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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 October 2017 (19.10.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/179581 A1

(51) International Patent Classification:
G06Q 50/22 (2012.01)

(21) International Application Number:
PCT/JP2017/014847

(22) International Filing Date:
11 April 2017 (11.04.2017)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/321,103 11 April 2016 (11.04.2016) US

(71) Applicant: **QUANTUM BIOSYSTEMS INC.** [JP/JP]; 2-3-11, Nihonbashihoncho, Chuo-ku, Tokyo, 1030023 (JP).

(72) Inventors; and
(71) Applicants : **VAKILI, Masoud** [CA/US]; 598 Magdalena Ave, Los Altos, California, 94024 (US). **CHRISTOFFERSON, Kurt** [US/US]; 364 Los Alamos Road, Santa Rosa, California, 95409 (US). **OLDHAM, Mark** [US/US]; 738 Glenmere Way, Emerald Hills, California, 94062 (US).

(74) Agents: **INABA, Yoshiyuki** et al.; TMI ASSOCIATES, 23rd Floor, Roppongi Hills Mori Tower, 6-10-1, Roppongi, Minato-ku, Tokyo, 1066123 (JP).

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

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

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

(54) **Title**: SYSTEMS AND METHODS FOR BIOLOGICAL DATA MANAGEMENT

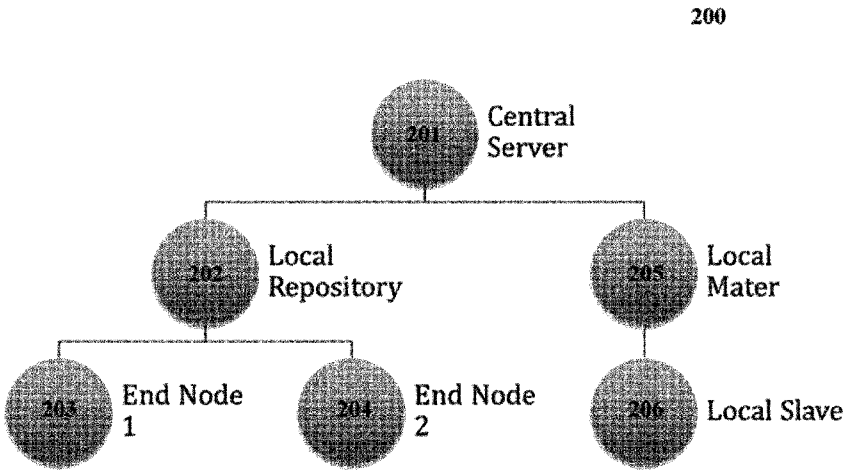

FIG. 2

(57) **Abstract**: Systems and methods for biological data management may preserve alternative interpretations of data and may implement multi-level encryption and privacy management. Systems and methods for biological data management may include a cell-level architecture, a bank-and-bloc-level architecture, and/or a multi-tiered architecture. Systems and methods for biological data management may incorporate definitions, rules, and directives and/or employ a two-dimensional or three-dimensional data structure.

WO 2017/179581 A1

Description

Title of Invention: SYSTEMS AND METHODS FOR BIOLOGICAL DATA MANAGEMENT

CROSS-REFERENCE

[0001] This application claims priority to U.S. Provisional Patent Application No. 62/321,103, filed April 11, 2016, which is entirely incorporated herein by reference.

Background Art

[0002] New research continues to increase our understanding of genetic information and raise challenges about how to manage such information. A more complete understanding of genetic maps with a higher level of resolution may render valuable results in healthcare and other disciplines.

[0003] As an example, one of the challenges in managing genetic deoxyribonucleic acid (DNA) data is that there are highly conserved regions of code, which remain unchanged over time, yet do not seem to code proteins. Research indicates, however, that they may play important roles in gene expression regulation, alternative splicing, and distal enhancers. An efficient way to save regions that are utilized infrequently, while sustaining fast access to more frequently used regions of a genetic sequence, is therefore desirable.

Summary of Invention

[0004] Recognized herein is a need for data management schemes that can accommodate alternative interpretations of data and hence may have access to lower-level data measured by various devices. Also recognized herein is a need to sense, store, and manage genetic data with greater flexibility and greater completeness, as well as a need to flexibly and efficiently create, add to, maintain, and query these data sets at different levels while handling error scenarios.

[0005] Provided herein are systems and methods for efficiently and securely managing genetic data, including reading and interpreting raw data, storing and interpreting the genetic data, and maintaining privacy and confidentiality of the data.

[0006] Some systems and methods may provide definitions and rules, and issue appropriate directives for issues related to healthcare, food safety, and/or other pathogen handling situations. A multi-tier network architecture in an information handling environment may be utilized.

[0007] Parallelism may be used as required by the task and type of biological data interpretation. Information may be initially stored in a distributed storage of semi-structured data, allowing for scanning, reducing, and reorganizing information as needed into structured, columnar, or relational databases.

[0008] Systems and methods may stage and perform different queries concurrently, allowing information to be stored in repositories, and may be encrypted at rest. Information may be transmitted across a distributed system, between repositories, between servers, or between servers and clients in an secure and flexible fashion.

[0009] Systems and methods can store biological data in one or more storage devices according to a relationship between a size of data or units of data and a size of unit storage blocks or banks of one or more storage devices.

[0010] Systems and methods may support access controls, which may be user, role, application, process, or location based.

[0011] Systems and methods may relate to mapping and storing genetic data, (e.g., polynucleotide data) in one or more memory devices at a memory cell level, at a memory block level, at a memory bank level, or at another memory partition level.

[0012] An aspect of the present disclosure provides a biological data management system, comprising: (a) an end-user module comprising a sequencing device, the sequencing device configured to generate base data; (b) a local repository in network communication with the end-user module, the local repository programmed or configured to (i) receive the base data, (ii) convert the base data into sequence data, (iii) produce abbreviated data based on the sequence data, and (iv) compare the abbreviated data with a database of existing abbreviations; and (c) a central server in network communication with the local repository, the central server configured to update the database of the existing abbreviations.

[0013] In some embodiments, the local repository is further programmed or configured to flag abbreviations and communicate the flagged abbreviations to the central server. In some embodiments, the central server is further programmed or configured to receive a flagged abbreviation and perform further analysis on the flagged abbreviation. In some embodiments, the central server is further programmed or configured to generate a directive and communicate the directive to the local repository upon the analysis of the flagged abbreviation. In some embodiments, the abbreviation is a variance, hash, or a checksum.

[0014] Another aspect of the present disclosure provides a method for storing biological data, comprising: (a) determining a size of the biological data to identify a storage unit size suitable to store the biological data; (b) identifying a memory location in a memory device having a block size compatible with the storage unit size; and (c) storing the biological data in an erasable block at the memory location of the memory device.

[0015] In some embodiments, each erasable block comprises a section for storing the biological data and a section for storing metadata related to the biological data. In some embodiments, the section for storing metadata comprises a longer lifetime. In some

embodiments, the section for storing metadata comprises a controller different from a controller of the section for storing sequence data. In some embodiments, the section for storing metadata is configured for more frequent access than the section for storing sequence data.

[0016] Another aspect of the present disclosure provides a biological data management system, comprising: (a) a first memory device configured to store biological data for infrequent access; and (b) a second memory device having a block size, the second memory device being in communication with the first memory device and configured to store biological data for frequent access; wherein the second memory device is faster than the first memory device, and wherein the block size is selected to store the biological data according to a size of the biological data.

[0017] In some embodiments, the biological data is an n-mer sequence, and the block size is n times a number of bits required to store a monomer of the n-mer. In some embodiments, the biological data is an n-mer sequence, and the block size is at least n times a number of bits required to store a monomer of the n-mer. In some embodiments, the second memory device comprises a flash memory device. In some embodiments, the second memory device comprises a block that is a flash memory erase block.

[0018] Another aspect of the present disclosure provides a method for storing sequence base data in a multi-level cell (MLC) memory device, the MLC memory device comprising memory cells, each of the memory cells configured to store two bits, the method comprising, in a memory cell: (a) setting the two bits to 00 to represent a base of a first type; (b) setting the two bits to 01 to represent a base of a second type; (c) setting the two bits to 10 to represent a base of a third type; or (d) setting the two bits to 11 to represent a base of a fourth type.

[0019] In some embodiments, the sequence base data represents one or more polynucleotides, each of the polynucleotides comprising one or more bases, each of the one or more bases being one of at least four possible bases. In some embodiments, the polynucleotide is a DNA or an RNA.

[0020] Another aspect of the present disclosure provides a method for storing biological data in a memory device, the memory device comprising blocks, each of the blocks comprising a block size, the method comprising: (a) determining a size of the biological data; (b) determining a block size of at least a subset of the blocks; (c) compressing the biological data based on the block size to produce compressed biological data; and (d) storing the biological data in the at least a subset of the blocks.

[0021] The method of claim 19, wherein the memory device comprises a flash memory device, and wherein the block size is an erase block size.

[0022] In some embodiments, the block size is greater than or equal to a size of the

compressed biological data. In some embodiments, the erase block stores the biological data and metadata of the biological data.

[0023] Another aspect of the present disclosure provides a method for storing sequence base data in a memory device, the memory device comprising memory cells, each of the memory cells configured to store at least three bits, the method comprising, in a memory cell: (a) setting three of the at least three bits to 000 to represent a base of a first type; (b) setting three of the at least three bits to 001 to represent a base of a second type; (c) setting three of the at least three bits to 010 to represent a base of a third type; (d) setting three of the at least three bits to 011 to represent a base of a fourth type; (e) setting three of the at least three bits to 100 to represent a base of a fifth type; (f) setting three of the at least three bits to 101 to represent a base of a sixth type; (g) setting three of the at least three bits to 110 to represent a base of a seventh type; and (h) setting three of the at least three bits to 111 to represent a base of an eighth type.

[0024] In some embodiments, the sequence base data represents one or more polynucleotides, each of the polynucleotides comprising one or more bases, each of the one or more bases being one of four different native bases, a methylated base, an oxidated base, or an abasic location. In some embodiments, the polynucleotide is a DNA or an RNA. In some embodiments, the memory device comprises a flash memory, a phase-change memory, or a resistive memory.

[0025] Another aspect of the present disclosure provides a method for storing sequence base data in a memory device, the sequence base data comprising two probable bases to represent each of a plurality of bases measured, the memory device comprising memory cells, each of the memory cells configured to store a plurality of bits, the method comprising: storing in a first bit of the plurality of bits a most probable base of the sequence base data; storing in a second bit of the plurality of bits a second most probable base of the sequence base data; and storing in a remainder of the plurality of bits a relative probability of the most probable base and the second most probable base.

[0026] In some embodiments, the method further comprises, using a first cell of the memory cells to identify the most probable base; using a second cell of the memory cells to identify the second most probable base; and using one or more other cells of the memory cells to store the relative probability. In some embodiments, the method further comprises storing in a third cell of the memory cells a probability of the second most probable base.

[0027] Another aspect of the present disclosure provides a method for storing sequence base data in a memory device comprising memory cells each configured to store at least three bits, the method comprising, in a memory cell: (a) providing a first bit indication comprising three bits of the at least three bits to represent a base of a first type; (b)

providing a second bit indication comprising three bits of the at least three bits to represent a base of a second type; (c) providing a third bit indication comprising three bits of the at least three bits to represent a base of a third type; (d) providing a fourth bit indication comprising three bits of the at least three bits to represent a base of a fourth type; (e) providing a fifth bit indication comprising three bits of the at least three bits to represent a methylated base; (f) providing a sixth bit indication comprising three bits of the at least three bits to represent an oxidated base; and (g) providing a seventh bit indication comprising three bits of the at least three bits to represent an abasic site.

[0028] In some embodiments, the memory device comprises a flash memory, a phase-change memory, or a resistive memory.

[0029] Another aspect of the present disclosure provides a method for encrypting biological sequence data, the method comprising: (a) identifying a normal level of variance in the biological sequence data; and (b) introducing a second level of variation into the biological sequence data, the second level of variation comparable to the normal level of variance, such that the biological sequence data is indistinguishable with respect to the normal level of variance.

[0030] In some embodiments, the method further comprises communicating the introduced level of variance using an encryption method.

[0031] Another aspect of the present disclosure provides a method for encrypting biological sequence data of a subject, the method comprising: (a) encrypting information related to the subject using a first encryption scheme; and (b) encrypting the biological sequence data using a second encryption scheme, which second encryption scheme is different from the first encryption scheme.

[0032] In some embodiments, the second encryption scheme comprises a less extensive encryption than the first encryption scheme. In some embodiments, the second encryption scheme comprises chaffing and winnowing. In some embodiments, the first encryption scheme uses a public key infrastructure and the second encryption scheme uses the public key infrastructure. In some embodiments, the first encryption scheme uses a first public key infrastructure and the second encryption scheme uses a second public key infrastructure different from the first public key infrastructure.

[0033] Another aspect of the present disclosure provides a method for storing sequence base data, the method comprising: providing a two-dimensional table structure in computer memory, the two-dimensional table structure configured to store information representing potential bases; storing information representing the most probable measured bases of the sequence base data in a first dimension of the two-dimensional table structure; storing information representing other potential bases of the sequence base data in a second dimension of the two-dimensional table structure; and storing probabilities corresponding to an intersection of the first dimension and the second

dimension in the two-dimensional table structure.

[0034] In some embodiments, the potential bases comprise a set of each of four possible bases and at least one of a methylated base, an oxidated base, and an abasic site. In some embodiments, the method further comprises providing a second two-dimensional table structure in computer memory, the second two-dimensional table structure configured to store information representing potential bases; and storing in the second two-dimensional table structure the most probable measured bases of the sequence base data and the second most probable measured bases of the sequence base data.

[0035] Another aspect of the present disclosure provides a method for managing biological data, the method comprising: providing an application server programmed or configured to (i) receive raw measured biological data from a sensor and (ii) generate processed biological data from the raw measured biological data; receiving, at the application server, from a local repository, definitions and rules related to the processed biological data; and issuing, by the application server, directives based on the definitions and rules related to the processed biological data.

[0036] In some embodiments, the processed biological data comprises a portion of the processed biological data for which related definitions and rules are not found in the local repository, and the method further comprises sending at least the portion of the processed biological data to the local repository. In some embodiments, the method further comprises sending at least the portion of the processed biological data from the local repository to a central server. In some embodiments, the method further comprises sending directives from the central server to the local repository. In some embodiments, the method further comprises sending new definitions and rules from the central server to the local repository.

[0037] Another aspect of the present disclosure provides a method for storing sequence base data, the method comprising: for a base location, storing information representing a most probable base of the sequence base data in a first location of a storage device, and storing a probability of a number of occurrences of the most probable base in a second location of the storage device.

[0038] Another aspect of the present disclosure provides a method for storing sequence base data comprising at least four possible bases, the method comprising: (a) providing a three-dimensional table structure in computer memory, which three-dimensional table structure is configured to store the sequence base data, wherein (i) a first dimension of the three-dimensional table structure stores information representing most probable measured bases of the genetic sequence base data; (ii) a second dimension of the three-dimensional table structure stores information representing potential bases of the genetic sequence base data; and (iii) a third dimension of the three-dimensional table structure stores information representing a base count probability for each of the at

least four possible bases of the sequence base data; (b) storing probabilities corresponding to an intersection of the first dimension, the second dimension, and the third dimension in the three-dimensional table structure.

[0039] Another aspect of the present disclosure provides a method for protecting biological data related to a subject, the method comprising: encrypting personal identification information of the subject using a first encryption scheme; encrypting phenotypes of the subject using a second encryption scheme; encrypting the biological data using a third encryption scheme, wherein the second encryption scheme or the third encryption scheme is different from the first encryption scheme; and storing the encrypted personal identification information, the encrypted phenotypes, and the encrypted biological data in computer memory.

[0040] In some embodiments, (i) the second encryption scheme is different from the first encryption scheme, and (ii) the third encryption scheme is different from the first encryption scheme, and (iii) the third encryption scheme is different from the second encryption scheme. In some embodiments, the method further comprises storing gene expression data of the subject. In some embodiments, the method further comprises storing geographic data of the subject.

[0041] Another aspect of the present disclosure provides a method for storing genetic data of a subject, the method comprising: storing personal identification information of the subject in a first storage segment with a first level of limitation of access; storing phenotype data of the subject in a second storage segment with a second level of limitation of access; and storing the genetic data of the subject in a third storage segment with a third level of limitation of access.

[0042] In some embodiments, the second level of limitation of access or the third level of limitation of access is different from the first level of limitation of access. In some embodiments, (i) the second level of limitation of access is different from the first level of limitation of access, and (ii) the third level of limitation of access is different from the first level of limitation of access, and (iii) the third level of limitation of access is different from the second level of limitation of access.

[0043] Additional aspects and advantages of the present disclosure will become readily apparent to those skilled in this art from the following detailed description, wherein only illustrative embodiments of the present disclosure are shown and described. As will be realized, the present disclosure is capable of other and different embodiments, and its several details are capable of modifications in various obvious respects, all without departing from the disclosure. Accordingly, the drawings and description are to be regarded as illustrative in nature, and not as restrictive.

INCORPORATION BY REFERENCE

[0044] All publications, patents, and patent applications mentioned in this specification are

herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference.

Brief Description of Drawings

[0045] The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings (also "figure" and "FIG." herein), of which:

[0046] [fig.1]FIG. 1 illustrates an example of a conductance-time profile of a sensor.

