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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
07 December 2017 (07.12.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/210456 A1

(51) International Patent Classification:
G06F 19/18 (2011.01) *G06F 19/24* (2011.01)

(21) International Application Number:
PCT/US2017/035496

(22) International Filing Date:
01 June 2017 (01.06.2017)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

62/344,131	01 June 2016 (01.06.2016)	US
62/410,261	19 October 2016 (19.10.2016)	US

(71) Applicant: **LINEAGEN, INC.** [US/US]; 2677 E. Parleys Way, Salt Lake City, Utah 84109 (US).

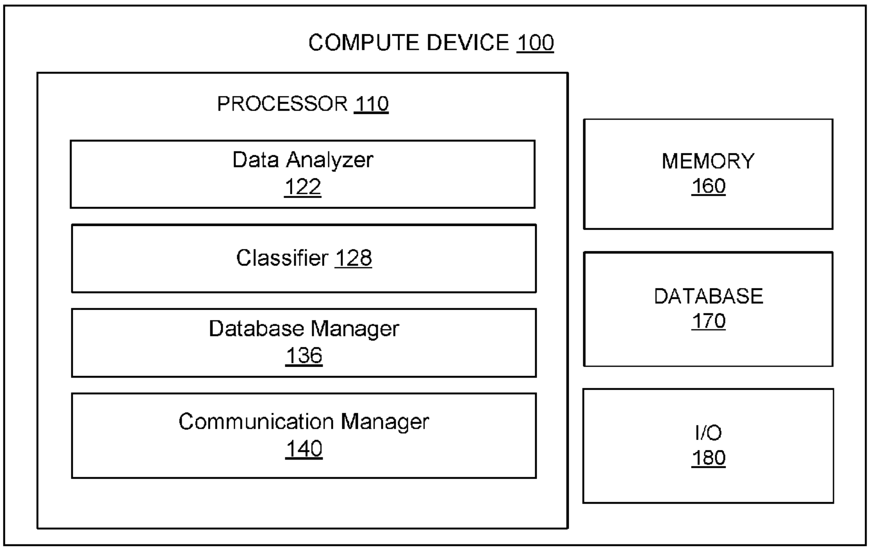
(72) Inventors: **VENKATASUBRAMANIAN, Suresh**; 4603 S. Wallace Lane, Salt Lake City, Utah 84117 (US). **HO, Karen S.**; 4603 S. Wallace Lane, Salt Lake City, Utah 84117 (US).

(74) Agent: **SUD, Dhruv** et al.; Cooley LLP, 1299 Pennsylvania Avenue, NW, Suite 700, Washington, District of Columbia 20004 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, 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,

(54) Title: SYSTEMS, DEVICES AND METHODS FOR IMPROVED ANALYSIS AND STORAGE OF GENOTYPIC AND PHENOTYPIC DATA

FIG. 1



(57) Abstract: An apparatus includes a database storing: a set of sequences, each associated with a different individual; location information for each region from a set of regions; and for each individual, a characteristic. The apparatus also includes a processor for generating, for each region, a first element matrix for a first element at a first element location in each sequence. The processor also generates a region matrix based on the first element matrix and estimates a first correlation between the region matrix and the characteristic. The processor also generates a second element matrix for a second element at a second element location in each sequence, and updates the region matrix based on the second element matrix to define an updated region matrix. The processor also estimates a second correlation between the updated region matrix and the characteristic, and classifies an individual based on the first correlation and/or the second correlation.

[Continued on next page]

SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

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

Declarations under Rule 4.17:

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

Published:

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

SYSTEMS, DEVICES AND METHODS FOR IMPROVED ANALYSIS AND STORAGE OF GENOTYPIC AND PHENOTYPIC DATA

CROSS-REFERENCE TO RELATED APPLICATIONS

[0100] This application claims priority to U.S. Provisional Application No. 62/344,131 filed June 1, 2016 and titled “SYSTEMS, DEVICES AND METHODS FOR CORRELATING GENOTYPIC AND PHENOTYPIC DATA”; and to U.S. Provisional Application No. 62/410,261 filed October 19, 2016 and titled “SYSTEMS, DEVICES AND METHODS FOR CORRELATING GENOTYPIC AND PHENOTYPIC DATA”, the entire disclosures of which are incorporated herein by reference in their entireties.

BACKGROUND

[0101] With advances in genomic sequencing, the storage and computing needs of genomics are increasingly prohibitive as ever more genomes are sequenced. However, storage is only part of the problem; the analytical data generated with genomic analysis is also uniquely large. For instance, analyzing a genome can require comparing millions of base pairs, and storing the results of the analysis at each step. When millions of such genomes are analyzed, there are million-squared analyses to store in some form. Conventional storage and computing approaches are ill-equipped to handle such extensive data sets. Effective means of genomic data analysis, such as correlation between genotypic and phenotypic data for an individual, can lead to direct changes in medical management and clinical care for individuals with specific genotypes

[0102] Accordingly, there is a need for systems and methods for efficient storage, manipulation and interpretation of genomic data.

SUMMARY

[0103] An apparatus includes a database storing a set of sequences, each sequence including a set of elements and associated with a different individual. The database also stores an indication of location information for each region from a set of regions associated with the set of sequences.

The database also stores, for each individual, an indication of a characteristic. The apparatus also includes a processor configured to, for each region, generate a first element matrix associated with a first element of that region at a first element location in each sequence. The processor is also configured to generate a region matrix based on the first element matrix and estimate a first correlation between the region matrix and the indication of the characteristic. The processor is also configured to generate a second element matrix associated with a second element of that region at a second element location in each sequence, and to update the region matrix based on the second element matrix to define an updated region matrix. The processor is also configured to estimate a second correlation between the updated region matrix and the indication of the characteristic, and to classify at least one individual to an individual type from a set of individual types based on at least one of the first correlation or the second correlation.

BRIEF DESCRIPTION OF THE DRAWINGS

[0104] FIG. 1 is a schematic diagram of an apparatus for analysis and storage of an individual's genotypic and phenotypic data, according to an embodiment.

[0105] FIG. 2 is a flowchart of a method for data analysis, according to an embodiment.

[0106] FIG. 3 illustrates an example mapping of a candidate seizure propensity region on chromosome 4, according to an embodiment.

[0107] FIG. 4 is a bar graph that shows both reported congenital heart disease and the size and relative locations of 4p deletions in 34 WHS patients.

[0108] FIG. 5 shows an exemplary balanced error rate method (BER) data plot of two chromosomal candidate regions corresponding to *MSX1* (Region 1) and *CC2D2A* (Region 2).

DETAILED DESCRIPTION

[0109] Aspects disclosed herein are beneficial for potential storage and computational inefficiencies associated with storing genotypic and phenotypic analysis information for genomes with millions of base pairs by discarding analytical information not be deemed significant. In some aspects, benefits of the approach disclosed herein are directed to efficiency of storage management, since when replicated across millions of elements/base pairs, significant savings in storage can be realized. Additionally, benefits of the approaches disclosed herein are directed to improved speed of genomic data analysis by reducing the amount of genomic data under consideration.

[0110] In some instances, genomic processing system/device is used to process genomic data, and particularly for correlating phenotypic and genotypic data. It is understood that the genomic processing system/device can perform some or all of the functionality disclosed herein, and can encompass some or all of the structural aspects (e.g., various devices, systems, subsystems, computing means, apparatus, sequencers, analyzers, etc.) disclosed herein. The components of the genomic processing system/device can interconnect in any suitable manner to achieve the functionality disclosed herein such as, for example, a wired or wireless network that connects the output of a sequencer to a computing apparatus. In some embodiments, the genomic processing system and/or at least one component thereof includes a processor (e.g., executing one or more modules) and a memory for performing the functionality disclosed herein.

[0111] In some instances, a method includes receiving a set of sequences, where each sequence from the set of sequences includes a set of elements. Each sequence from the set of sequences is associated with a different individual from a set of individuals. The method also includes receiving an indication of location information for each region from a set of regions associated with the set of sequences, and receiving, for each individual, an indication of a characteristic of that user. The method also includes, for each region from the set of regions, generating a first element matrix associated with a first element of that region at a first element location in each sequence. The method also includes generating a region matrix based on the first element matrix, and estimating a first correlation between the region matrix and the indication of the characteristic based on a first predetermined criterion. The method further includes generating a second element matrix associated with a second element of that region at a second element location in each sequence, and updating the region matrix based on the second element matrix to define an updated region matrix. The method further includes estimating a second correlation between the updated region matrix and the indication of the characteristic based on a second predetermined criterion. The method also includes classifying at least one individual to an individual type from a set of individual types based on at least one of the first correlation and the second correlation, and transmitting an indication of the individual type.

[0112] Embodiments disclosed herein are directed to a genomic processing system/device (and methods thereof) for classifying and/or correlating genomic information associated with a set of subjects to one or more phenotypes. FIG. 1 illustrates a compute device 100 configured for data analysis. The compute device 100 can be, for example, a server, a compute device, a data storage device, and/or the like. The compute device, or process associated with the compute device 100, can include, for example, computer software (stored in and/or executed at hardware) such as a web application, a database application, a cache server application, a queue server application, an application programming interface (API) application, an operating system, a file system, etc.; computer hardware such as a network appliance, a storage device (e.g., disk drive, memory module), a processing device (e.g., computer central processing unit (CPU)), computer graphic processing unit (GPU)), a networking device (e.g., network interface card), etc.; and/or combinations of computer software and hardware. In some instances, although not shown in FIG. 1, the compute device 100 can be operatively coupled to one or more other devices, such as a genomic sequencer.

