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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
21 January 2016 (21.01.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/011378 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01) *G06F 19/22* (2011.01)

(21) International Application Number:
PCT/US2015/040951

(22) International Filing Date:
17 July 2015 (17.07.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/026,300 18 July 2014 (18.07.2014) US

(71) Applicant: **LIFE TECHNOLOGIES CORPORATION** [US/US]; 5791 Van Allen Way, Carlsbad, CA 92008 (US).

(72) Inventors: **GOTTIMUKKALA, Rajesh, Kumar**; 5791 Van Allen Way, Carlsbad, CA 92008 (US). **HYLAND, Fiona, C., Laird**; 5791 Van Allen Way, Carlsbad, CA 92008 (US).

(74) Agents: **D'AVIGNON-AUBUT, Christian** et al.; Life Technologies Corporation, c/o Legal Department, IP Docketing, 5791 Van Allen Way, Carlsbad, CA 92008 (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, IR, IS, JP, KE, KG, KN, KP, KR, 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))*

— *of inventorship (Rule 4.17(iv))*

Published:

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

(54) **Title**: SYSTEMS AND METHODS FOR DETECTING STRUCTURAL VARIANTS

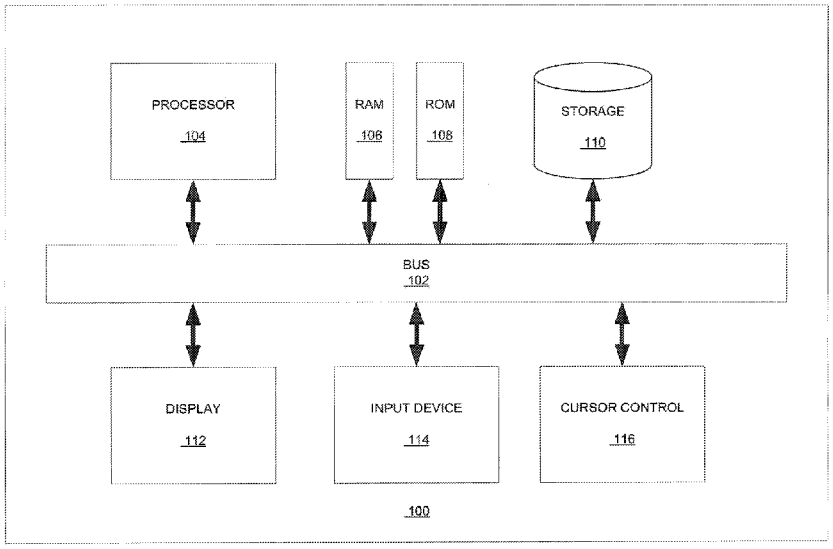

FIG. 1

(57) **Abstract**: Systems and method for identifying gene fusions can obtain sequencing information for a plurality of amplicons from a nucleic acid sample. The sequencing information can include a plurality of reads that are initially partially mapped to a reference sequence. Fragments may be generated by splitting the partially mapped reads into mapped and unmapped fragments, and the fragments may be remapped to the reference sequence. Gene fusions can be identified based on reads where the first fragment maps to a first gene and the second fragment maps to a second gene.

WO 2016/011378 A1

SYSTEMS AND METHODS FOR DETECTING STRUCTURAL VARIANTS

[0001] This application claims priority to U.S. Provisional Application No. 62/026,300, filed July 18, 2014, which is incorporated by reference herein in its entirety.

TECHNICAL FIELD

[0002] The present disclosure generally relates to the field of nucleic acid sequencing including systems and methods for detecting gene fusions.

INTRODUCTION

[0003] Upon completion of the Human Genome Project, one focus of the sequencing industry has shifted to finding higher throughput and/or lower cost nucleic acid sequencing technologies, sometimes referred to as "next generation" sequencing (NGS) technologies. In making sequencing higher throughput and/or less expensive, one goal is to make the technology more accessible. This can be reached through the use of sequencing platforms and methods that provide sample preparation for samples of significant complexity, sequencing larger numbers of samples in parallel (for example through use of barcodes and multiplex analysis), and/or processing high volumes of information efficiently and completing the analysis in a timely manner. Various methods, such as, for example, sequencing by synthesis, sequencing by hybridization, and sequencing by ligation are evolving to meet these challenges.

[0004] Ultra-high throughput nucleic acid sequencing systems incorporating NGS technologies typically produce a large number of short sequence reads. Sequence processing methods should desirably assemble and/or map a large number of reads quickly and efficiently, such as to minimize use of computational resources. For example, data arising from sequencing of a mammalian genome can result in tens or hundreds of millions of reads that typically need to be assembled before they can be further analyzed to determine their biological, diagnostic and/or therapeutic relevance.

[0005] Exemplary applications of NGS technologies include, but are not limited to: genomic variant detection, such as insertions/deletions, copy number variations, single

nucleotide polymorphisms, genomic resequencing, gene expression analysis, genomic profiling, and the like.

[0006] Structural variants, such as large scale deletions, insertions, inversions, genomic rearrangements, gene fusions, and the like, can be associated with various genetic disorders and cancers. Structure variants can often lead to significant disruptions in the production of proteins essential for the proper function of a cell. For example, genomic rearrangements and gene fusions can lead to mRNA coding for chimeric proteins, having a first part from one protein and a second part from another protein. Often, these chimeric proteins no longer function like either the first or second protein and can lead to disruption of regularity pathways. In cancer cells, the disrupted regulatory pathways may be involved in the regulation of apoptosis, cell growth, or the like and, as a result of the gene fusion, enable the cancer cells to grow unchecked.

[0007] From the foregoing it will be appreciated that a need exists for systems and methods that can detect gene fusions using nucleic acid sequencing data.

DRAWINGS

[0008] For a more complete understanding of the principles disclosed herein, and the advantages thereof, reference is now made to the following descriptions taken in conjunction with the accompanying drawings, in which:

[0009] Figure 1 is a block diagram that illustrates an exemplary computer system, in accordance with various embodiments.

[0010] Figure 2 is a schematic diagram of an exemplary system for reconstructing a nucleic acid sequence, in accordance with various embodiments.

[0011] Figure 3 is a schematic diagram of an exemplary genetic analysis system, in accordance with various embodiments.

[0012] Figure 4 is a diagram illustrating an exemplary gene fusion, in accordance with various embodiments.

[0013] Figure 5 is a flow diagram illustrating an exemplary method of detecting gene fusions, in accordance with various embodiments.

[0014] Figure 6 is a diagram illustrating an exemplary synthetic control, in accordance with various embodiments.

[0015] It is to be understood that the figures are not necessarily drawn to scale, nor are the objects in the figures necessarily drawn to scale in relationship to one another. The figures are depictions that are intended to bring clarity and understanding to various embodiments of apparatuses, systems, and methods disclosed herein. Wherever possible, the same reference numbers will be used throughout the drawings to refer to the same or like parts. Moreover, it should be appreciated that the drawings are not intended to limit the scope of the present teachings in any way.

DESCRIPTION OF VARIOUS EMBODIMENTS

[0016] Embodiments of systems and methods for detecting gene fusions are described and illustrated herein.

[0017] The section headings used herein are for organizational purposes only and are not to be construed as limiting the described subject matter in any way.

[0018] In this detailed description of the various embodiments, for purposes of explanation, numerous specific details are set forth to provide a thorough understanding of the embodiments disclosed. One skilled in the art will appreciate, however, that these various embodiments may be practiced with or without these specific details. In other instances, structures and devices are shown in block diagram form. Furthermore, one skilled in the art can readily appreciate that the specific sequences in which methods are presented and performed are illustrative and it is contemplated that the sequences can be varied and still remain within the scope of the various embodiments disclosed herein.

[0019] All literature and similar materials cited in this application, including but not limited to, patents, patent applications, articles, books, treatises, and internet web pages are expressly incorporated by reference in their entirety for any purpose. Unless described otherwise, all technical and scientific terms used herein have a meaning as is

commonly understood by one of ordinary skill in the art to which the various embodiments described herein belongs.

[0020] It will be appreciated that there is an implied "about" prior to the temperatures, concentrations, times, number of bases, coverage, etc. discussed in the present teachings, such that slight and insubstantial deviations are within the scope of the present teachings. In this application, the use of the singular includes the plural unless specifically stated otherwise. Also, the use of "comprise", "comprises", "comprising", "contain", "contains", "containing", "include", "includes", and "including" are not intended to be limiting. It is to be understood that both the foregoing general description and the following detailed description are exemplary and explanatory only and are not restrictive of the present teachings.

[0021] As used herein, "a" or "an" also may refer to "at least one" or "one or more." Also, the use of "or" is inclusive, such that the phrase "A or B" is true when "A" is true, "B" is true, or both "A" and "B" are true.