[0047] [fig.2]FIG. 2 illustrates an example of a schematic of a biological data management system.

[0048] [fig.3]FIG. 3 illustrates an example of a diagram of a distributed network for biological data management.

[0049] [fig.4]FIG. 4 illustrates an example of a schematic of a biological data management system where the central server is sitting in a central location.

[0050] [fig.5]FIG. 5 illustrates an example of a flow chart illustrating processes that can be executed by an application server.

[0051] [fig.6]FIG. 6 illustrates an example of a flow chart illustrating processes that can be executed by a local repository.

[0052] [fig.7]FIG. 7 illustrates an example of a base probability matrix for a 21-mer reading by a sensor.

[0053] [fig.8]FIG. 8 illustrates an example of additional dimensions of data kept for a read.

[0054] [fig.9]FIG. 9 illustrates examples of various sample identifiers.

[0055] [fig.10]FIG. 10 illustrates three examples of syntaxes.

[0056] [fig.11]FIG. 11 illustrates an example of a transitional syntax.

[0057] [fig.12]FIG. 12 illustrates an example of an application server input.

[0058] [fig.13]FIG. 13 illustrates an example of an application server output.

[0059] [fig.14]FIG. 14 illustrates an example of a distributed file system.

[0060] [fig.15]FIG. 15 illustrates an example of an architecture for segmented access control.

[0061] [fig.16A]FIGs. 16A, 16B, 16C, and 16D illustrate examples of a tiered storage access schemes.

[fig.16B]FIGs. 16A, 16B, 16C, and 16D illustrate examples of a tiered storage access schemes.

[fig.16C]FIGs. 16A, 16B, 16C, and 16D illustrate examples of a tiered storage access schemes.

[fig.16D]FIGs. 16A, 16B, 16C, and 16D illustrate examples of a tiered storage access

schemes.

[0062] [fig.17]FIG. 17 illustrates an example of a computer system programmed or otherwise configured to manage biological data.

Description of Embodiments

[0063] While various embodiments of the invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions may occur to those skilled in the art without departing from the invention. It should be understood that various alternatives to the embodiments of the invention described herein may be employed.

[0064] The term "subject," as used herein, generally refers to an animal, such as a mammalian species (e.g., human) or avian (e.g., bird) species, or other organism, such as a plant. The subject can be a vertebrate, a mammal, a mouse, a primate, a simian, or a human. Animals may include, but are not limited to, farm animals, sport animals, or pets. A subject can be a healthy individual, an individual that has or is suspected of having a disease or a pre-disposition to the disease, or an individual that is in need of therapy or suspected of needing therapy. A subject can be a patient.

[0065] The "genome," as used herein, generally refers to an entirety of an organism's hereditary information. A genome may be encoded either in deoxyribonucleic acid (DNA) or in ribonucleic acid (RNA). A genome may comprise coding regions that code for proteins or non-coding regions. A genome may comprise sequences of any or all chromosomes of an organism. For example, the human genome has a total of 46 chromosomes. The sequence of all of these chromosomes may collectively constitute a human genome.

[0066] The term "genetic variant," as used herein, generally refers to an alteration, variant, or polymorphism in a nucleic acid sample or genome of a subject. Such alteration, variant, or polymorphism may be with respect to a reference genome, which may be a reference genome of the subject or other individual. Polymorphisms may comprise single nucleotide polymorphisms (SNPs). In some examples, one or more polymorphisms comprise one or more single nucleotide variations (SNVs), insertions or deletions (indels), repeats, small insertions, small deletions, small repeats, structural variant junctions, variable length tandem repeats, and/or flanking sequences. Genetic variants may comprise copy number variants (CNVs), transversions, or other types of rearrangements. A genomic alteration may comprise a base change, an insertion or deletion (indel), a substitution, a repeat, a copy number variation, or a transversion.

[0067] The term "polynucleotide," as used herein, generally refers to a molecule comprising one or more nucleic acid subunits. A polynucleotide may comprise one or more

subunits selected from adenosine (A), cytosine (C), guanine (G), thymine (T), and uracil (U), or variants thereof. A nucleotide may comprise A, C, G, T, U, or variants thereof. A nucleotide may comprise any subunit that can be incorporated into a nucleic acid strand. Such a subunit may comprise an A, C, G, T, U, or any other subunit that is specific to one or more complementary A, C, G, T, or U, or complementary to a purine (e.g., A, G, or a variant thereof) or a pyrimidine (e.g., C, T, or U, or a variant thereof). A subunit may enable individual nucleic acid bases or groups of bases (e.g., AA, TA, AT, GC, CG, CT, TC, GT, TG, AC, CA, or uracil-counterparts thereof) to be resolved. In some examples, a polynucleotide may comprise deoxyribonucleic acid (DNA), ribonucleic acid (RNA), or derivatives thereof. A polynucleotide may be single stranded or double stranded.

[0068] Systems and methods described herein may relate to genetic data management. Genetic data management may comprise to network architectures, reports, definitions and rules, directives and actions, storage devices and storage management, privacy, encryption, or compression.

[0069] Various types of sensors may be used to measure different genetic attributes. Some sensors may record and report different levels of resolution. Some sensors may provide native base sequence. In some cases, the sensors may detect chemical modifications such as methylation, amination/deamination, oxidation, and/or any other modifications and abasic (AP) sites in DNA and RNA.

[0070] The sensors may be configured to detect various types of signals, such as optical signals, electrical signals, or a combination thereof. Optical signals may include fluorescence, luminescence, chemiluminescence, bioluminescence, incandescence, lasers, light emitting diodes (LEDs), visible light, infrared radiation, near-infrared radiation, or combinations thereof. Electrical signals may include electrical current, voltage, differential impedance, tunneling current, resistance, capacitance, conductance, or combinations thereof. Some solutions for genetic detection may alter native molecules to detect them. Some detection methods, such as polymerase chain reaction (PCR), may rely on amplification, in which many copies of an original genetic polymer may be produced.

[0071] Amplification processes, in turn, may introduce apparent mutation errors that may render results inaccurate. Other error sources, such as electronic noise, phase errors, spectral deconvolution errors, fluidic diffusion errors, quantitation errors, position in a read, sequence context, spatial and spectral optical cross-talk, may also be present, which makes various sensors or detectors differ in terms of signal quality, types of error, measurement accuracy, or alternative interpretation of sensed or measured data.

[0072] In managing these different types of genetic data, it may be important to manage information about the source of the data, how they were measured, and the sensors,

detection systems, hardware, consumables, chemistry methods, or software version used for measurement. Each set of data may comprise characteristic errors and uncertainties that may need to be accounted for in various situations.

[0073] Another issue in managing genetic data may be managing data storage. Different storage techniques and devices may be employed. Various types of specific storage media may be used, which may be designated in connection with a nature, quality, or quantity of the genetic data. Various types of genetic data, such as DNA or RNA sequences, may be stored in multi-cell storage devices. Blocks of memory may be used in various ways with respect to characteristics of the genetic data. For example, there may be a relationship between a size of a memory block and a type and size of data stored in the memory block.

Data Collection

[0074] One or more biological sensors may detect raw data of molecular chains. Each raw data read may be converted into a native formatted record of the read. For example, if a sensor senses and measures electrical conductance, the sensor may produce a time series of conductance over time as a chain passes through the sensor, as shown in FIG. 1.

[0075] Conductance raw data may be later interpreted into nucleotide base data or records in the case of deoxyribonucleic acid (DNA) or ribonucleic acid (RNA).

[0076] Raw data from a sensor may be passed to an application server. Data may depend on a sensor type and may be derived from an electric property, such as conductance, capacitance, current (e.g., tunneling current), voltage, resistance, or any combination thereof. Data may comprise optical data, such as optical data derived from fluorescence (e.g., chemifluorescence) or absorbance, such as by fluorescent label tagging or modification of subunits (e.g., nucleic acid bases).

[0077] Transfer of data from a sensor to an application server may be performed using a wireless module integrated with a sensor through a wireless protocol, such as wireless fidelity (Wi-Fi), Bluetooth, or near field communication (NFC). Transfer of data may be performed using a wired connection, such as universal serial bus (USB).

[0078] The application server may comprise a desktop computer, a laptop computer, or a mobile device such as a mobile phone (e.g., iPhone or Android phone) or a tablet (e.g., iPad or Android tablet).

[0079] The application server may have instruction sets that receive the raw signal data and produce base data using certain base-calling routines. These routines may be programmed and updated on the application server based on the capabilities and characteristics of the sensor or other global directives, as described elsewhere herein.

[0080] The sensor updates can be received or pushed from the sensor manufacturer, for instance, to improve signal measurement or to alter hardware or firmware.

[0081] As shown in FIG. 2, an application server, or central server 201, may comprise, or have access to, a dedicated database of definitions and rules that the application server or central receives from a local repository 202. The definitions and rules may be updated as needed. The definitions and rules may identify various situations and actions. For instance, there may be pathogen signatures or sequences or any other data associated with a specific pathogen that may be detected by the local sensor. As such, the definitions and rules may be custom-made and may be dynamic. The application server 201 may be in communication with a local master 205, which may serve as a resource for data that cannot be interpreted or concluded by the application server. The local master 205 may be in communication with a local slave 206, which may stay in the same facility but may serve a limited function with quick access to the local master. The local repository 202 may be in communication with end node 1 203 and end node 2 204, which may be measurement devices.

[0082] As an application server performs a measurement, it may compare its results with definitions and rules it has access to, and may subsequently suggest directives accordingly.

[0083] If no definitions or rules are available for a particular situation, the application server may communicate this situation with its local repository 202.

[0084] A local repository may comprise a server that is in network connection with one or more application servers, as shown in FIG. 3. The local repository 301 may comprise, or may have access to, a larger database and more definitions and rules, or more updated ones.

[0085] For example, the local repository may be in network connection with a central server 302. The central server may be in network connection with a number of local repositories 302 which may in turn be in network connections with local application servers 303.

[0086] As illustrated in FIG. 4, the central server may be located at a central location, such as a national laboratory or a health organization facility.

[0087] A role of the central server may comprise communicating or updating definitions and rules along with directives to a number of local repositories or receiving reports from them.

[0088] There may be several scenarios depending on the viewpoint from a certain machine. In some instances, one or more operations as shown in FIG. 5 may be performed with respect to the application server:

Sensor measures signals from a polynucleotide measurement 501;

Sensor communicates signal data to the application server 502;

Application server receives signal data and generates base data 503;

Application server identifies sequence data based on base data 504;

Application server analyzes sequence data with respect to definitions and rules received from a local repository 505;

Application server provides a message to the user based on the analysis 506;

Application server communicates sequence data to a local repository 507, if needed.

[0089] FIG. 6 illustrates possible operations performed by a local repository that may correspond to the set of operations described in FIG. 5 when an application server communicates sequence data to a local repository:

Local repository receives base data from the application server 601;

Local repository checks definitions and rules 602;

Local repository communicates abnormalities related to the base data to the central server 603;

Local repository receives global and regional updates from a central server 604;

Local repository updates definitions and rules 605;

Local repository communicates with Application Server new definitions and rules 606;

Central server communicates directives to the local repository; and

Local repository communicates directives to the application server.

[0090] The application server may be in direct or network communication with the local repository. The local repository may periodically send updates to the application server that the local repository has received from the central server.

[0091] The central server may be located at a central laboratory or a health center, and may analyze sequence data communicated by the local repositories. The central server may have access to a database of sequences.

Example: Pathogens

[0092] A database of sequences may comprise a database of pathogen sequences. The central server may have faster access to recent pathogen sequences reported by using a faster memory and communication pipeline.

[0093] When a local repository receives information that may relate to a possibility of a new pathogen or a harmful known pathogen, the local repository may look for definitions and rules provided by the central server that may be related to the received sequence in a dedicated database. Based on a comparison of the received sequence data with sequences in the dedicated database with specific definitions and rules, the local repository may take appropriate options accordingly. For instance, the local repository may find specific rules and then pass specific directives to the application server.

[0094] Alternatively, if the local repository's definitions and rules meet a certain set of criteria, it may communicate the received sequence to the central server.

[0095] The central server may have access to a larger database, such as a comprehensive central database of recent and/or older breakouts. The central server may continuously

update the central database based on what the central server collects from a plurality of local repositories.

[0096] The central server may be accessed by a central laboratory or a health center, where health or safety professionals have access and are alerted about events with specific predetermined thresholds.

[0097] Various decisions may be made by an authority running the central server. These decisions may comprise automatic or semi-automatic decisions. For instance, if the central lab determines that a certain sequence is not dangerous, the central lab may communicate to the local repositories a decision to ignore such instances. Alternatively, if there is an indication of a more serious situation, the central server may add the flagged sequence to a directive dedicated to such instances and keep the directive for faster access in a memory. Some subsequent instances reported to the central laboratory with a same or similar pattern may receive the same directive. The directive may comprise a decision regarding a medication, a quarantine, a rest, etc.

[0098] When a central lab has addressed and categorized a situation, the central lab may then establish definition and rules related to the situation. These definitions and rules and directives may then be communicated to local repositories of relevance. For instance, if a geographic outbreak is concluded, the central server may update any or all of the local repositories that are in connections to end users and application servers related to the area, while putting other areas in a vicinity of the area on alert.

[0099] In relation to food safety, a plurality of sensors in different locations may measure sequences from various types of food. The sensors at these locations may measure sequences and may search for pathogen candidates. Each sensor may be in communication with an application server. A sensor may measure signals from a sequence and send raw data to the application server.

[0100] The application server may comprise a set of definitions and rules. When the application server receives raw data from a sensor, the application server may run a program to produce base reads from the raw data and sequence contigs from the base reads. After the sequence contigs have been produced, the application server may run a program that compares the base data or sequence data with pre-established definitions and rules. These definitions may be in a database that the application server has access to. The definitions may be stored remotely on a dedicated server. There may be a subset of definitions that are designated as particularly important or crucial. For example, there may be a set of recent or current pathogen information. These particularly important or crucial data may be stored in a faster access memory or storage that the application server may have access to readily. In some situations, the application server may be instructed by a directive or a rule to search for a specific pattern. For example, this specific pattern may be related to current breakouts or

reports from other sensors that may have indicated a pathogen in a similar type of food (e.g., produce).

[0101] The application server may be in network communication with a local repository. A local repository may serve a number of application servers with definitions and rules and may provide directives to the application server. The local repository therefore may periodically sends updates to the application servers.

[0102] If an application server does not find a proper definition or rule for a specific case, the application server may send the sequence data or other biological data to the local repository. The local repository may then search a broader database to which it may have access for definitions or rules. This database may be shared amongst one or more local repositories. The database may have a larger collection of known pathogens, for example, or may have some pathogens related to historical outbreaks that have not been observed for some period of time. Alternatively, such pathogens may not have been observed in a vicinity of the sensor location but the local repository may have access to a database that records the pathogens and therefore may be aware of them.

[0103] In special cases, the local repository may take any of multiple options. For instance, the local repository may look up definitions and rules related to the pathogen and communicate it along with certain directives to the application server. Alternatively, the local repository may communicate the data to a central server.

[0104] A local repository can have its own definition and rules which it receives from a central server. A central server can be in network communications with a number of local repositories. Accordingly, the central server can update definitions and rules at a local repository on a regular basis.

[0105] If a local repository cannot find any definition or rules for a particular case, the local repository may opt to communicate the data to a central server. A rule may require the local repository to report any base data, sequence data, or biological data that may indicate a special case.

[0106] A central repository may be located in, used in, or used by a central laboratory comprising researchers or health professionals. For instance, a national or international health center may be in control of the central repository. When a special case has been detected and communicated from a sensor to the central server, the central server may have access to a large set of definitions or rules to handle the situations. Optionally, upon reaching certain predetermined thresholds or at user discretion, researchers or health professionals may assess a situation to determine a severity of the situation.

[0107] A single sample may produce a plurality of gigabytes of raw analog conductance information representing millions of reads of sequence information. The initial interpretation process may consume these analog readings and may filter out background noise when no molecules are passing through the molecular sensors or when con-

taminants are causing unreliable or invalid results. The interpretation process may interpret and translate data into base sequence strings. Each base determination may be associated with one or more dimensions of data. For example, a dimension, or vector, may indicate a probability rating for what base it is reading, as shown in FIG. 7.

[0108] FIG. 7 shows a base probability matrix for a 21-mer reading by a sensor capable of sensing abasic (AP) sites or one of five possible bases. The determined base sequence 310 may represent a highest probability base at each location in the read. The possibilities of abasic sites or bases may comprise:

A = Adenine

B = abasic site

C = Cytosine

G = Guanine

T = Thymine

U = Uracil

[0109] Each column shows a probability of a specific nucleotide base at each location in the sequence. The sensor end node or an application server may interpret the probability for each possible base at each location. For example, this figure shows Cytosine (C) as the most probable base at the 16th base location.

[0110] FIG. 8 illustrates how additional dimensions of data may be kept for a read. In this illustration, the modification table shows, at each base location, if the base is methylated, oxidized, or acylated. In this example, the third and fourth bases comprise a 5'-C-phosphate-G-3' (CpG) pair that is methylated. The Cytosine (C) is also believed to be oxidized. The associated base probability table shows the determined base sequence. The distance table, or transition location table, contains the distances, in number of bases, between transitions to a new base giving the determined length of the homopolymers. The example shows a run of approximately two Thymine (T) bases before transitioning to an Adenine (A). It also shows two Adenine (A) bases before transitioning to a Guanine (G) later in the sequence. Storing dimensions of data for a read may address the type of sensor with intrinsic uncertainty regarding the number of same-type bases in a sequence or a sub-sequence.

