116975929-chr7:116975932; however, in this particular case, the functional annotation depends on the exact position of the variant. Translating genomic positions directly into their functional analogues would lead to a splice site annotation from chr7:116975929delG whereas the equivalent chr7:116975931delG is frame shift.

Consistent annotation requires implementing rules (or performing simulations) that consider insertion and deletion variants in both genomic and functional contexts. Taken together, the process of applying EIR-assisted confidence scores and functional annotations can be reduced to the following steps:

1. Determine if the variant is known to be disease causing by consulting a relevant database(s);
2. If the variant is not known to be disease causing then by definition it is novel. If the variant is a substitution, determine its clinical impact directly from its genomic coordinate. Otherwise calculate the equivalent insertion/deletion region (EIR) using methods described in Krawitz et al., 2010;
3. If the variant EIR length is equal to one, use this information to assess the likelihood that the variant is a false positive (e.g., the result of a sequence artifact). If it is determined that the variant is real, continue to the next step, otherwise stop.
4. Annotate the variant EIR with all proportional functional information.
5. Attempt to push the variant completely out of the functional region by retrieving the extreme lower or upper position of the variant EIR. Choosing the correct extreme position depends on the orientation of the variant relative to its associated functional region or regions.
6. If the variant can be pushed completely out of the functional region, don't report or report as being unknown or benign, otherwise determine the variant's clinical significance.

Computers and Software

Other embodiments are within the scope and spirit of the invention. For example, due to the nature of software, functions described above can be implemented using software, hardware, firmware, hardwiring, or combinations of any 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.

As one skilled in the art would recognize as necessary or best-suited for performance of the methods of the invention and sequence assembly in general, computer system 200 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.

In an exemplary embodiment shown in FIG. 3, system 200 can include a sequencer 201 with data acquisition module 205 to obtain sequence read data. Sequencer 201 may optionally include or be operably coupled to its own, e.g., dedicated, sequencer computer 233 (including an input/output mechanism 237, one or more of processor 241 and memory 245). Additionally or alternatively, sequencer 201 may be operably coupled to a server 213 or computer 249 (e.g., laptop, desktop, or tablet) via network 209. Computer 249 includes one or more processor 259 and memory 263 as well as an input/output mechanism 254. Where methods of the invention employ a client/server architecture, an steps of methods of the invention may be performed using server 213, which includes one or more of processor 221 and memory 229, capable of obtaining data, instructions, etc., or providing results via interface module 225 or providing results as a file 217. Server 213 may be engaged over network 209 through computer 249 or terminal 267, or server 213 may be directly connected to terminal 267, including one or more processor 275 and memory 279, as well as input/output mechanism 271.

System 200 or machines according to the invention may further include, for any of I/O 249, 237, or 271 a video display unit (e.g., a liquid crystal display (LCD) or a cathode ray tube (CRT)). Computer systems or machines according to the invention can also include 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), a touchscreen, an accelerometer, a microphone, a cellular radio frequency antenna, and a network interface device, which can be, for example, a network interface card (NIC), Wi-Fi card, or cellular modem.

Memory 263, 245, 279, or 229 according to the invention can include a 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 and/or within the processor during execution thereof by the computer system, 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.

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. The term "machine-readable medium" shall accordingly be taken to include, but not be limited to, solid-state memories (e.g., subscriber identity module (SIM) card, secure digital card (SD card), micro SD card, or solid-state drive (SSD)), optical and magnetic media, and any other tangible storage media.

Incorporation by Reference

References and citations to other documents, such as patents, patent applications, patent publications, journals, books, papers, web contents, have been made throughout this disclosure. All such documents are hereby incorporated herein by reference in their entirety for all purposes.

Equivalents

The invention may be embodied in other specific forms without departing from the spirit or essential characteristics thereof. The foregoing embodiments are therefore to be considered in all respects illustrative rather than limiting on the invention described herein.

What is claimed is:

1. A method for identifying a novel mutation associated with a disease, the method comprising:

obtaining nucleic acid from a subject having a disease;

identifying at least one mutation in the nucleic acid; and

comparing the mutation to a database of mutations known to be associated with the disease, wherein mutations that do not match to the database are identified as novel mutations.

