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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 September 2016 (29.09.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/154154 A2

(51) International Patent Classification: Not classified

(21) International Application Number:
PCT/US2016/023490

(22) International Filing Date:
21 March 2016 (21.03.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/137,103 23 March 2015 (23.03.2015) US

(71) Applicant: **EDICO GENOME CORPORATION** [US/US]; 3344 North Torrey Pines Court, Plaza Level, La Jolla, CA 92037 (US).

(72) Inventors: **VAN ROOYEN, Pieter**; 3344 North Torrey Pines Court, Plaza Level, La Jolla, CA 92037 (US). **STONE, Gavin**; 3344 North Torrey Pines Court, Plaza Level, La Jolla, CA 92037 (US). **LANCOVICI, Lucian**; 3344 North Torrey Pines Court, Plaza Level, La Jolla, CA 92037 (US).

(74) Agent: **CATANIA, Michael**; Clause Eight IPS, P.O. Box 131270, Carlsbad, CA 92013 (US).

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

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

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

(54) Title: METHOD AND SYSTEM FOR GENOMIC VISUALIZATION

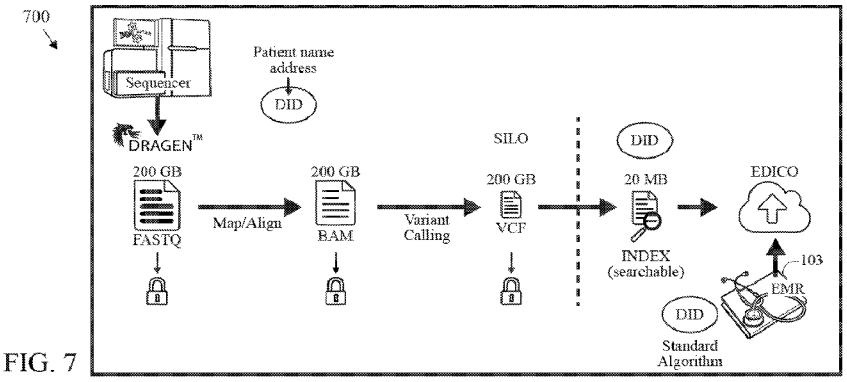

FIG. 7

(57) Abstract: A method (200) and system (100) for correlating genome data with EMR/PHR data is disclosed herein. The method (200) includes identifying a plurality of sources (102) of genome data. The method (200) also includes generating an index file for each of the plurality of genome files. The method (200) also includes transmitting each index file to a central depository (101). The method (200) also includes identifying electronic medical record (EMR) and/or personal health record (PHR) data at each source of the plurality of sources of genome data. The method (200) also includes correlating each genome file of the plurality of genome files with a corresponding EMR/PHR data.

WO 2016/154154 A2

Title

Method And System For Genomic Visualization

(EGC-006PCT)

Technical Field

The present invention generally relates to genomic visualization.

Background Art

[0001] Genomic visualization tools have been devised to assist researchers, laboratories, and other users to visually display and understand genomic data. The genomic data is often in the form of individual samples having chromosomal data (including measurements of at least one event at a particular location on the chromosomes). An event here would indicate some measurement related to the genome. Examples of such measurements include the expression of a gene, an exon at a particular location, the number of copies of a portion of the genome that have been gained or lost, the extent of methylation of the genome at a particular location, the affinity of certain promoters to bind to a particular area on the genome, etc. In some cases, users may calculate a frequency of event based on a frequency of occurrence of the event in the selected sample. For example, it may be desirable to calculate the frequency of aberration, such as the frequency of a gain or loss of chromosomal copies when compared to a reference sample in a selected population of samples. In other circumstances, it may be desirable to review an annotation regarding specific information as related to a particular chromosomal region of the chromosome. Such information might include items such as what genes are present in a location and if there are known copy number polymorphisms in that area (including a list of such polymorphisms). Other items might include information pertaining to the presence of miroRNAs and potential Single Nucleotide Polymorphism (SNP)s in the area, etc.

[0002] Genomic data are available from public or private databases and academic or commercial diagnostic laboratories. Genomic data can also be obtained by sequencing the entire genome of an individual, or a portion thereof. Suitable methods of DNA sequencing include Sanger sequencing, polony sequencing, pyrosequencing, ion semiconductor sequencing, single molecule sequencing, and the like. Sequenced genomic data can be provided as electronic text files, html files, xml files and various other regular databases formats.

[0003] Existing systems available for visualization of chromosomal or genomic annotations, such as the University of California of Santa Cruz browser and the Ensemble Genome Browser, display various annotations for a specific region of the genome. Ensemble is a joint project between the European Molecular Biology Laboratory, the European Bioinformatics Institute and the Wellcome Trust Sanger Institute.

[0004] The molecular data to be processed in a bioinformatics based platform typically concerns genomic data, such as Deoxyribonucleic acid (DNA) data. For example, a well-known method for generating DNA data involves DNA sequencing. DNA sequencing can be performed manually, such as in a lab, or may be performed by an automated sequencer, such as at a core sequencing facility, for the purpose of determining the genetic makeup of a sample of an individual's DNA. The person's genetic information may then be used in comparison to a referent, e.g., a reference genome, so as to determine its variance therefrom. Such variant information may then be subjected to further processing and used to determine or predict the occurrence of a diseased state in the individual.

[0005] Manual or automated DNA sequencing may be employed to determine the sequence of nucleotide bases in a sample of DNA, such as a sample obtained from a subject. Using various different bioinformatics techniques these sequences may then be assembled together to generate the genomic sequence of the subject, and/or mapped and aligned to genomic positions

relative to a reference genome. This sequence may then be compared to a reference genomic sequence to determine how the genomic sequence of the subject varies from that of the reference. Such a process involves determining the variants in the sampled sequence and presents a central challenge to bioinformatics methodologies. Genomic data includes sequences of the DNA bases adenine (A), guanine (G), cytosine (C) and thymine (T). Genomic data includes sequences of the RNA bases adenine (A), guanine (G), cytosine (C) and uracil (U). Genomic data also includes epigenetic information such as DNA methylation patterns, histone deacetylation patterns, and the like.

[0006] "Phenotypic traits" are an organism's observable characteristics, including but not limited to its morphology, development, biochemical or physiological properties, behavior, and products of behavior (such as a bird's nest). Phenotypic traits also include diseases, such as various cancers, heart disease, Age-related Macular Degeneration, and the like.

[0007] Non-limiting general definitions for terms utilized in the pertinent art are set forth below.

[0008] Allele is any two or more alternative forms of the same gene that have the same relative position on homologous chromosomes.

[0009] BAM format is a binary alignment map format, which is the binary version of SAM.

[00010] Chromosome is a strand of DNA that is encoded with genes.

[00011] DNA is deoxyribonucleic acid, which contains the genetic code. It consists of two nucleotide chains in a double helix and joined by hydrogen bonds between complimentary bases of adenine and thymine, and cystosine and guanine.

[00012] Exome is part of the genome formed by exons, the sequences which when transcribed remain within the mature RNA after the introns are removed by RNA splicing.

[00013] Genome is the full set of chromosomes, the genetic material of an organism, and includes genes and non-coding sequences of DNA/RNA.

[00014] Hypertext Transfer Protocol ("HTTP") is a set of conventions for controlling the transfer of information via the Internet from a web server computer to a client computer, and also from a client computer to a web server, and Hypertext Transfer Protocol Secure ("HTTPS") is a communications protocol for secure communication via a network from a web server computer to a client computer, and also from a client computer to a web server by at a minimum verifying the authenticity of a web site.

[00015] Internet is the worldwide, decentralized totality of server computers and data-transmission paths which can supply information to a connected and browser-equipped client computer, and can receive and forward information entered from the client computer.

[00016] Nucleic acid library is a plurality of polynucleotide molecules that are prepared, assemble and/or modified for a specific process.