[0022] Further, unless otherwise required by context, singular terms shall include pluralities and plural terms shall include the singular. Generally, nomenclatures utilized in connection with, and techniques of, cell and tissue culture, molecular biology, and protein and oligo- or polynucleotide chemistry and hybridization described herein are those well known and commonly used in the art. Standard techniques are used, for example, for nucleic acid purification and preparation, chemical analysis, recombinant nucleic acid, and oligonucleotide synthesis. Enzymatic reactions and purification techniques are performed according to manufacturer's specifications or as commonly accomplished in the art or as described herein. The techniques and procedures described herein are generally performed according to conventional methods well known in the art and as described in various general and more specific references that are cited and discussed throughout the instant specification. *See, e.g.,* Sambrook *et al., Molecular Cloning: A Laboratory Manual* (Third ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. 2000). The nomenclatures utilized in connection with, and the

laboratory procedures and techniques described herein are those well known and commonly used in the art.

[0023] In various embodiments, a "system" sets forth a set of components, real or abstract, comprising a whole where each component interacts with or is related to at least one other component within the whole.

[0024] In various embodiments, a "biomolecule" may refer to any molecule that is produced by a biological organism, including large polymeric molecules such as proteins, polysaccharides, lipids, and nucleic acids (DNA and RNA) as well as small molecules such as primary metabolites, secondary metabolites, and other natural products.

[0025] In various embodiments, the phrase "next generation sequencing" or NGS refers to sequencing technologies having increased throughput as compared to traditional Sanger- and capillary electrophoresis-based approaches, for example with the ability to generate hundreds of thousands of relatively small sequence reads at a time. Some examples of next generation sequencing techniques include, but are not limited to, sequencing by synthesis, sequencing by ligation, and sequencing by hybridization. More specifically, the Personal Genome Machine (PGM) and Proton of Life Technologies Corp. provides massively parallel sequencing with enhanced accuracy. The PGM and Proton Systems and associated workflows, protocols, chemistries, etc. are described in more detail in U.S. Patent Application Publication No. 2009/0127589 and No. 2009/0026082, the entirety of each of these applications being incorporated herein by reference.

[0026] In various embodiments, the phrase "sequencing run" refers to any step or portion of a sequencing experiment performed to determine some information relating to at least one biomolecule (e.g., nucleic acid molecule).

[0027] In various embodiments, the phase "base space" refers to a representation of a sequence of nucleotides. The phase "flow space" refers to a representation of an incorporation event or non-incorporation event for a particular nucleotide flow. For example, flow space can be a series of values representing a nucleotide incorporation

events (such as a one, "1") or a non-incorporation event (such as a zero, "0") for that particular nucleotide flow. Nucleotide flows having a non-incorporation event can be referred to as empty flows. Nucleotide flows having a nucleotide incorporation event can be referred to as positive flows. It should be understood that zeros and ones are convenient representations of a non-incorporation event and a nucleotide incorporation event; however, any other symbol or designation could be used alternatively to represent and/or identify incorporation and non-incorporation events. In particular, when multiple nucleotides are incorporated at a given position, such as for a homopolymer stretch, the value can be proportional to the number of nucleotide incorporation events and thus the length of the homopolymer stretch (e.g., greater than one).

[0028] In various embodiments, DNA (deoxyribonucleic acid) may be referred to as a chain of nucleotides consisting of 4 types of nucleotides; A (adenine), T (thymine), C (cytosine), and G (guanine), and that RNA (ribonucleic acid) is comprised of 4 types of nucleotides; A, U (uracil), G, and C. Certain pairs of nucleotides specifically bind to one another in a complementary fashion (called complementary base pairing). That is, adenine (A) pairs with thymine (T) (in the case of RNA, however, adenine (A) pairs with uracil (U)), and cytosine (C) pairs with guanine (G). When a first nucleic acid strand binds to a second nucleic acid strand made up of nucleotides that are complementary to those in the first strand, the two strands bind to form a double strand. In various embodiments, "nucleic acid sequencing data," "nucleic acid sequencing information," "nucleic acid sequence," "genomic sequence," "genetic sequence," or "fragment sequence," or "nucleic acid sequencing read" denotes any information or data that is indicative of the order of the nucleotide bases (e.g., adenine, guanine, cytosine, and thymine/uracil) in a molecule (e.g., whole genome, whole transcriptome, exome, oligonucleotide, polynucleotide, fragment, etc.) of DNA or RNA. It should be understood that the present teachings contemplate sequence information obtained using all available varieties of techniques, platforms or technologies, including, but not limited to: capillary electrophoresis, microarrays, ligation-based systems, polymerase-based systems (such as Illumina HiSeq, MiSeq, and Genome Analyzer), hybridization-based systems, direct or indirect nucleotide identification systems, pyrosequencing (such as 454 Life Science GS FLX and GS Junior), ion- or pH-based detection systems (such as Ion

Torrent), electronic signature-based systems (such as Oxford Nanopore GridION and MinION), etc.

[0029] In various embodiments, a “polynucleotide", "nucleic acid", or "oligonucleotide" refers to a linear polymer of nucleosides (including deoxyribonucleosides, ribonucleosides, or analogs thereof) joined by internucleosidic linkages. Typically, a polynucleotide comprises at least three nucleotides. Usually oligonucleotides range in size from a few monomeric units, e.g. 3-4, to several hundreds of monomeric units. Whenever a polynucleotide such as an oligonucleotide is represented by a sequence of letters, such as "ATGCCTG," it will be understood that the nucleotides are in 5'-3' order from left to right and that "A" denotes deoxyadenosine, "C" denotes deoxycytidine, "G" denotes deoxyguanosine, and "T" denotes thymidine, unless otherwise noted. The letters A, C, G, and T may be used to refer to the bases themselves, to nucleosides, or to nucleotides comprising the bases, as is standard in the art.

[0030] In various embodiments, a “structural variant” refers to a variation in the structure of a chromosome. Structural variants can include deletions, duplications, copy-number variants, insertions, gene fusions, inversions and translocations. Many of structural variants are associated with genetic diseases, however more are not.

MULTIPLEX AMPLIFICATION METHODS:

[0031] In various embodiments, target nucleic acids generated by the amplification of multiple target-specific sequences from a population of nucleic acid molecules can be sequenced. In some exemplary embodiments, the amplification can include hybridizing one or more target-specific primer pairs to the target sequence, extending a first primer of the primer pair, denaturing the extended first primer product from the population of nucleic acid molecules, hybridizing to the extended first primer product the second primer of the primer pair, extending the second primer to form a double stranded product, and digesting the target-specific primer pair away from the double stranded product to generate a plurality of amplified target sequences. In some embodiments, the amplified target sequences can be ligated to one or more adapters. In some embodiments, the adapters can include one or more nucleotide barcodes or tagging sequences. In some

embodiments, the amplified target sequences once ligated to an adapter can undergo a nick translation reaction and/or further amplification to generate a library of adapter-ligated amplified target sequences. Exemplary methods of multiplex amplification are described in U.S. Application No. 13/458,739 filed November 12, 2012 and titled "Methods and Compositions for Multiplex PCR", now published as US 2012/0295819 A1.

[0032] In various exemplary embodiments, a method of performing multiplex PCR amplification includes contacting a plurality of target-specific primer pairs having a forward and reverse primer, with a population of target sequences to form a plurality of template/primer duplexes; adding a DNA polymerase and a mixture of dNTPs to the plurality of template/primer duplexes for sufficient time and at sufficient temperature to extend either (or both) the forward or reverse primer in each target-specific primer pair via template-dependent synthesis thereby generating a plurality of extended primer product/template duplexes; denaturing the extended primer product/template duplexes; annealing to the extended primer product the complementary primer from the target-specific primer pair; and extending the annealed primer in the presence of a DNA polymerase and dNTPs to form a plurality of target-specific double-stranded nucleic acid molecules.

COMPUTER-IMPLEMENTED SYSTEM

[0033] Figure 1 is a block diagram that illustrates an exemplary computer system 100, upon which embodiments of the present disclosure may be implemented. In various embodiments, computer system 100 can include a bus 102 or other communication mechanism for communicating information, and a processor 104 coupled with bus 102 for processing information. In various embodiments, computer system 100 can also include a memory 106, which can be a random access memory (RAM) or other dynamic storage device, coupled to bus 102 for determining base calls, and instructions to be executed by processor 104. Memory 106 also can be used for storing temporary variables or other intermediate information during execution of instructions to be executed by processor 104. In various embodiments, computer system 100 can further include a read only memory (ROM) 108 or other static storage device coupled to bus 102

for storing static information and instructions for processor 104. A storage device 110, such as a magnetic disk or optical disk, can be provided and coupled to bus 102 for storing information and instructions.

[0034] In various embodiments, computer system 100 can be coupled via bus 102 to a display 112, such as a cathode ray tube (CRT), a liquid crystal display (LCD), or other display or mechanism, for displaying or otherwise providing information to a computer user. An input device 114, including alphanumeric and other keys, can be coupled to bus 102 for communicating information and command selections to processor 104. Another type of user input device is a cursor control 116, such as a mouse, a trackball, cursor direction keys, and the like. for communicating direction information and command selections to processor 104 and for controlling cursor movement on display 112. This input device typically has two degrees of freedom in two axes, a first axis (i.e., x) and a second axis (i.e., y), that allows the device to specify positions in a plane.