[0111] Other dimensions may include an overall length and a base location as a distance from the beginning of the read. Some sequencing techniques start at one end of an oligonucleotide (oligo) and perform sequencing by synthesis (SBS). Such processes may involve looking for base incorporation after each round (e.g., one at a time). As such, there may be a possibility of generating phase errors each time a base is incorporated. For instance, if there is a clonal population, incorporation of the bases may be non-uniform across the population. Certain members may incorporate more than one base, while others may not incorporate a base. As such, confidence may decrease

farther along the sequence read. A fourth dimension may incorporate a distance, in number of bases, base paired ends, or base transitions from the primer cleaved end of a sequence being analyzed.

[0112] Raw data reads may be kept for further analysis. For example, one may want to improve sensitivity by detecting polymeric creep, phototoxicity, a presence of contaminants affecting the sensors, or atomic structural changes to tips of nano gateways. The uncertainty in base call may be specific to the make and model of sensor used.

[0113] For instance, the interpretation process controller may pass each filtered conductance recording to a single interpretation worker process or thread. Each raw reading may be interpreted without concern for locking, since there may be no shared data. Synchronization may be unnecessary, since the processes downstream of interpretation may execute multiple times on the growing interpreted sample data set until the interpretation reaches its finished state with an acceptable degree of confidence.

[0114] Further, the system may incorporate sensors from different vendors to use various technologies to sense a sequence. In some cases, the raw information may not available. Instead, reads may be available from the sample where the probabilities and induced errors are specific to the technologies used. Each technology may have strengths and weaknesses, and may have various levels of sensitivity. Each technology may have various resolutions to various aspects or dimensions of reading DNA or RNA sequences. Some technologies may be highly sensitive to transitioning from one base to the next, but less sensitive to a particular base of interest. In this case, it may be desirable to conduct further analysis on the base reads.

[0115] Some technologies may be particularly good at base determinations, but less strong at determining base movement or transition. This situation may result in a high probability that it is looking at a particular base, but provide less certainty regarding the number of bases and when they repeat. Yet another technology may read each base along an oligo (e.g., one at a time) with an additive error model, such that the farther away from the starting marker, the less certain of the base being sensed.

[0116] Hence, various embodiments support interpreting sequence base data in various styles and formats for files and records when stored in non-volatile memory. For example, the data from a sample in an eXtensible Markup Language (XML) or JavaScript Object Notation (JSON) file may be stored on a distributed file system.

[0117] The file may comprise reads stored as a single base value for each nucleotide in the chain. The reads may be stored as a probability value. Alternatively, the reads may be stored as a complete probability matrix for each possible base at each nucleotide location. A possible syntax may comprise using one or more attributes to describe the meta-data syntax for what is stored in the read record.

[0118] There are various examples of semi-structured read formats with which various em-

bodiments are capable of interpreting and working with, based on various factors involved in collecting the sample. Examples of such factors may include sample preparation, make and/or model of the sensors, or analysis of the data. Sample files may comprise a simple and basic schema comprising a unique sample identifier with one or more base reads.

[0119] FIG. 9 shows examples of a sequence read, a base format read, and syntax. Part A shows a read comprising the determined base sequence. Part B shows an example of the same base format read including probability data for each base. The syntax for this second example comprises each word describing a single base. For example, the word "C67.74" describes the third base as a Cytosine (C) with a probability of over 67%.

[0120] The third example, shown in Part C, shows the same base format read with each word describing a single base location. In this example, each word describes a base, a probability, and any modifications. For example, the word "Cf67.74" describes the third base as Cytosine (C) with a 67% probability. Modifications may be recorded into each word by adding a lower case letter after the base. In this example, a lack of following lower case letters indicates that the base was not methylated, oxidized, nor acylated. The lower case letters "a" through "h" can be translated into the numbers 1 through 8 to hold a bit mask of the modification table. Methylation equals the most significant bit (MSB) (4), oxidation is (2), and acylation is the least significant bit (LSB) (1). Hence the Cytosine (C) base, modified by "f", shows the Cytosine was methylated and oxidized.

[0121] In accordance with the systems and methods described herein, it is possible to maintain secondary and tertiary possible base values, any modifications to those bases, and any other sensor-recorded dimension of data. FIG. 10 represents three examples of syntax for storing (A) each of six tracked base or AP site possibilities; (B) the highest two most probable bases or AP site possibilities; or (C) only maintaining an array of base location probabilities if the probability exceeds a certain predetermined threshold. In the first example shown in Part A, the file stores probabilities for each of the six bases and probability values for the third base location in the read as cytosine (C) having the highest probability at over 67% and an abasic site having the lowest probability at under 2%. If only the two highest probable base values are maintained, that base location may be seen as a primary cytosine (C) base and alternatively a thymine (T) base with a probability of approximately 14%, as shown in Part B.

[0122] Storing probabilities only if they exceed a predetermined threshold may be accomplished with a length/value syntax, shown in Part C. A base location with two base possibilities that exceed the threshold of 15% may result in a lead number "2" as the first character of the word "2C64.46", which also provides the length of the array of bases kept for that base location. Cytosine (C) is the highest probability at 64%, and

guanine also exceeds the threshold at 15%.

[0123] A transitional syntax for sensors that record a distance dimension between base transitions may also be used, as shown in FIG. 11.

[0124] The application server may collect millions of reads from a sample. It may then identify longer aligned sequence, or contig, data from analysis of the reads. For further evaluation, the application server may perform an alignment of the base reads against a reference. Alternatively, the reads may be grouped with several other reads and used in a de novo assembly. The application server may be extensible such that it may call other processes that accept only a subset of the information stored in the semi-structured format of the reads. For example, the interface to the alignment processes may accept a FASTA formatted syntax or a FASTQ formatted syntax for the reads. In this situation, the read may be translated into a format understood by the alignment processes.

[0125] For instance, the example read described in FIG. 12, when translated into a FASTQ format, may look similar to the following four lines:

@10032QB:11578:1.1:20151221:09:42:37
ATCGTCGAGBAGTTACAAGCT
+10032QB:11578:1.1:20151221:09:42:37
'*&*'+%+)&(%'(&&)&&&&(

[0126] The bases and a corresponding Phread quality score may be sent. The reads may be interpreted and contigs may be returned from the consensus algorithms of the alignment processes. A sample may contain millions of reads. Reads may be either aligned against a reference sequence or assembled de novo. This translation of base reads into a different syntax may lose some context or resolution of the base reads. In an example shown in FIG. 13, the indicated sensors are able to capture transition distances and chemical modifications in addition to the base sequence and probability or quality score sent and returned by the programs that align the reads into contigs. The application server may take the alignments and, when the consensus is determined, reapply some lost context or resolution back into the sequence contigs, such that the contigs are stored in a similar semi-structured syntax as the reads. For example, for a contig derived from base reads that contain chemical modifications, the application server may reapply any modifications not used to sequence the reads.

[0127] The application server may analyze sequence contig data with respect to definitions and rules received from a local repository. An installation may be distributed with end nodes, servers, and/or repositories that are networked and cooperating to manage and act upon sequence data acquisition. In an aspect, the application server may incorporate rules to discover and act upon genetic sequence information with high efficiency. Sequence discovery may be directed to find a pathogen. In other cases, one

may want to discover contigs for certain gene expressions. Various embodiments allow one, such as a microbiologist, to administer a database of sequence definitions for the pathogens or genes. Rule definitions may be assigned to, or associated with, a specific directive or set of directives.

[0128] The central controls and rules management module may process these rules. In some cases, they may translate the rule or further modify it, such that it runs on specific downstream servers and nodes. Many rules will be distributed themselves.

[0129] For example, a rule may comprise a simple sequence, a matching method, a weighting, one or more regression adjustments, or directives to bundle the sample information into a National Center for Biotechnology (NCBI) compliant BioSample and to notify a department head.

[0130] The instantiation of the system in this example may include a basic sensor, a local node, and/or a local server. Rules may be adjusted to a specific piece of equipment where it executes. An application server may attempt to discover a sequence from each individual read or contigs. The discover portion of the rule may be better served by modifying the higher level rule to more effectively discover the sequence based on a make or a model of the sensor used. The rule at a high level may be to align a sequence to a contig with less than a predetermined number of variances based on the type of sequencing equipment used. In some cases, a global method and valuation may be used, while with other sequencing equipment a local method and valuation may be applied. Alternatively, the sequence to contig mapping may have a threshold variance level based on a flowgram, e.g. if the sensor used was a Roche 454.

[0131] In an embodiment, rules may be distributed and may comprise cooperation with dedicated application servers. This may allow for more accurate results with fewer false results without adversely affecting overall performance of the end sequencing equipment. For example, an installation may have a plurality of sensor nodes testing food samples:

These read signals are sent to an application server for interpretation into base reads and subsequently contigs.

This initial application server executes a rule with a simple lower processing cost sequence alignment algorithm on each base read against an array of pathogen signatures.

If a threshold for a number of close matches or score is met for one or more of the pathogens, then the directive may include:

extending the sampling at the sensor; and/or

bundling the complete sample and forwarding it to a dedicated pathogen testing application server for a more rigorous interpretation of the sensor measurements.

The pathogen testing application server may then apply its own directives based upon

its findings.

[0132] This embodiment may ensure the information is protected, both when the information is being communicated across networks and when the information is stored in a repository.

[0133] For data in transit, encryption schemes such as secure socket layer (SSL) or transport layer security (TLS) may be applied. Data may be produced at the sensors. These end node sensors may support connections to local application servers, which analyze the raw data into base reads. The application server may further analyze the base reads into contigs or sequences. Alternatively, the application server may communicate the reads to another application server to create the base reads and sequences. Communications between sensors and application servers, between cooperating application servers, between application servers and repositories, and between application servers and services may support secure sockets layer (SSL) or transport layer security (TLS) connections. This may include servers that associate base reads and sequences with other meta data, such as names or geographic locations, and apply rules and directives.

[0134] For data at rest (e.g., not in transit), various mechanisms may be used to protect the data. Data may be stored in a plurality of locations. Sample data may be stored in a file system. Each sample may comprise a semi-structured data file. A process may perform marshalling, unmarshalling, and/or removal of sample files.

[0135] Derived contig or sequence data may be stored in a similar way as a plurality of semi-structured files. Contig data may be kept in a distributed file system, since the contig data may comprise a large data set, may be continuously mined and analyzed to test hypotheses, and may require a repository that can support access with high parallelism. As with sample files, a process may perform marshalling, unmarshalling, and/or removal of contig files. These files may be anonymized. The encryption and compression mechanisms may be tuned for lower central processing unit (CPU) costs of access and higher throughput in reading.

[0136] When sequences are stored into a repository, only an identifier may be associated with the contigs. They may be de-identified with respect to the subject, location, contact information, or study corresponding to the sample. The identity data may be stored in a separate repository from the sequence. Likewise, base reads from samples may be associated only with an unique identifier. If raw data is retained, it too may only be associated with an identifier. Identity data may be placed in a separate database. The identity data may be kept in a relational database. A sample-identity and contig-identity reference table may be maintained to allow the linkage to re-identify a pair of a sample and a contig if access controls allow. A different set of access controls may be applied to the anonymized samples. Both the identity data and the sequence data may be encrypted at rest.

[0137] Sample data, contigs, and sequences may represent relatively static data sets. Upon being added to a repository, they may be seldom updated. They may represent as much as petabytes (e.g., millions of gigabytes) of data. Analytical processing of these extremely large data sets may be enabled through the use of a distributed file system storing protected semi-structured data sets that may be accessed and reduced through processes, such as MapReduce or Spark, into working transactional or columnar databases.

[0138] For instance, FIG. 14 illustrates an example of a distributed file system where the information is retained in three separate storage systems - one each for samples 1401, contigs 1402, and working data 1403. Raw sample data 1401 may be interpreted and translated into a semi-structured format consisting of the molecular reads along with simple or basic meta-data concerning the sample. The basic meta-data may comprise a sample identifier. All other meta-data regarding the sample may be considered working information. Working information may be stored separately in a database with a reference to the sample identifier. Once processed, sample data may or may not be retained. If sample data is retained for long periods of time and is used or accessed for other purposes, it may be stored in a distributed file repository 1404. Alternatively, if sample data is retained for long periods of time but is not commonly accessed and used for other purposes, it may be archived.

[0139] Sample data may be further interpreted, aligned, or assembled into sets of contigs or sequences. These contigs may be stored in a distributed file system 1404, in a semi-structured format, such as XML or JSON, with an assigned a contig identifier. In a similar manner as sample data, other meta-data regarding the contig may be working information and may be stored separately in a database with a reference to the contig identifier.

[0140] Contigs also may have working data. Working data may comprise additional data captured and used other than the reads and derived contigs. This may include information regarding the process involved in capturing the information, such as a make, model, or serial number of the equipment used; sample preparation information; source information; a location at which the sample was obtained; and protected health information such as names and contact information of a patient.

[0141] These sample data and contig data files may be compressed to increase capacity, with the understanding that in doing so, there is a computational cost incurred when reading the files. These files may be encrypted. As the information within these files may be anonymous, an embodiment uses an encryption algorithm that employs a high-performant (e.g., secure) decrypting counterpart. Hardware cryptographic accelerators may be employed to minimize encryption and decryption costs.

[0142] Working data may comprise additional information stored in order to re-identify or

work with samples and contigs. The working data also may include a phenotype schema with associations between identities, sequences, and phenotypes 1405. Working data also may be encrypted. However, whereas performance may be an important factor in deciding which algorithms to use, security may be an important factor for the working data. Further, fine grain security and access, such as record-level access, may be implemented for working data.

[0143] The sample storage and the contig/sequence distributed storage may encrypt the semi-structured files using a symmetric key. Application server processes responsible for marshalling and unmarshalling the files may maintain a list of ciphers for files in a secure wallet. Additionally, hosts upon which the application server processes are running may include an accelerator, such as an Intel Advanced Encryption Standard - New Instructions (AES-NI).

[0144] Among the benefits of the embodiment may be that the repository is modeled to maintain and provide necessary tools to access and mine a large collection of bioinformatic information that the repository is capable of storing over a long period of time in an anonymous context. The anonymous contigs and optionally initial sample data may be retained and may be securely made available to researchers in improving understanding of genetics.

[0145] In some embodiments, a physician may be able to access a patient medical record comprising both the genetic contigs linked to the associated working information. In this example, the physician is within an application that provides two different types of accesses: a performant access to specific contig and sequence sets and a secure access to the working data linked to the contigs and sequences.

Example 1: Research

[0146] In research contexts, raw data of samples from a plurality of sensors of various manufacturers are sent to an application server. The application server interprets the raw data and determines the base sequences of a portion of or all of the reads in the raw data. The application server then either performs the alignment analysis itself or formats the reads into a syntax understood by an external alignment analysis server tool to which it calls out. The resulting contigs are returned from the external server to the application server.

[0147] In some cases, the application server re-applies information from the sample reads back into the contigs. The re-constituted contigs are tagged with an identifier and transmitted to the contig repository, where they are saved as semi-structured files in the application server's distributed file system. Additional information, such as source, identity, location, and/or address, related to the contigs are inserted into the repository's working database.

[0148] Additional meta information may be incorporated in the semi-structured files, such as

taxonomy, to allow for efficient storage in the distributed file system or to reduce the data during an extraction. The repository of contigs grows over time.

[0149] A researcher hypothesizes on relationships between specific genetic signatures and a cause or probability of some expression of one or more phenotypes. The contig repository is mined. Specific signatures and their associated identifier are extracted as independent variables and loaded into a database for testing the researcher's theory.

[0150] Signatures may then be mapped to phenotypes obtained from external sources.

[0151] Hypotheses that prove useful may be saved and incorporated into an application server in a separate database 1406 of gene signature associations to gene expressions and phenotypes.

[0152] Semi-structured files are encrypted, as is the database. Access is controlled to the level of the sample and contig identifier.

[0153] Sample and contig information may be retrieved without working information with a different level of security. For example, a researcher may be allowed access to all the contigs in the system, but not to any contig with its associated working information.

[0154] Access control is abstracted and may support concepts such as group and role security. Fine-grain security with abstract controls provides effective security and privacy over time. As an example, employees of a medical group may access an embodiment that stores bioinformatic information on a portion of or all of the patient members of the medical group. Over time, the doctors responsible for a particular patient may change. Doctors may have access to only the bioinformatic information of patients for whom they are currently responsible.

[0155] Access is granted through strong public/private key management systems and provides support for nonrepudiation.

[0156] A management program may manage the nodes and users of the system. The management program may incorporate certificate authority services for issuing keys and maintaining the certificate revocation list. Processes running in the end node sensors, application servers, and distributed file system manager have public/private key pairs that allow them to act upon the information. Users also have generated key pairs. A user may have multiple key pairs associated to his account to support authentication from a plurality of different computers, tablets, or other computing devices.

[0157] The concept of roles or groups is supported. Accessing stored data is controlled by roles, while a currently active user may belong to one or more roles.

[0158] This architecture and abstraction of access controls for data at rest has the added benefits of ensuring a portion of or all sequence information is secured and made available only authorized entities over the life of the data records. FIG. 15 shows an exemplary architecture illustrating segmented access control.

[0159] Access control is capable of being fine grained, e.g., to the individual sample level.