[00017] Phenotype is the composite of an organism's observable characteristics or traits, such as its morphology, development, biochemical or physiological properties, phenology, behavior, and products of behavior. A phenotype results from the expression of an organism's genes as well as the influence of environmental factors.

[00018] SAM is sequence alignment map format is a text format of mapping sequence reads (sequence information from a fragment whose physical genomic position is unknown) with a matching sequence in a reference genome.

[00019] Single Nucleotide Polymorphism ("SNP") is a DNA sequence variation occurring when a single nucleotide in the genome differs between members of a species (or between paired chromosomes in an individual).

[00020] URL or Uniform Resource Locator is an address on the World Wide Web.

[00021] User Interface or UI is the junction between a user and a computer program. An interface is a set of commands or menus through which a user communicates with a program. A command driven interface is one in which

the user enter commands. A menu-driven interface is one in which the user selects command choices from various menus displayed on the screen.

[00022] Variant calling is a method of identifying factual differences between sequence reads of test samples and a reference sequence. Variant calling is used to identify somatic variants with a high degree of confidence.

[00023] Web-Browser is a complex software program, resident in a client computer, that is capable of loading and displaying text and images and exhibiting behaviors as encoded in HTML (HyperText Markup Language) from the Internet, and also from the client computer's memory. Major browsers include MICROSOFT INTERNET EXPLORER, NETSCAPE, APPLE SAFARI, MOZILLA FIREFOX, and OPERA.

[00024] Web-Server is a computer able to simultaneously manage many Internet information-exchange processes at the same time. Normally, server computers are more powerful than client computers, and are administratively and/or geographically centralized. An interactive-form information-collection process generally is controlled from a server computer, to which the sponsor of the process has access.

[00025] There is a need for distributing genomic data from a source to a recipient in a secure and efficient means.

Summary of the Invention

[00026] One aspect of the present invention is a method for correlating genome data with EMR/PHR data. The method includes identifying a plurality of sources of genome data. Each source of the plurality of sources comprises a plurality of genome files. The method also includes indexing and encrypting each of the plurality of genome files utilizing a processor at a source site for the source. Each of the plurality of genome files is identified with a unique DID. The method also includes generating an index file for each of the plurality of genome files. The method also includes transmitting each index

file to a central depository. Each index file is stored as part of a plurality of index files. The method also includes identifying electronic medical record (EMR) and/or personal health record (PHR) data at each source of the plurality of sources of genome data. Each EMR/PHR data has a unique DID, and each EMR/PHR data matches a genome file of the plurality of genome files. The method also includes correlating each genome file of the plurality of genome files with a corresponding EMR/PHR data.

[00027] Another aspect of the present invention is a system for searching correlated genome data and EMR data. The system comprises a central depository site, a plurality of sources for genome data, a plurality of sources for EMR/PHR data, and search browser. The central depository site comprises a plurality of index files. Each of the plurality of index files represents encrypted genome files. Each of plurality of sources for genome data comprises a database of encrypted genome files. Each of the plurality of encrypted genome files has a unique DID. Each of the plurality of sources for EMR/PHR data comprises a database of EMR/PHR files. Each of the plurality of EMR/PHR files has a unique DID. Each EMR/PHR file matches a genome file having the same unique DID. The browser is for searching the plurality of index files.

[00028] Yet another aspect of the present invention is a method for privacy controlled genomic visualization. The method includes indexing and encrypting each of the plurality of genome files utilizing a processor at a source site. Each of the plurality of genome files is identified with a unique DID. The method also includes generating an index file for each of the plurality of genome files. The method also includes transmitting each index file to a brokering server. Each index file is stored as part of a plurality of searchable index files. The method also includes identifying electronic medical record (EMR) and/or personal health record (PHR) data at each source of a plurality of sources of genome data. Each EMR/PHR data has a unique DID. Each EMR/PHR data matches a genome file of the plurality of

genome files. The method also includes matching each genome file of the plurality of genome files with a corresponding EMR/PHR data. The method also includes searching the plurality of searchable index files at a browser for the brokering server. The owner of an encrypted genome file controls access to the encrypted genome file and tracks the encrypted genome file.

[00029] Yet another aspect of the present invention is a method for privacy controlled genomic visualization. The method includes searching a plurality of index files for a specific genome level. Each of the plurality of index files represents encrypted genome data for an owner of the data. The specific level is selected from a chromosome level, an exome level, a gene level, an allele panel, or at an individual SNP or allele level. The method also includes identifying a set of index files of the plurality of index files for review. The method also includes requesting permission for access to the encrypted genome data and EMR/PHR data from an owner of the data for each index file of the set of index files. The method also includes receiving permission from the owner of the data. The method also includes receiving the genome data and the EMR/PHR data.

Brief Description of the Drawings

[00030] FIG. 1 is a block diagram of a system for genomic visualization with privacy control.

[00031] FIG. 2 is a flow chart for a method for correlating genome data with EMR/PHR data.

[00032] FIG. 3 is a flow chart for a method for privacy controlled genomic visualization.

[00033] FIG. 4 is a flow chart for a method for privacy controlled genomic visualization.

[00034] FIG. 5 is a communications sequence diagram for genomic visualization.

[00035] FIG. 6 is a block diagram for a method for genomic visualization.

[00036] FIG. 7 is a block diagram for a method for processing of the genomic data at the genomic data site.

[00037] FIG. 7A is an illustration of an index file for a patient.

Best Mode(s) For Carrying Out The Invention

[00038] A system 100 for genomic visualization with privacy control is shown in FIG. 1. The system 100 includes a central depository site 101, a source of genomic data 102, a source of EMR data 103, and owner 104 (the patient) of the genomic data and EMR data, and a plurality of search entities 105a-c. The central depository site 101 functions as a brokerage of data between sources that have generated genomic data 102 and EMR data 103, and multiple searchers that utilize the information for academic, commercial and other purposes. The search entities 105 are researchers like universities, biotechnology companies, hospitals, and the like. The owner 104 preferably controls access to the unencrypted data and also tracks the data as it is distributed to search entities 105a-c.

[00039] The EMR data of a patient includes general health records, medical procedure records, allergies, illnesses, and the like of the patient.

[00040] The genomic data is preferably encrypted and indexed and stored locally. Thus, the central site 101 is not a warehouse of data, requiring an enormous storage data facility. The central site 101 only maintains a plurality of index files that can be easily searched. Further, the processing of the genomic data is performed at the genomic data site 102. A general process for processing the genomic data involves processing sequence data to generate a sequenced data file, processing the sequenced data file to generate an aligned data file, and processing the aligned data file to generate a variant called data file (VCF). One specific process for processing the genomic data involves processing sequence data to generate a FASTQ file, processing the FASTQ file to generate a binary sequence alignment map (BAM) file, and processing

the BAM file to generate a variant call file (VCF). A more detail description is set forth in Van Rooyen et al., U.S. Patent Publication Number 20140371110 for *Bioinformatics Systems, Apparatuses, and Methods Executed On An Integrated Circuit Processing Platform*, which is hereby incorporated by reference in its entirety. The VCF is indexed into an index file and then encrypted. A DID (De-identified Identifier, e.g., a unique ID token that includes no identifying information like a patient's email, phone number, date-of-birth, zip, etc.) number is provided to each index file to maintain privacy and anonymity. The algorithm utilized to generate the DID for the index file is the same one used for the EMR data, and therefore the files can be matched based on the DID number.

[00041] The index files are transmitted from each genomic data site 102 to the central site 101. The index files are searchable at a chromosome level, exome level, gene level, allele panel, or at an individual SNP or allele level.