[0035] In various exemplary embodiments, a computer system 100 can perform at least portions of the methods in accordance with the present disclosure. Consistent with certain implementations of the present teachings, results can be provided by computer system 100 in response to processor 104 executing one or more sequences of one or more instructions contained in memory 106. Such instructions can be read into memory 106 from another computer-readable medium, such as storage device 110. Execution of the sequences of instructions contained in memory 106 can cause processor 104 to perform the processes described herein. Alternatively hard-wired circuitry can be used in place of or in combination with software instructions to implement the present teachings. Thus implementations of the present teachings are not limited to any specific combination of hardware circuitry and software.

[0036] In various embodiments, the term "computer-readable medium" as used herein refers to any media that participates in providing instructions to processor 104 for execution. Such a medium can take many forms, including but not limited to, non-volatile media, volatile media, and transmission media. Examples of non-volatile media can include, but are not limited to, optical or magnetic disks, such as storage device 110.

Examples of volatile media can include, but are not limited to, dynamic memory, such as memory 106. Examples of transmission media can include, but are not limited to, coaxial cables, copper wire, and fiber optics, including the wires that comprise bus 102.

[0037] Common forms of non-transitory computer-readable media include, for example, a floppy disk, a flexible disk, hard disk, magnetic tape, or any other magnetic medium, a CD-ROM, any other optical medium, punch cards, paper tape, any other physical medium with patterns of holes, a RAM, PROM, and EPROM, a FLASH-EPROM, any other memory chip or cartridge, or any other tangible medium from which a computer can read.

[0038] In accordance with various embodiments, instructions configured to be executed by a processor to perform a method are stored on a computer-readable medium. The computer-readable medium can be a device that stores digital information. For example, a computer-readable medium includes a compact disc read-only memory (CD-ROM) as is known in the art for storing software. The computer-readable medium is accessed by a processor suitable for executing instructions configured to be executed.

NUCLEIC ACID SEQUENCING PLATFORMS

[0039] Nucleic acid sequence data can be generated using various techniques, platforms or technologies, including, but not limited to: capillary electrophoresis, microarrays, ligation-based systems, polymerase-based systems, hybridization-based systems, direct or indirect nucleotide identification systems, pyrosequencing, ion- or pH-based detection systems, electronic signature-based systems, etc.

[0040] Various embodiments of nucleic acid sequencing platforms, such as a nucleic acid sequencer, can include components as displayed in the block diagram of Figure 2. According to various embodiments, sequencing instrument 200 can include a fluidic delivery and control unit 202, a sample processing unit 204, a signal detection unit 206, and a data acquisition, analysis and control unit 208. Various embodiments of instrumentation, reagents, libraries and methods used for next generation sequencing are described in U.S. Patent Application Publication No. 2009/0127589 and No. 2009/0026082 are incorporated herein by reference. Various embodiments of instrument

200 can provide for automated sequencing that can be used to gather sequence information from a plurality of sequences in parallel, such as substantially simultaneously.

[0041] In various embodiments, the fluidics delivery and control unit 202 can include reagent delivery system. The reagent delivery system can include a reagent reservoir for the storage of various reagents. The reagents can include RNA-based primers, forward/reverse DNA primers, oligonucleotide mixtures for ligation sequencing, nucleotide mixtures for sequencing-by-synthesis, optional ECC oligonucleotide mixtures, buffers, wash reagents, blocking reagent, stripping reagents, and the like. The implemented primers can be designed for particular targets or may be universal primers. Additionally, the reagent delivery system can include a pipetting system or a continuous flow system which connects the sample processing unit with the reagent reservoir.

[0042] In various embodiments, the sample processing unit 204 can include a sample chamber, such as flow cell, a substrate, a micro-array, a multi-well tray, a through hole, or the like. The sample processing unit 204 can include multiple lanes, multiple channels, multiple wells, or other means of processing multiple sample sets substantially simultaneously. Additionally, the sample processing unit can include multiple sample chambers to enable processing of multiple runs simultaneously. In particular embodiments, the system can perform signal detection on one sample chamber while substantially simultaneously processing another sample chamber. Additionally, the sample processing unit can include an automation system for moving or manipulating the sample chamber.

[0043] In various embodiments, the signal detection unit 206 can include an imaging or detection sensor. For example, the imaging or detection sensor can include a CCD, a CMOS, an ion or chemical sensor, such as an ion sensitive layer overlying a CMOS or FET, a current or voltage detector, or the like. The signal detection unit 206 can include an excitation system to cause a probe, such as a fluorescent dye, to emit a signal. The excitation system can include an illumination source, such as arc lamp, a laser, a light emitting diode (LED), or the like. In particular embodiments, the signal detection unit

206 can include optics for the transmission of light from an illumination source to the sample or from the sample to the imaging or detection sensor. Alternatively, the signal detection unit 206 may provide for electronic or non-photon based methods for detection and consequently not include an illumination source. In various embodiments, electronic-based signal detection may occur when a detectable signal or species is produced during a sequencing reaction. For example, a signal can be produced by the interaction of a released byproduct or moiety, such as a released ion, such as a hydrogen ion, interacting with an ion or chemical sensitive layer. In other embodiments a detectable signal may arise as a result of an enzymatic cascade such as used in pyrosequencing (see, for example, U.S. Patent Application Publication No. 2009/0325145, the entirety of which being incorporated herein by reference) where pyrophosphate is generated through base incorporation by a polymerase which further reacts with ATP sulfurylase to generate ATP in the presence of adenosine 5´ phosphosulfate wherein the ATP generated may be consumed in a luciferase mediated reaction to generate a chemiluminescent signal. In another example, changes in an electrical current can be detected as a nucleic acid passes through a nanopore without the need for an illumination source.

[0044] In various embodiments, a data acquisition analysis and control unit 208 can monitor various system parameters. The system parameters can include temperature of various portions of instrument 200, such as sample processing unit or reagent reservoirs, volumes of various reagents, the status of various system subcomponents, such as a manipulator, a stepper motor, a pump, or the like, or any combination thereof.

[0045] It will be appreciated by one skilled in the art that various embodiments of instrument 200 can be used to practice variety of sequencing methods including ligation-based methods, sequencing by synthesis, single molecule methods, nanopore sequencing, and other sequencing techniques.

[0046] In various embodiments, the sequencing instrument 200 can determine the sequence of a nucleic acid, such as a polynucleotide or an oligonucleotide. The nucleic acid can include DNA or RNA, and can be single stranded, such as ssDNA and RNA, or

double stranded, such as dsDNA or a RNA/cDNA pair. In various embodiments, the nucleic acid can include or be derived from a fragment library, a mate pair library, a ChIP fragment, or the like. In particular embodiments, the sequencing instrument 200 can obtain the sequence information from a single nucleic acid molecule or from a group of substantially identical nucleic acid molecules.

[0047] In various embodiments, sequencing instrument 200 can output nucleic acid sequencing read data in a variety of different output data file types/formats, including, but not limited to: *.fasta, *.csfasta, *seq.txt, *qseq.txt, *.fastq, *.sff, *prb.txt, *.sms, *.srs, *.bam, and/or *.qv.

SYSTEM AND METHODS FOR IDENTIFYING SEQUENCE VARIATION

[0048] Figure 3 is a schematic diagram of a system for identifying variants, in accordance with various embodiments.

[0049] As depicted herein, variant analysis system 300 can include a nucleic acid sequence analysis device 304 (e.g., a nucleic acid sequencer, real-time PCR instrument, digital PCR (dPCR) instrument, quantitative PCR (qPCR) instrument, microarray scanner, and the like), an analytics computing device 302, and a display 310 and/or a client device terminal 308.

[0050] In various embodiments, the analytics computing device 302 can be communicatively connected to the nucleic acid sequence analysis device 304, and client device terminal 308 via a network connection 326 that can be either a wired physical network connection (e.g., Internet, LAN, WAN, VPN, etc.) or a wireless network connection (e.g., Wi-Fi, WLAN, etc.).

[0051] In various embodiments, the analytics computing device 302 can be a workstation, mainframe computer, distributed computing node (such as, part of a "cloud computing" or distributed networking system), personal computer, mobile device, server, or the like. In various embodiments, the nucleic acid sequence analysis device 304 can be a nucleic acid sequencer, real-time PCR instrument, digital PCR (dPCR) instrument, quantitative PCR (qPCR) instrument, microarray scanner, or the like. It should be

understood, however, that the nucleic acid sequence analysis device 304 can be any type of instrument that can generate nucleic acid sequence data from samples (e.g., containing nucleic acid molecules) obtained from an individual.