Each sample may be tagged with a unique identifier.

[0160] For jobs that are not crucial in nature, a low-level sequencer or biological sensor may be used. A low-level sequencer or biological sensor may not require a large permanent storage device. Examples of such a device may include measurement or data acquisition modules. Such a device may have measuring hardware, a processor, and/or a system memory for handling system functions. Each of these components may have its own buffer memory for handling its own functions.

[0161] A low-level sequencer may require a communication link to relay its raw data to higher-level device such as an application server, a local repository, or a local server.

[0162] The communication link may comprise a near-field communication protocol, such as Bluetooth or near field communication (NFC), or a wireless protocols such as Wi-Fi. The communication link may comprise a cabled (e.g., wired) communication provisions such as USB. In some cases, the communication link may comprise a satellite or a cellular communication module.

[0163] A low-level sequencer may be integrated with an application server that may be operating on a mobile device such as a mobile smartphone to perform some of these aforementioned functions. For instance, the low-level sequencer may comprise measurement hardware and use mobile device capabilities and applications as a local memory, processor, and communication link.

[0164] Alternatively, a mid-level sequencer may be used in more critical circumstances. Examples of such critical circumstances may include monitoring of patients and point-of-care applications where an initial diagnosis is needed.

[0165] A mid-level sequencer may perform more accurate measurements of a polynucleotide. The accuracy may be set according to what is needed for a reliable accurate judgment of a sequence.

[0166] A mid-level sequencer may use a memory device and a communication component. Hence, the mid-level sequencer may include measurement and data acquisition modules with measurement hardware, a processor, and a system memory for handling system functions. Each of these components may comprise its own buffer memory for handling its own functions.

[0167] The additional memory device may comprise a flash memory (e.g., multi-level cell flash memory) capable of storing bits of data. The data in a mid-level sequencer may be base data, in which case a multi-level cell flash memory may be suitable to store the data locally. A port such as a USB port may be used to transfer the data, e.g., in cases where there is a lot of data such that a wired connection may be desirable for high bandwidth or throughput purposes.

[0168] In an embodiment, a multi-level cell device such as a flash memory is used as a relatively fast way of storing and accessing genetic sequence data. In a flash memory

storage device, a large number of cells may be used to store data based on floating gate field-effect transistors (FETs) that are capable of holding a charge. Cells may be programmed individually by charging the floating gate of each FET.

[0169] One advantage of this embodiment is due to the fact that flash memory cells may be erased in blocks, via block erase operations, thereby erasing all charge of all of a plurality of floating gates in a single operation.

[0170] This embodiment may also have a characteristic that individual cells are not erase-addressable. However, in this embodiment, an erasable block of the flash memory is used to store genetic data related to a sequence of bases, nucleotides, or otherwise contiguous genetic data. In case one needs to replace this erasable block, a user may typically wish to erase all of the data in the erasable block at once, rather than a portion of the erasable block. This embodiment therefore may allow flexibility of optimizing cost versus speed for genetic data storage.

[0171] In a flash memory storage device, cells may start to fail after a number of program and erase cycles, after which point, reading or writing may fail. This fact can be used advantageously for genetic data storage. Since the number of erase cycles of a flash memory may be limited, the data may be kept safe for a longer time than some other usage scenarios.

[0172] There may be specific relationships between erase block size and sequence or otherwise genetic data size. This may ensure integrity of the data related to the whole sequence.

[0173] As a specific example, a sequence of bases consisting of 128 kilo base pairs (kbp) is stored in an erase block of 128 cells:

CTT...GAG (128k bases)

= = = . . . = = = (128k cell erase block)

[0174] For native DNA and RNA bases, a two-bit multi-level cell (MLC)may be dedicated to each base. For instance, for the case involving DNA, one uses:

A(00) C(01) G(10) T(11)

which means, both the first and the second bits are off when the base is an A, the second bit is on when the base is a C, the first bit is on when the base is G, and finally both the first and the second bits are on when a base is T. A similar scheme may be used for RNA.

[0175] Each erase block may be designed or configured to store multiple sequences. Alternatively, a larger sequence may be stored on a specific number of erase blocks with similar or same properties and life cycle.

[0176] Differently-sized erase blocks may be used for differently- sized sequences. For instance, flash memory devices of a smaller erase block size may be used to store oligo data or hybridization data, while flash memory devices of a larger erase block size may

be used to store genes and mutations or reference genes. Flash memory devices of a large block size may be used to store genome data.

[0177] An advantage of using flash memory for faster access may be compromised by life cycle issues. A copy of flash memory content may be mirrored on a storage server with slower access but longer life cycle. A test may then be devised to probe the integrity of data in each block size. Occasionally the data in each block may be tested against the mirror data in the server. Should the flash memory erase block data show any sign of degradation, that block of the flash memory devise may be decommissioned.

[0178] This embodiment may be advantageous at least since the longer life cycle storage device may be, for instance, a remote hard disk drive (HDD) storage server in the cloud.

[0179] In a further example, an erase block of a flash memory storage device may be used to store sequence data plus some metadata:

CTT...GAG (96k bases) - Metadata (64k bit = 32 k cell MLC)

= = = . . . = = = (128k cell erase block)

[0180] Examples of metadata may include any information related to the origin of the sequence, such as name of a patient, other information related to a patient, or the sequence itself.

[0181] A shorthand of the biological data may optimize the size of the data with respect to the storage device architecture, for example, by using a compression or the biological data. The size of the compressed data may be fine-tuned for better storage device compatibility.

[0182] A hash table may be made of different biological data. Each hash may correspond to a one category or genre. For instance, in case of proliferation of pathogen data, one may build a hash for each pathogen and use a hash table. Whenever a new sample is measured, performing a hash of the new sample may readily find a match within the hash table. This is a fast and efficient way of obtaining information about the pathogen.

[0183] A multi-level cell (MLC) storage cell may store two bits. The two bits may be used to store information about a base of a polynucleotide. For example, for a DNA base, the following bit configurations may be used:

00 A

01 C

10 G

11 T

[0184] In this way, all native four bases may be represented using a single memory cell. This approach may be advantageous for ensuring integrity of data.

[0185] In another example, an MLC storage cell may store three bits. The three bits may be used to store information about a base of a polynucleotide with additional information

indicating methylation or oxidation status. For example, for a DNA base, the following bit configurations may be used:

000 Native A

001 Native C

010 Native G

011 Native T

100 Oxidated A

101 Methylated C

110 Abasic

111 Other information

[0186] In this manner, multi-cell memory devices such as flash memory and phase change memory may be used.

[0187] In case of data degradation in a storage device with blocks with multiple cells, loss of data may be avoided by providing a warning, by refresh cycles, or by automatic or instigated dumping of data into a storage server, e.g., a HDD, or into a cloud storage server.

[0188] Erase blocks in a flash memory device may be used for ease of access and storage management. When all the data on an erase block corresponds to a biological unit, for example, a DNA or RNA sequence, memory access may be economized and data may have more integrity. This may lead to power optimization in large-scale operations where many sequences areas or genetic data may be accessed and may be operated on in a short time.

[0189] Data integrity may be preserved through this embodiment by keeping all the data relevant to a certain genetic unit, such as a gene or a contig, in a certain unit or units of memory. In addition, other benefits such as processing, optimization, and reducing generated heat may be achieved. It is envisioned that data management, data compression, memory access, temperature control, and data integrity may have a positive net effect on the entire ecosystem of biological data management, whether local or global.

[0190] A memory block, such as a flash memory erase block, may be chosen to be compatible with the size of the genetic data. Toward this end, customized compression and variance analysis may be performed to make the compressed size of the genetic data more optimal to the size of a memory block or a memory bank. The optimization may be performed in terms of data loss and data preservation. For example, in case a memory unit size, such as a block size or bank size, is larger than the size of the biological unit data, the rest of the memory space may be used to store additional information about the biological unit data. For example, an erase block in a flash memory may be used to save gene information, while additional information about the

gene, such as gene expressions, may be saved on the remaining space of the block.

[0191] Access to biological data may be managed through a tiered storage access scheme, as shown in FIG. 16A. An application may be on a local repository or central server. First tier access may be achieved through using a fast memory. In crucial cases, a random access memory (RAM) 1601 may be used to access certain data that needs to be frequently accessed. In less crucial systems, the fast memory may comprise a flash memory 1602 in or adjacent to a local HDD or a cloud-based storage unit.

[0192] The decision to retain certain biological data may be based on a hit-or-miss architecture. When a certain number of hits are registered, a processor may access the biological data and may escalate it to faster memory (e.g., by copying or moving the biological data). For example, upon detecting a report of instances of a pathogen, a local repository or a central server may decide to bring a copy of the pathogen to local memory. Further upon identifying specific regions of the biological data unit that may be of importance, a copy of the specific region may be maintained in faster memory and the rest of the data unit may be kept at a lower level in a slower memory, for example HDD, cloud, or equivalent 1603. FIGs. 16B, 16C, and 16D provide additional examples of storage architectures. FIG. 16B shows an example of an architecture suitable for providing super fast data access and decision making, in which a processor can be configured to communicate with a RAM, a flash memory, and/or an HDD or equivalent. FIG. 16C shows an example of an architecture suitable for providing fast genetic access and decision making, in which a processor can be configured to communicate with a flash memory and/or an HDD or equivalent. FIG. 16D shows an example of an architecture suitable for providing genetic archiving, in which a processor can be configured to communicate with an HDD or equivalent.

Example 2: Privacy Encryption

[0193] An example is provided of an encryption technique applied to genetic sequence data for an imaginary person by the name of Michael Smith and a 16-mer sequence related to him. The 16-mer may be a part of a larger sequence, gene, or genome related to the person.

Michael Smith - ... t t g c g a t g t c t a a t g g ... (subject sequence)

[0194] In this example, the name "Michael Smith" is encrypted using a 24-bit cypher for the purpose of illustration. The encrypted name and corresponding syntax are expressed as:

Encrfn ("Michael Smith", cypher1) =

EnCt2568e6c561c2b3a78926b5dbb3adea5ba827c065e568e6c561c2b3a78926b5dbbJIGwNtmg0ACHd+Q9e1ZHTMJV2DqVe3XSDb77IwEmS

[0195] This approach may ensure privacy of the name, as long as the cypher is secure. This type of encryption and subsequent decryption and cypher protection is potentially com-

putationally intensive and costly. It may be appreciated that, in this example, the name of a person, which may be comprise few bytes, may grow by a few hundred bytes if extensive encryption is used.

[0196] To ensure privacy of the sequence, it may be assumed that there is reference sequence containing:

t t g c g a a g t c t a a t g g … (reference sequence)

[0197] The bold and underlined base is assumed to be the only varied base in the population.

[0198] Then, it may be assumed the original sequence taken from Michael Smith contains the following:

…t t g c g a t g t c t a a t g g … (subject sequence)

[0199] According this embodiment, this sequence is stored as:

…t t g c g a a* g t c t a a t g g … (subject sequence representation)

[0200] where * may be a number from 0 to 3, thereby giving:

a0 = a

a1 = c

a2 = g

and

a3 = t

[0201] In the case of Michael Smith, this number is taken to be 3, shifting an "a" to a "t".

[0202] This example shows that the sequence

…t t g c g a a(0123) g t c t a a t g g …

may represent the entire population with an expense of a two-bit character, in this case (0,1,2,3).

[0203] Since the rest of the sequence is identical for the entire population, according to this embodiment, complete privacy of the sequence may be achieved with an expense of a 2-bit key.

[0204] In this example, a portion of a oligo or contig is presented where only one base is variable compared to a reference oligo or contig.

[0205] In this example, to encrypt this sequence, the reference sequence is assumed plus a 2-bit code (123) that may shift one base by 1-3 places according to an encryption scheme, e.g.:

a c(1) g(2) t(3)

[0206] If the encrypted variable base was a "g", for example, the shift function in the encryption code may give:

a(2) c(3) g t(1)

[0207] Similar schemes may be used without departing from the scope of this embodiment.

Computer control systems

[0208] The present disclosure provides computer control systems that are programmed to

implement methods of the disclosure. FIG. 17 shows a computer system 1701 that is programmed or otherwise configured to manage biological data. The computer system 1701 can regulate various aspects of data management of the present disclosure, such as, for example, the collection, storage, encryption of biological data, communication between servers, servers and repositories with respect to definitions and rules, and management definitions and rules. The computer system 1701 can be an electronic device of a user or a computer system that is remotely located with respect to the electronic device. The electronic device can be a mobile electronic device.

[0209] The computer system 1701 includes a central processing unit (CPU, also "processor" and "computer processor" herein) 1705, which can be a single core or multi core processor, or a plurality of processors for parallel processing. The computer system 1701 also includes memory or memory location 1710 (e.g., random-access memory, read-only memory, flash memory), electronic storage unit 1715 (e.g., hard disk), communication interface 1720 (e.g., network adapter) for communicating with one or more other systems, and peripheral devices 1725, such as cache, other memory, data storage and/or electronic display adapters. The memory 1710, storage unit 1715, interface 1720 and peripheral devices 1725 are in communication with the CPU 1705 through a communication bus (solid lines), such as a motherboard. The storage unit 1715 can be a data storage unit (or data repository) for storing data. The computer system 1701 can be operatively coupled to a computer network ("network") 1730 with the aid of the communication interface 1720. The network 1730 can be the Internet, an internet and/or extranet, or an intranet and/or extranet that is in communication with the Internet. The network 1730 in some cases is a telecommunication and/or data network. The network 1730 can include one or more computer servers, which can enable distributed computing, such as cloud computing. The network 1730, in some cases with the aid of the computer system 1701, can implement a peer-to-peer network, which may enable devices coupled to the computer system 1701 to behave as a client or a server.

[0210] The CPU 1705 can execute a sequence of machine-readable instructions, which can be embodied in a program or software. The instructions may be stored in a memory location, such as the memory 1710. The instructions can be directed to the CPU 1705, which can subsequently program or otherwise configure the CPU 1705 to implement methods of the present disclosure. Examples of operations performed by the CPU 1705 can include fetch, decode, execute, and writeback.

[0211] The CPU 1705 can be part of a circuit, such as an integrated circuit. One or more other components of the system 1701 can be included in the circuit. In some cases, the circuit is an application specific integrated circuit (ASIC).

[0212] The storage unit 1715 can store files, such as drivers, libraries and saved programs. The storage unit 1715 can store user data, e.g., user preferences and user programs.

The computer system 1701 in some cases can include one or more additional data storage units that are external to the computer system 1701, such as located on a remote server that is in communication with the computer system 1701 through an intranet or the Internet.

[0213] The computer system 1701 can communicate with one or more remote computer systems through the network 1730. For instance, the computer system 1701 can communicate with a remote computer system of a user (e.g., a laboratory or hospital). Examples of remote computer systems include personal computers (e.g., portable PC), slate or tablet PC's (e.g., Apple (Registered trademark) iPad, Samsung (Registered trademark) Galaxy Tab), telephones, Smart phones (e.g., Apple (Registered trademark) iPhone, Android-enabled device, Blackberry (Registered trademark)), or personal digital assistants. The user can access the computer system 1701 via the network 1730.

[0214] Methods as described herein can be implemented by way of machine (e.g., computer processor) executable code stored on an electronic storage location of the computer system 1701, such as, for example, on the memory 1710 or electronic storage unit 1715. The machine executable or machine readable code can be provided in the form of software. During use, the code can be executed by the processor 1705. In some cases, the code can be retrieved from the storage unit 1715 and stored on the memory 1710 for ready access by the processor 1705. In some situations, the electronic storage unit 1715 can be precluded, and machine-executable instructions are stored on memory 1710.

[0215] The code can be pre-compiled and configured for use with a machine having a processer adapted to execute the code, or can be compiled during runtime. The code can be supplied in a programming language that can be selected to enable the code to execute in a pre-compiled or as-compiled fashion.

[0216] Aspects of the systems and methods provided herein, such as the computer system 1701, can be embodied in programming. Various aspects of the technology may be thought of as "products" or "articles of manufacture" typically in the form of machine (or processor) executable code and/or associated data that is carried on or embodied in a type of machine readable medium. Machine-executable code can be stored on an electronic storage unit, such as memory (e.g., read-only memory, random-access memory, flash memory) or a hard disk. "Storage" type media can include any or all of the tangible memory of the computers, processors or the like, or associated modules thereof, such as various semiconductor memories, tape drives, disk drives and the like, which may provide non-transitory storage at any time for the software programming. All or portions of the software may at times be communicated through the Internet or various other telecommunication networks. Such communications, for example, may enable loading of the software from one computer or processor into another, for

example, from a management server or host computer into the computer platform of an application server. Thus, another type of media that may bear the software elements includes optical, electrical and electromagnetic waves, such as used across physical interfaces between local devices, through wired and optical landline networks and over various air-links. The physical elements that carry such waves, such as wired or wireless links, optical links or the like, also may be considered as media bearing the software. As used herein, unless restricted to non-transitory, tangible "storage" media, terms such as computer or machine "readable medium" refer to any medium that participates in providing instructions to a processor for execution.