[0113] As shown in FIG. 1, the compute device 100 includes a processor 110 and a memory 160. In some embodiments, and as illustrated in FIG. 1, the compute device 100 can also include a database 170. In some implementations (not shown), the database 170 can include multiple databases. In some implementations (not shown), part or the entirety of the database 170 can be external to the compute device 100. In some embodiments, and as illustrated in FIG. 1, the compute device can also include an I/O component 180 configured for interfacing with a user of the compute device 100, with another compute device, and/or the like.

[0114] The memory 180 can be, for example, a Random-Access Memory (RAM) (e.g., a dynamic RAM, a static RAM), a flash memory, a removable memory, and/or so forth. In some instances, instructions associated with performing the operations described herein (e.g., user classification) can be stored within the memory 160 and/or the database 170 and executed at the processor 110. The processor 110 includes a data analyzer 122, a classifier 128, a database manager 136, a communication manager 140, and/or other module(s)/component(s) (not shown in FIG. 1). In some instances, the communication manager 166 is configured to manage connectivity of the compute device 100 with other devices (not shown), with other networks (not shown), and/or the like. In some instances, the database manager 136 is configured to manage access, update, delete, and/or edit the contents of the database 170.

[0115] Each module/component in the processor 110 can be any combination of hardware-based module/component (e.g., a field-programmable gate array (FPGA), an application specific integrated circuit (ASIC), a digital signal processor (DSP)), software-based module (e.g., a module of computer code stored in the memory 160 and/or in the database 170, and/or executed at the processor 110), and/or a combination of hardware- and software-based modules. Each module/component in the processor 110 is capable of performing one or more specific functions/operations as described herein. In some implementations, the modules/components included and executed in the processor 160 can be, for example, a process, application, virtual machine, and/or some other hardware or software module/component. The processor 110 can be any suitable processor configured to run and/or execute those modules/components. For example, the processor 110 can include a general purpose processor, a field-programmable gate array (FPGA), an application specific integrated circuit (ASIC), a digital signal processor (DSP), and/or the like.

[0116] In other implementations, the processor 110 can include more or less modules/components than those shown in FIG. 1. For example, the processor 110 can include more than one user classifier to implement different classification approaches. In some embodiments, the compute device 100 can include more modules/components than those shown in FIG. 1.

[0117] As used herein, a module or component can be, for example, any assembly and/or set of operatively-coupled electrical components associated with performing a specific function, and can include, for example, a memory, a processor, electrical traces, optical connectors, hardware executing software and/or the like. As used herein, the singular forms "a," "an" and "the" include plural referents unless the context clearly dictates otherwise. Thus, for example, the term "a database" is intended to mean a single database or a combination of databases.

[0118] The operation of the various modules/components is explained herein with reference to a single compute device 100 with a single processor 110 for simplicity, though it is understood that unless explicitly stated otherwise, aspects of the modules/components described herein are extendible to multiple compute devices, multiple processors in an compute device, and/or multiple processors across multiple compute devices (e.g., connected by a network).

[0119] In some instances, the memory 160 and/or the database 170 is configured to store a set of sequences such as, for example, DNA or RNA sequences of a set of individuals/patients. In some instances, the set of sequences can be either DNA sequences or RNA sequences. In some instances, each sequence includes a set of elements and is associated with a different individual from a set of individuals. In instances where each sequence is a DNA sequence, the set of elements can include adenine, cytosine, guanine, and thymine. In instances where each sequence is an RNA sequence, the set of elements can include adenine, cytosine, guanine, and uracil. In some instances, the memory 160 and/or the database 170 can be configured to store individual/subject/patient information associated with a set of individuals such as, for example, both genomic/genotypic information associated with the set of individuals, and characteristic/phenotypic information associated with the set of individuals. In some instances, the genotypic information includes, for each individual, genetic structural information. In some instances, the structural information includes structural variation information. In some instances, the structural variation information includes information on a region of interest that includes one or more deletions and/or duplications, such as copy number variation (CNV) information. In some instances, the memory 160 and/or the database 170 is configured to store an indication of location information for each region from a set of regions (e.g., regions of interests) that are associated with the set of sequences. In some instances, the indication of location information for a region includes a startpoint and an endpoint for that region with respect to each sequence from the set of sequences. In some instances, the CNV information includes, for each CNV, a chromosomal identifier (e.g., a chromosome ID), a deletion startpoint for the CNV, and a deletion endpoint for the CNV. In some instances, the genotypic information includes, for each individual, single nucleotide variant SNV) information. In some instances, the indication of location information for an SNV includes the chromosome coordinate location of the SNV. In some instances, the memory of 160 and/or the database 170 is configured to store chromosome coordinate location for each SNV.

[0120] In some instances, the memory 160 and/or the database 170 is configured to store, for each patient/individual, an indication of a characteristic of that individual. In some instances, the characteristic includes a phenotype (e.g., disease or no disease) associated with that individual. For example, in some instances, the phenotypic information includes information on whether the particular individual manifests one or more observable characteristics. For example, the phenotypic information can include, among others, a specification of TRUE of FALSE for a characteristic, indicating whether the individual manifests that characteristic.

[0121] In some instances, the observable characteristic can include a disorder such as, for example, Wolf-Hirschhorn syndrome. In some instances, the observable characteristic can include a medical condition with multiple genetic causes and/or unknown causes, such as, for example, epilepsy. In some instances, the observable characteristic can include a response to a specific drug, such as a favorable or unfavorable response, a lack of response, a response classified as a side effect of the drug (e.g., an allergic reaction), and/or the like.

[0122] In some instances, the individual's information is received by the compute device 100, and stored to the memory 160 and/or in the database 170 (e.g., by the database manager 136). In some instances, the processor 110 can be configured to receive the individual's information, and to store the individual's information to the memory 160 and/or in the database 170 (directly or via the database manager 136). In some instances, the processor 110 (e.g., via the data analyzer 122) is configured to receive a specification, selection and/or a subset of the phenotypic information. Employing the same example disclosed above for simplicity, the data analyzer 122 can receive the specification of the TRUE/FALSE information for the set of individuals, indicating whether the individual manifests that characteristic/phenotype.

[0123] The processor 110 (e.g., via the data analyzer 122) can be further configured to compute, for each chromosome ID, the earliest deletion startpoint (also referred to as a MIN value) and the latest deletion endpoint (also referred to as a MAX value) of a deletion associated with the chromosome ID and associated with an entry. Similarly stated, for each chromosome ID and entry, the data analyzer 122 can identify and associate a MIN value, a MAX value, as well as the original TRUE/FALSE phenotypic information (collectively referred to as an "entry" hereon, and the entries for the set of individuals is referred to as a "set of entries").

[0124] In some instances, the processor 110 is configured to, for each region, generate a first element matrix associated with a first element/base pair of that region at a first element location in each sequence. For example, in some instances, the processor 110 (e.g., via the data analyzer 122) can be configured to generate combined genotypic/phenotypic information for the set of entries as described herein. For each base pair within a location under analysis (e.g., between the startpoint/MIN and endpoint/MAX value locations), a 2x2 matrix (i.e., an element matrix for the element at that location) of counts is generated, with one axis specifying whether the deletion exists or is absent at the specific element/base pair associated with that matrix, and the other axis specifying whether the phenotypic information for that individual is TRUE or FALSE. In other words, the 2x2 element matrix can include four different counts evaluated across the individuals for the set of entries: a) the deletion exists, the phenotype is TRUE; b) the deletion exists, the phenotype is FALSE; c) the deletion does not exist, the phenotype is TRUE; and d) the deletion does not exist, the phenotype is FALSE.

[0125] In some instances, the processor 110 (e.g., via the data analyzer 122) is configured to, generate a second element matrix associated with a second element/base pair of that region at a second element location in each sequence, the second element location being different than the first element location. In this manner, the processor 110 (e.g., via the data analyzer 122) can sequentially or simultaneously generate a set of element matrices (e.g., one for each base pair under analysis) based on the set of entries. Each matrix can identify the four different count values for its associated base pair for the set of entries.

[0126] In some instances, the processor 110 (e.g., via the data analyzer 122) can be configured to filter the element matrices using any suitable method. In some instances, the element matrices are sequentially generated, and the processor is configured to filter each element matrix as it is generated. In some instances, multiple element matrices are substantially simultaneously generated, and the processor is configured to filter one or more element matrix at a given time. In some embodiments, the data analyzer 122 filters the set of matrices to select matrices where the value changes. In some instances, the data analyzer 122 selects matrices where the value of one or more of the four counts of the matrix changes from one matrix to the other. In some instances, the data analyzer 122 selects matrices where the count associated with the highest value changes. For example, when a first matrix has the highest count for "the deletion exists, the phenotype is TRUE", and a subsequent second matrix has the highest count for both "the deletion exists, the phenotype is TRUE" and "the deletion does not exist, the phenotype is TRUE", then the first matrix and/or the second matrix can be selected by the data analyzer 122. In some instances, a sweep line technique is employed for selecting matrices to calculate and/or store.

[0127] In this manner, potential storage inefficiencies associated with storing matrix information for millions of base pairs can be avoided, since the information in the element matrix may not be deemed significant. For example, there may be no need to store element matrices for areas of a chromosome that do not have any deletions across the entries. Benefits of the approach disclosed herein are directed to efficiency of storage management, since when replicated across millions of elements/base pairs, significant savings in storage can be realized. Further, subsequent analysis of the element matrices as disclosed herein is sped up as well, due to the use of a reduced set of element matrices as input.