[00042] FIG. 2 illustrates a flow chart for a method 200 for correlating genome data with EMR/PHR data. At block 201, sources of genome data are identified. Each source comprises a plurality of genome files. At block 202, each of the plurality of genome files indexed and encrypted utilizing a processor at a source site for the source. Each of the plurality of genome files is identified with a unique DID. At block 203, an index file is generated for each of the plurality of genome files. At block 204, each index file is transmitted to a central depository site. Each index file is stored as part of a plurality of index files. At block 205, the electronic medical record (EMR) and/or personal health record (PHR) data is identified at each source of the plurality of sources of genome data. Each EMR/PHR data has a unique DID, and each EMR/PHR data matches a genome file of the plurality of genome files. At block 206, each genome file of the plurality of genome files is correlated with a corresponding EMR/PHR data.

[00043] FIG. 3 illustrates a flow chart for a method 300 for privacy controlled genomic visualization. At block 301, each of the plurality of genome files is

indexed and encrypted utilizing a processor at a source site. Each of the plurality of genome files is identified with a unique DID. At block 302, an index file is generated for each of the plurality of genome files. At block 303, each index file is transmitted to a brokering server. Each index file is stored as part of a plurality of searchable index files. At block 304, electronic medical record (EMR) and/or personal health record (PHR) data is identified at each source of a plurality of sources of genome data. Each EMR/PHR data has a unique DID. Each EMR/PHR data matches a genome file of the plurality of genome files. At block 305, each genome file of the plurality of genome files is matched with a corresponding EMR/PHR data. At block 306, the plurality of searchable index files is search at a browser for the brokering server. The owner of an encrypted genome file controls access to the encrypted genome file and tracks the encrypted genome file.

[00044] FIG. 4 illustrates a flow chart for a method 400 for privacy controlled genomic visualization. At block 401, a plurality of index files is searched for a specific genome level. Each of the plurality of index files represents encrypted genome data for an owner of the data. The specific level is selected from a chromosome level, an exome level, a gene level, an allele panel, or at an individual SNP or allele level. At block 402, a set of index files of the plurality of index files is identified for review. At block 403, permission for access to the encrypted genome data and EMR/PHR data is requested from an owner of the data for each index file of the set of index files. At block 404, permission is received from the owner of the data. At block 405, the genome data and the EMR/PHR data are received by the searcher.

[00045] FIG. 5 illustrates a communication sequence diagram 500 for genomic visualization with privacy control. A search entity 105 searches the index files available at the central search site 101 using a central site browser. The search results are returned to the search entity 105. The search entity then requests permission for the unencrypted genome files which are represented by some or all of the index files. Since the index files only provide information on a

variation, with no identifying information, the search entity must now receive permission from the owner of the genome to gain access to the more detailed information. The central search site 101 acts as a broker and presents the request to the owner. Preferably, the requests involve details behind the research so that the owner will know what his or her information is to be used for by the search entity 105. The owner then grants permission to the central search site 105 for access to the unencrypted genome data and the EMR data that matches the genome data. The central search site 101 then requests that the EMR data file be sent from the EMR data site 103, and that the genome data be sent from the genomic data site 102. The central search site 101 unencrypts the data and transfers the EMR data and the genome data to the search entity 105. The owner 104 tracks the data sent to the search entity, enabling privacy control by the owner 104.

[00046] FIG. 6 is a block diagram for a method 600 for genomic visualization.

[00047] FIG. 7 is a block diagram for a method 700 for processing of the genomic data at the genomic data site. FIG. 7A is an illustration of an index file for a patient.

[00048] The networks utilized with the present invention may be one or more of a wireless network, a wired network or any combination of wireless network and wired network. The networks utilized may include one or more of an Internet network, a wireless local area network ("LAN"), a cellular network, a fiber optics network, a passive optical network, a cable network, a satellite network (e.g., operating in Band C, Band Ku or Band Ka), a Global System for Mobile Communication, a Personal Communication Service, a Personal Area Network Wi-Fi, Fixed Wireless Data, IEEE 802.11a, 802.11b, 802.15.1, 802.11n and 802.11g or any other wired or wireless network for transmitting and receiving a data signal. The network may utilize one or more protocols of one or more network elements to which it is communicatively coupled. The network may translate to or from other protocols to one or more protocols of devices connected to the network. The invention may utilized a

plurality of interconnected networks, such as, for example, a service provider network, the Internet, a broadcaster's network, a cable television network, a corporate network, and a home net.

[00049] Each of the interface descriptions preferably discloses use of at least one communication protocol to establish handshaking or bi-directional communications. These protocols preferably include but are not limited to XML, HTTP, TCP/IP, Serial, UDP, FTP, Web Services, WAP, SMTP, SMPP, DTS, Stored Procedures, Import/Export, Global Positioning Triangulation, IM, SMS, MMS, GPRS and Flash. The storage of data may be network accessible storage and may be local, remote, or a combination thereof. The storage of data may utilize a redundant array of inexpensive disks, tape, disk, a storage area network, an internet small computer systems interface a common Internet File System, network attached storage, a network file system, or other computer accessible storage. The databases used with the system preferably include but are not limited to MSSQL, Access, MySQL, Progress, Oracle, DB2, Open Source DBs and others. Operating system used with the system preferably include Microsoft 2010, XP, Vista, 200o Server, 2003 Server, 2008 Server, Windows Mobile, Linux, Android, Unix, I series, AS 400 and Apple OS.

[00050] The underlying protocol at a server is preferably Internet Protocol Suite (Transfer Control Protocol/Internet Protocol ("TCP/IP")), and the transmission protocol to receive a file is preferably a file transfer protocol ("FTP"), Hypertext Transfer Protocol ("HTTP"), Secure Hypertext Transfer Protocol ("HTTPS") or other similar protocols. The transmission protocol ranges from SIP to MGCP to FTP and beyond. The protocol at the server is preferably HTTPS.

[00051] It is further noted that the software described herein may be tangibly embodied in one or more physical media, such as, but not limited to, a compact disc ("CD"), a digital versatile disc ("DVD"), a floppy disk, a hard

drive, read only memory ("ROM"), random access memory ("RAM"), as well as other physical media capable of storing software, or combinations thereof.

[00052] Numerous references were made regarding servers, services, interfaces, portals, platforms, or other systems formed from computing devices. It should be appreciated that the use of such terms is deemed to represent one or more computing devices having at least one processor configured to execute software instructions stored on a computer readable tangible, non-transitory medium. For example, a server can include one or more computers operating as a web server, database server, or other type of computer server in a manner to fulfill described roles, responsibilities, or functions. The genomic visualization system may utilize various computing devices including servers, graphical user interfaces, databases, engines, controllers, or other types of computing devices operating individually or collectively. One skilled in the pertinent art will appreciate that the computing devices comprise a processor configured to execute software instructions stored on a tangible, non-transitory computer readable storage medium (e.g., hard drive, solid state drive, RAM, flash, ROM, etc.). The software instructions preferably configure the computing device to provide the roles, responsibilities, or other functionality as discussed below with respect to the invention. In preferred embodiments, the servers, databases, or interfaces preferably exchange data using standardized protocols or algorithms, possibly based on HTTP, HTTPS, AES, public-private key exchanges, web service APIs, known financial transaction protocols, or other electronic information exchanging methods. Data exchanges preferably are conducted over the Internet, LAN, a packet-switched network, WAN, VPN, or other type of packet switched network. One skilled in the pertinent art will appreciate that the form of a computer program product stored by one or more computer-readable storage media having computer-readable program code, or instructions, embodied in or on the storage media. Any suitable computer readable storage media may be utilized, including hard disks, CD-ROMs,

optical storage devices, magnetic storage devices, flash devices and/or any combination thereof. In addition, various signals representing data or events as described herein may be transferred between a source and a destination in the form of electromagnetic waves traveling through signal-conducting media such as metal wires, optical fibers, and/or wireless transmission media--e.g. air and/or space. Data may move between various entities in any of the embodiments of the invention via electronic transmission or manual means. Electronic transmission may utilize email, SMS or any other suitable method. Manual exchange may utilize floppy disks, USB drives, CDs, DVDs or any other suitable mechanism.