[0217] Hence, a machine readable medium, such as computer-executable code, may take many forms, including but not limited to, a tangible storage medium, a carrier wave medium or physical transmission medium. Non-volatile storage media include, for example, optical or magnetic disks, such as any of the storage devices in any computer(s) or the like, such as may be used to implement the databases, etc. shown in the drawings. Volatile storage media include dynamic memory, such as main memory of such a computer platform. Tangible transmission media include coaxial cables; copper wire and fiber optics, including the wires that comprise a bus within a computer system. Carrier-wave transmission media may take the form of electric or electromagnetic signals, or acoustic or light waves such as those generated during radio frequency (RF) and infrared (IR) data communications. Common forms of computer-readable media therefore include for example: a floppy disk, a flexible disk, hard disk, magnetic tape, any other magnetic medium, a CD-ROM, DVD or DVD-ROM, any other optical medium, punch cards paper tape, any other physical storage medium with patterns of holes, a RAM, a ROM, a PROM and EPROM, a FLASH-EPROM, any other memory chip or cartridge, a carrier wave transporting data or instructions, cables or links transporting such a carrier wave, or any other medium from which a computer may read programming code and/or data. Many of these forms of computer readable media may be involved in carrying one or more sequences of one or more instructions to a processor for execution.

[0218] The computer system 1701 can include or be in communication with an electronic display 1735 that comprises a user interface (UI) 1740 for providing, for example, genetic data, including for example, base sequence strings, or reads in various syntaxes, sequence alignments. Examples of UIs include, without limitation, a graphical user interface (GUI) and web-based user interface.

[0219] Methods and systems of the present disclosure can be implemented by way of one or more algorithms. An algorithm can be implemented by way of software upon execution by the central processing unit 1705. The algorithm can, for example, encrypt data, translate genetic reads, analyze, interpret, align, and assemble various data

including but not limited to sequence data, working data, meta data, sample data, contig data.

[0220] While preferred embodiments of the present invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. It is not intended that the invention be limited by the specific examples provided within the specification. While the invention has been described with reference to the aforementioned specification, the descriptions and illustrations of the embodiments herein are not meant to be construed in a limiting sense. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the invention. Furthermore, it shall be understood that all aspects of the invention are not limited to the specific depictions, configurations or relative proportions set forth herein which depend upon a variety of conditions and variables. It should be understood that various alternatives to the embodiments of the invention described herein may be employed in practicing the invention. It is therefore contemplated that the invention shall also cover any such alternatives, modifications, variations or equivalents. It is intended that the following claims define the scope of the invention and that methods and structures within the scope of these claims and their equivalents be covered thereby.

Claims

[Claim 1] A biological data management system, comprising:
(a) an end-user module comprising a sequencing device, the sequencing device configured to generate base data;
(b) a local repository in network communication with the end-user module, the local repository programmed or configured to (i) receive the base data, (ii) convert the base data into sequence data, (iii) produce abbreviated data based on the sequence data, and (iv) compare the abbreviated data with a database of existing abbreviations; and
(c) a central server in network communication with the local repository, the central server configured to update the database of the existing abbreviations.

[Claim 2] The biological data management system of claim 1, wherein the local repository is further programmed or configured to flag abbreviations and communicate the flagged abbreviations to the central server.

[Claim 3] The biological data management system of claim 2, wherein the central server is further programmed or configured to receive a flagged abbreviation and perform further analysis on the flagged abbreviation.

[Claim 4] The biological data management system of claim 3, wherein the central server is further programmed or configured to generate a directive and communicate the directive to the local repository upon the analysis of the flagged abbreviation.

[Claim 5] The biological data management system of claim 1, wherein the abbreviation is a variance, hash, or a checksum.

[Claim 6] A method for storing biological data, comprising:
(d) determining a size of the biological data to identify a storage unit size suitable to store the biological data;
(e) identifying a memory location in a memory device having a block size compatible with the storage unit size; and
(f) storing the biological data in an erasable block at the memory location of the memory device.

[Claim 7] The method of claim 6, wherein each erasable block comprises a section for storing the biological data and a section for storing metadata related to the biological data.

[Claim 8] The method of claim 7, wherein the section for storing metadata comprises a longer lifetime.

[Claim 9] The method of claim 7, wherein the section for storing metadata

comprises a controller different from a controller of the section for storing sequence data.

[Claim 10] The method of claim 7, wherein the section for storing metadata is configured for more frequent access than the section for storing sequence data.

[Claim 11] A biological data management system, comprising:
(g) a first memory device configured to store biological data for in-frequent access; and
(h) a second memory device having a block size, the second memory device being in communication with the first memory device and configured to store biological data for frequent access; wherein the second memory device is faster than the first memory device, and wherein the block size is selected to store the biological data according to a size of the biological data.

[Claim 12] The biological data management system of claim 11, wherein the biological data is an n-mer sequence, and wherein the block size is n times a number of bits required to store a monomer of the n-mer.

[Claim 13] The biological data management system of claim 11, wherein the biological data is an n-mer sequence, and wherein the block size is at least n times a number of bits required to store a monomer of the n-mer.

[Claim 14] The biological data management system of claim 11, wherein the second memory device comprises a flash memory device.

[Claim 15] The biological data management system of claim 14, wherein the second memory device comprises a block that is a flash memory erase block.

[Claim 16] A method for storing sequence base data in a multi-level cell (MLC) memory device, the MLC memory device comprising memory cells, each of the memory cells configured to store two bits, the method comprising, in a memory cell:
(i) setting the two bits to 00 to represent a base of a first type;
(j) setting the two bits to 01 to represent a base of a second type;
(k) setting the two bits to 10 to represent a base of a third type; or
(l) setting the two bits to 11 to represent a base of a fourth type.

[Claim 17] The method of claim 16, wherein the sequence base data represents one or more polynucleotides, each of the polynucleotides comprising one or more bases, each of the one or more bases being one of at least four possible bases.

[Claim 18] The method of claim 17, wherein the polynucleotide is a DNA or an

RNA.

[Claim 19] A method for storing biological data in a memory device, the memory device comprising blocks, each of the blocks comprising a block size, the method comprising:
(m) determining a size of the biological data;
(n) determining a block size of at least a subset of the blocks;
(o) compressing the biological data based on the block size to produce compressed biological data; and
(p) storing the biological data in the at least a subset of the blocks.

[Claim 20] The method of claim 19, wherein the memory device comprises a flash memory device, and wherein the block size is an erase block size.

[Claim 21] The method of claim 19, wherein the block size is greater than or equal to a size of the compressed biological data.

[Claim 22] The method of claim 20, wherein the erase block stores the biological data and metadata of the biological data.

[Claim 23] A method for storing sequence base data in a memory device, the memory device comprising memory cells, each of the memory cells configured to store at least three bits, the method comprising, in a memory cell:
(q) setting three of the at least three bits to 000 to represent a base of a first type;
(r) setting three of the at least three bits to 001 to represent a base of a second type;
(s) setting three of the at least three bits to 010 to represent a base of a third type;
(t) setting three of the at least three bits to 011 to represent a base of a fourth type;
(u) setting three of the at least three bits to 100 to represent a base of a fifth type;
(v) setting three of the at least three bits to 101 to represent a base of a sixth type;
(w) setting three of the at least three bits to 110 to represent a base of a seventh type; and
(x) setting three of the at least three bits to 111 to represent a base of an eighth type.

[Claim 24] The method of claim 23, wherein the sequence base data represents one or more polynucleotides, each of the polynucleotides comprising one or more bases, each of the one or more bases being one of four different

native bases, a methylated base, an oxidated base, or an abasic location.

[Claim 25] The method of claim 24, wherein the polynucleotide is a DNA or an RNA.

[Claim 26] The method of claim 23, wherein the memory device comprises a flash memory, a phase-change memory, or a resistive memory.

[Claim 27] A method for storing sequence base data in a memory device, the sequence base data comprising two probable bases to represent each of a plurality of bases measured, the memory device comprising memory cells, each of the memory cells configured to store a plurality of bits, the method comprising:
storing in a first bit of the plurality of bits a most probable base of the sequence base data;
storing in a second bit of the plurality of bits a second most probable base of the sequence base data; and
storing in a remainder of the plurality of bits a relative probability of the most probable base and the second most probable base.

[Claim 28] The method of claim 27, further comprising:
using a first cell of the memory cells to identify the most probable base;
using a second cell of the memory cells to identify the second most probable base; and
using one or more other cells of the memory cells to store the relative probability.

[Claim 29] The method of claim 27, further comprising storing in a third cell of the memory cells a probability of the second most probable base.

[Claim 30] A method for storing sequence base data in a memory device comprising memory cells each configured to store at least three bits, the method comprising, in a memory cell:
(y) providing a first bit indication comprising three bits of the at least three bits to represent a base of a first type;
(z) providing a second bit indication comprising three bits of the at least three bits to represent a base of a second type;
(aa) providing a third bit indication comprising three bits of the at least three bits to represent a base of a third type;
(bb) providing a fourth bit indication comprising three bits of the at least three bits to represent a base of a fourth type;
(cc) providing a fifth bit indication comprising three bits of the at least three bits to represent a methylated base;
(dd) providing a sixth bit indication comprising three bits of the at least

three bits to represent an oxidated base; and

(ee) providing a seventh bit indication comprising three bits of the at least three bits to represent an abasic site.

[Claim 31] The method of claim 29, wherein the memory device comprises a flash memory, a phase-change memory, or a resistive memory.

[Claim 32] A method for encrypting biological sequence data, the method comprising:

(ff) identifying a normal level of variance in the biological sequence data; and

(gg) introducing a second level of variation into the biological sequence data, the second level of variation comparable to the normal level of variance, such that the biological sequence data is indistinguishable with respect to the normal level of variance.

[Claim 33] The method of claim 32, further comprising communicating the introduced level of variance using an encryption method.

[Claim 34] A method for encrypting biological sequence data of a subject, the method comprising:

(hh) encrypting information related to the subject using a first encryption scheme; and

(ii) encrypting the biological sequence data using a second encryption scheme, which second encryption scheme is different from the first encryption scheme.

[Claim 35] The method of claim 34, wherein the second encryption scheme comprises a less extensive encryption than the first encryption scheme.

[Claim 36] The method of claim 35, wherein the second encryption scheme comprises chaffing and winnowing.

[Claim 37] The method of claim 35, wherein the first encryption scheme uses a public key infrastructure and the second encryption scheme uses the public key infrastructure.

[Claim 38] The method of claim 35, wherein the first encryption scheme uses a first public key infrastructure and the second encryption scheme uses a second public key infrastructure different from the first public key infrastructure.

[Claim 39] A method for storing sequence base data, the method comprising:

providing a two-dimensional table structure in computer memory, the two-dimensional table structure configured to store information representing potential bases;

storing information representing the most probable measured bases of

the sequence base data in a first dimension of the two-dimensional table structure;

storing information representing other potential bases of the sequence base data in a second dimension of the two-dimensional table structure; and

storing probabilities corresponding to an intersection of the first dimension and the second dimension in the two-dimensional table structure.

[Claim 40] The method of claim 39, wherein the potential bases comprise a set of each of four possible bases and at least one of a methylated base, an oxidated base, and an abasic site.

[Claim 41] The method of claim 39, further comprising providing a second two-dimensional table structure in computer memory, the second two-dimensional table structure configured to store information representing potential bases; and storing in the second two-dimensional table structure the most probable measured bases of the sequence base data and the second most probable measured bases of the sequence base data.

[Claim 42] A method for managing biological data, the method comprising:

providing an application server programmed or configured to (i) receive raw measured biological data from a sensor and (ii) generate processed biological data from the raw measured biological data;

receiving, at the application server, from a local repository, definitions and rules related to the processed biological data; and

issuing, by the application server, directives based on the definitions and rules related to the processed biological data.

[Claim 43] The method of claim 42, wherein the processed biological data comprises a portion of the processed biological data for which related definitions and rules are not found in the local repository, and wherein the method further comprises sending at least the portion of the processed biological data to the local repository.

[Claim 44] The method of claim 43, further comprising sending at least the portion of the processed biological data from the local repository to a central server.

[Claim 45] The method of claim 44, further comprising sending directives from the central server to the local repository.

[Claim 46] The method of claim 45, further comprising sending new definitions and rules from the central server to the local repository.

[Claim 47] A method for storing sequence base data, the method comprising:
for a base location, storing information representing a most probable base of the sequence base data in a first location of a storage device, and storing a probability of a number of occurrences of the most probable base in a second location of the storage device.

[Claim 48] A method for storing sequence base data comprising at least four possible bases, the method comprising:
(jj) providing a three-dimensional table structure in computer memory, which three-dimensional table structure is configured to store the sequence base data, wherein (i) a first dimension of the three-dimensional table structure stores information representing most probable measured bases of the genetic sequence base data; (ii) a second dimension of the three-dimensional table structure stores information representing potential bases of the genetic sequence base data; and (iii) a third dimension of the three-dimensional table structure stores information representing a base count probability for each of the at least four possible bases of the sequence base data;
(kk) storing probabilities corresponding to an intersection of the first dimension, the second dimension, and the third dimension in the three-dimensional table structure.

[Claim 49] A method for protecting biological data related to a subject, the method comprising:
encrypting personal identification information of the subject using a first encryption scheme;
encrypting phenotypes of the subject using a second encryption scheme;
encrypting the biological data using a third encryption scheme, wherein the second encryption scheme or the third encryption scheme is different from the first encryption scheme; and
storing the encrypted personal identification information, the encrypted phenotypes, and the encrypted biological data in computer memory.

[Claim 50] The method of claim 49, wherein (i) the second encryption scheme is different from the first encryption scheme, and (ii) the third encryption scheme is different from the first encryption scheme, and (iii) the third encryption scheme is different from the second encryption scheme.

[Claim 51] The method of claim 49, further comprising storing gene expression data of the subject.

[Claim 52] The method of claim 50, further comprising storing geographic data of

the subject.

[Claim 53] A method for storing genetic data of a subject, the method comprising:
storing personal identification information of the subject in a first storage segment with a first level of limitation of access;
storing phenotype data of the subject in a second storage segment with a second level of limitation of access; and
storing the genetic data of the subject in a third storage segment with a third level of limitation of access.

[Claim 54] The method of claim 53, wherein the second level of limitation of access or the third level of limitation of access is different from the first level of limitation of access.

[Claim 55] The method of claim 54, wherein (i) the second level of limitation of access is different from the first level of limitation of access, and (ii) the third level of limitation of access is different from the first level of limitation of access, and (iii) the third level of limitation of access is different from the second level of limitation of access.

[Fig. 1]

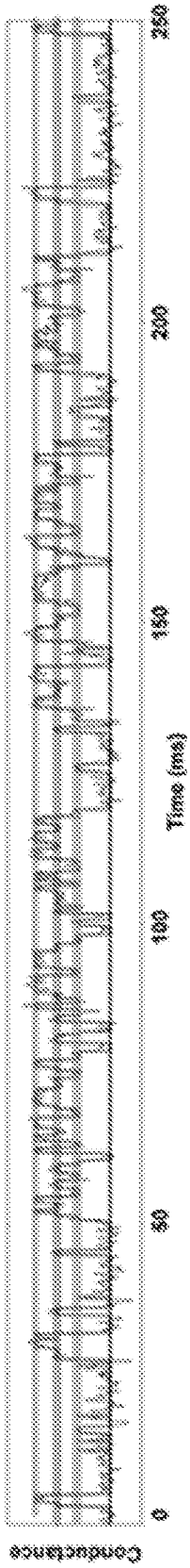

FIG. 1

[Fig. 2]

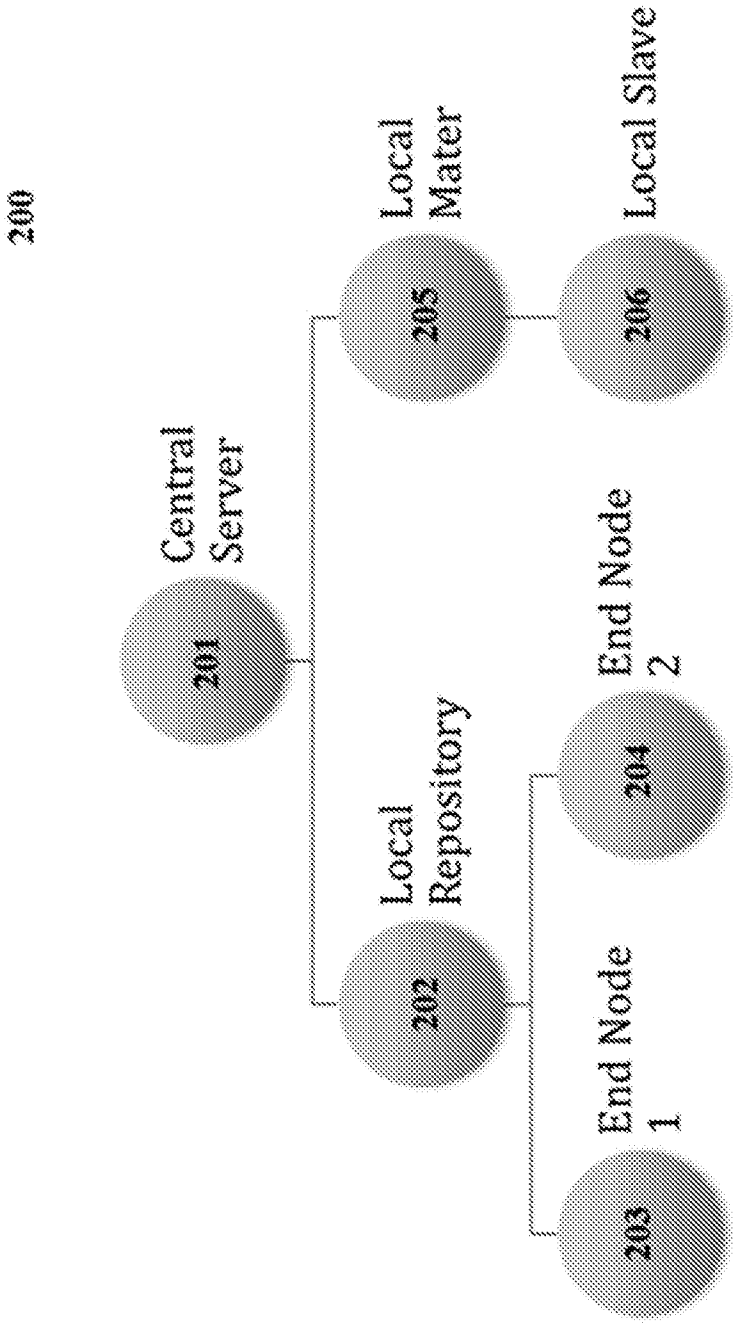
200
Central Server
201
Local Repository
202
End Node 1
203
End Node 2
204
Local Mater
205
Local Slave
206