[0128] In some instances, the processor 110 (e.g., via the data analyzer 122) can be configured to populate the counts in an element matrix at specific end-point locations as follows: for the MIN value location, the count corresponding to "deletion does not exist, the phenotype is TRUE" is incremented; for the deletion startpoint for the CNV, the count corresponding to "deletion exists, the phenotype is TRUE" is incremented and the count corresponding to "deletion does not exist, the phenotype is TRUE" is decremented or not incremented; and for the location corresponding to one after the deletion endpoint for the CNV, the count corresponding to "deletion does not exist, the phenotype is TRUE" is incremented and the count corresponding to "deletion exists, the phenotype is TRUE" is decremented or not incremented.

[0129] In some instances, the processor 110 (e.g., via the classifier 128) is configured to generate a region matrix (also sometimes referred to as a "sum matrix") for a region based on the first element matrix and (as explained in detail herein) and other element matrices for that region), and estimate a first correlation between the region matrix and the indication of the characteristic based on a first predetermined criterion such as, for example, a minimum correlation threshold, a correlation range of values, and/or the like. The first correlation may be estimated in any suitable way, such as, for example, Pearson Product Moment Correlation, Spearman rank Order Correlation, Kendall rank order Correlation, Point-Biserial Correlation, and/or the like.

[0130] In some instances, the processor 110 (e.g., via the classifier 128) is further configured to update the region matrix based on the second element matrix to define an updated region matrix, and estimate a second correlation between the updated region matrix and the indication of the characteristic based on a second predetermined criterion. The second correlation may be estimated in any suitable way, such as, for example, Pearson Product Moment Correlation, Spearman rank Order Correlation, Kendall rank order Correlation, Point-Biserial Correlation, and/or the like. In some instances, the processor 110 (e.g., via the classifier 128) is further configured to estimate the second correlation by performing one or more statistical analyses on the updated region matrix. In some instances, for example, the statistical analyses includes a Fisher Exact Test, a test of the balanced error rate (BER), and/or the like.

[0131] As an example of region/sum matrix generation and analysis, after the entries have been processed, the processor 110 (e.g., via the classifier 128) is configured to generate a sum/region matrix, which maintains a running sum of the four counts: a) the deletion exists, the phenotype is TRUE; b) the deletion exists, the phenotype is FALSE; c) the deletion does not exist, the phenotype is TRUE; and d) the deletion does not exist, the phenotype is FALSE. The sum matrix can, for example, sum each of the entries in a number of matrices across a specified region to further analyze that region. The processor 110 (e.g., via the classifier 128) is further configured to update the counts of the sum matrix iteratively, based on each successive matrix of the set of matrices. In this manner, aspects of the approach laid out herein can be directed to treating the sum matrix as a contingency table, and performing statistical tests thereon. The processor 110 (e.g., via the classifier 128), at each iteration, can be configured to conduct one or more statistical tests on the sum matrix to determine the extent and/or degree to which the deletion correlates with, corresponds to, and/or is otherwise associated with the phenotypic information. The one or more statistical tests can include, but are not limited to, a Fisher Exact Test, a test of the balanced error rate (BER), and/or the like. In some instances, the processor 110 (e.g., via the classifier 128) is configured to identify all matrices that meet a prespecified criterion. For example, the processor 110 (e.g., via the classifier 128) can deem all matrices that meet a user-specified level of statistical significance to be of interest, e.g., measured by looking at contingency tables, by performing bootstrapping, and/or the like. In this manner, genetic intervals between the locations corresponding to the matrices of interest can be deletions that correlate with the phenotype.

[0132] In some instances, the processor 110 (e.g., via the classifier 128) is further configured to classify at least one individual to an individual type (e.g., exhibits significant deletions and has disease, exhibits a threshold level of deletions, has a predetermined likelihood of disease or greater, and/or the like) from a set of individual types based on at least one of the first correlation or the second correlation. In some instances, the processor 110 is further configured to transmit an indication of the individual type, such as to an interface of the compute device 100, to a device associated with a user, to another device connected to the compute device 100 via a wired and/or wireless network, and/or the like.

[0133] In some instances, the characteristic of each individual from the set of individuals includes a phenotype associated with that individual, and each sequence from the set of sequences is a biological sequence susceptible to having one or more deletions. Further, in some instances, the individual type is a first individual type and the second element matrix includes at least one count associated with a) a presence of a deletion at the second element, and b) the indication of the characteristic being TRUE for the set of individuals. In such instances, the processor can be further configured to, for each region, generate a third element matrix associated with a third element of that region at a third element location in each sequence. The third element matrix can include a count associated with a) a presence of a deletion at the third element, and b) the indication of the characteristic being TRUE for the set of individuals.

[0134] In other instances, the second element matrix includes at least one count associated with a) a presence of a deletion at the second element, and b) the indication of the characteristic being FALSE for the set of individuals. In such instances, the processor can be further configured to, for each region, generate a third element matrix associated with a third element of that region at a third element location in each sequence. The third element matrix can include a count associated with a) the presence of a deletion at the third element, and b) the indication of the characteristic being FALSE for the set of individuals.

[0135] In still other instances, the second element matrix includes at least one count associated with a) an absence of a deletion at the second element, and b) the indication of the characteristic being FALSE for the set of individuals. In such instances, the processor can be further configured to, for each region, generate a third element matrix associated with a third element of that region at a third element location in each sequence. The third element matrix can include a count associated with a) an absence of a deletion at the third element, and b) the indication of the characteristic being FALSE for the set of individuals.

[0136] In still other instances, the second element matrix includes at least one count associated with a) an absence of a deletion at the second element, and b) the indication of the characteristic being TRUE for the set of individuals. In such instances, the processor can be further configured to, for each region, generate a third element matrix associated with a third element of that region at a third element location in each sequence. The third element matrix can include a count associated with a) the absence of a deletion at the third element, and b) the indication of the characteristic being TRUE for the set of individuals.

[0137] In some instances, the processor 110 can be further configured to, when the count of the third element matrix is different from the count of the second element matrix, update an updated region matrix based on the third element matrix to define a second updated region matrix, and to estimate a third correlation between the second updated region matrix and the indication of the characteristic based on a third predetermined criterion. The processor 110 can be further configured to reclassify the at least one individual to a second individual type based on one or more of the first correlation, the second correlation, and the third correlation, and to transmit an indication of the second individual type. The processor 110 can be further configured to, when the count of the third element matrix is the same as the count of the second element matrix, discard the third matrix, and maintain the classification of the individual to the first individual type. In this manner, element matrices that do not result in changes of a specific count value can be discarded, thereby providing efficiency of storage and downstream computation of other element matrices.

[0138] In some instances, the processor 110 is configured to generate a third element matrix associated with a third element of that region at a third element location in each sequence. In some instances, the third element matrix includes one or more counts associated with an absence of a deletion at the third element location for the set of sequences. In such instances, when the one or more counts of the third element matrix have a value below a predetermined threshold (e.g., less than 5), the processor 110 is further configured to discard the third matrix, and maintain the classification of the individual to the first individual type. In this manner, when certain counts of element matrices are not sufficient to warrant further analysis, the element matrix can be discarded.

[0139] FIG. 2 illustrates a method 200 executable by a compute device, such as the compute device 100 or a structurally and/or functionally similar variant thereof. Explained with reference to the compute device 100, in some instances, the method 200 includes, at 210, receiving a set of sequences. Each sequence from the set of sequences can include a set of elements. Each sequence from the set of sequences can be associated with a different individual from a set of individuals. In some instances, each sequence is a biological sequence, and is either a DNA sequence or an RNA sequence. In some instances, each sequence is a DNA sequence and the group of elements include at least one of adenine, cytosine, guanine, or thymine. In some instances, each sequence is an RNA sequence and the group of elements includes at least one of adenine, cytosine, guanine, or uracil.

[0140] The method 200 further includes, at 212, receiving an indication of location information for each region from a set of regions associated with the set of sequences. In some instances, the indication of location information for each region includes a startpoint and an endpoint for that region with respect to each sequence.

[0141] The method 200 further includes, at 214, receiving, for each individual from the set of individuals, an indication of a characteristic of that user. In some instances, the characteristic of each individual from the set of individuals includes a phenotype associated with that individual.

[0142] The method 200 further includes, at 216, for each region from the set of regions, generating a first element matrix associated with a first element of that region at a first element location in each sequence of the set of sequences (substep 216a). The step 216 further includes generating a region matrix based on the first element matrix (substep 216b) and estimating a first correlation between the region matrix and the indication of the characteristic based on a first predetermined criterion (substep 216c). The step 216 further includes generating a second element matrix associated with a second element of that region at a second element location in each sequence from the set of sequences (substep 216d) and updating the region matrix based on the second element matrix to define an updated region matrix (substep 216e).

[0143] The step 216 further includes, estimating a second correlation between the updated region matrix and the indication of the characteristic based on a second predetermined criterion (substep 216f), and classifying at least one individual to an individual type from a set of individual types based on at least one of the first correlation and the second correlation (substep 216g). In some instances, the estimating at 216f further includes performing one or more statistical analyses on the updated region matrix. The step 216 further includes transmitting an indication of the first individual type (substep 216h).

[0144] In some instances, the characteristic of each individual from the set of individuals includes a phenotype associated with that individual, and each sequence from the set of sequences is a biological sequence susceptible to having one or more deletions. Further, in some instances, the individual type is a first individual type.