[00053] An exemplary hardware configuration of a computing system utilized with the invention preferably includes at least one processor or central processing unit (CPU). The CPUs are preferably interconnected via a system bus to a RAM, a ROM, input/output (I/O) adapter, user interface adapter, a communication adapter for connecting the system to a data processing network, the Internet, an Intranet, a LAN, or the like, and a display adapter for connecting the bus to a display device.

[00054] Any combination of one or more computer readable medium(s) may be utilized with the invention. The computer readable medium may be a computer readable signal medium or a computer readable storage medium. A computer readable storage medium may be, for example, but not limited to, an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, or device, or any suitable combination of the foregoing. More specific examples (a non-exhaustive list) of the computer readable storage medium include an electrical connection having one or more wires, a portable computer diskette, a hard disk, a RAM, a ROM, an erasable programmable read-only memory, an optical fiber, a portable CD-ROM, an optical storage device, a magnetic storage device, or any suitable combination of the foregoing. A computer readable storage medium may be any tangible

medium that can contain, or store a program for use by or in connection with a system, apparatus, or device running an instruction.

[00055] Computer program code for carrying out operations for aspects of the invention may be written in any combination of one or more programming languages, including an object oriented programming language such as Java, Smalltalk, C++ or the like and conventional procedural programming languages, such as the "C" programming language or similar programming languages. The program code may run entirely on the user's computer, partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through any type of network, including a LAN or a WAN or the connection may be made to an external computer through the Internet using an Internet Service Provider.

[00056] There are many processing stages for data from DNA (or RNA) sequencing, which can vary depending on the sequencing technology and the application. Processing steps include: Signal processing on electrical measurements from the sequencer; Image processing on optical measurements from the sequencer; Base calling using processed signal or image data to determine the most likely nucleotide sequence and confidence scores; Filtering sequenced reads with low quality or polyclonal clusters; Detecting and trimming adapters, key sequences, barcodes, and low quality read ends; De novo sequence assembly, utilizing De Bruijn graphs and/or sequence graphs; De Bruijn and sequence graph construction, editing, trimming, cleanup, repair, coloring, annotation, comparison, transformation, splitting, splicing, analysis, subgraph selection, traversal, iteration, recursion, searching, filtering, import, export; Mapping reads to a reference genome; Aligning reads to candidate mapping locations in a reference genome; Local assembly of reads mapped to a reference region; Sorting reads by aligned position; Marking duplicate reads, including PCR or optical duplicates; Re-alignment of multiple overlapping

reads for indel consistency; Base quality score recalibration; Variant calling (single sample or joint); Structural variant analysis; Copy number variant analysis; Somatic variant calling (tumor sample only, matched tumor/normal, or tumor / unmatched normal); RNA splice junction detection; RNA alternative splicing analysis; RNA transcript assembly; RNA transcript expression analysis; RNA differential expression analysis; RNA variant calling; DNA/RNA difference analysis; DNA methylation analysis and calling; Variant quality score recalibration; Variant filtering; Variant annotation from known variant databases; Sample contamination detection and estimation; Phenotype prediction, disease testing; Treatment response prediction, custom treatment design; Ancestry and mutation history analysis; Population DNA analysis, genetic marker identification; Encoding genomic data into standard formats (e.g. FASTA, FASTQ, SAM, BAM, VCF, BCF); Decoding genomic data from standard formats; Querying, selecting or filtering genomic data subsets; General compression and decompression for genomic files (gzip, BAM compression); Specialized compression and decompression for genomic data (CRAM); Genomic data encryption and decryption; Statistics calculation, comparison, and presentation from genomic data; Genomic result data comparison, accuracy analysis and reporting; Genomic file storage, archival, retrieval, backup, recovery, and transmission; Genomic database construction, querying, access management, data extraction.

[00057] A more detailed description of a system for analysis of biological and chemical materials is set forth in van Rooyen et al., U.S. Patent Publication Number 20140371110 for *Bioinformatics Systems, Apparatuses, and Methods Executed On An Integrated Circuit Processing Platform*, which is hereby incorporated by reference in its entirety. A more detailed description of a system for analysis of biological and chemical materials is set forth in van Rooyen et al., U.S. Patent Publication Number 20140309944 for *Bioinformatics Systems, Apparatuses, and Methods Executed On An Integrated Circuit Processing Platform*, which is hereby incorporated by

reference in its entirety. A more detailed description of a system for analysis of biological and chemical materials is set forth in van Rooyen et al., U.S. Patent Publication Number 20140236490 for *Bioinformatics Systems, Apparatuses, and Methods Executed On An Integrated Circuit Processing Platform*, which is hereby incorporated by reference in its entirety. A more detailed description of a system for analysis of biological and chemical materials is set forth in van Rooyen et al., U.S. Patent Number 9014989 for *Bioinformatics Systems, Apparatuses, and Methods Executed On An Integrated Circuit Processing Platform*, which is hereby incorporated by reference in its entirety. A more detailed description of a system for analysis of biological and chemical materials is set forth in U.S. Patent Publication Number 20150339437, for *Dynamic Genome Reference Generation For Improved NGS Accuracy And Reproducibility*, filed February 24, 2015, which is hereby incorporated by reference in its entirety. A description of a GFET is set forth in Hoffman et al., U.S. Patent Application Number 14/963253, filed on December 9, 2015, for *Chemically Sensitive Field Effect Transistor*, which is hereby incorporated by reference in its entirety.

Claims

1. A method for correlating genome data with EMR/PHR data, the method comprising:

identifying a plurality of sources of genome data, each of source of the plurality of sources comprising a plurality of genome files;

indexing and encrypting each of the plurality of genome files utilizing a processor at a source site for the source, each of the plurality of genome files identified with a unique DID;

generating an index file for each of the plurality of genome files;

transmitting each index file to a central depository, wherein each index is stored as a plurality of index files;

identifying electronic medical record (EMR) and/or personal health record (PHR) data at each source of the plurality of sources of genome data, wherein each EMR/PHR data has a unique DID, wherein each EMR/PHR data matches a genome file of the plurality of genome files; and

correlating each genome file of the plurality of genome files with a corresponding EMR/PHR data.

2. The method according to claim 1 wherein the EMR/PHR data comprises phenotypic data.

3. The method according to claim 1 wherein the plurality of index files is searchable at a chromosome level, exome level, gene level, allele panel, or at an individual SNP or allele level.

4. The method according to claim 1 wherein each genome file of the plurality of genome files is tracked and controlled by an owner of the genome file.

5. The method according to claim 1 wherein generating an index file comprises processing sequence data to generate a sequenced data file, processing the sequenced

data file to generate an aligned data file, and processing the aligned data file to generate a called data file (VCF).

6. The method according to claim 1 wherein generating an index file comprises processing sequence data to generate a FASTQ file, processing the FASTQ file to generate a binary sequence alignment map (BAM) file, and processing the BAM file to generate a variant call file (VCF).

7. The method according to claim 6 further comprising encrypting the VCF.

8. The method according to claim 1 wherein the plurality of sources comprises a plurality of genomic data sites.

9. The method according to claim 1 further searching the plurality of index files for a specific anomaly.

10. The method according to claim 1 further comprising brokering the plurality of index files and the matching EMR data.

11. A system for searching correlated genome data and EMR data, the system comprising:

a central depository site comprising a plurality of index files, each of the plurality of index files representing encrypted genome files;

a plurality of sources for genome data, each of the plurality of sources comprising a database of encrypted genome files, each of the plurality of encrypted genome files having a unique DID;

a plurality of sources for EMR/PHR data, each of plurality of sources comprising a database of EMR/PHR files, each of the plurality of EMR/PHR files

having a unique DID, wherein an EMR/PHR file matching a genome file have the same unique DID; and

a browser for searching the plurality of index files.