FIG. 2

[Fig. 3]

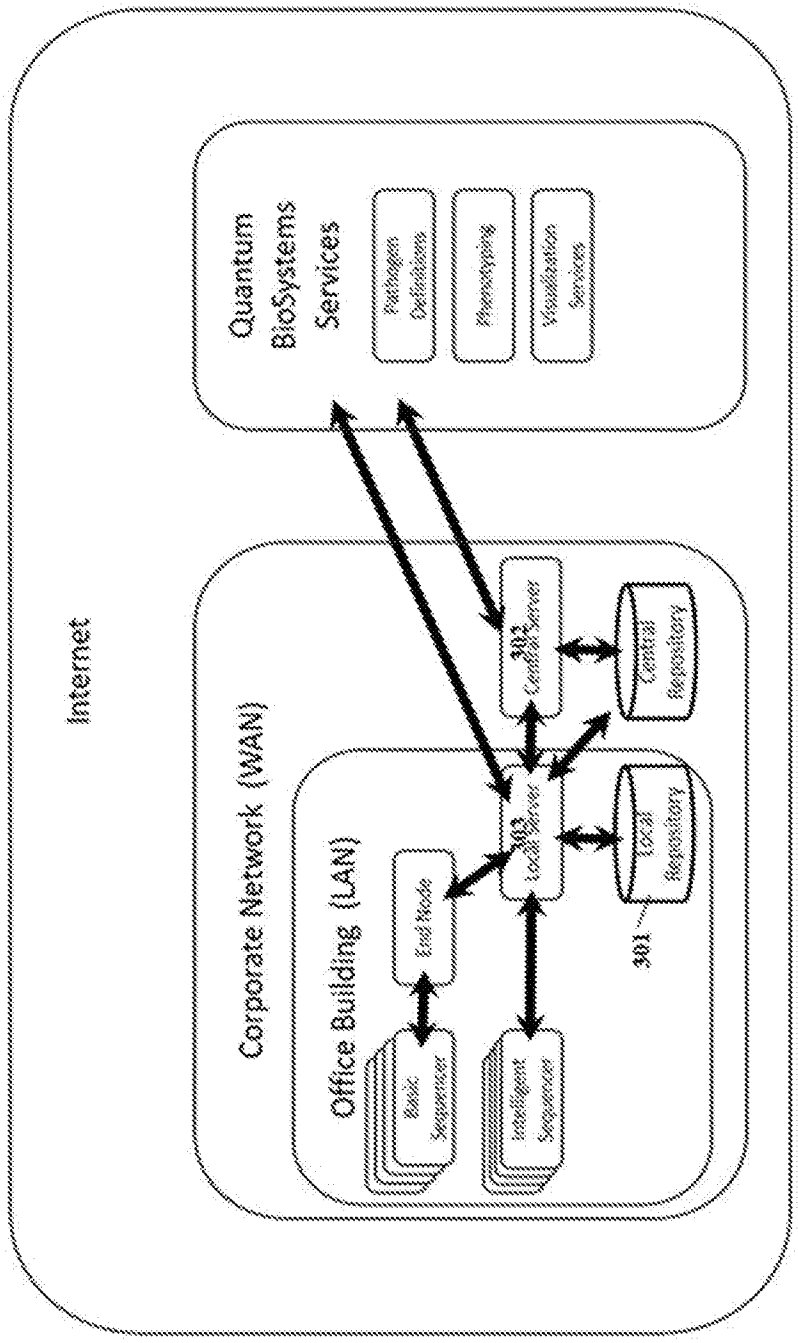
Internet
Quantum BioSystems Services
Pathogen Definitions
Prescriptive
Visualization Services
Corporate Network (WAN)
Office Building (LAN)
Central Repository
Local Repository
End Node
Basic Sequencer
Intelligent Sequencer
301

FIG. 3

[Fig. 4]

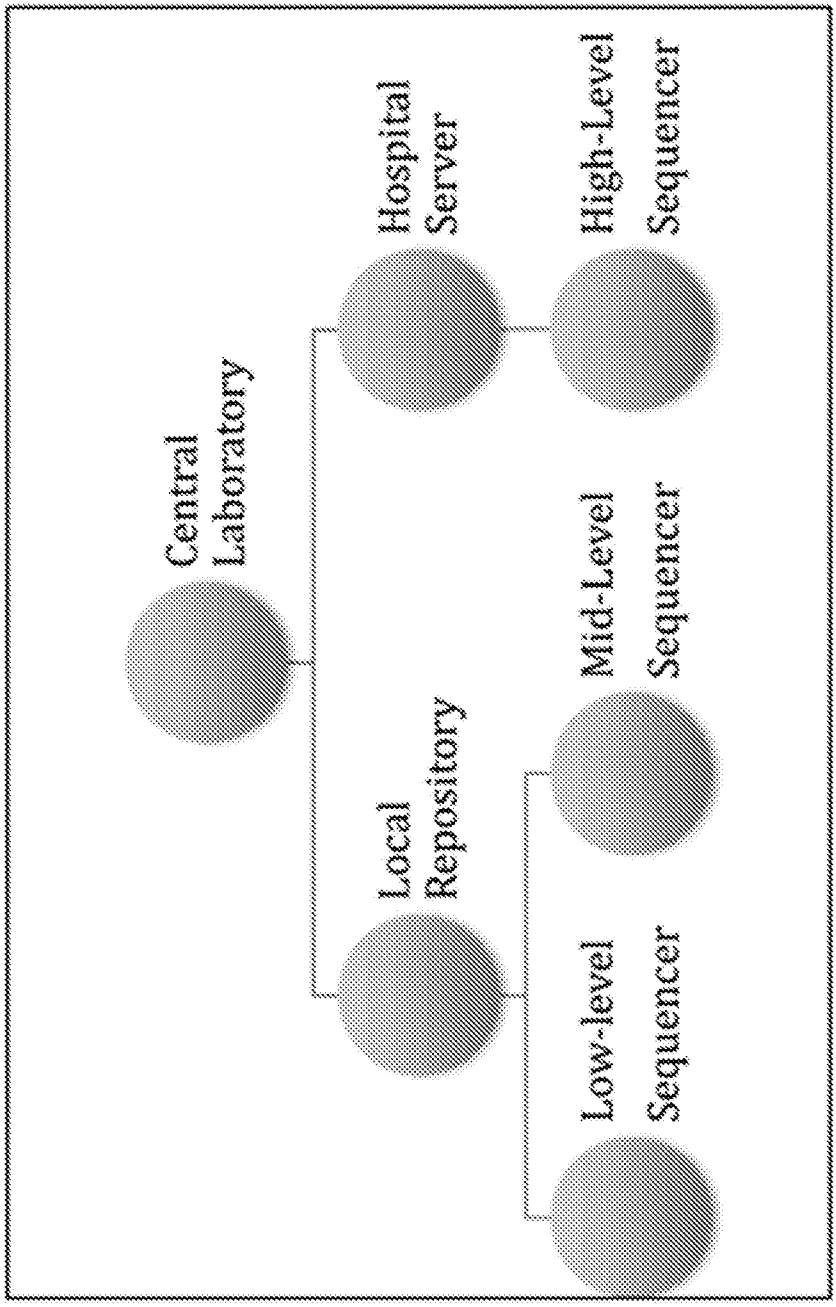
Central Laboratory
Hospital Server
High-Level Sequencer
Local Repository
Mid-Level Sequencer
Low-level Sequencer

FIG. 4

[Fig. 5]

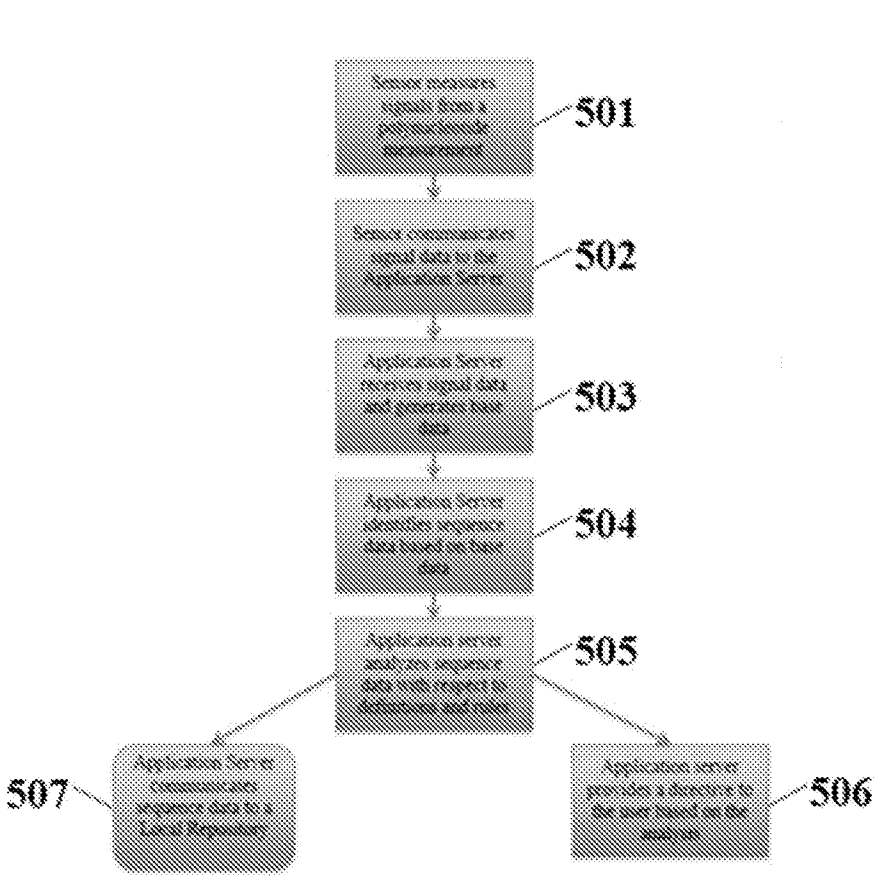

FIG. 5

[Fig. 6]

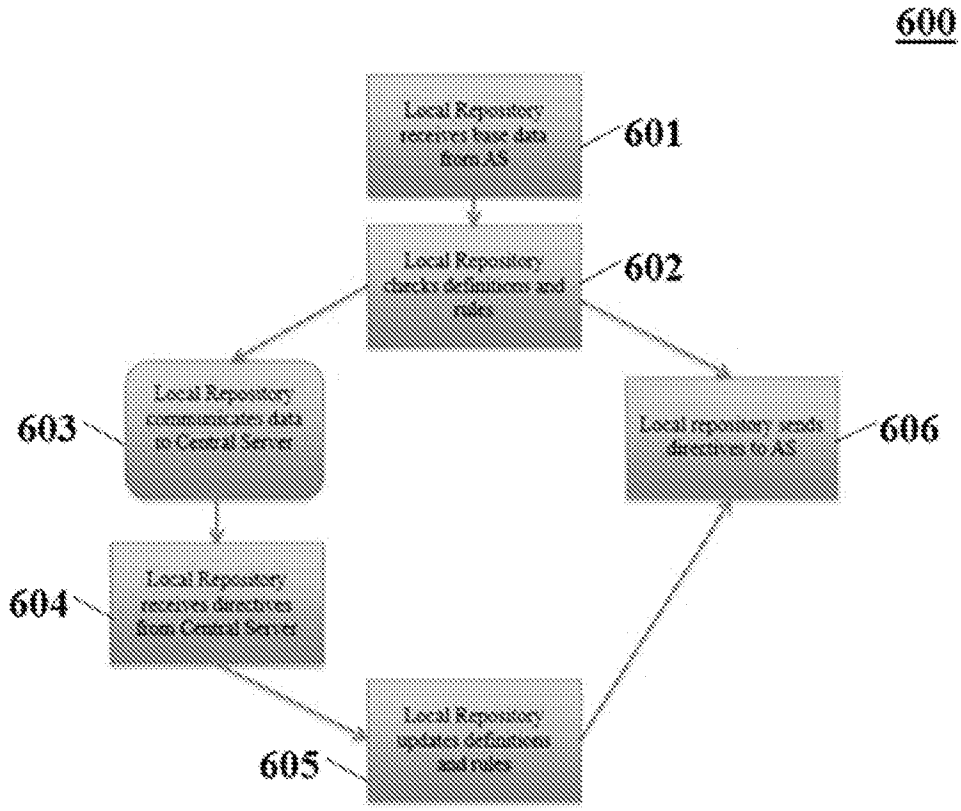

FIG. 6

[Fig. 7]

	A	B	C	G	T	U
A	68.84	1.55	7.70	0.60	0.77	20.54
T	5.26	0.76	8.32	1.05	84.24	0.36
C	5.91	1.87	67.74	7.76	14.24	2.48
G	5.67	1.55	4.00	84.84	2.88	1.06
T	10.00	0.56	1.48	4.24	74.49	1.23
C	1.21	0.74	89.94	5.08	2.81	0.22
G	10.55	0.77	16.04	58.70	3.19	10.76
A	87.89	0.49	2.59	5.22	1.97	1.83
G	2.63	0.82	7.02	81.54	0.60	7.38
S	12.40	64.38	10.21	0.01	11.78	1.21
A	77.30	1.95	2.57	6.66	4.72	6.79
G	11.74	0.07	12.09	55.56	9.16	11.37
T	1.67	1.99	8.91	1.23	71.45	14.76
T	2.69	0.74	0.34	5.54	76.60	14.10
A	66.97	0.56	7.18	3.30	6.82	15.18
C	5.51	1.64	64.46	15.82	5.09	7.48
A	81.36	0.31	7.66	6.86	1.49	2.31
A	65.57	1.36	13.81	5.47	8.11	5.68
G	3.24	1.98	2.73	61.25	1.34	29.46
C	7.37	1.63	68.22	10.07	7.46	5.26
T	1.04	0.11	8.48	2.77	79.20	8.40

FIG. 7

[Fig. 8]

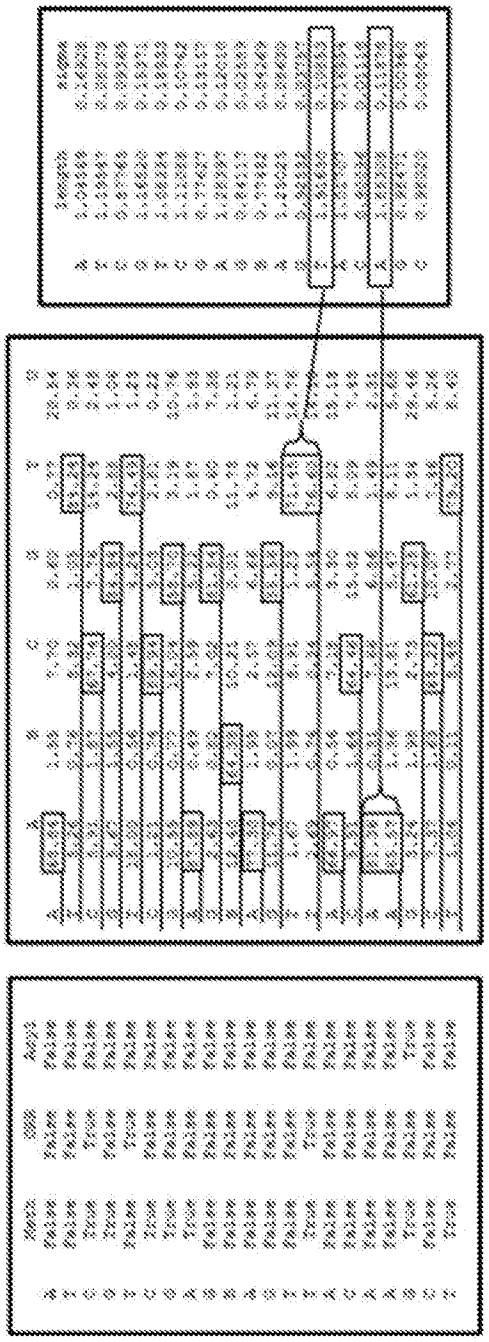

FIG. 8

[Fig. 9]