[0145] In some instances, the second element matrix includes a count associated with a) a presence of a deletion at the second element, and b) the indication of the characteristic being TRUE for the set of individuals, and the method 200 further includes generating a third element matrix associated with a third element of that region at a third element location in each sequence from the set of sequences. In some instances, the third element matrix includes a count associated with a) a presence of a deletion at the third element, and b) the indication of the characteristic being TRUE for the set of individuals.

[0146] In other instances, the second element matrix includes a count associated with a) a presence of a deletion at the second element, and b) the indication of the characteristic being FALSE for the set of individuals, and the method 200 further includes generating a third element matrix associated with a third element of that region at a third element location in each sequence from the set of sequences. In some instances, the third element matrix includes a count associated with a) a presence of a deletion at the third element, and b) the indication of the characteristic being FALSE for the set of individuals.

[0147] In still other instances, the second element matrix includes a count associated with a) an absence of a deletion at the second element, and b) the indication of the characteristic being FALSE for the set of individuals, and the method 200 further includes generating a third element matrix associated with a third element of that region at a third element location in each sequence from the set of sequences. In some instances, the third element matrix includes a count associated with a) an absence of a deletion at the third element, and b) the indication of the characteristic being FALSE for the set of individuals.

[0148] In still other instances, the second element matrix includes a count associated with a) an absence of a deletion at the second element, and b) the indication of the characteristic being TRUE for the set of individuals, and the method 200 further includes generating a third element matrix associated with a third element of that region at a third element location in each sequence from the set of sequences. In some instances, the third element matrix includes a count associated with a) an absence of a deletion at the third element, and b) the indication of the characteristic being TRUE for the set of individuals.

[0149] In such instances, when the count of the third element matrix is different from the count of the second element matrix, the method 200 can further include updating the updated region matrix based on the third element matrix to define a second updated region matrix, and estimating a third correlation between the second updated region matrix and the indication of the characteristic based on a third predetermined criterion. In some instances, the method 200 can further include, reclassifying the at least one individual to a second individual type from the set of individual types based on one or more of the first correlation, the second correlation, and the third correlation, and transmitting an indication of the second individual type. In some instances, when the count of the third element matrix is the same as the count of the second element matrix, the method 200 can further include discarding the third element matrix, and maintaining the classification of the at least one individual to the first individual type.

[0150] In some instances, the third element matrix includes one or more counts associated with an absence of a deletion at the third element location for the set of sequences, and when the one or more counts have a value below a predetermined threshold, the method 200 can further include discarding the third element matrix, and maintaining the classification of the at least one individual to the individual type.

[0151] FIG. 3 illustrates an example mapping of a candidate seizure propensity region on chromosome 4, according to example embodiments, as can be performed by the compute device 100. Bars show deletion sizes and locations of small 4p terminal or interstitial deletions in the 4p region that help define a 197 kbp seizure susceptibility region. The smallest region of overlap between three patients with seizures is shown by “CANDIDATE SEIZURE REGION”. This region is supported by patients (patient numbers labelled on Y-axis) as well as from the literature who have deletions excluding the seizure region and lack seizures (black solid line indicates no seizures) and patients who have deletions including the seizure region who have seizures (dotted line indicates a seizure phenotype). Patient data from the literature are indicated along the Y-axis by citation followed by the number of the patient as assigned in the citation in parentheses. Izumi 2010 is sourced from Izumi K, Okuno H, Maeyama K, Sato S, Yamamoto T, Torii C, Kosaki R, Takahashi T, Kosaki K, Am J Med Genet A 2010;152A:1028–32. Zollino 2014 (3 and 4) labels the size and location of the deletion shared by siblings, patients 3 and 4, in Zollino et al. See Zollino M, Orteschi D, Ruiter M, Pfundt R, Steindl K, Cafiero C, Ricciardi S, Contaldo I, Chieffo D, Ranalli D, Acquafondata C, Murdolo M, Marangi G, Asaro A, Battaglia D, Epilepsia 2014;55:849–57. Coordinates are given in base pairs (bps) along the X-axis. Ellipses (...) indicate that the deletion extends further than shown.

[0152] EXAMPLE 1 - Identification of muscle segment homeobox gene 1 (MSX1) as a candidate susceptibility gene for congenital heart disease in individuals with Wolf-Hirschhorn syndrome.

[0153] Wolf-Hirschhorn syndrome (WHS) is a well described contiguous gene deletion syndrome caused by varying sized deletions of the short arm of chromosome 4 with highly variable phenotypic features and disability. High-resolution genotype-phenotype correlation (e.g., using the compute device 100) was used to define genetic loci within the 4p region that are likely causative for individual features, and recently described a novel candidate gene associated with seizures in these individuals. See, for example, Ho KS, South ST, Lortz A, et al. J Med Genet 2016;53:256–263, which is incorporated herein by reference in its entirety.

[0154] Briefly, a custom, 2.8M-probe, chromosomal microarray platform was used to finely map CNVs (see, WO 2014/055915, which is incorporated herein by reference in its entirety). To score the phenotypic data, parent-reported answers from a questionnaire administered to families associated with the 4p-family support group were used. This questionnaire is designed to capture information on more than 20 different features. Correlations between genotypes and phenotypes were observed and candidate loci were identified using high-resolution genotype-phenotype correlation (e.g., using the compute device 100) as noted above in order to identify potentially pathogenic genes in identified regions.

[0155] Forty eight families completed the questionnaire and had high resolution chromosomal microarray analysis (CMA) to define deletion breakpoints. Seventy one percent of patients reported some form of congenital heart defect. Heart defects included right ventricular enlargement, PDA, ASD, VSD, bicuspid aortic valve, long QT, tricuspid atresia, hypoplastic right heart, pulmonary valve stenosis, murmurs, extremities turning blue, and thickening of pulmonary valve. The average age was 11.2 years with an age range of 0.9 years to 38 years. The female to male ratio was 28:20. Thirty four of the forty eight patients had pure 4p deletions (no other CNV was identified on CMA), and these individuals were analyzed.

[0156] In the same cohort of thirty-four individuals with WHS, with deletions of 4p ranging from 1.3-32.5 MBp in size and containing 28-200 genes, deletion breakpoints were further correlated on a custom, ultra-high resolution chromosomal microarray with over 20 other specific phenotypic features of WHS (FIG. 4). A statistical technique, as described above, was used to analyze this dataset for non-obvious correlations between specific clinical features and genomic regions to identify candidate genes of likely pathogenicity (FIG. 5).

[0157] Congenital heart defects are highly prevalent in the WHS cohort (71%) and the statistical approach suggested muscle segment homeobox 1 (*MSX1*) as a relevant candidate gene within a 3 Mbp region associated with congenital heart defects (p=0.0048 Fisher's exact test, two-tailed). Within this region, *MSX1* is the likely candidate gene supported by its role in cardiac development (Table 1). Some known studies have shown that *MSX1* acts along with Wnt and BMP signaling in early vertebrate cardiac development (Rao J, et al. *Cell Stem Cell.* 2016 Mar 3;18(3):341-353), and dual disruption of both *MSX1* and its paralog *MSX2* in mice results in cardiac outflow tract abnormalities (Chen YH, et al. *BMC Dev Biol.* 2008 Jul 30;8:75). In some known studies, *MSX1* appeared hyper-methylated in a human fetus with double outlet right ventricle, VSD, and hypoplasia of the ascending aorta (Serra-Juhe C, et al. *Epigenetics.* 2015;10(2):167-177). Aortic valve dysplasia was reported in a woman with a *de novo* duplication of a 3.8 Mbp containing the *MSX1* gene (Hitz MP, et al. *PLoS Genet.* 2012 Sep;8(9):e1002903). Some known independent genome-wide association studies and some known replication studies identified SNPs in *MSX1* associated with atrial and ventricular septal defects (Li FF, et al. PLoS One. 2015 Nov 10;10(11):e0142666; Cordell HJ, et al. *Nat Genet.* 2013 Jul;45(7):822-824).

Table. 1 Candidate Genes

Region 1	Region 2

MSX1 msh homeobox 1	*CC2D2A* coiled coil and C2 domain containing 2A
Literature supporting *MSX1* role in aortic development	Joubert syndrome PMID: 23692786
MSX1 deleted in 20 individuals with CHD and 4 without CHD *MSX1* not deleted in 3 individuals with CHD and 7 individuals without CHD	
P <= 0.0048	

[0158] *MSX1* had been proposed as a candidate gene involved in the oligodontia and cleft lip/palate associated with WHS. The analysis did not support *MSX1* as the best candidate gene association for either, rather finding FGF pathway members involved in both.

[0159] Although disclosed herein in an exemplary order, it is understood that these steps can be performed in any suitable order, in parallel, in series, and/or the like.

[0160] While various embodiments have been described herein, it should be understood that they have been presented by way of example, and not limitation. Where methods described above indicate certain events occurring in certain order, the ordering of certain events may be modified. Additionally, certain of the events may be performed concurrently in a parallel process when possible, as well as performed sequentially as described herein.