12. The system according to claim 11 wherein the EMR/PHR data comprises phenotypic data.

13. The system according to claim 11 wherein the plurality of index files is searchable at a chromosome level, exome level, gene level, allele panel, or at an individual SNP or allele level.

14. The system according to claim 11 wherein each genome file of the plurality of genome files is tracked and controlled by an owner of the genome file.

15. The system according to claim 11 wherein the plurality of sources comprises a plurality of genomic data sites.

16. The system according to claim 11 wherein generating an index file comprises processing sequence data to generate a sequenced data file, processing the sequenced data file to generate an aligned data file, and processing the aligned data file to generate a called data file (VCF).

17. The system according to claim 11 wherein each index file is generated by processing sequence data to generate a FASTQ file, processing the FASTQ file to generate a binary sequence alignment map (BAM) file, and processing the BAM file to generate a variant call file (VCF).

18. The system according to claim 17 wherein the VCF is encrypted.

19. The system according to claim 11 further comprising a privacy control engine configured to allow an owner of an encrypted genome file to control access to the encrypted genome file and track the encrypted genome file.

20. A method for privacy controlled genomic visualization, the method comprising:

indexing and encrypting each of the plurality of genome files utilizing a processor at a source site, each of the plurality of genome files identified with a unique DID;

generating an index file for each of the plurality of genome files;

transmitting each index file to a brokering server, wherein each index file is stored as part of a plurality of searchable index files;

identifying electronic medical record (EMR) and/or personal health record (PHR) data at each source of a plurality of sources of genome data, wherein each EMR/PHR data has a unique DID, wherein each EMR/PHR data matches a genome file of the plurality of genome files; and

matching each genome file of the plurality of genome files with a corresponding EMR/PHR data;

searching the plurality of searchable index files at a browser for the brokering server;

wherein an owner of an encrypted genome file controls access to the encrypted genome file and tracks the encrypted genome file.

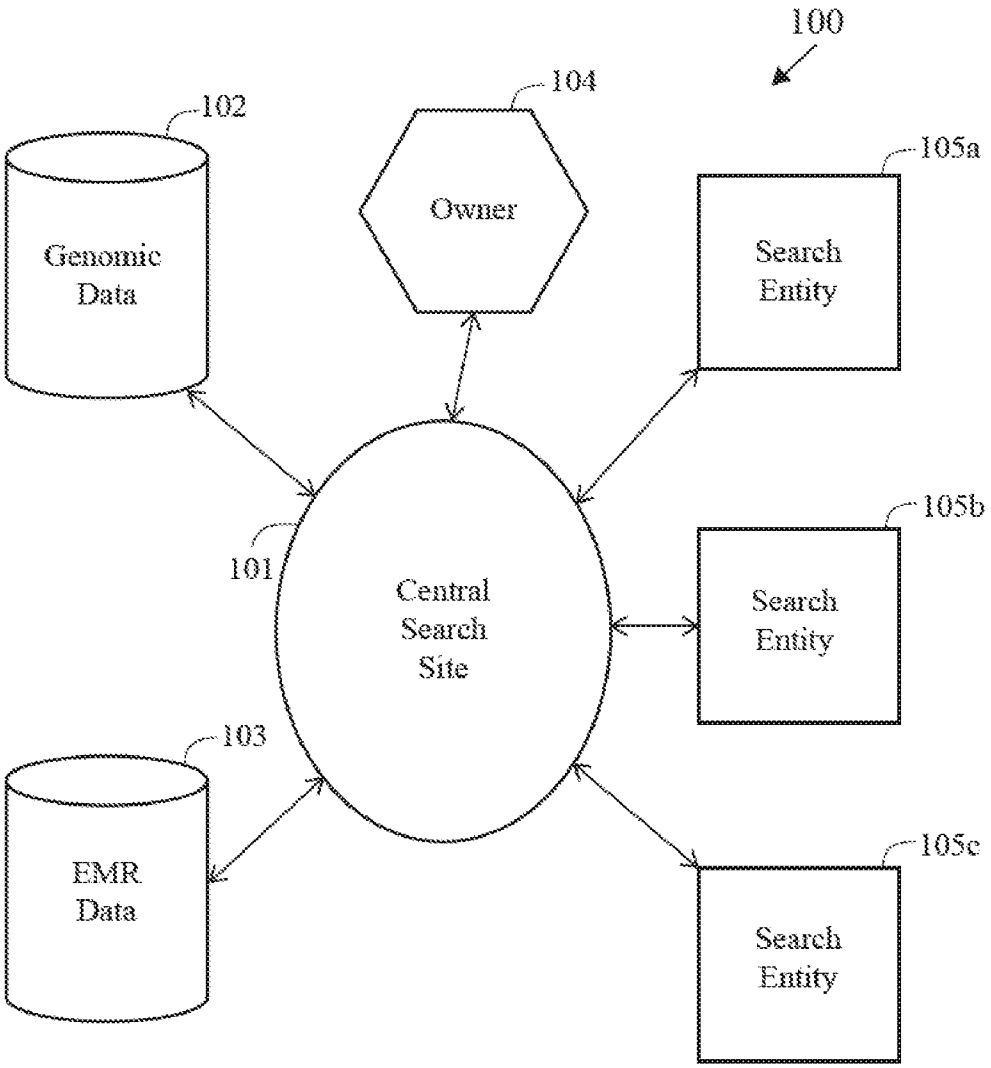
100
102
Genomic Data
104
Owner
105a
Search Entity
101
Central Search Site
105b
Search Entity
103
EMR Data
105c
Search Entity

FIG. 1

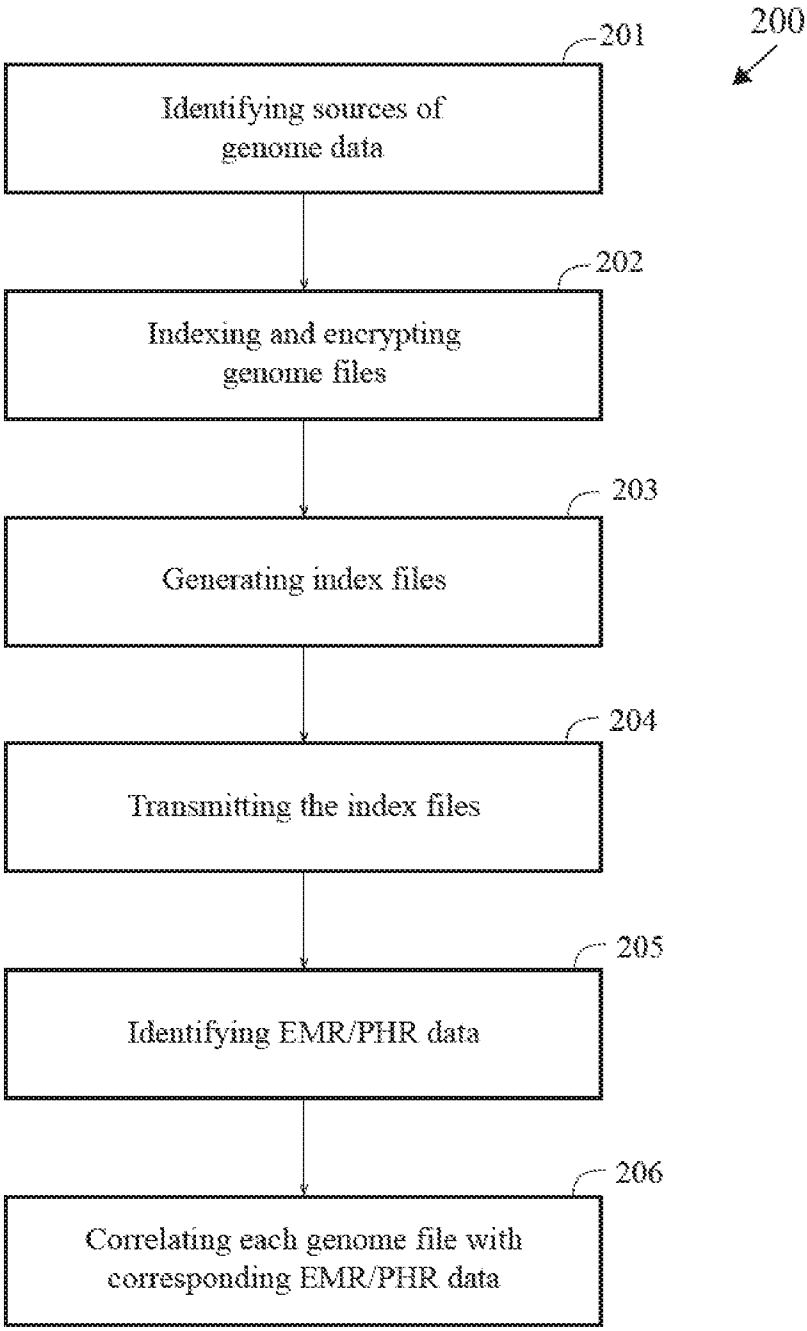
200
201
Identifying sources of genome data
202
Indexing and encrypting genome files
203
Generating index files
204
Transmitting the index files
205
Identifying EMR/PHR data
206
Correlating each genome file with corresponding EMR/PHR data