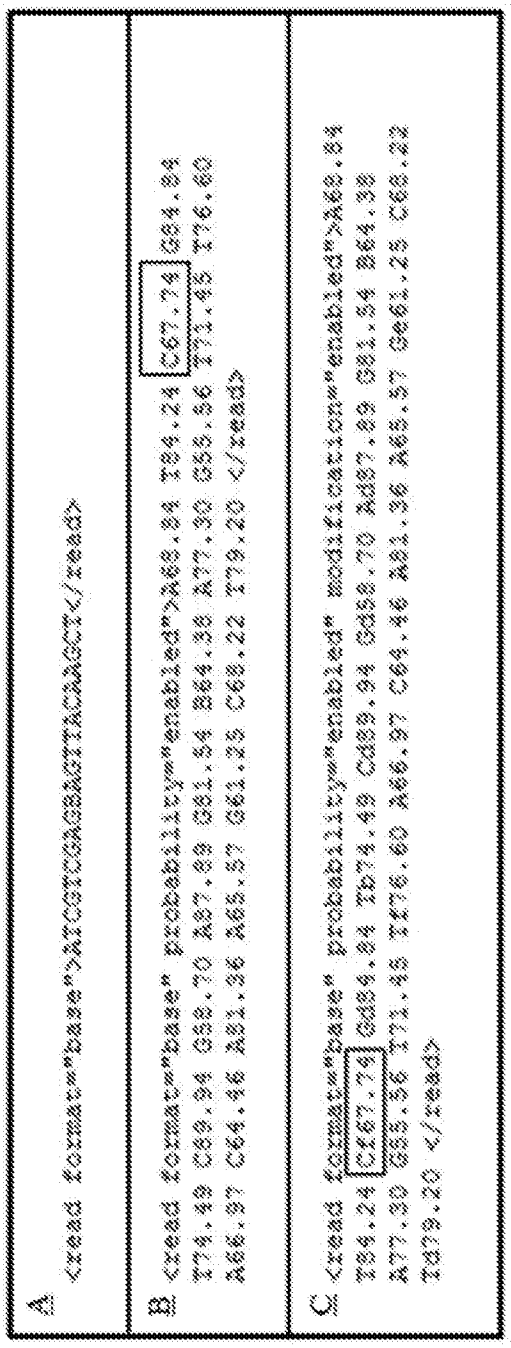
A
<read format="base">ATCGTCGAGGAGTTACAAGCT</read>
B
<read format="base" probability="enabled">A68.84 T84.24 C67.74 G84.84
T74.49 C89.94 G58.70 A87.69 G81.54 B64.38 A77.30 G55.56 T71.45 T76.60
A66.87 C64.46 A81.36 A63.57 G61.25 C68.22 T79.20 </read>
C
<read format="base" probability="enabled" modification="enabled">A68.84
T84.24 Cf67.74 Gd89.84 Tb74.49 Cd89.94 Gd58.70 Ad87.69 G81.54 B64.38
A77.30 G55.56 T71.45 Tf76.60 A66.87 C64.46 A81.36 A63.57 Ge61.25 C68.22
T079.20 </read>

FIG. 9

[Fig. 10]

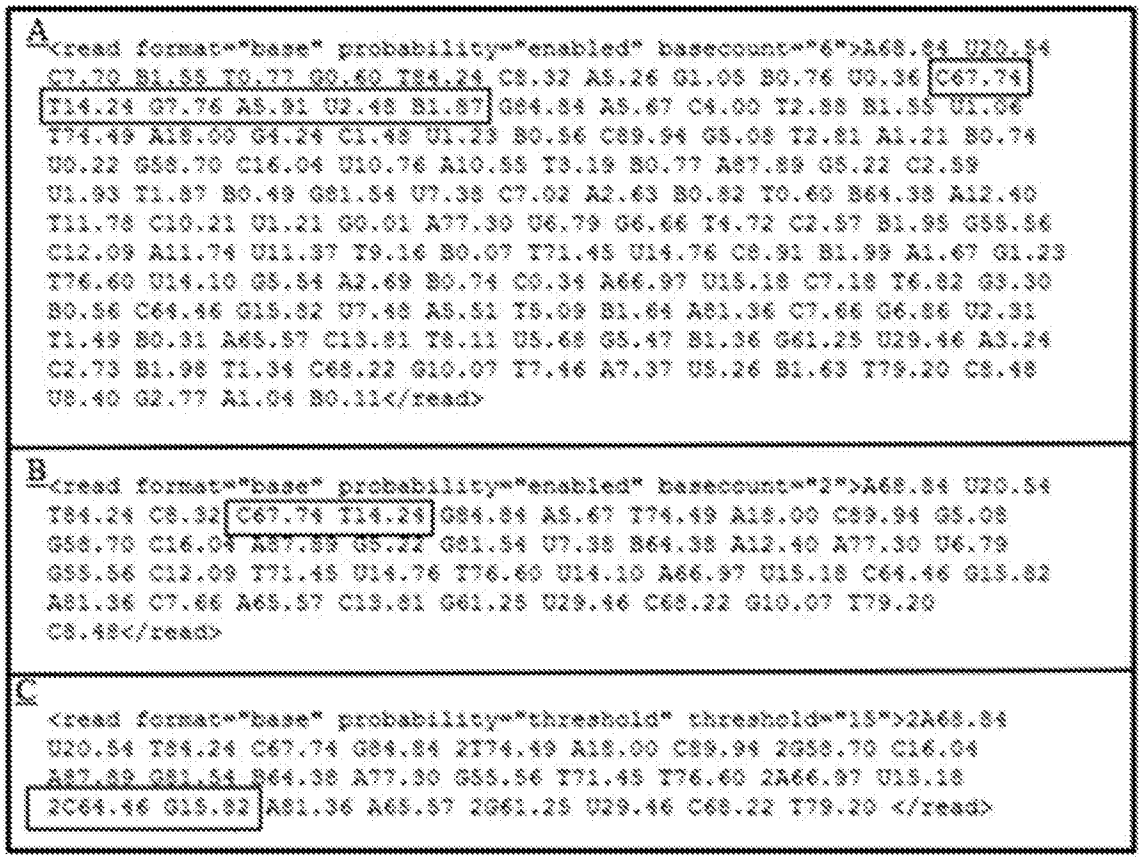
A
<read format="base" probability="enabled" basecount="6">A68.84 G20.54
C67.74 G84.84 A5.67 T74.49 A18.00 C89.94 G5.08
G58.70 C16.04 A87.89 G5.22 G61.54 G7.38 B64.38 A12.40 A77.30 G6.79
G55.56 C12.09 T71.45 G14.76 T76.60 G14.10 A66.97 G18.18 C64.46 G13.82
A81.36 C7.66 A65.57 C13.81 G61.25 G29.46 C68.22 G10.07 T79.20
C8.48</read>
C
<read format="base" probability="threshold" threshold="13">2A68.84
G20.54 T84.24 C67.74 G84.84 2T74.49 A18.00 C89.94 2G58.70 C16.04
A87.89 2G61.54 B64.38 A77.30 G55.56 T71.45 T76.60 2A66.97 G18.18
3C64.46 G13.82 A81.36 A65.57 2G61.25 G29.46 C68.22 T79.20 </read>

FIG. 10

[Fig. 11]

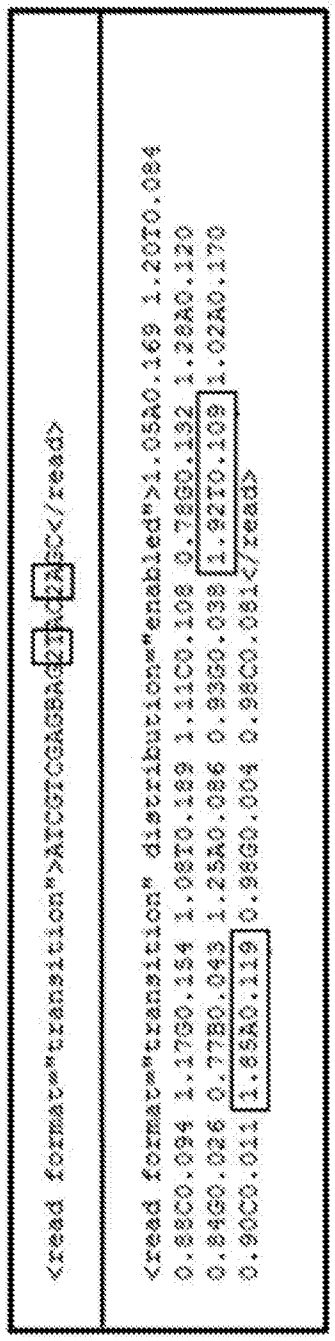
<read format="transaction">ATCGTCGAGGAGTCA</read>
<read format="transaction" distribution="enabled">1.05A0.169 1.2020.084
0.88C0.094 1.1730.154 1.0830.188 1.11C0.108 0.7860.132 1.28A0.120
0.84G0.026 0.7780.043 1.25A0.086 0.93G0.038 1.82T0.109 1.02A0.170
0.90C0.011 1.85A0.119 0.96G0.004 0.98C0.081</read>

FIG. 11

[Fig. 12]

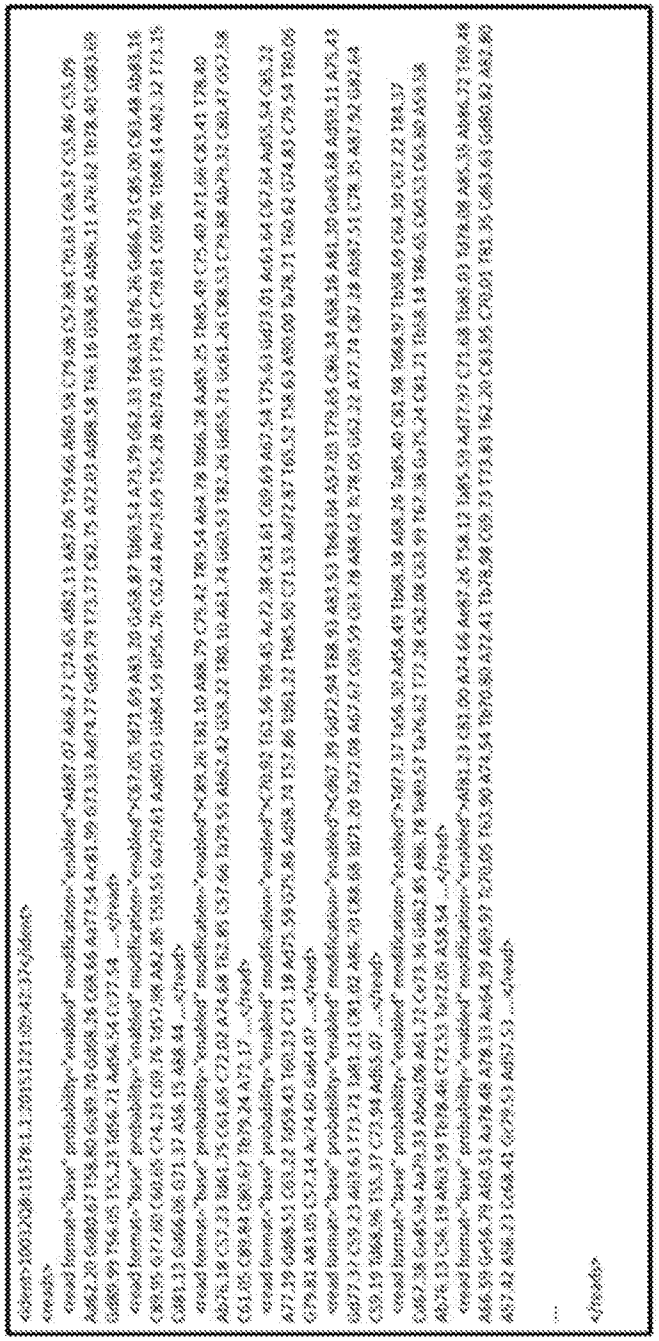

FIG. 12

[Fig. 13]

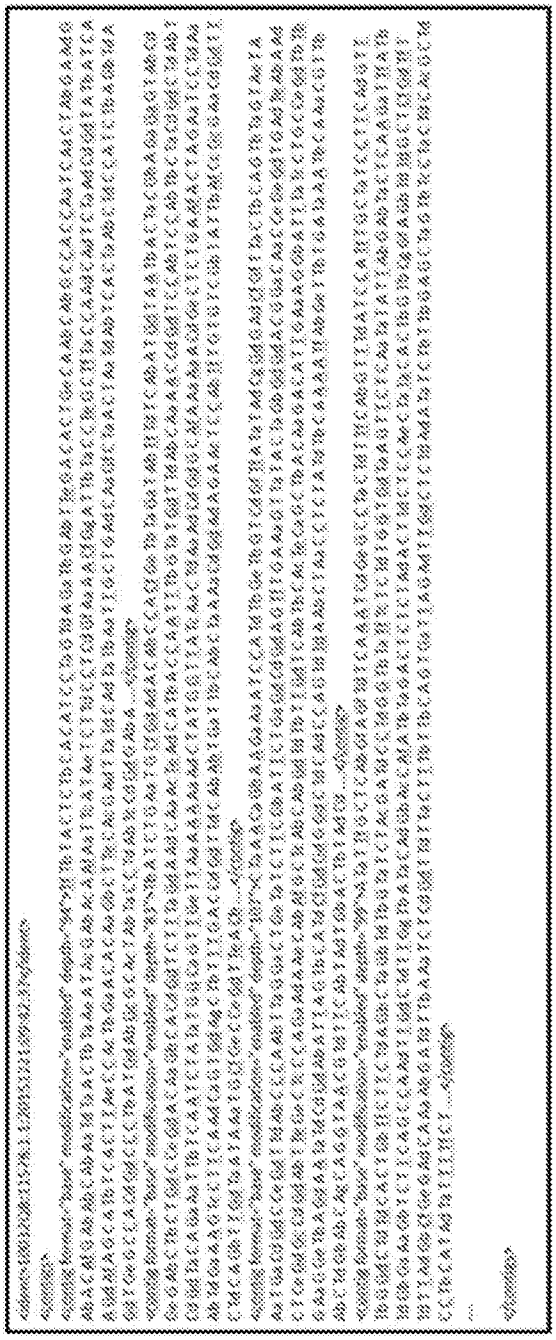

FIG. 13

[Fig. 14]

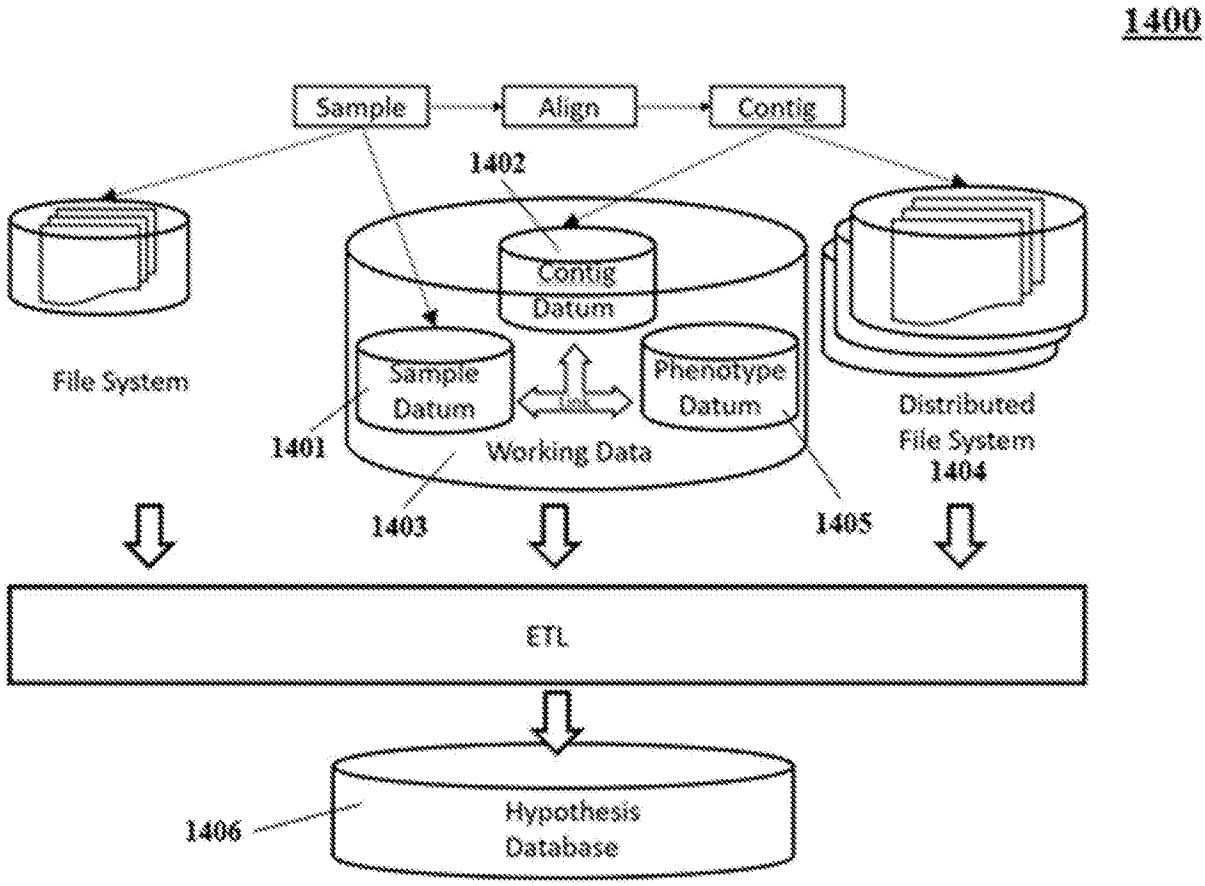
1400
Sample
Align
Contig
1402
Contig Datum
Sample Datum
Phenotype Datum
Working Data
File System
Distributed File System
1401
1403
1404
1405
ETL
Hypothesis Database
1406