[0052] The analytics computing device 302 can be configured to host an optional pre-processing module 312, a mapping module 314, and a fusion detection module 316.

[0053] Pre-processing module 312 can be configured to receive from the nucleic acid sequence analysis device 304 and perform processing steps, such as conversion from f space to base space, color space to base space, or from flow space to base space, determining call quality values, preparing the read data for use by the mapping module 314, and the like. Quality values can be a per base or per flow estimate of the accuracy of the sequencing.

[0054] The mapping module 314 can be configured to align (i.e., map) a nucleic acid sequence read to a reference sequence. In some examples, the length of the sequence read will be substantially less than the length of the reference sequence. In reference sequence mapping or alignment, sequence reads are assembled against an existing backbone sequence (e.g., reference sequence, etc.) to build a sequence that is similar but not necessarily identical to the backbone sequence. Once a backbone sequence is found for an organism, comparative sequencing or re-sequencing can be used to characterize the genetic diversity within the organism's species or between closely related species. In various embodiments, the reference sequence can be a whole/partial genome, whole/partial exome, and the like.

[0055] In various embodiments, the sequence read and reference sequence can be represented as a sequence of nucleotide base symbols in a base space. In various embodiments, the sequence read and reference sequence can be represented as one or more colors in a color space. In various embodiments, the sequence read and reference sequence can be represented as nucleotide base symbols with signal or numerical quantitation components in a flow space.

[0056] In various embodiments, the alignment of the sequence read and reference sequence can include a limited number of mismatches between the bases that comprise the sequence read and the bases that comprise the reference sequence. In some examples, the sequence read can be aligned to a portion of the reference sequence in order to minimize the number of mismatches between the sequence read and the reference sequence.

[0057] The fusion detection module 316 can include a remapping engine 318, an evaluation engine 322, and an optional post processing engine 324. In various embodiments, fusion detection module 316 can be in communication with the mapping module 314. That is, structural variant module 316 can request and receive data and information (through, e.g., data streams, data files, text files, and the like.) from mapping module 314.

[0058] The remapping engine 318 can be configured to receive mapped reads from the mapping module 314, and identify reads that are partially aligned (partially mapped). For example, the reads can be aligned at the beginning or the end of the read, and may be aligned not greater than 60% of the read length, not greater than 50% of the read length, or not greater than 40% of the read length. Additionally, the remapping engine 318 can split the reads separating the mapped portion from the unmapped portion to generate a set of read fragments. The read fragments can be mapped to the reference genome, and a set of candidate fusions can be generated by combining the locations where the first fragment and the second fragment of a read are mapped.

[0059] Evaluation engine 322 can be configured to receive evidence from the remapping engine 318. The evaluation engine 322 can obtain counts of the reads supporting candidate fusions. Optionally, the evaluation engine 322 can use annotations of the genomic loci to consolidate and filter the candidate fusions. Based on the evidence received from the remapping engine 318, the evaluation engine can classify the variant and determine a confidence value. In various embodiments, the evaluation engine 322 can filter the candidate fusions based on the number of counts, annotation of both loci of

a candidate fusion with gene names, annotation of at least one locus with an exon number, and the like.

[0060] Post processing engine 324 can be configured to receive the gene fusions identified by the evaluation engine 318 and perform additional processing steps, such as selecting and formatting the data for display on display 310 or use by client device 308.

[0061] Client device 308 can be a thin client or thick client computing device. In various embodiments, client device 308 can have a web browser (e.g., INTERNET EXPLORER™, FIREFOX™, SAFARI™, etc) that can be used to communicate information to and/or control the operation of the pre-processing module 312, mapping module 314, remapping engine 318, evaluation engine 322, and post processing engine 324 using a browser to control their function. The web browser may execute software (e.g., web applications, applets, and the like) configured to operate nucleic acid sequence analysis device 304. For example, the client device 308 can be used to configure the operating parameters (e.g., match scoring parameters, annotations parameters, filtering parameters, data security and retention parameters, etc.) of the various modules, depending on the requirements of the particular application. Similarly, client device 308 can also be configured to display the results of the analysis performed by the structural variant module 316 and the nucleic acid sequencer 304. Client device 308 may execute any other suitable software (e.g., stand-alone applications) configured to operate nucleic acid sequence analysis device 304.

[0062] It should be understood that the various data stores disclosed as part of system 300 can represent hardware-based storage devices (e.g., hard drive, flash memory, RAM, ROM, network attached storage, etc.) or instantiations of a database stored on a standalone or networked computing device(s) (e.g., virtual data storage). It should also be appreciated that the various data stores and modules or engines shown as being part of the system 300 can be combined or collapsed into a single module, engine, and/or data store, or expanded in multiple modules, engines, and/or data stores depending on the configuration of the particular application or system architecture implemented. Moreover, in various embodiments, the system 300 can comprise additional modules,

engines, components or data stores as needed by the particular application or system architecture.

[0063] In various embodiments, the system 300 can be configured to process the nucleic acid reads in color space. In various embodiments, system 300 can be configured to process the nucleic acid reads in base space. In various embodiments, system 300 can be configured to process the nucleic acid sequence reads in flow space. It should be understood to one of ordinary skill in the art, however, that the system 300 disclosed herein can process or analyze nucleic acid sequence data in any schema or format as long as the schema or format can convey the base identity and position of the nucleic acid sequence.

[0064] In various embodiments, system 300 may be used to identify candidates for gene fusions. For example, a gene may have experienced a fusion event, such a translocation, interstitial deletion, chromosomal inversion, and the like. The gene resulting from this combination may introduce complexities to the sequencing analysis.

[0065] Figure 4 is a diagram showing an exemplary gene fusion 400. Prior to a gene fusion event, exemplary gene 402 can be transcribed into an RNA 404 containing exons 406 and 408 and an intron 410. Splicing of the RNA 404 can remove intron 410 and produce an mRNA 412 including exons 406 and 408.

[0066] Similarly, exemplary gene 414 can be transcribed into an RNA 416 containing exons 418 and 420 and an intron 422. Splicing of the RNA 416 can remove intron 422 and produce an mRNA 424 including exons 418 and 420. A fusion event 426 can combine portions of gene 402 and gene 414 into fusion gene 428. Fusion gene 428 can be transcribed into an RNA 430 containing exons 432 and 434 and an intron 436. In various embodiments, exon 432 can correspond to exon 406 of gene 402, and exon 434 can correspond to exon 420 of gene 414. Splicing of the RNA 430 can remove intron 436 and produce an mRNA 438 including exons 432 and 434. Presence of mRNA 438 in a sample can be indicative of the presence of fusion gene 428. Additionally, the fusion event 426 that produces fusion gene 428 can disrupt the production of mRNA 412 and 424, as well as any proteins encoded by mRNA 412 and 424.

[0067] In various embodiments, detection techniques may aid in the sequencing of genes that have experienced a fusion event. For example, the detection of particular mapping conditions, such as, for example, partially aligned reads for a beginning or end of a read, may indicate a gene that has experienced a fusion event. Based on the detection of such conditions, additional processing may be performed to identify the fused gene portions, or candidate gene portions, involved in the fusion event.

[0068] Figure 5 is a flow diagram illustrating an exemplary method 500 of detecting gene fusions. At 502, reads can be mapped to a reference, such as a reference genome or transcriptome. For example, a plurality of amplicons may be generated in the presence of a primer pool. In various embodiments, the reads can be generated by sequencing amplicons generated from a multiplex amplification. The amplification can include a first set of primers corresponding to 3' end of a first plurality of exons and a second set of primers corresponding to a 5' end of a second plurality of exons. In various embodiments, amplicons can be generated from a plurality of known exon-exon junctions in a gene. Additionally, amplicons can be generated from gene fusions, where an exon from a first gene is joined to a portion of a second gene, as discussed herein.

[0069] In various embodiments, the amplicons may undergo a variety of sequencing reactions that result in signals emission such that a plurality of reads may be generated based on the detected signals. The sequence reads and reference sequence can be represented as a sequence of nucleotide base symbols in a base space, one or more colors in a color space, nucleotide base symbols with signal or numerical quantitation components in a flow space, or as any other suitable representation.

[0070] In various embodiments, mapping may include aligning a nucleic acid sequence read to a reference genome. In some examples, the length of the sequence read will be substantially less than the length of the reference genome. In reference sequence mapping or alignment, sequence reads may be assembled against an existing backbone sequence (e.g., reference sequence) to build a sequence that is similar but not necessarily identical to the backbone sequence.

[0071] In various embodiments, the alignment of the sequence read and reference genome can include a limited number of mismatches between the bases that comprise the sequence read and the bases that comprise the reference sequence. In some examples, the sequence read can be aligned to a portion of the reference sequence in order to minimize the number of mismatches between the sequence read and the reference sequence.