FIG. 2

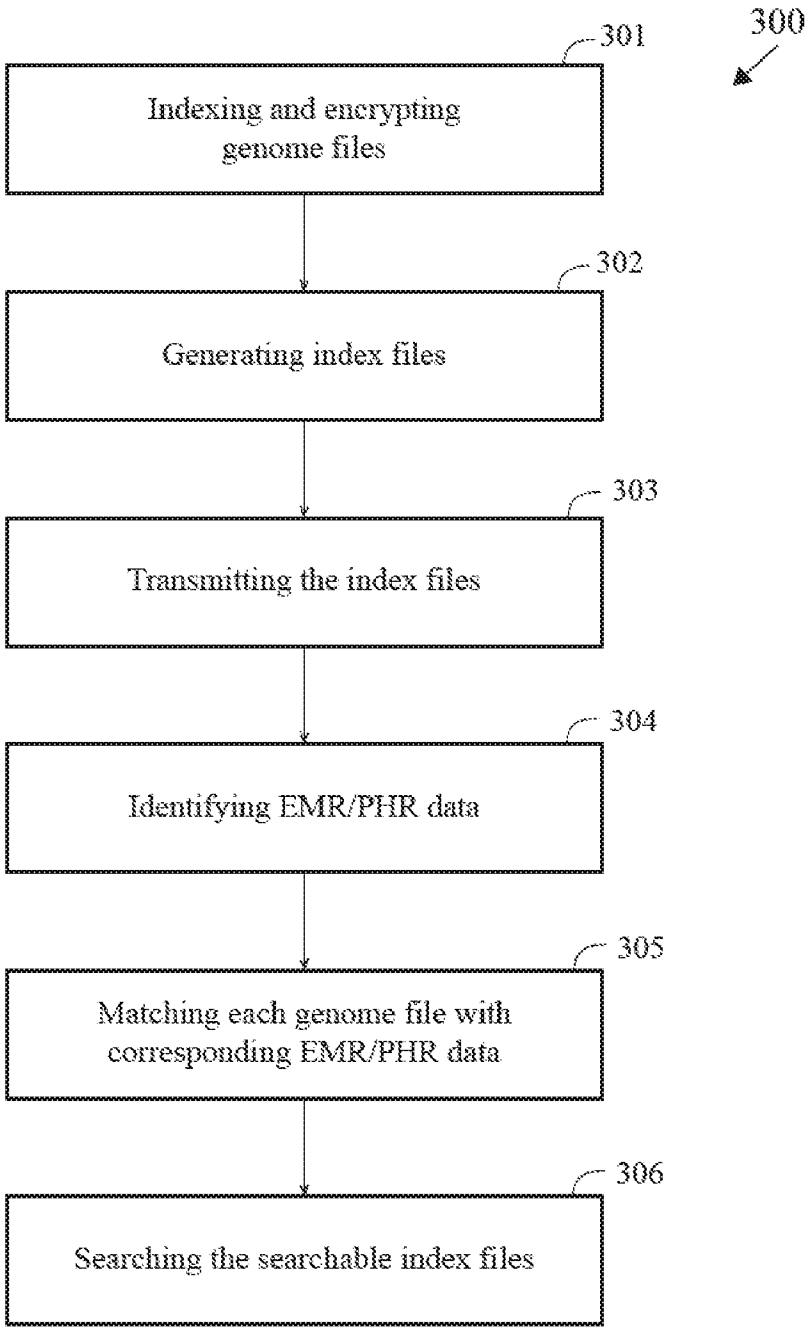
300
301
Indexing and encrypting genome files
302
Generating index files
303
Transmitting the index files
304
Identifying EMR/PHR data
305
Matching each genome file with corresponding EMR/PHR data
306
Searching the searchable index files

FIG. 3

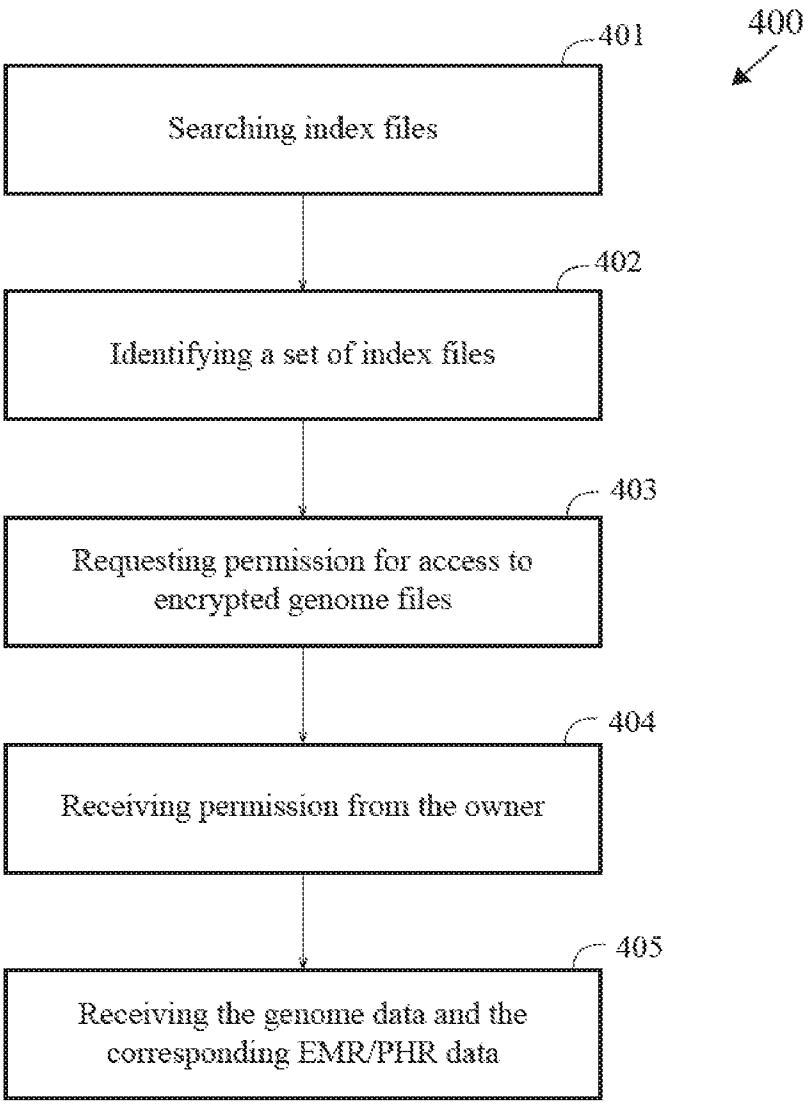
400
401
Searching index files
402
Identifying a set of index files
403
Requesting permission for access to encrypted genome files
404
Receiving permission from the owner
405
Receiving the genome data and the corresponding EMR/PHR data