[0161] Some embodiments described herein relate to a computer storage product with a non-transitory computer-readable medium (also can be referred to as a non-transitory processor-readable medium) having instructions or computer code thereon for performing various computer-implemented operations. The computer-readable medium (or processor-readable medium) is non-transitory in the sense that it does not include transitory propagating signals per se (e.g., a propagating electromagnetic wave carrying information on a transmission medium such as space or a cable). The media and computer code (also can be referred to as code) may be those designed and constructed for the specific purpose or purposes. Examples of non-transitory computer-readable media include, but are not limited to: magnetic storage media such as hard disks, floppy disks, and magnetic tape; optical storage media such as Compact Disc/Digital Video Discs (CD/DVDs), Compact Disc-Read Only Memories (CD-ROMs), and holographic devices; magneto-optical storage media such as optical disks; carrier wave signal processing modules; and hardware devices that are specially configured to store and execute program code, such as Application-Specific Integrated Circuits (ASICs), Programmable Logic Devices (PLDs), Read-Only Memory (ROM) and Random-Access Memory (RAM) devices. Other embodiments described herein relate to a computer program product, which can include, for example, the instructions and/or computer code discussed herein.

[0162] Examples of computer code include, but are not limited to, micro-code or micro-instructions, machine instructions, such as produced by a compiler, code used to produce a web service, and files containing higher-level instructions that are executed by a computer using an interpreter. For example, embodiments may be implemented using imperative programming languages (e.g., C, Fortran, etc.), functional programming languages (Haskell, Erlang, etc.), logical programming languages (e.g., Prolog), object-oriented programming languages (e.g., Java, C++, etc.) or other suitable programming languages and/or development tools. Additional examples of computer code include, but are not limited to, control signals, encrypted code, and compressed code.

[0163] While various embodiments have been described above, it should be understood that they have been presented by way of example, and not limitation. Where methods described above indicate certain events occurring in certain order, the ordering of certain events can be modified. Additionally, certain of the events may be performed concurrently in a parallel process when possible, as well as performed sequentially as described above.

What is claimed is:

1. An apparatus, comprising:

a database configured to store:

a set of sequences, each sequence from the set of sequences including a set of elements, each sequence from the set of sequences associated with a different individual from a set of individuals;

an indication of location information for each region from a set of regions associated with the set of sequences; and

for each individual from the set of individuals, an indication of a characteristic of that individual; and

a processor operatively coupled to the database and configured to, for each region from the set of regions:

generate a first element matrix associated with a first element of that region at a first element location in each sequence from the set of sequences;

generate a region matrix based on the first element matrix;

estimate a first correlation between the region matrix and the indication of the characteristic based on a first predetermined criterion;

generate a second element matrix associated with a second element of that region at a second element location in each sequence from the set of sequences;

update the region matrix based on the second element matrix to define an updated region matrix;

estimate a second correlation between the updated region matrix and the indication of the characteristic based on a second predetermined criterion;

classify at least one individual to an individual type from a set of individual types based on at least one of the first correlation or the second correlation; and

transmit an indication of the individual type.

2. The apparatus of claim 1, wherein each sequence from the set of sequences is a biological sequence selected from a deoxyribonucleic acid (DNA) sequence and a ribonucleic acid (RNA) sequence.

3. The apparatus of claim 1, wherein each sequence from the set of sequences is a deoxyribonucleic acid (DNA) sequence and includes a group of elements including at least one of adenine, cytosine, guanine, or thymine.

4. The apparatus of claim 1, wherein each sequence from the set of sequences is a ribonucleic acid (RNA) sequence and includes a group of elements including at least one of adenine, cytosine, guanine, or uracil.

5. The apparatus of claim 1, wherein the characteristic of each individual from the set of individuals includes a phenotype associated with that user.

6. The apparatus of claim 5, wherein the phenotype is Wolf-Hirschhorn syndrome, epilepsy, congenital heart disease, or seizure.

7. The apparatus of claim 1, wherein the indication of location information for each region from the set of regions includes a startpoint and an endpoint for that region with respect to each sequence of the set of sequences.

8. The apparatus of claim 1, wherein the characteristic of each individual from the set of individuals includes a phenotype associated with that individual and each sequence from the set of sequences is a biological sequence susceptible to having one or more deletions, wherein the individual type is a first individual type of the set of individual types, the second element matrix including a count associated with a) a presence of a deletion at the second element, and b) the indication of the characteristic being TRUE for the set of individuals,

the processor further configured to, for each region from the set of regions:

generate a third element matrix associated with a third element of that region at a third element location in each sequence of the set of sequences, the third element matrix including a count associated with a) a presence of a deletion at the third element, and b) the indication of the characteristic being TRUE for the set of individuals;

when the count of the third element matrix is different from the count of the second element matrix:

update the updated region matrix based on the third element matrix to define a second updated region matrix;

estimate a third correlation between the second updated region matrix and the indication of the characteristic based on a third predetermined criterion;

reclassify the at least one individual to a second individual type from the set of individual types based on one or more of the first correlation, the second correlation, and the third correlation; and

transmit an indication of the second individual type; and

when the count of the third element matrix is the same as the count of the second element matrix:

discard the third element matrix; and

maintain the classification of the at least one individual to the first individual type from the set of individual types based on at least one of the first correlation or the second correlation.

9. The apparatus of claim 1, wherein the characteristic includes a phenotype associated with the user, wherein each sequence from the set of sequences is a biological sequence capable of having one or more deletions, wherein the individual type is a first individual type from the set of individual types, the second element matrix including a count associated with a) a presence of a deletion at the second element, and b) the indication of the characteristic being FALSE for the set of individuals,

the processor further configured to, for each region from the set of regions:

generate a third element matrix associated with a third element of that region at a third element location in each sequence of the set of sequences, the third element matrix including a count associated with a) the presence of a deletion at the third element, and b) the indication of the characteristic being FALSE for the set of individuals;

when the count of the third element matrix is different from the count of the second element matrix:

update the updated region matrix based on the third element matrix to define a second updated region matrix;

estimate a third correlation between the second updated region matrix and the indication of the characteristic based on a third predetermined criterion;

reclassify the at least one individual to a second individual type of a set of individual types based on one or more of the first correlation, the second correlation, and the third correlation; and

transmit an indication of the second individual type; and

when the count of the third element matrix is the same as the count of the second element matrix:

discard the third element matrix; and

maintain the classification of the at least one individual to the first individual type from the set of individual types based on at least one of the first correlation and the second correlation.

10. The apparatus of claim 1, wherein the characteristic includes a phenotype associated with the user, wherein each sequence from the set of sequences is a biological sequence capable of having one or more deletions, wherein the individual type is a first individual type of the set of individual types, the second element matrix including a count associated with a) an absence of a deletion at the second element, and b) the indication of the characteristic being FALSE for the set of individuals,

the processor further configured to, for each region from the set of regions:

generate a third element matrix associated with a third element of that region at a third element location in each sequence of the set of sequences, the third element matrix including a count associated with a) an absence of a deletion at the third element, and b) the indication of the characteristic being FALSE for the set of individuals;

when the count of the third element matrix is different from the count of the second element matrix:

update the updated region matrix based on the third element matrix to define a second updated region matrix;

estimate a third correlation between the second updated region matrix and the indication of the characteristic based on a third predetermined criterion;

reclassify the at least one individual to a second individual type of a set of individual types based on one or more of the first correlation, the second correlation, and the third correlation; and

transmit an indication of the second individual type; and

when the count of the third element matrix is the same as the count of the second element matrix:

discard the third element matrix; and

maintain the classification of the at least one individual to the first individual type from the set of individual types based on at least one of the first correlation and the second correlation.

11. The apparatus of claim 1, wherein the characteristic includes a phenotype associated with the user, wherein each sequence from the set of sequences is a biological sequence capable of having one or more deletions, wherein the individual type is a first individual type of the set of individual types, the second element matrix including a count associated with a) an absence of a deletion at the second element, and b) the indication of the characteristic being TRUE for the set of individuals,

the processor further configured to, for each region from the set of regions:

generate a third element matrix associated with a third element of that region at a third element location in each sequence of the set of sequences, the third element matrix including a count associated with a) the absence of a deletion at the third element, and b) the indication of the characteristic being TRUE for the set of individuals;

when the count of the third element matrix is different from the count of the second element matrix:

update the updated region matrix based on the third element matrix to define a second updated region matrix;

estimate a third correlation between the second updated region matrix and the indication of the characteristic based on a third predetermined criterion;

reclassify the at least one individual to a second individual type of a set of individual types based on one or more of the first correlation, the second correlation, and the third correlation; and

transmit an indication of the second individual type; and when the count of the third element matrix is the same as the count of the second element matrix:

discard the third element matrix; and

maintain the classification of the at least one individual to the first individual type from the set of individual types based on at least one of the first correlation and the second correlation.

12. The apparatus of claim 1, wherein the characteristic includes a phenotype associated with the individual and each sequence from the set of sequences is a biological sequence capable of having one or more deletions, the processor further configured to, for each region from the set of regions:

generate a third element matrix associated with a third element of that region at a third element location in each sequence of the set of sequences, the third element matrix including one or more counts associated with an absence of a deletion at the third element location for the set of sequences;

when the one or more counts have a value below a predetermined threshold:

discard the third element matrix; and

maintain the classification of the at least one individual as the individual type from the set of individual types based on at least one of the first correlation or the second correlation.

13. The apparatus of claim 1, wherein the processor is further configured to estimate the second correlation by performing one or more statistical analyses on the updated region matrix.

14. A method, comprising:

receiving a set of sequences, each sequence from the set of sequences including a set of elements, each sequence from the set of sequences associated with a different individual from a set of individuals;

receiving indication of location information for each region from a set of regions associated with the set of sequences;

receiving, for each individual of the set of individuals, an indication of a characteristic of that user;

for each region of the set of regions:

generating a first element matrix associated with a first element of that region at a first element location in each sequence of the set of sequences;

generating a region matrix based on the first element matrix;

estimating a first correlation between the region matrix and the indication of the characteristic based on a first predetermined criterion;

generating a second element matrix associated with a second element of that region at a second element location in each sequence from the set of sequences;

updating the region matrix based on the second element matrix to define an updated region matrix;

estimating a second correlation between the updated region matrix and the indication of the characteristic based on a second predetermined criterion;

classifying at least one individual to an individual type from a set of individual types based on at least one of the first correlation and the second correlation; and

transmitting an indication of the first individual type.