[0072] At 504, a subset of reads can be identified that are partially mapped to the reference genome. Specifically, the reads can have a mapped portion and an unmapped portion. The mapped portion may be near the beginning of the read with the unmapped portion near the end of the read, or the unmapped portion can be near the beginning of the read and the mapped portion can be near the end of the read. For example, the mapped portion may be within a threshold distance (e.g., threshold absolute number or percentage number of bases) from the beginning or end of the read. In various embodiments, the mapped portion can be not greater than 50% of the read length, such as not greater than 50% of the read length, not greater than 40% of the read length, or any other suitable percentage. The mapped portion may be mapped within a threshold distance from the reference sequence, such that the mapped portion includes a threshold number of mismatches with the reference sequence.

[0073] At 506, the reads of the subset can be split into the mapped portion and the unmapped portion, such that a read of the subset generates two read fragments. For example, the reads may be soft clipped such that a first read fragment includes the mapped portion of the read and a second read fragment includes the unmapped portion of the read. In an exemplary embodiment, each read fragment may be associated with a key (e.g., R1, R2, R3, and the like) identifying the partially mapped read from which the fragment was generated.

[0074] At 508, the read fragments generated from the partially mapped reads can be aligned to the reference genome. In an example, a first fragment of a partially mapped read (e.g., identified as R1) will map to a first locus within the reference genome and the second fragment of the partially mapped read (e.g., identified as R1) will map to a second locus within the reference genome. A locus may be a mapped location for each of the

read fragments on the reference genome. In some examples, the location may correspond to a known gene (e.g., with a gene name) and/or a known portion of a gene (e.g., known exon of a known gene). Similar to previous mappings/alignments, each fragment mapping may be within a threshold distance from the reference genome.

[0075] At 510, a candidate list of fusions can be generated based on an ordered pair of loci for the read fragments. For example, a subset of read fragments that each map to a loci on the reference genome may be selected. The candidate list of fusions may be a list of partially mapped reads where each fragment generated from the partially mapped read is selected for the subset. For example, the subset of read fragments may be analyzed such that fragments that contain matching keys (identifying the partially mapped read from which the fragment was generated) may be added to the candidate list. Based on the selection, read fragments from the same partially mapped read may be selected when each fragment from that partially mapped read is mapped to the reference genome.

[0076] The following data set illustrates an exemplary candidate fusion list, for instance as a database table. The entries may include each loci for the read fragments with accompanying data and a count for the particular loci combination.

[chr2:29446338-29446396, chr2:42491846-42491869] 1689

[chr4:39259398-39259419, chr6:170871271-170871321] 64

[chr2:29446338-29446396, chr2:42492057-42492089] 70

[chr12:128904278-128904299, chr6:170871271-170871321] 31

[chr19:55857939-55857959, chr1:156104280-156104320] 37

[chr2:29446327-29446396, chr2:42491846-42491869] 55

[chr1:156104594-156104644, chrX:137308905-137308925] 70

[0077] At 512, the candidate fusions can be evaluated. For example, one or more filters may be applied to the candidate fusion list. In an embodiment, the number of reads with fragments mapping to the particular loci pair can be counted. For example, reads that generate fragments that map to the same (or substantially some) loci can be counted. An exemplary filter may comprise a threshold number of counted pairs.

[0078] In an exemplary embodiment, the loci can be annotated, such as with gene names and exon designations. For example, the reference genome may include accompanying metadata associated with the loci, such as gene names and exon designations. One or more filters may include at least one loci being annotated with a gene name, both loci being annotated with a gene name, at least one loci being annotated with an exon number, both loci being annotated with an exon number, and any suitable combination of filters. In some embodiments, a filter may include maximum distance from an (known) exon boundary (e.g., max number of bases or percentages of bases in the read fragment).

[0079] The following data set may illustrate a filtered and annotated candidate fusion list, for instance as a database table. The entries may include each loci for the read fragments with accompanying annotation data, if known, (e.g., gene name, exon number, sequence location, and the like), and a count for the particular loci combination.

[LMNA, SUV420H2E8] 45

[TBP, TMEM132C] 49

[TBP, WDR19] 91

[ALKE21, EML4E5] 1774

[ALKE21, EML4] 238

[LMNA, chrX:137308905-137308925] 105

[0080] In various embodiments, the remaining candidate fusions after filtering and consolidating can be reported to a user. In some embodiments, one or more databases may be updated with data based on the filtered candidate fusion list. For example, a database that stores known gene fusions may be updated based on candidate fusion list. The database updates may be entries annotated with known fusion data (e.g., gene name, exon number, and other identifying information). In some examples, annotation data from the candidate fusion list may include previously unknown information about a known gene fusion (e.g., a distance from an exon boundary for the gene fusion, and the like), and the database may be updated with the previously unknown information. In some embodiments, one or more fusions from the candidate fusion list may not have been known (e.g., may not have been stored in a database of known gene fusions). In this

example, the database of known gene fusions may be updated with the unknown gene fusion and any accompanying annotation data from the candidate fusion list.

[0081] In some embodiments, one or more data files may be generated such that a user may visualize the gene fusion. For example, the annotation data for the candidate gene fusions may be used to identify known genes, and a visualization of the gene fusion event based on the known genes may be generated. The visualization may be based on software executing on a computing device that interfaces with one or more displays such that a visualization (such as an animation) may be displayed.

[0082] In some embodiments, the breakpoints for candidate gene fusions may be unknown prior to sequencing. In some embodiments, a limited number of breakpoints may be known for the gene fusion (e.g., one of two known breakpoints). In some embodiments, the known information about the breakpoints may be limited (e.g., an exact location for the breakpoints may not be known). For instance, gene names may be known, but the precise coordinates for the breakpoints may be unknown. In these examples, the primer pool may be designed based on the limited knowledge for breakpoints. For instance, when one breakpoint is known a first primer may be designed based on the known breakpoint and a second primer may be a universal primer.

[0083] The exemplary workflow of FIG. 5 may include technical advantages over previous gene fusion detection methodologies. For example, some methodologies may require precise knowledge of fusion breakpoints in order to design a primer pool for sequencing. Other methodologies may not require such knowledge, but may be inefficient, slow, inaccurate, or require large amounts of processing due to excessive data that needs to be analyzed. Techniques according to various disclosed embodiments can provide an enhanced gene fusion detection methodology that does not require *a priori* knowledge of breakpoints, or at least can leverage only limited knowledge of breakpoints, and/or that efficiently detects gene fusions without excessive processing.

[0084] For example, one or more of identifying the subset of reads, soft clipping the subset of reads, generating the candidate fusion list based on matching keys, and/or filtering the candidate fusion list, as described herein, may reduce the number of false

positives in the analysis, and subsequently reduce the amount of processing needed to detect the candidate gene fusions. In some embodiments, the filtering may reduce the size of the candidate fusion list by 90%, 80%, 70%, and the like. In other examples, for instance where data is known about the candidate gene fusions, the filtering may reduce the size of the list by a smaller percentage.

[0085] Figure 6 is a diagram showing an exemplary synthetic nucleotide controls. In various embodiments, the synthetic nucleotide control is a synthetic RNA control that can be used for fusion transcripts. The synthetic nucleotide can be spiked into the sample, prior to multiplex amplification. The "natural" spike-in sequence can be identical or substantially identical to the sequence of the fusion transcript. The "scrambled" spike-in sequence can be identical or substantially identical to the sequence of the fusion transcript only under the corresponding primer binding sites; the remaining internal sequence can be scrambled to distinguish it from the natural fusion transcript. For instance, the remaining internal sequence can be an arbitrary sequence of bases.

[0086] In various embodiments, a fusion assay may include a number of primer pairs that do not produce an amplicon. For example only a limited number of fusion species may be present in a positive sample, and some RNA samples may not have any of the targeted fusions species. Therefore, the data can produce "negative" evidence, in that a positive indicator is not triggered, and it may be challenging to determine if the primers are present and functional. The "scrambled" spike in sequence can provide a quality control mechanism for the multiplex amplification.

[0087] A scrambled spike in sequence ("scramblicon") can have the correct primer binding sequences for an RNA fusion amplicon, but between the two primer binding sites the sequence is scrambled such that the GC content and the length is maintained. These synthetic templates can clearly be distinguished from the native fusion species. If "native" templates were used to validate the RNA primers, it may be difficult to distinguish between the templates and the presence of the fusion in the sample. Additionally, a small amount of the "native" templates could contaminate samples and result in false positive results.

[0088] In various embodiments, the methods of the present teachings may be implemented in a software program and applications written in conventional programming languages such as C, C++, and the like.

[0089] While the present disclsoure sets forth various embodiments, it is not intended that the scope of the disclosure and claims be limited to such embodiments. On the contrary, those of ordinary skill in the art will appreciate that various alternatives, modifications, and equivalents are encompassed.