2. The method according to claim 1, wherein identifying comprising:

sequencing the nucleic acid; and

comparing the sequence of the nucleic acid from the sample to a reference sequence.

3. The method according to claim 2, wherein sequencing is sequencing-by-synthesis.

4. The method according to claim 3, wherein sequencing-by-synthesis is single molecule sequencing-by-synthesis.

5. The method according to claim 2, wherein the reference sequence is a consensus human sequence or a sequence from a non-diseased sample.

6. The method according to claim 1, wherein prior to identifying, the method further comprises attaching a barcode sequence to the nucleic acid.

7. The method according to claim 1, wherein the disease is cystic fibrosis.

8. The method according to claim 7, wherein the subject is the subject is Hispanic.

9. The method according to claim 1, further comprising determining of the novel mutation is causative of the disease.

10. The method according to claim 9, wherein determining comprises:

annotating the mutation with appropriate functional information;

identifying a lower and upper boundary for a position of the mutation; and

attempting to push the mutation completely out of a functional region, wherein mutations that can be pushed completely out of the functional region are not causative of the disease.

11. A method for identifying a novel mutation associated with a disease, the method comprising:

obtaining nucleic acid from a subject having a disease;

sequencing the nucleic acid;

comparing the sequence of the nucleic acid from the sample to a reference sequence, thereby identifying mutations in the nucleic acid; and

comparing the mutation to a database of mutations known to be associated with the disease, wherein mutations that do not match to the database are identified as novel mutations.

12. The method according to claim 11, wherein sequencing is sequencing-by-synthesis.

13. The method according to claim 12, wherein sequencing-by-synthesis is single molecule sequencing-by-synthesis.

14. The method according to claim 11, wherein the reference sequence is a consensus human sequence or a sequence from a non-diseased sample.

15. The method according to claim 11, wherein prior to sequencing, the method further comprises attaching a barcode sequence to the nucleic acid.

16. The method according to claim 11, wherein the disease is cystic fibrosis.

17. The method according to claim 11, wherein the subject is the subject is Hispanic.

18. The method according to claim 11, further comprising determining of the novel mutation is causative of the disease.

19. The method according to claim 18, wherein determining comprises:

annotating the mutation with appropriate functional information;

identifying a lower and upper boundary for a position of the mutation; and

attempting to push the mutation completely out of a functional region, wherein mutations that can be pushed completely out of the functional region are not causative of the disease.

20. A method for determining if a mutation is causative of a disease, the method comprising:

conducting an assay to obtain a nucleic acid sequence from a subject having a disease;

identifying at least one novel mutation in the sequence;

annotating the mutation with appropriate functional information;

identifying a lower and upper boundary for a position of the mutation; and

attempting to push the mutation completely out of a functional region, wherein mutations that can be pushed completely out of the functional region are not causative of the disease.