15. The method of claim 14, wherein each sequence from the set of sequences is a biological sequence selected from a deoxyribonucleic acid (DNA) sequence and a ribonucleic acid (RNA) sequence.

16. The method of claim 14, wherein each sequence from the set of sequences is a deoxyribonucleic acid (DNA) sequence and includes a group of elements including at least one of adenine, cytosine, guanine, or thymine.

17. The method of claim 14, wherein each sequence from the set of sequences is a ribonucleic acid (RNA) sequence and includes a group of elements including at least one of adenine, cytosine, guanine, or uracil.

18. The method of claim 14, wherein the characteristic of each individual from the set of individuals includes a phenotype associated with that individual.

19. The method of claim 17, wherein the phenotype is Wolf-Hirschhorn syndrome, epilepsy, congenital heart disease, or seizure.

20. The method of claim 14, wherein the indication of location information for each region from the set of regions includes a startpoint and an endpoint for that region with respect to each sequence of the set of sequences.

21. The method of claim 14, wherein the characteristic of each individual from the set of individuals includes a phenotype associated with that individual and each sequence from the set of sequences is a biological sequence susceptible to having one or more deletions, wherein the individual type is a first individual type of the set of individual types, the second element matrix including a count associated with a) a presence of a deletion at the second element, and b) the indication of the characteristic being TRUE for the set of individuals,

the method further comprising, for each region from the set of regions:

generating a third element matrix associated with a third element of that region at a third element location in each sequence from the set of sequences, the third element matrix including a count associated with a) a presence of a deletion at the third element, and b) the indication of the characteristic being TRUE for the set of individuals;

when the count of the third element matrix is different from the count of the second element matrix:

updating the updated region matrix based on the third element matrix to define a second updated region matrix;

estimating a third correlation between the second updated region matrix and the indication of the characteristic based on a third predetermined criterion;

reclassifying the at least one individual to a second individual type from the set of individual types based on one or more of the first correlation, the second correlation, and the third correlation; and

transmitting an indication of the second individual type; and

when the count of the third element matrix is the same as the count of the second element matrix:

discarding the third element matrix; and maintaining the classification of the at least one individual to the first individual type of a set of individual types based on at least one of the first correlation and the second correlation.

22. The method of claim 14, wherein the characteristic includes a phenotype associated with the user, wherein each sequence from the set of sequences is a biological sequence capable of having one or more deletions, wherein the individual type is a first individual type from the set of individual types, the second element matrix including a count associated with a) a presence of a deletion at the second element, and b) the indication of the characteristic being FALSE for the set of individuals,

the method further comprising, for each region from the set of regions:

generate a third element matrix associated with a third element of that region at a third element location in each sequence from the set of sequences, the third element matrix including a count associated with a) a presence of a deletion at the third element, and b) the indication of the characteristic being FALSE for the set of individuals;

when the count of the third element matrix is different from the count of the second element matrix:

updating the updated region matrix based on the third element matrix to define a second updated region matrix;

estimating a third correlation between the second updated region matrix and the indication of the characteristic based on a third predetermined criterion;

reclassifying the at least one individual to a second individual type from the set of individual types based on at least one of the first correlation, the second correlation, or the third correlation; and

transmitting an indication of the second individual type; and

when the count of the third element matrix is the same as the count of the second element matrix:

discarding the third element matrix; and

maintaining the classification of the at least one individual to the first individual type from the set of individual types based on one or more of the first correlation and the second correlation.

23. The method of claim 14, wherein the characteristic includes a phenotype associated with the user, wherein each sequence from the set of sequences is a biological sequence capable of having one or more deletions, wherein the individual type is a first individual type from the set of individual types, the second element matrix including a count associated with a) an absence of a deletion at the second element, and b) the indication of the characteristic being FALSE for the set of individuals,

the method further comprising, for each region from the set of regions:

generating a third element matrix associated with a third element of that region at a third element location in each sequence of the set of sequences, the third element matrix including a count associated with a) an absence of a deletion at the third element, and b) the indication of the characteristic being FALSE for the set of individuals;

when the count of the third element matrix is different from the count of the second element matrix:

updating the updated region matrix based on the third element matrix to define a second updated region matrix;

estimating a third correlation between the second updated region matrix and the indication of the characteristic based on a third predetermined criterion;

reclassifying the at least one individual to a second individual type from the set of individual types based on one or more of the first correlation, the second correlation, and the third correlation; and

transmitting an indication of the second individual type; and

when the count of the third element matrix is the same as the count of the second element matrix:

discarding the third element matrix; and

maintaining the classification of the at least one individual to the first individual type from the set of individual types based on at least one of the first correlation and the second correlation.

24. The method of claim 14, wherein the characteristic includes a phenotype associated with the user, wherein each sequence of the set of sequences is a biological sequence capable of

having one or more deletions, wherein the individual type is a first individual type from the set of individual types, the second element matrix including a count associated with a) an absence of a deletion at the second element, and b) the indication of the characteristic being TRUE for the set of individuals,

the method further comprising, for each region of the set of regions:

generating a third element matrix associated with a third element of that region at a third element location in each sequence of the set of sequences, the third element matrix including a count associated with a) an absence of a deletion at the third element, and b) the indication of the characteristic being TRUE for the set of individuals;

when the count of the third element matrix is different from the count of the second element matrix:

updating the updated region matrix based on the third element matrix to define a second updated region matrix;

estimating a third correlation between the second updated region matrix and the indication of the characteristic based on a third predetermined criterion;

reclassifying the at least one individual to a second individual type from the set of individual types based on one or more of the first correlation, the second correlation, and the third correlation; and

transmitting an indication of the second individual type; and

when the count of the third element matrix is the same as the count of the second element matrix:

discarding the third element matrix; and

maintaining the classification of the at least one individual to the first individual type from the set of individual types based on at least one of the first correlation and the second correlation.

25. The method of claim 14, wherein the characteristic includes a phenotype associated with the individual and each sequence from the set of sequences is a biological sequence capable of having one or more deletions, the processor further configured to, for each region from the set of regions:

generate a third element matrix associated with a third element of that region at a third element location in each sequence from the set of sequences, the third element matrix including one or more counts associated with an absence of a deletion at the third element location for the set of sequences;

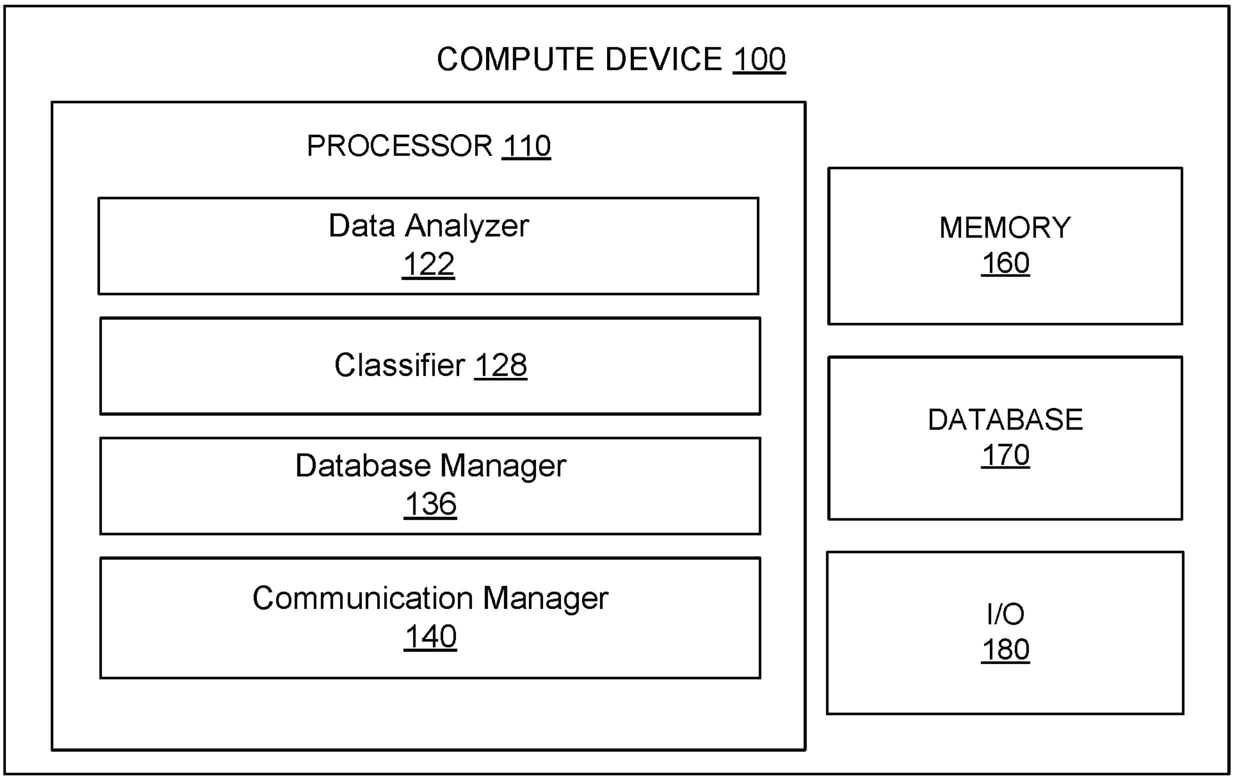
when the one or more counts have a value below a predetermined threshold:

discarding the third element matrix; and

maintaining the classification of the at least one individual to the individual type from the set of individual types based on at least one of the first correlation and the second correlation.