FIG. 4

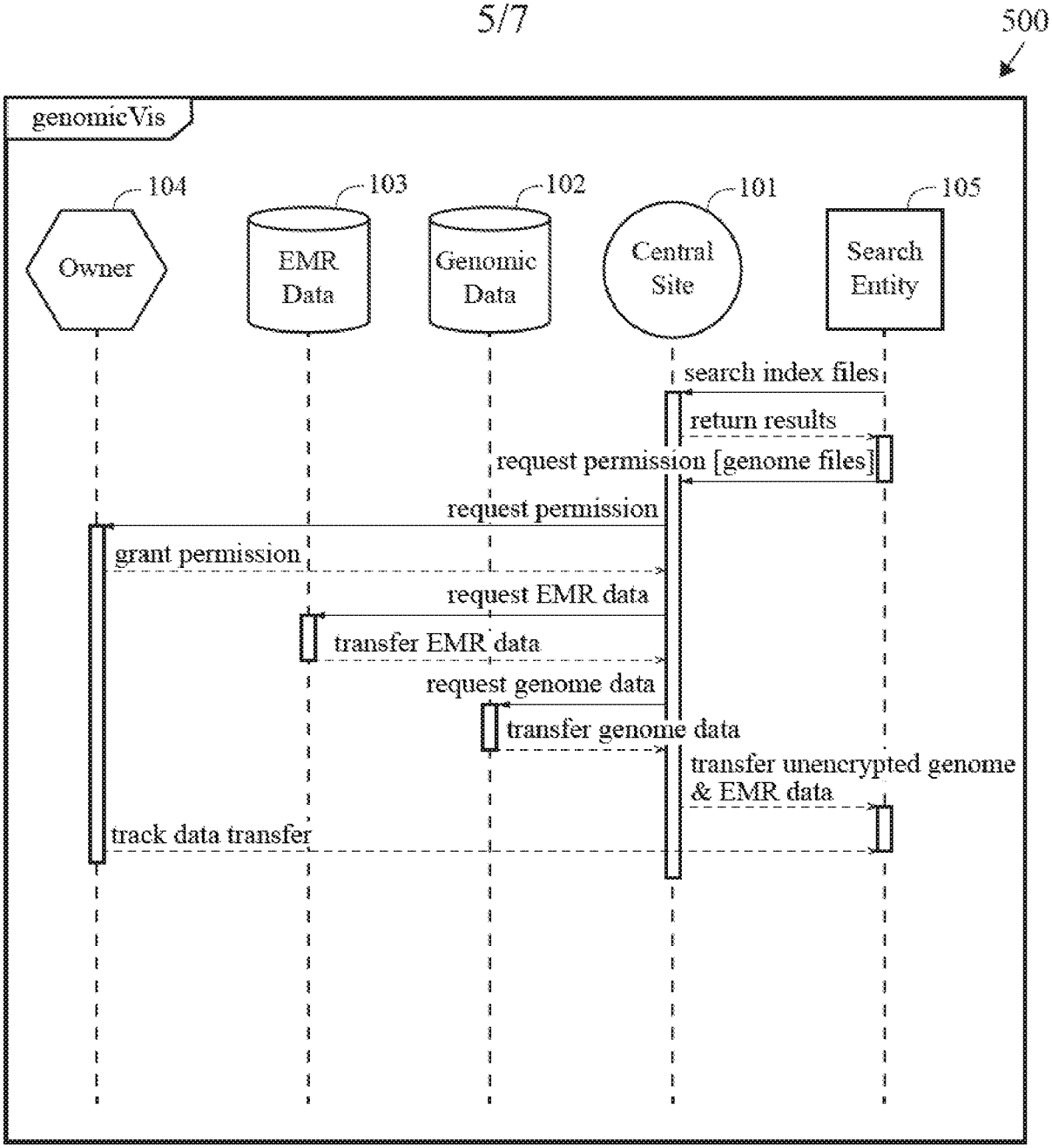
500
genomicVis
104
Owner
103
EMR Data
102
Genomic Data
101
Central Site
105
Search Entity
search index files
return results
request permission [genome files]
request permission
grant permission
request EMR data
transfer EMR data
request genome data
transfer genome data
transfer unencrypted genome & EMR data
track data transfer

FIG. 5

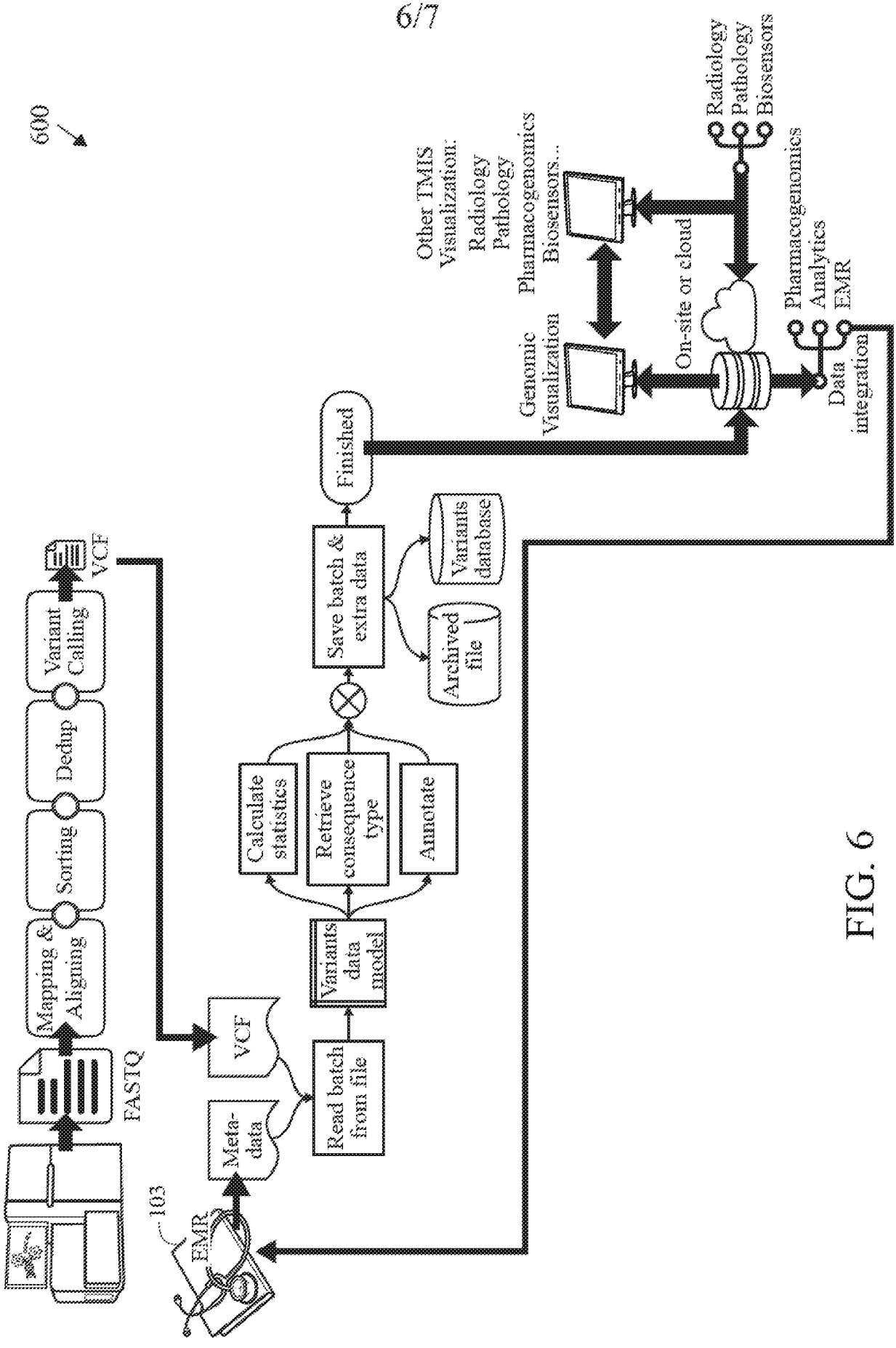
600
Other TMIS Visualization:
Radiology
Pathology
Pharmacogenomics
Biosensors...
Radiology
Pathology
Biosensors
On-site or cloud
Genomic Visualization
Pharmacogenomics
Analytics
EMR
Data integration
Finished
Save batch & extra data
Variants database
Archived file
Calculate statistics
Retrieve consequence type
Annotate
Variants data model
Read batch from file
VCF
Meta-data
103
EMR
FASTQ
Mapping & Aligning
Sorting
Dedup
Variant Calling
VCF