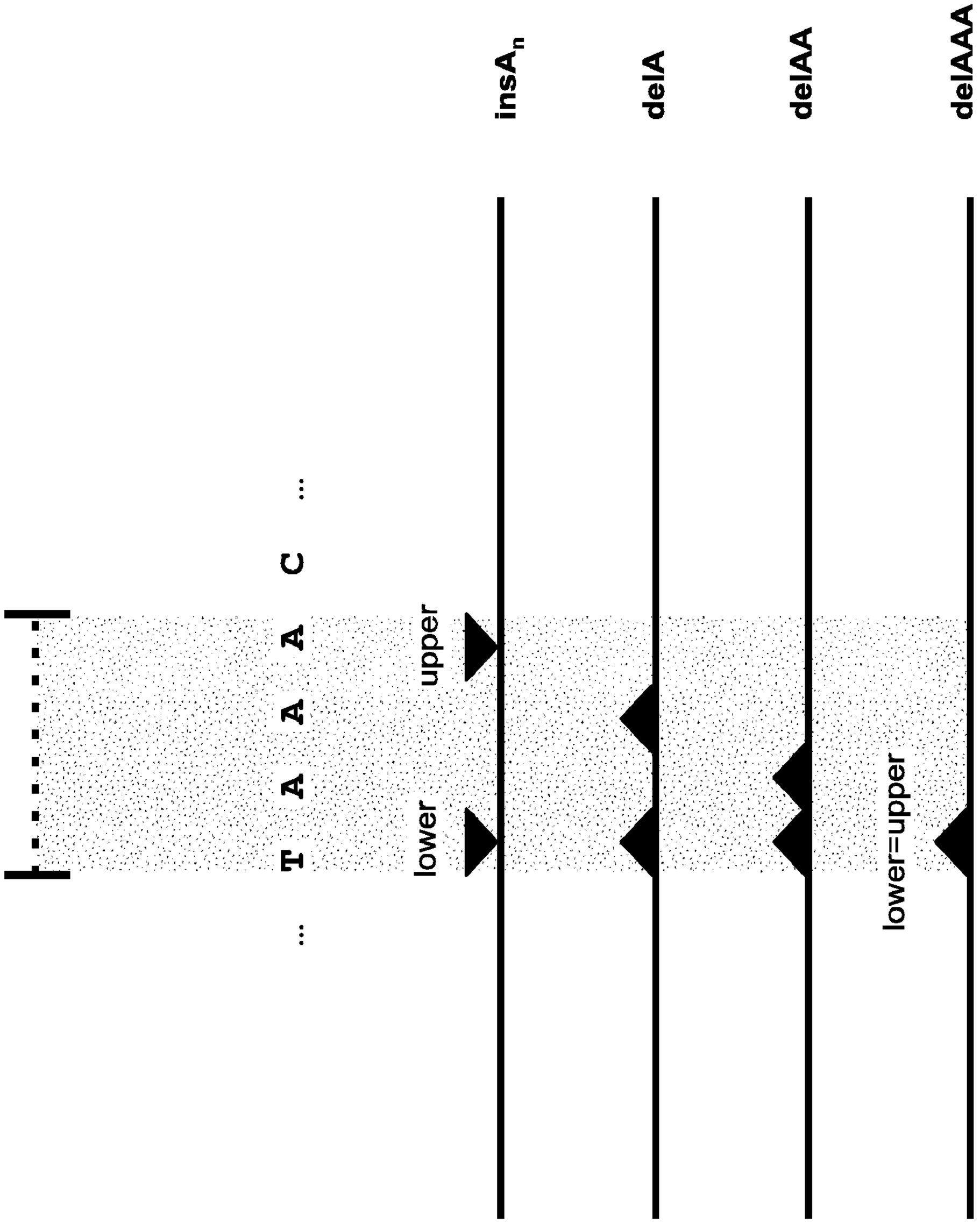
Lesion
insA$_n$
delA
delAA
delAAA
C ...
EIR
T A A A ...
lower
upper
lower=upper

FIG. 1

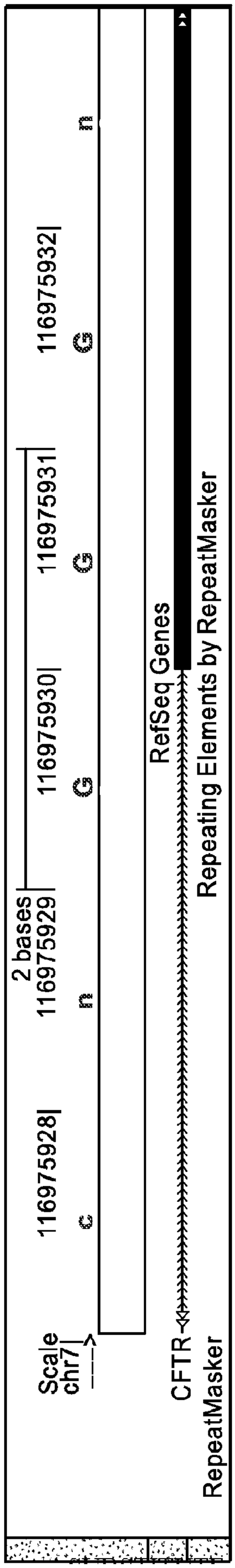
116975932
116975931
116975930
116975928
2 bases
116975929
n
G
G
c
RefSeq Genes
Repeating Elements by RepeatMasker
Scale
chr7
CFTR
RepeatMasker