26. The method of claim 14, the estimating the second correlation including performing one or more statistical analyses on the updated region matrix.

FIG. 1


COMPUTE DEVICE 100
PROCESSOR 110
Data Analyzer 122
Classifier 128
Database Manager 136
Communication Manager 140
MEMORY 160
DATABASE 170
I/O 180

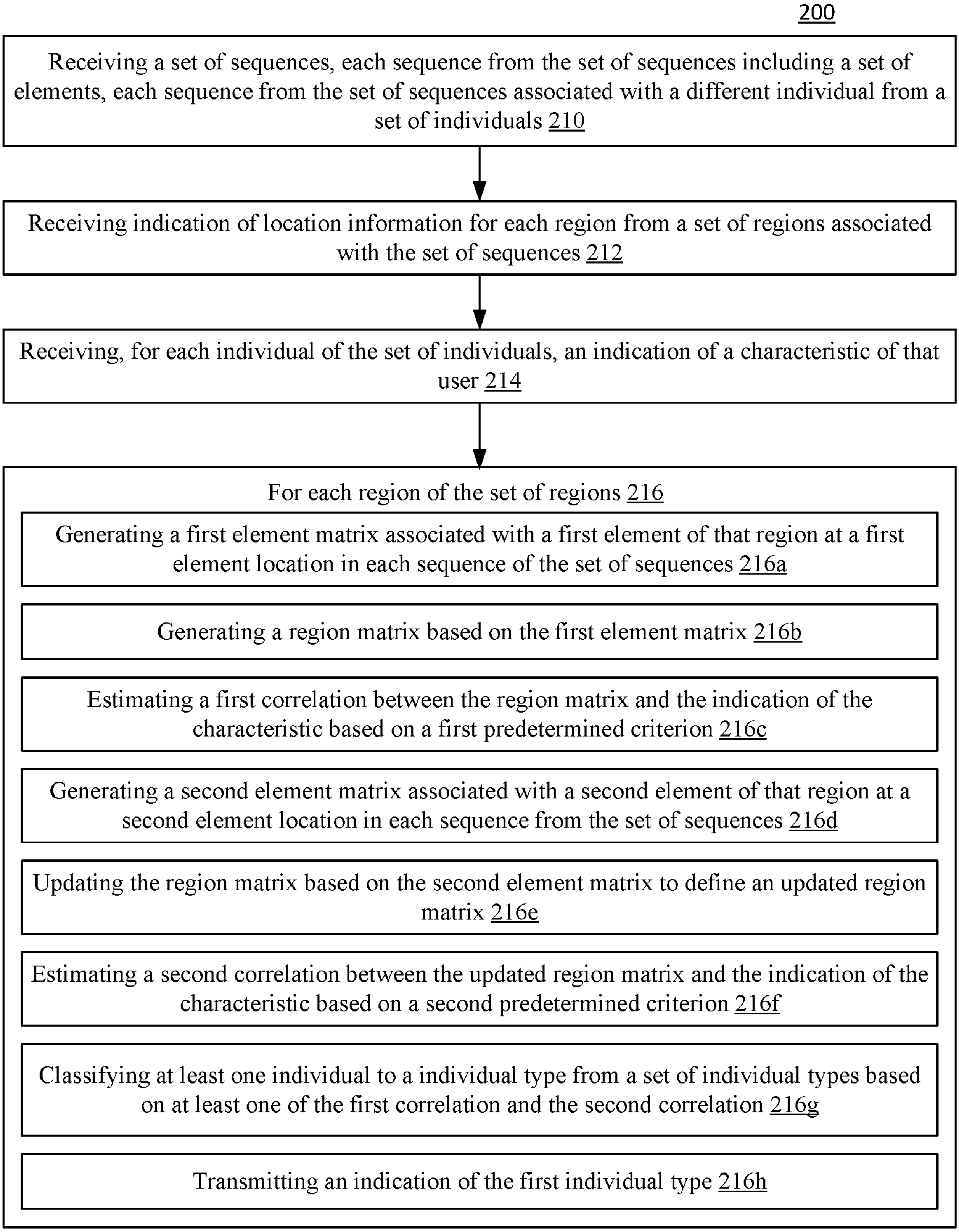

200
Receiving a set of sequences, each sequence from the set of sequences including a set of elements, each sequence from the set of sequences associated with a different individual from a set of individuals 210
Receiving indication of location information for each region from a set of regions associated with the set of sequences 212
Receiving, for each individual of the set of individuals, an indication of a characteristic of that user 214
For each region of the set of regions 216
Generating a first element matrix associated with a first element of that region at a first element location in each sequence of the set of sequences 216a
Generating a region matrix based on the first element matrix 216b
Estimating a first correlation between the region matrix and the indication of the characteristic based on a first predetermined criterion 216c
Generating a second element matrix associated with a second element of that region at a second element location in each sequence from the set of sequences 216d
Updating the region matrix based on the second element matrix to define an updated region matrix 216e
Estimating a second correlation between the updated region matrix and the indication of the characteristic based on a second predetermined criterion 216f
Classifying at least one individual to a individual type from a set of individual types based on at least one of the first correlation and the second correlation 216g
Transmitting an indication of the first individual type 216h

FIG. 2

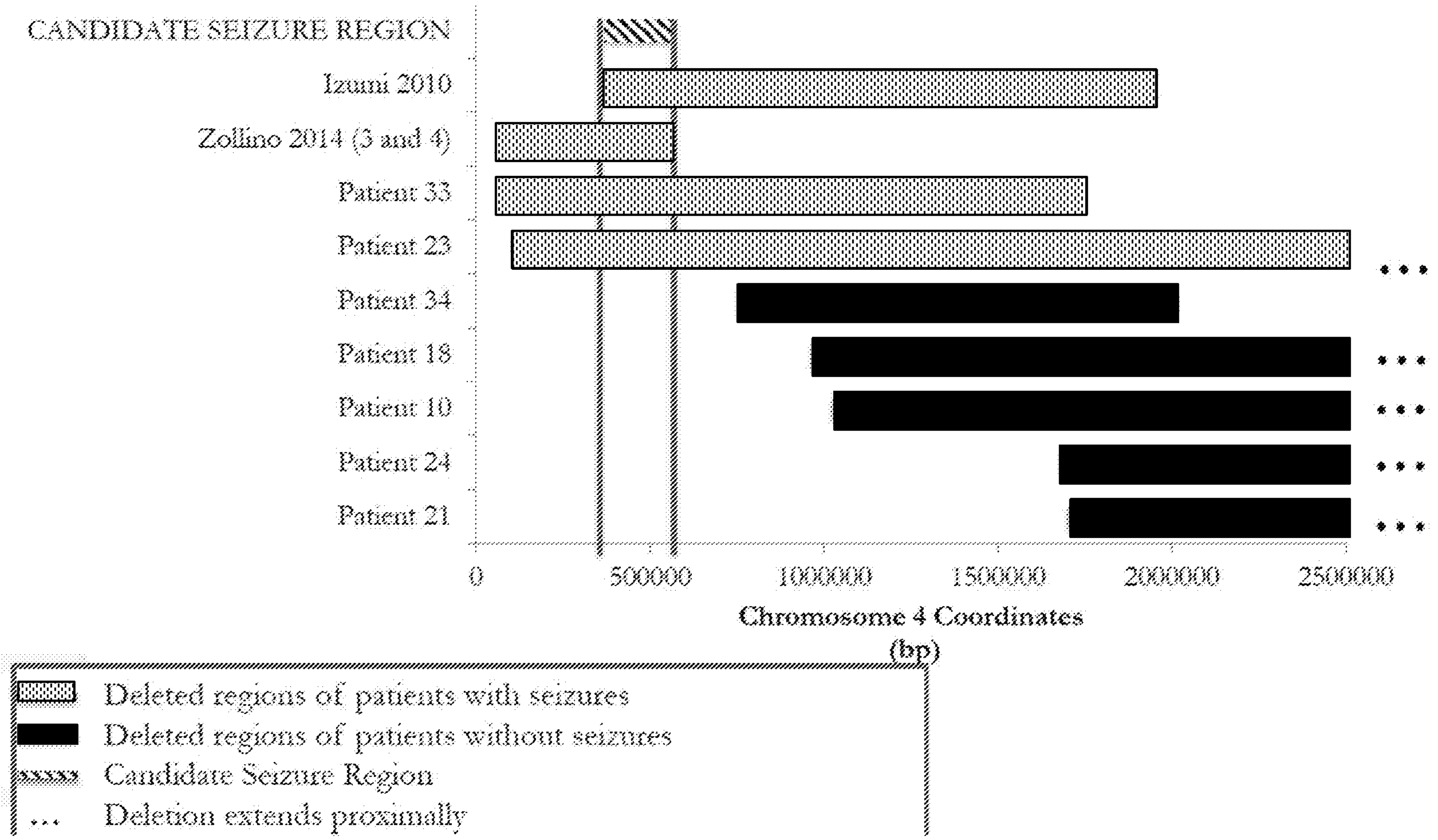

CANDIDATE SEIZURE REGION
Izumi 2010
Zollino 2014 (3 and 4)
Patient 33
Patient 23
Patient 34
Patient 18
Patient 10
Patient 24
Patient 21
0
500000
1000000
1500000
2000000
2500000
Chromosome 4 Coordinates (bp)
▨ Deleted regions of patients with seizures
■ Deleted regions of patients without seizures
▧ Candidate Seizure Region
... Deletion extends proximally