FIG. 6

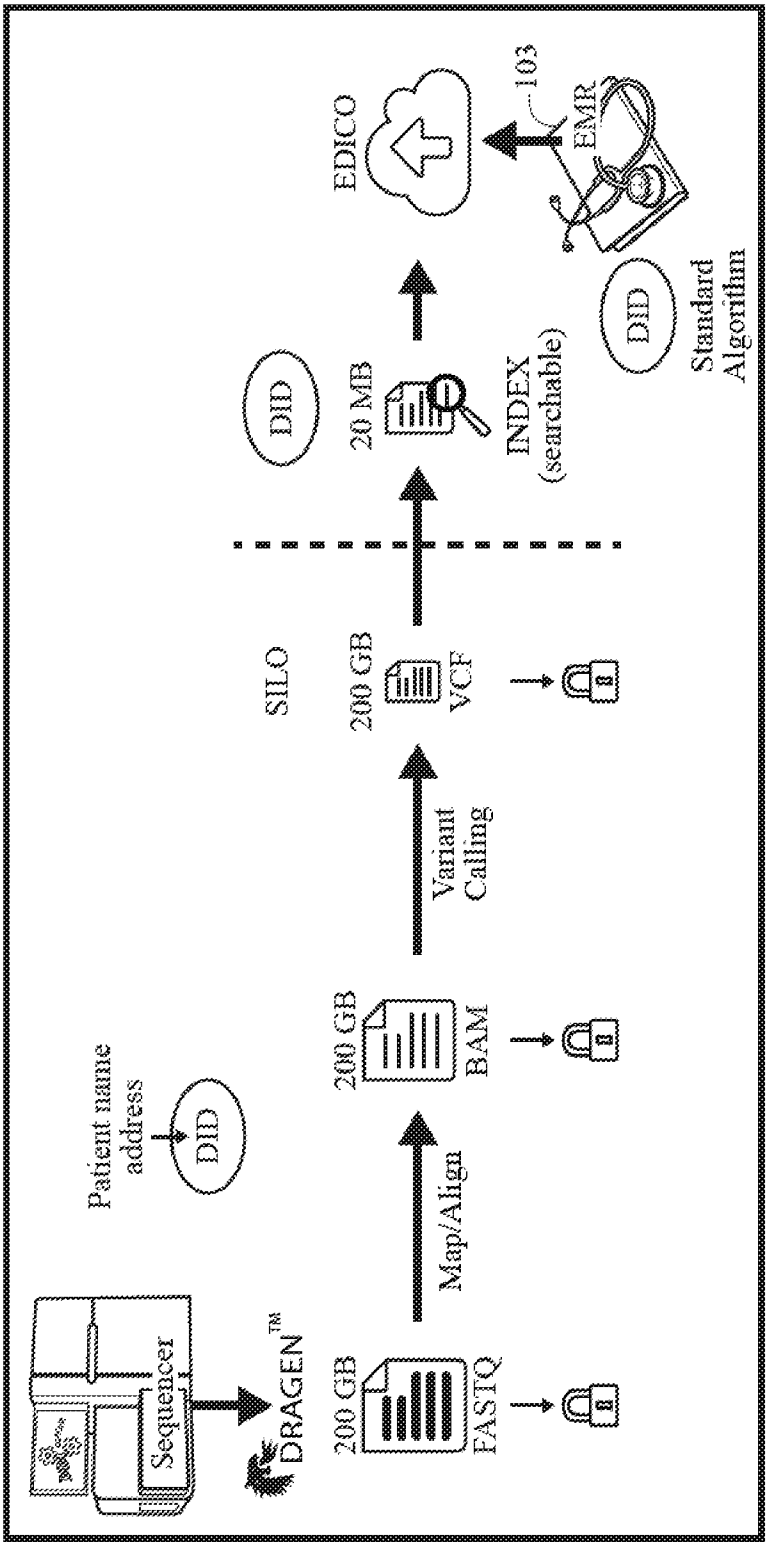
700
Sequencer
DRAGEN™
200 GB
FASTQ
Map/Align
200 GB
BAM
Variant Calling
200 GB
VCF
SILO
Patient name address
DID
DID
20 MB
INDEX (searchable)
DID
Standard Algorithm
EDICO
103
EMR

FIG. 7

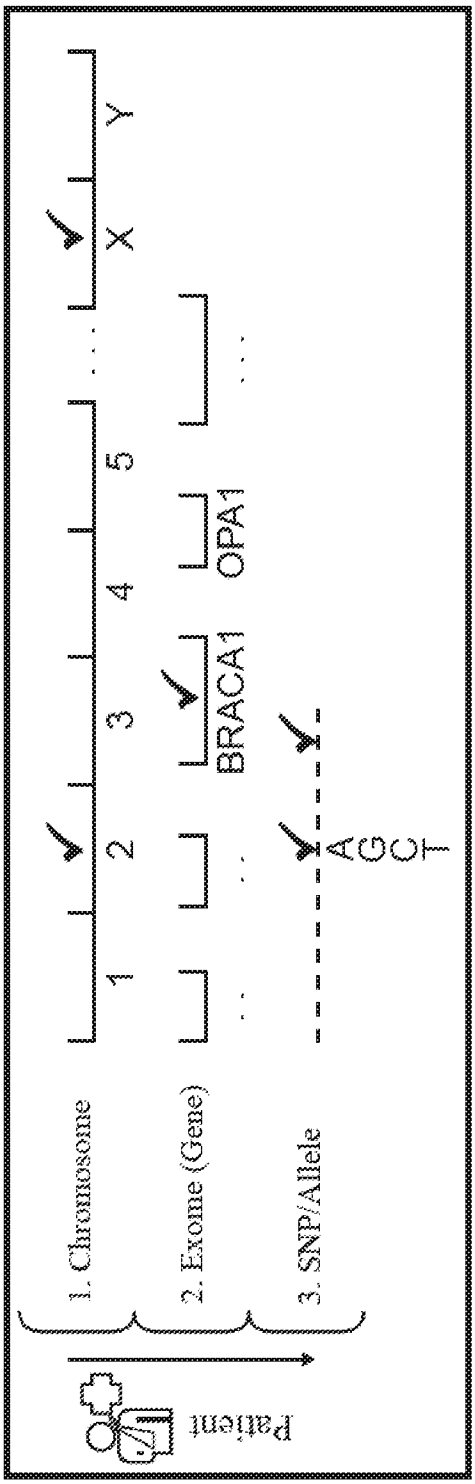
Patient
1. Chromosome
2. Exome (Gene)
3. SNP/Allele
1
2
3
4
5
X
Y
BRACA1
OPA1
A
G
C
T

FIG. 7A