FIG. 14

[Fig. 15]

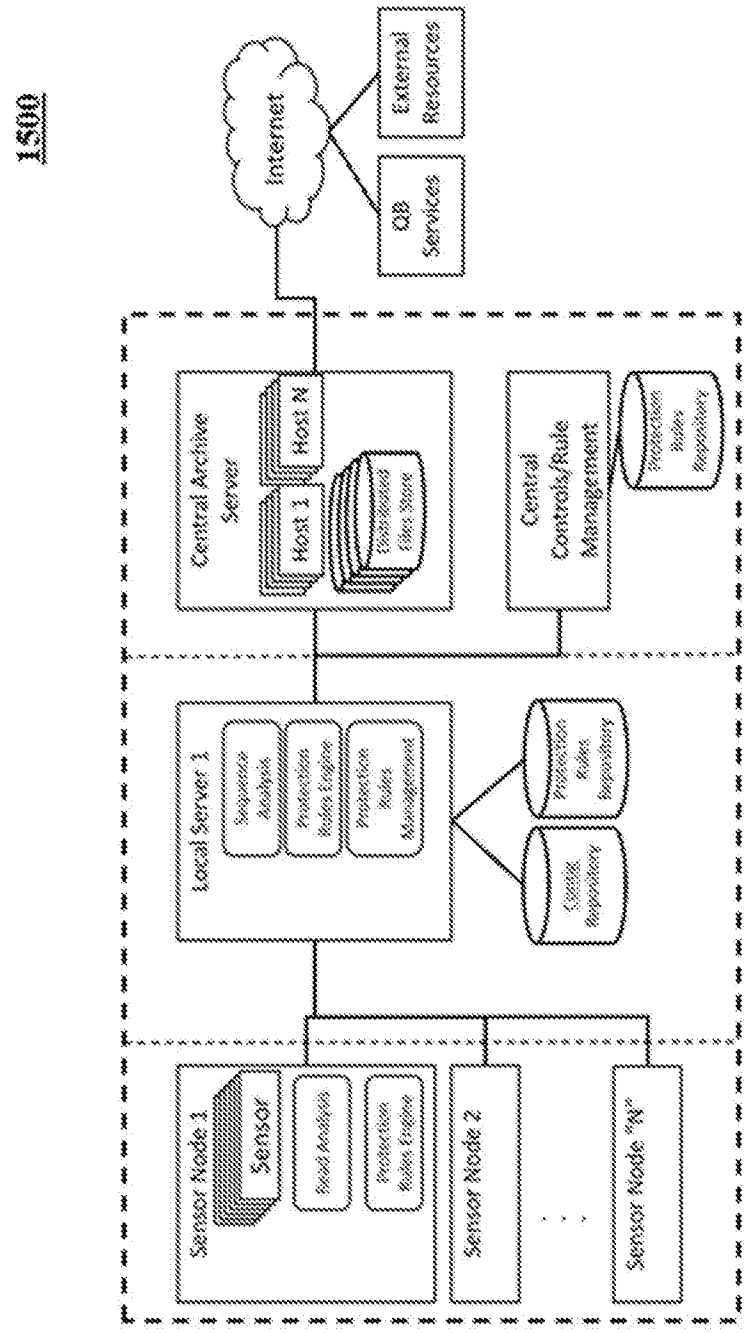
1500
Internet
External Resources
OB Services
Central Archive Server
Host N
Host 1
Central Controls/Rule Management
Local Server 1
Sensor Node 1
Sensor
Sensor Node 2
Sensor Node "N"

FIG. 15

[Fig. 16A]

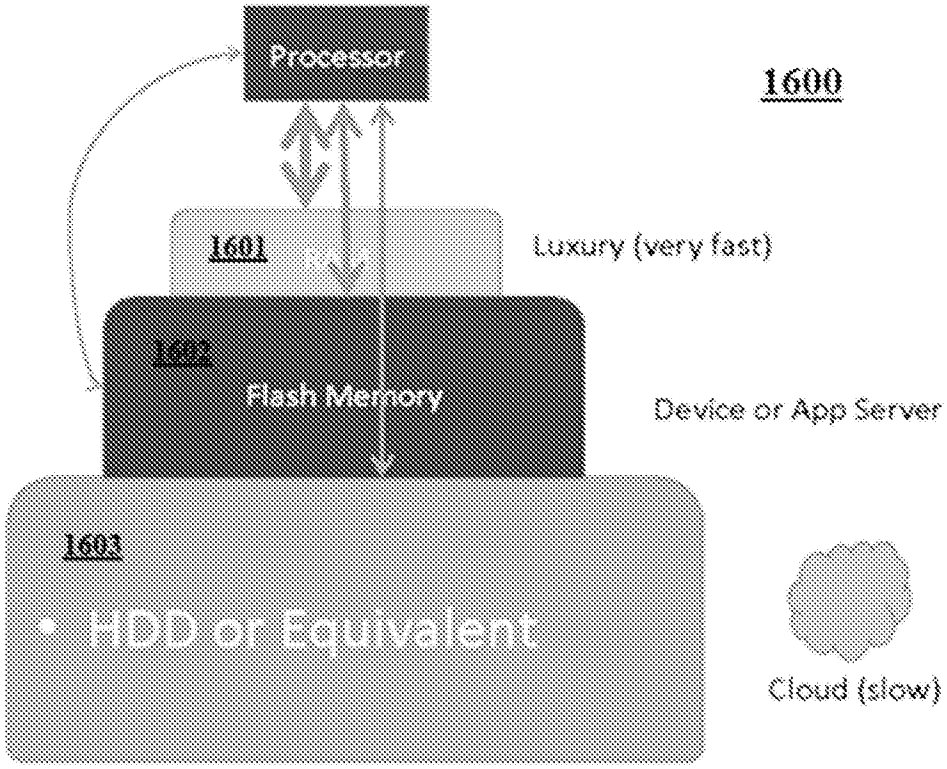

FIG. 16A

[Fig. 16B]

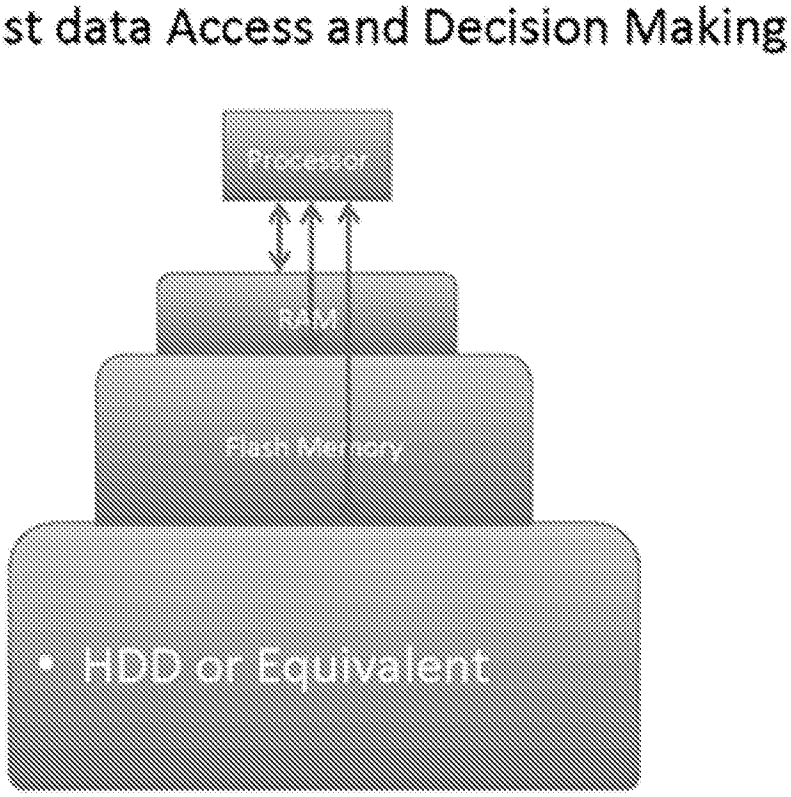

FIG. 16B

[Fig. 16C]

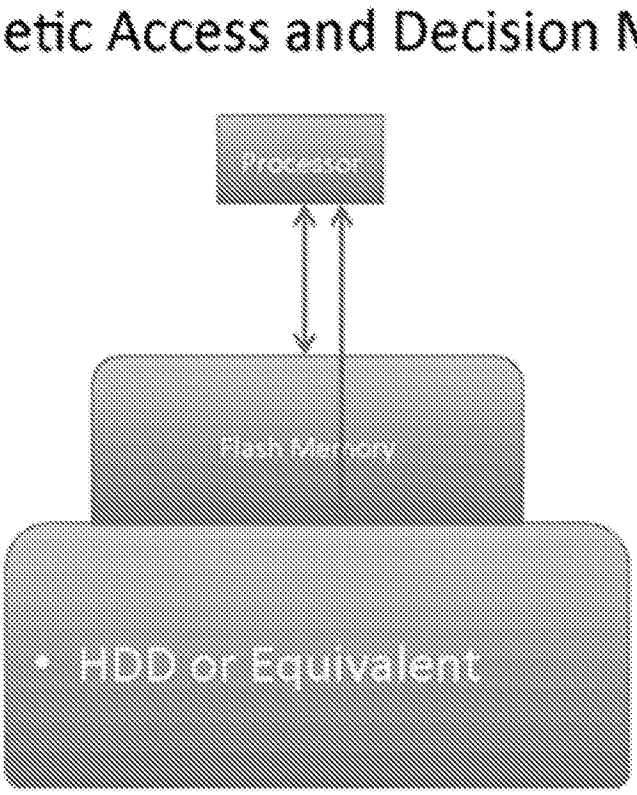

FIG. 16C

[Fig. 16D]

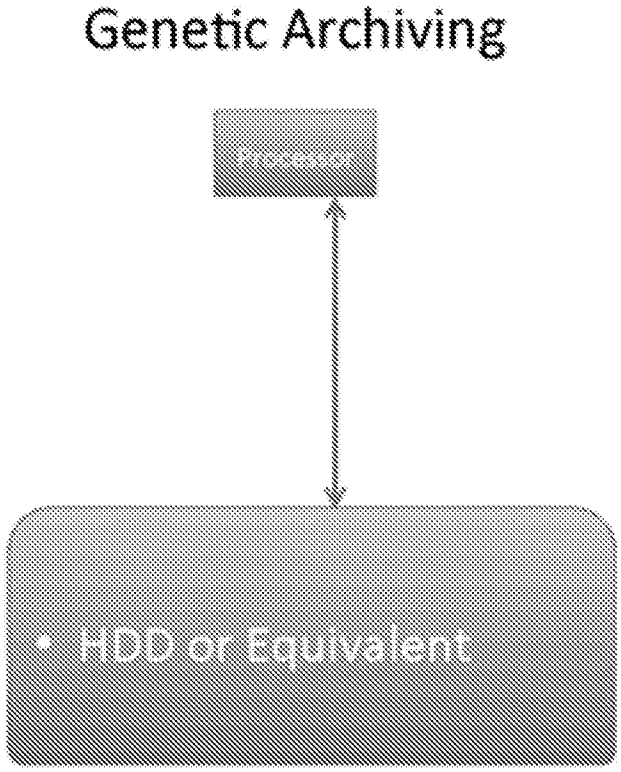

FIG. 16D

[Fig. 17]

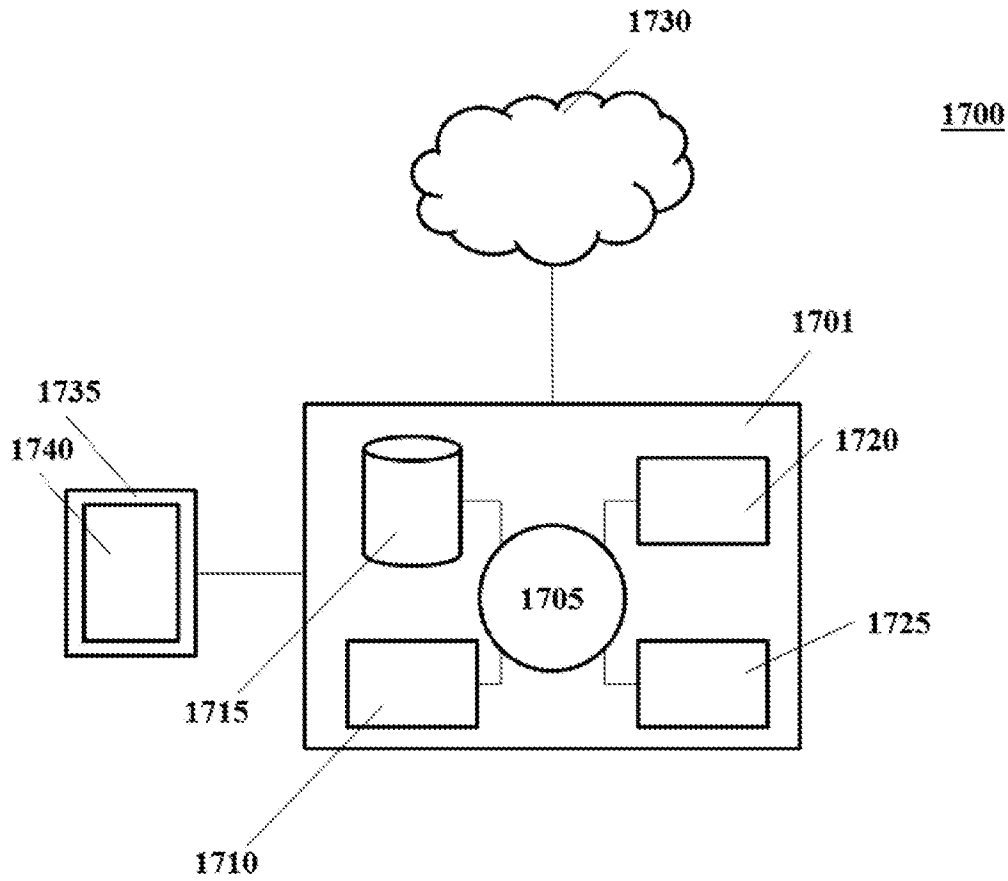

FIG. 17

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2017/014847

A. CLASSIFICATION OF SUBJECT MATTER

Int.Cl. G06Q50/22(2012.01)i

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)

Int.Cl. G06Q50/22

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

Published examined utility model applications of Japan 1922-1996
Published unexamined utility model applications of Japan 1971-2017
Registered utility model specifications of Japan 1996-2017
Published registered utility model applications of Japan 1994-2017

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X A	US 2015/0310228 A1 (NANTOMICS) 2015.10.29, paragraph [0059] & WO 2015/130954 A1 & EP 3111353 A1 & KR 10-2017-0019335 A & CN 106537400 A & JP 2017-513156 A	1-5 32, 33, 48
Y A	JP 2012-118709 A (BROTHER INDUSTRIES, LTD.) 2012.06.21, claim 1 (Family: none)	6-10 32, 33, 48
Y A	JP 4-289938 A (NIPPON TELEGRAPH AND TELEPHONE CORPORATION) 1992.10.14, paragraph [0003] (Family: none)	6-10 32, 33, 48
X Y A	JP 2004-303162 A (OMRON CORPORATION) 2004.10.28, paragraph [0023] (Family: none)	47, 53-55 7-10 32, 33, 48

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

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

Date of the actual completion of the international search	Date of mailing of the international search report
09.06.2017	20.06.2017

Name and mailing address of the ISA/JP	Authorized officer
Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan	YAMAUCHI, Hiroshi 5L 4064 Telephone No. +81-3-3581-1101 Ext. 3562

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2017/014847

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y A	JP 64-37640 A (MITSUBISHI ELECTRIC CORPORATION) 1989.02.08, claim 1 (Family: none)	11-15 32, 33, 48
Y A	JP 10-283230 A (NEC CORPORATION) 1998.10.23, claim 1 (Family: none)	11-15 32, 33, 48
X A	US 2007/0171714 A1 (MARVELL INTERNATIONAL LTD.) 2007.07.26, paragraph [0011] & JP 2009-524152 A & WO 2007/084751 A2 & EP 1984923 A2 & KR 10-2008-0098041 A & CN 101405811 A & TW 200739334 A	16-18, 23-31 32, 33, 48
X A	US 2007/0183198 A1 (OTSUKA, Takeshi) 2007.08.09, paragraph [0046] & JP 3825465 B2 & WO 2005/096220 A1 & EP 1720119 A1 & KR 10-2006-0126775 A & CN 1938720 A	19-22 32, 33, 48
X A	US 2016/0048690 A1 (MITSUBISHI SPACE SOFTWARE CO., LTD.) 2016.02.18, paragraph [0185] & JP 2014-191670 A & WO 2014/156400 A1 & EP 2980718 A1 & TW 201506653 A & CN 105190636 A & HK 1219324 A	34-38, 49-52 32, 33, 48
X A	JP 2008-146538 A (INTEC WEB & GENOME INFOMATICS CORP) 2008.06.26, paragraph [0056] (Family: none)	39-41 32, 33, 48
X A	US 2008/0077607 A1 (GATAWOOD, Joe M.) 2008.03.27, paragraphs [0048]-[0050] & WO 2006/052242 A1	42-46 32, 33, 48