FIG. 3

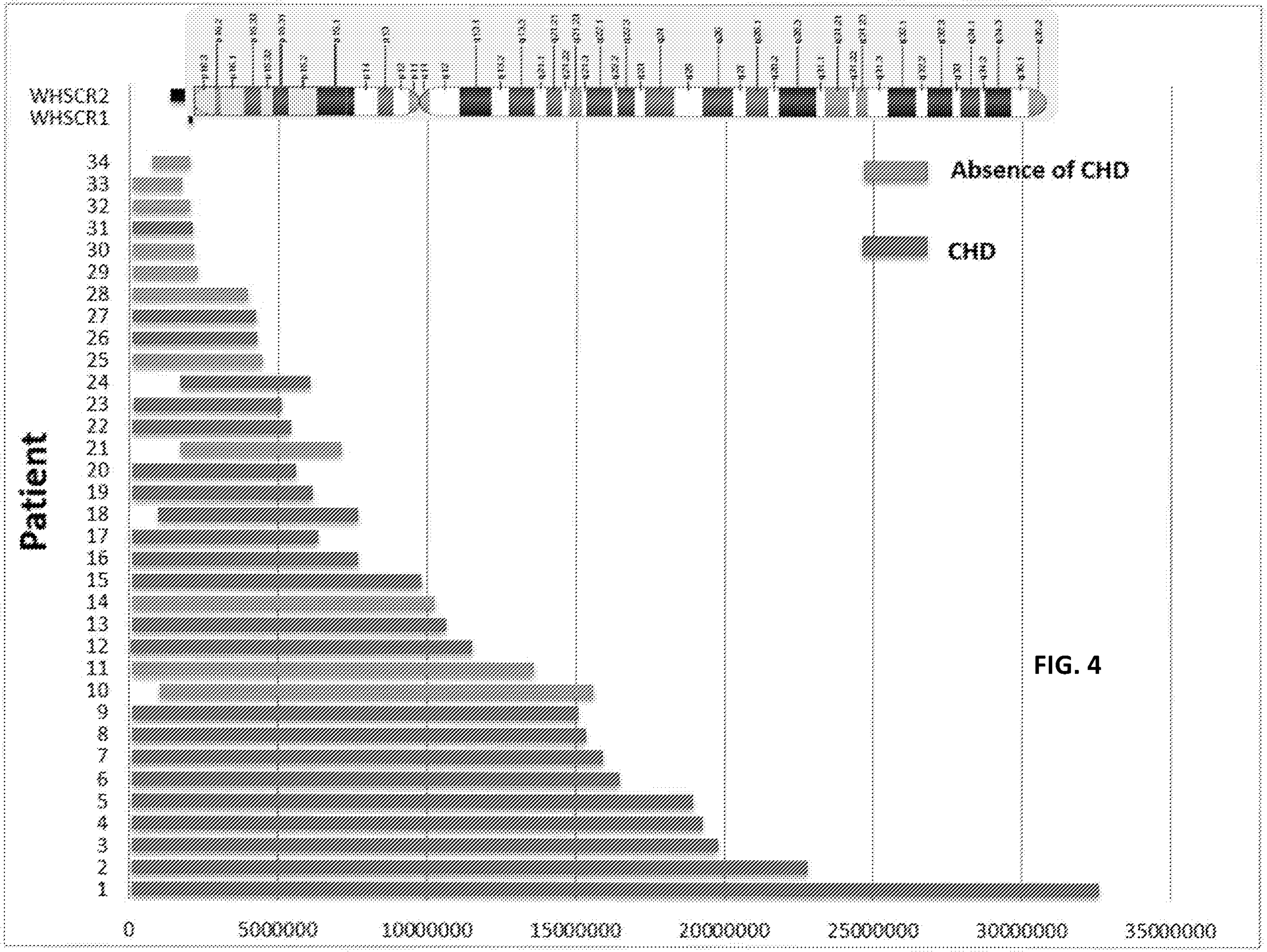

WHSCR2
WHSCR1
Absence of CHD
CHD
Patient
34
33
32
31
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
5000000
10000000
15000000
20000000
25000000
30000000
35000000

FIG. 4

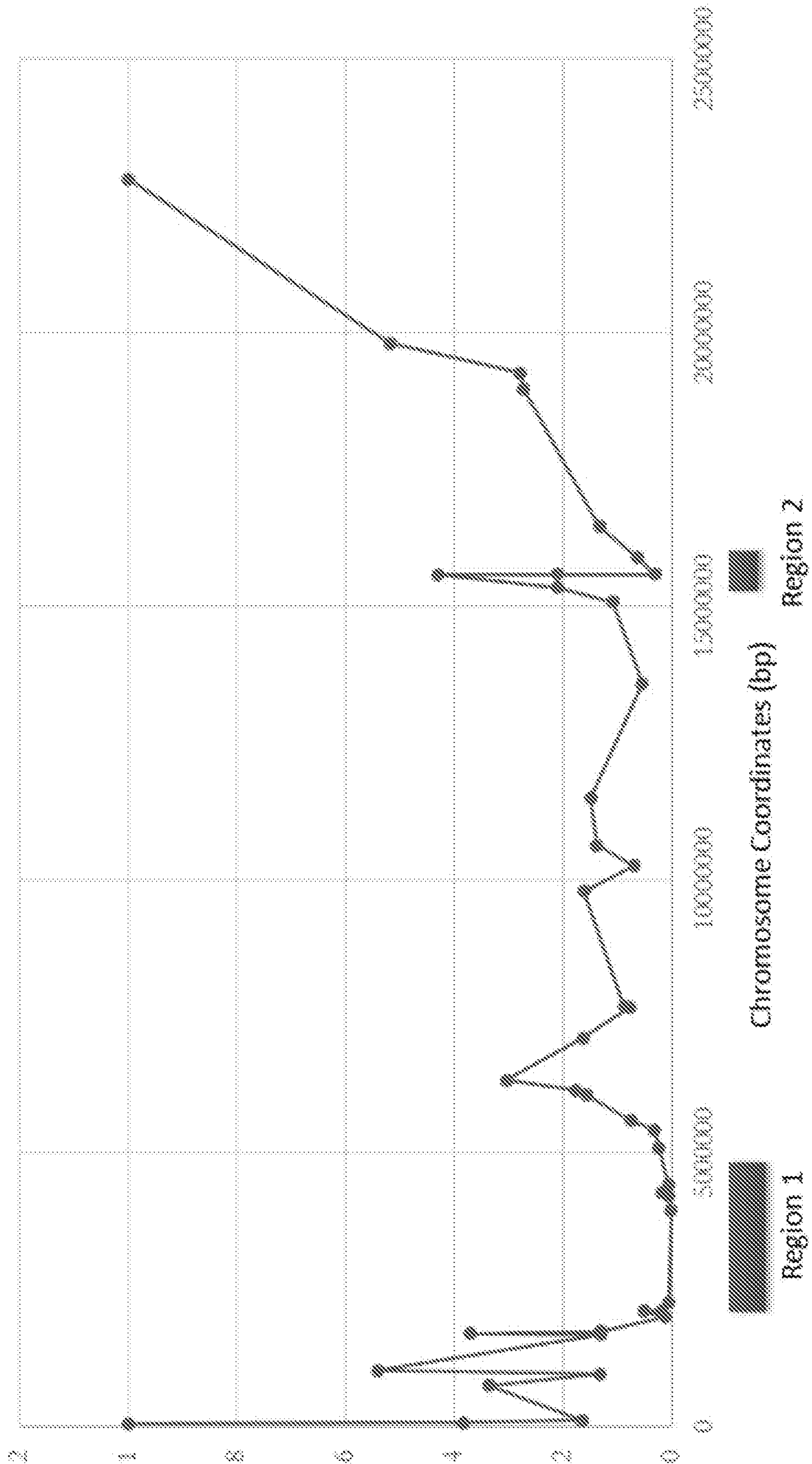

1.2
1
0.8
0.6
0.4
0.2
0
0
5000000
10000000
15000000
20000000
25000000
Chromosome Coordinates (bp)
Region 1
Region 2

FIG. 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 17/35496

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 19/18, G06F 19/24 (2017.01)
CPC - G06F19/20, C12Q2600/158, C12Q1/6883, G06F19/24, G06F19/12,

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)
See Search History Document

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2007/0166707 A1 (Schadt et al.), 19 July 2007 (19.07.2007), entire document, especially Abstract; Para [0053], [0067], [0175], [0188], [0326]-[0327], [0743], [0985]	1-26
Y	US 2016/0076046 A1 (CERES, INC.), 17 March 2016 (17.03.2016), entire document, especially Abstract; para [0423], [0569]-[0570]	1-26
A	US 2002/0137080 A1 (Usuka et al.), 26 September 2002 (26.09.2002), entire document	1-26
A	US 2005/0086035 A1 (Peccoud et al.), 21 April 2005 (21.04.2005), entire document	1-26
A	US 2013/0040826 A1 (Braun, III et al.), 14 February 2013 (14.02.2013), entire document	1-26
A	US 2010/0070186 A1 (Soper), 18 March 2010 (18.03.2010), entire document	1-26

□ 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 04 August 2017	Date of mailing of the international search report 29 AUG 2017
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774