FIG. 2

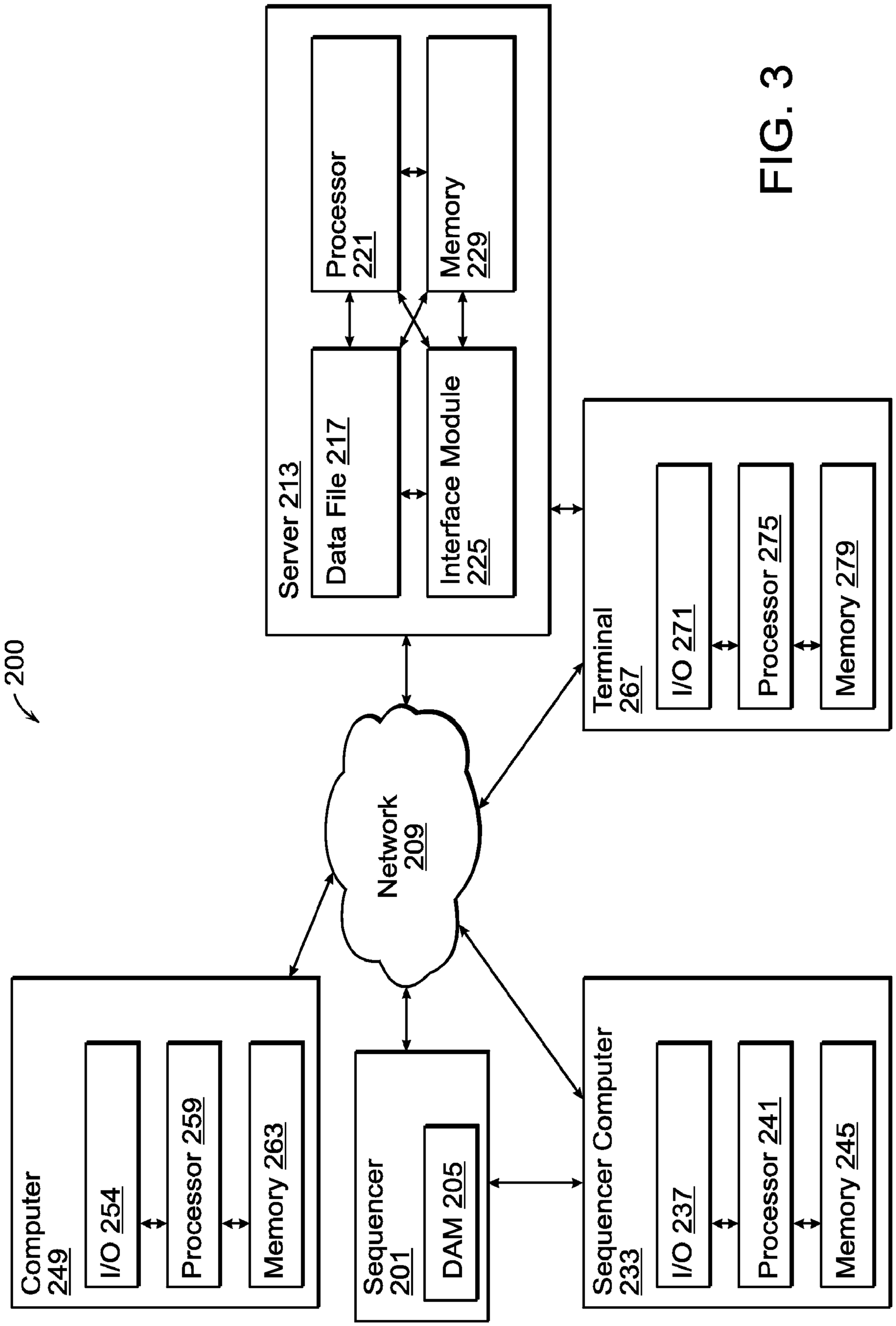
200
Server 213
Data File 217
Processor 221
Interface Module 225
Memory 229
Network 209
Terminal 267
I/O 271
Processor 275
Memory 279
Computer 249
I/O 254
Processor 259
Memory 263
Sequencer 201
DAM 205
Sequencer Computer 233
I/O 237
Processor 241
Memory 245

FIG. 3