[0090] Further, in describing various embodiments, the specification may have presented a method and/or process as a particular sequence of steps. However, to the extent that the method or process does not rely on the particular order of steps set forth herein, the method or process should not be limited to the particular sequence or practice of steps described. As one of ordinary skill in the art would appreciate, other sequences of steps may be possible. Therefore, the particular order and inclusion of any of the steps set forth in the specification should not be construed as limitations on the claims. In addition, the claims directed to the method and/or process should not be limited to the performance of their steps in the order written, and one skilled in the art can readily appreciate that the sequences may be varied and still remain within the spirit and scope of the various embodiments.

[0091] The embodiments described herein, can be practiced with other computer system configurations including hand-held devices, microprocessor systems, microprocessor-based or programmable consumer electronics, minicomputers, mainframe computers and the like. The embodiments can also be practiced in distributing computing environments where tasks are performed by remote processing devices that are linked through a network.

[0092] It should also be understood that the embodiments described herein can employ various computer-implemented operations involving data stored in computer systems. These operations are those requiring physical manipulation of physical quantities. Usually, though not necessarily, these quantities take the form of electrical or magnetic signals capable of being stored, transferred, combined, compared, and otherwise

manipulated. Further, the manipulations performed are often referred to in terms, such as producing, identifying, determining, or comparing.

[0093] Any of the operations that form part of the embodiments described herein are useful machine operations. The embodiments, described herein, also relate to a device or an apparatus for performing these operations. The systems and methods described herein can be specially constructed for the required purposes or it may be a general purpose computer selectively activated or configured by a computer program stored in the computer. In particular, various general purpose machines may be used with computer programs written in accordance with the teachings herein, or it may be more convenient to construct a more specialized apparatus to perform the required operations.

[0094] Certain embodiments can also be embodied as computer readable code on a computer readable medium. The computer readable medium is any data storage device that can store data, which can thereafter be read by a computer system. Examples of the computer readable medium include hard drives, network attached storage (NAS), read-only memory, random-access memory, CD-ROMs, CD-Rs, CD-RWs, magnetic tapes, and other optical and non-optical data storage devices. The computer readable medium can also be distributed over a network coupled computer systems so that the computer readable code is stored and executed in a distributed fashion.

WHAT IS CLAIMED IS:

1. A method for detecting gene fusions comprising:

amplifying a nucleic acid sample in the presence of a primer pool to produce a plurality of amplicons;

sequencing the amplicons by detecting a plurality of signals indicative of nucleotide incorporation events to generate a plurality of reads;

mapping the reads to a reference genome based on alignments between the reads and the reference genome;

identifying reads that partially map to the reference genome;

generating read fragments by splitting the partially mapped reads into a mapped region and an unmapped region;

mapping the read fragments to the reference genome;

identifying candidate fusions based on loci on the reference genome for the mapped region and the unmapped region of the read fragments.

2. The method of claim 1, wherein the primer pool includes a first set of primers corresponding to a first end of a first plurality of exons and a second set of primers corresponding to a second end of a second plurality of exons.

3. The method of claim 2, wherein one of the first set of primers and the second set of primers is designed based on a known breakpoint for a gene fusion and the other of the first set of primers and the second set of primers comprises a universal set of primers.

4. The method of claim 1, wherein the identified reads that partially map to the reference genome comprise a mapped portion that is not greater than 50% of a read length for the identified read or not greater than 40% of the read length.

5. The method of claim 1, wherein each of a generated first fragment and second fragment comprise a key associated with the read from which the fragments were generated.

6. The method of claim 5, further comprising selecting a subset of the generated read fragments that comprise a mapping to the reference genome and a corresponding read fragment with the same key that also comprises a mapping to the reference genome, wherein the subset of read fragments are identified as candidate fusions.

7. The method of claim 1, further comprising filtering the identified candidate fusions based on a count for each particular loci combination of the mapped portion and the unmapped portion of the read fragments.

8. The method of claim 7, further comprising:

annotating the candidate fusions with known information comprising one or more of a gene name and an exon identification; and

filtering the identified candidate fusions based on the count for each particular loci combination of the mapped portion and the unmapped portion of the read fragments and an availability of the annotated known information.

9. The method of claim 1, wherein the identified candidate fusions are filtered based on at least one of an availability of a gene name for at least one of the read fragments, an availability of a gene name for both of the read fragments, an availability of an exon identification for at least one of the read fragments, and an availability of an exon identification for both of the read fragments.

10. The method of claim 1, further comprising updating a database of fusion genes with information based on the identified candidate fusions.

11. A system for detecting gene fusions comprising:
a nucleic acid sequencing device configured to:

sequence a plurality of amplicons by detecting a plurality of signals indicative of nucleotide incorporation events to generate a plurality of reads, wherein the amplicons were produced by amplifying a nucleic acid sample in the presence of a primer pool; and

an analytics computing device comprising a processor configured to:

map the reads to a reference genome based on alignments between the reads and the reference genome;

identify reads that partially map to the reference genome;

generating read fragments by splitting the partially mapped reads into a mapped region and an unmapped region;

map the read fragments to the reference genome; and

identify candidate fusions based on loci on the reference genome for the mapped region and the unmapped region of the read fragments.

12. The system of claim 11, wherein the primer pool includes a first set of primers corresponding to a first end of a first plurality of exons and a second set of primers corresponding to a second end of a second plurality of exons.

13. The system of claim 12, wherein one of the first set of primers and the second set of primers is designed based on a known breakpoint for a gene fusion and the other of the first set of primers and the second set of primers comprises a universal set of primers.

14. The system of claim 11, wherein the identified reads that partially map to the reference genome comprise a mapped portion that is less than or equal to 50% of a read length for the identified read or less than or equal to 40% of the read length.

15. The system of claim 11, wherein each of the generated first fragments and second fragments comprise a key associated with the read from which the fragments were generated.

16. The system of claim 15, wherein the analytics computing device is further configured to select a subset of the generated read fragments that comprise a mapping to the reference genome and a corresponding read fragment with the same key that also comprises a mapping to the reference genome, wherein the subset of read fragments are identified as candidate fusions.

17. The system of claim 11, wherein the analytics computing device is further configured to filter the identified candidate fusions based on a count for each particular loci combination of the mapped portion and the unmapped portion of the read fragments.

18. The system of claim 17, wherein the analytics computing device is further configured to:

annotate the candidate fusions with known information comprising one or more of a gene name and an exon identification; and

filter the identified candidate fusions based on the count for each particular loci combination of the mapped portion and the unmapped portion of the read fragments and an availability of the annotated known information.

19. The system of claim 11, wherein the identified candidate fusions are filtered based on at least one of an availability of a gene name for at least one of the read fragments, an availability of a gene name for both of the read fragments, an availability of an exon identification for at least one of the read fragments, and an availability of an exon identification for both of the read fragments.

20. The system of claim 11, wherein the analytics computing device is further configured to update a database of fusion genes with information based on the identified candidate fusions.

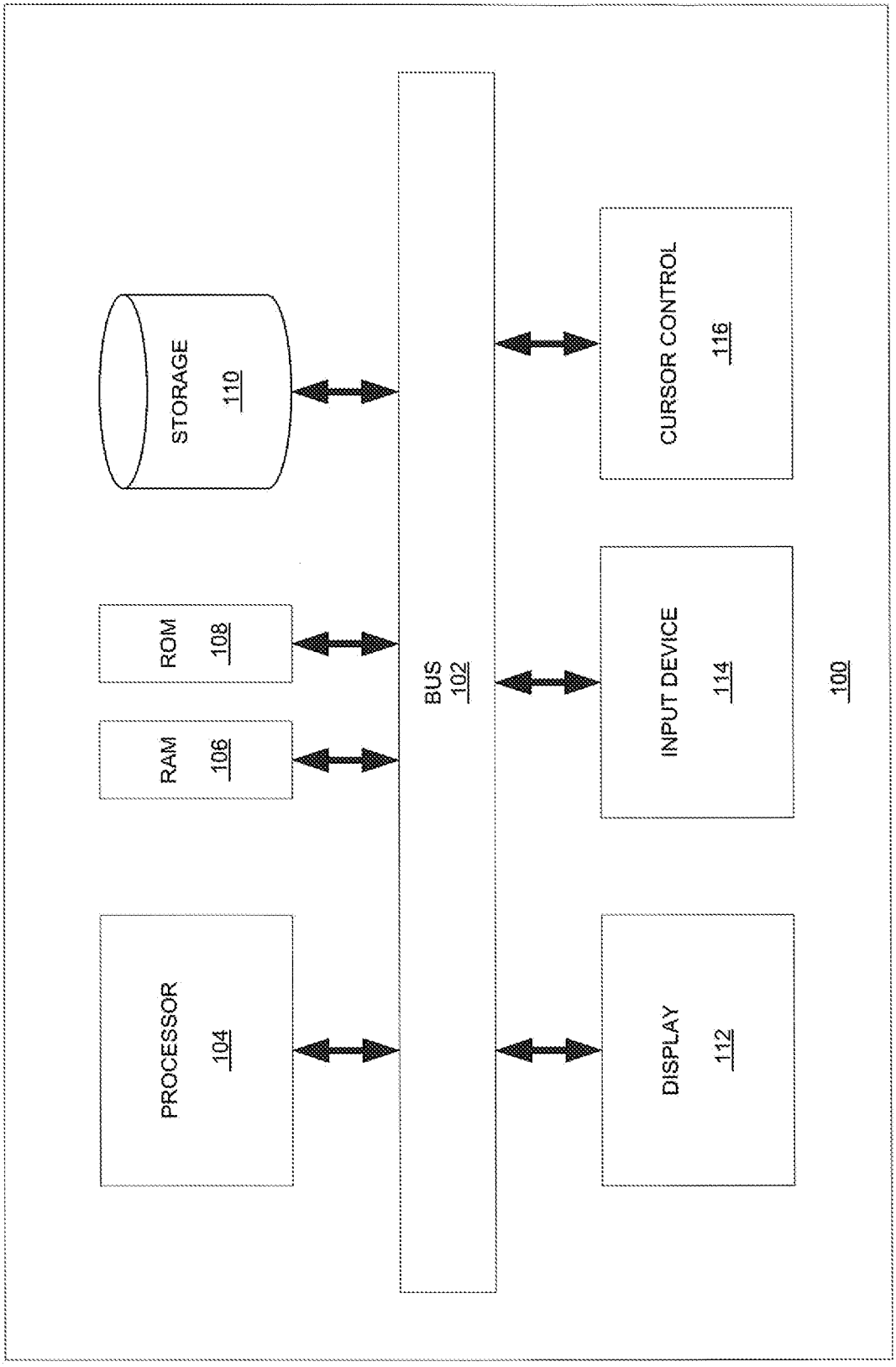
STORAGE
110
ROM
108
RAM
106
PROCESSOR
104
BUS
102
CURSOR CONTROL
116
INPUT DEVICE
114
DISPLAY
112
100

FIG. 1

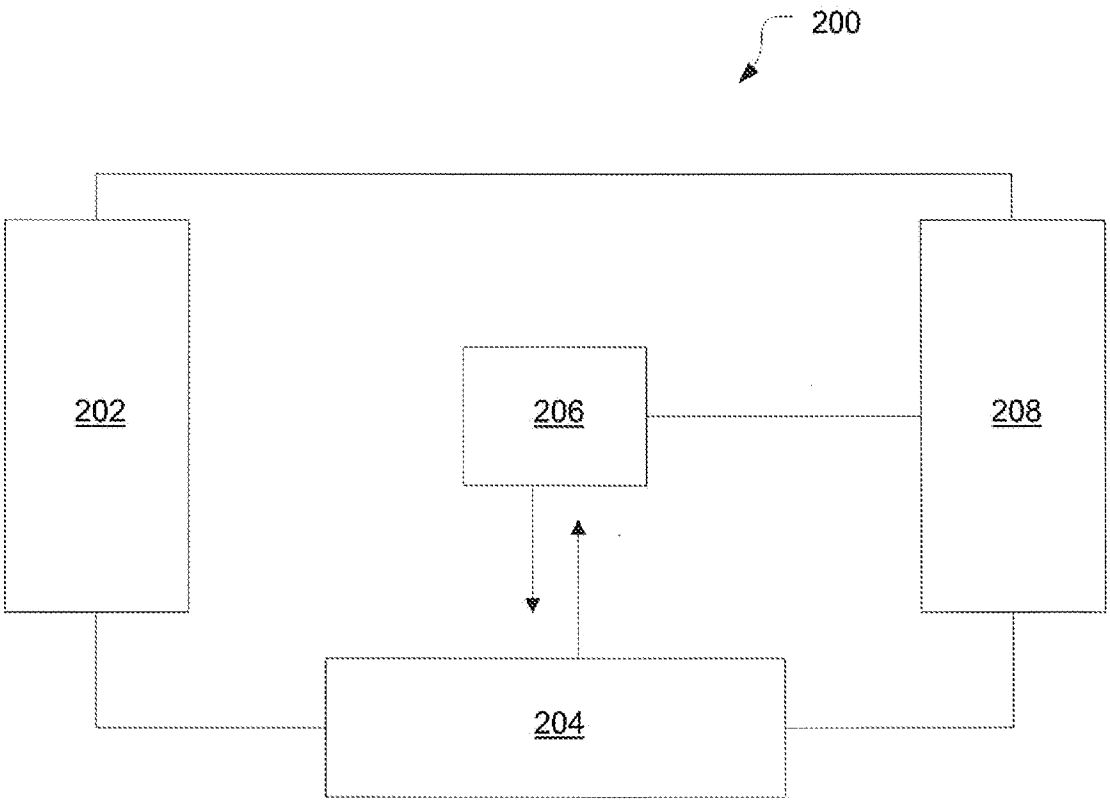

FIG. 2

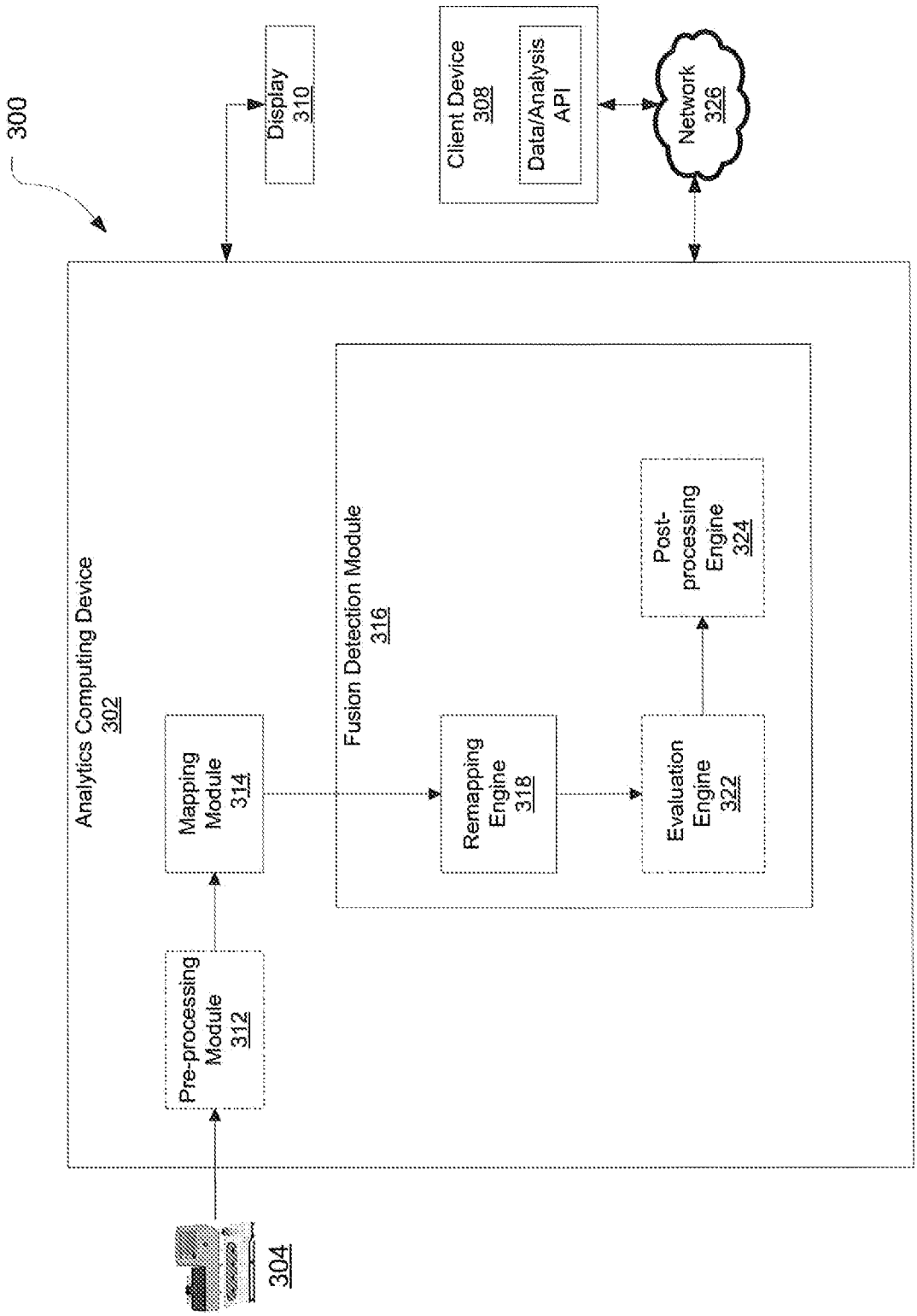
300
Display 310
Client Device 308
Data/Analysis API
Network 326
Analytics Computing Device 302
Fusion Detection Module 316
Mapping Module 314
Remapping Engine 318
Evaluation Engine 322
Post-processing Engine 324
Pre-processing Module 312
304

FIG. 3

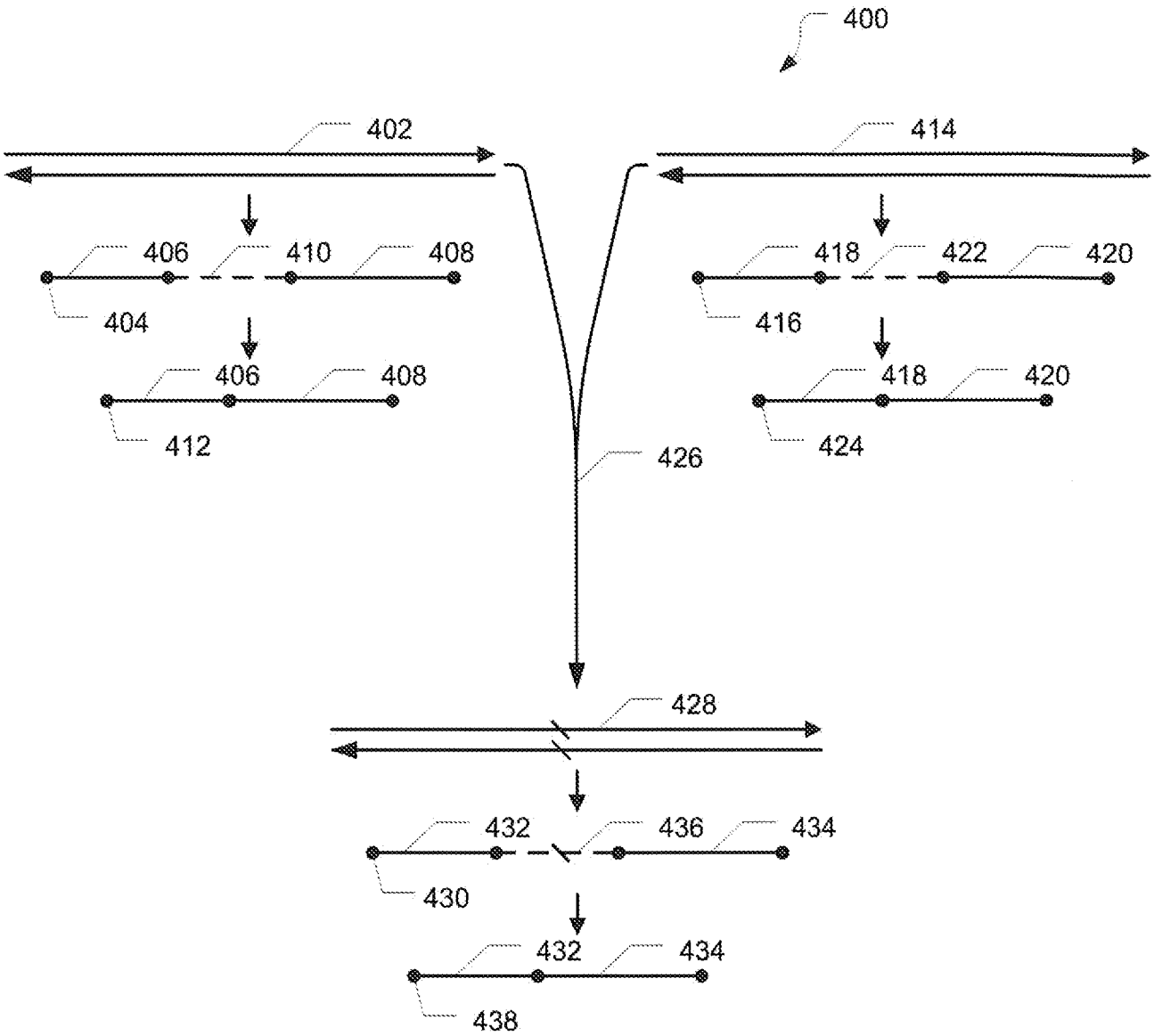
400
402
414
406
410
408
418
422
420
404
416
406
408
418
420
412
424
426
428
432
436
434
430
432
434
438

FIG. 4

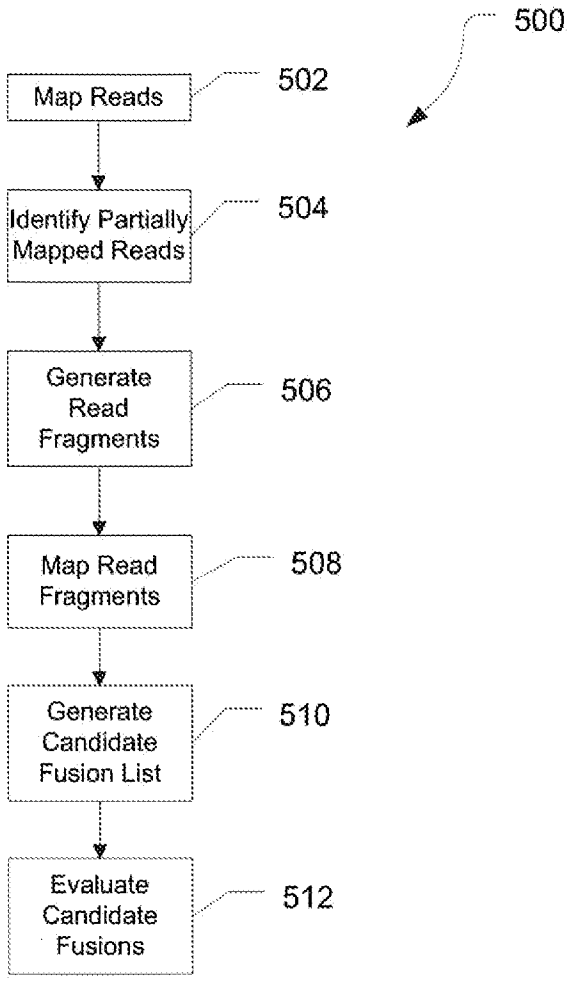
500
Map Reads
502
Identify Partially Mapped Reads
504
Generate Read Fragments
506
Map Read Fragments
508
Generate Candidate Fusion List
510
Evaluate Candidate Fusions
512

FIG. 5

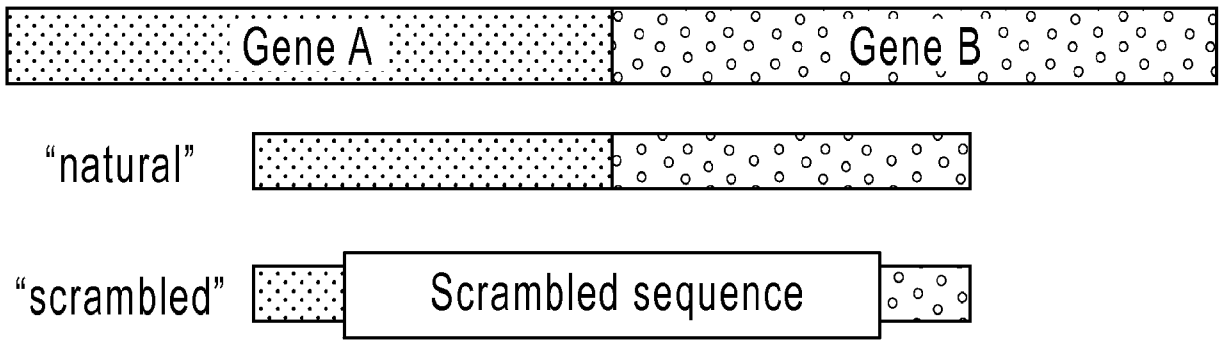

FIG. 6

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2015/040951

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12Q1/68 G06F19/22
ADD.

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)
C12Q G06F

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)

EPO-Internal, CHEM ABS Data, EMBASE, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EDGREN HENRIK ET AL: "Identification of fusion genes in breast cancer by paired-end RNA-sequencing", GENOME BIOLOGY, BIOMED CENTRAL LTD., LONDON, GB, vol. 12, no. 1, 19 January 2011 (2011-01-19), page R6, XP021091784, ISSN: 1465-6906, DOI: 10.1186/GB-2011-12-1-R6 abstract; figure 1 page r6, column 2 page r7, column 1 page r8, column 1 page r14 - column 2 ----- -/--	1-20

[X] Further documents are listed in the continuation of Box C.

[X] 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
2 October 2015	09/10/2015

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Celler, Jakub

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2015/040951

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2013/110410 A1 (HONG YOO JIN [KR] ET AL) 2 May 2013 (2013-05-02) abstract; figures 3,7 paragraphs [0003], [0007], [0008], [0009], [0018], [0104] -----	1-20

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2015/040951

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2013110410 A1	02-05-2013	CN 103087906 A	08-05-2013
		EP 2587396 A2	01-05-2013
		JP 5710572 B2	30-04-2015
		JP 2013094169 A	20-05-2013
		KR 20130047383 A	08-05-2013
		US 2013110410 A1	02-05-2013