



(51) International Patent Classification:

G16B 20/00 (2019.01)

(21) International Application Number:

PCT/US2018/066538

(22) International Filing Date:

19 December 2018 (19.12.2018)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/608,137 20 December 2017 (20.12.2017) US
62/624,195 31 January 2018 (31.01.2018) US

(71) Applicant: CHILDREN'S HOSPITAL MEDICAL CENTER [US/US]; 3333 Burnet Avenue, MI 7032, Cincinnati, OH 45229 (US).

(72) Inventors: ZHANG, Kejian; c/o Children's Hospital Medical Center, 3333 Burnet Avenue, MI 7032, Cincinnati, OH 45229 (US). VALENCIA, C., Alexander; c/o Children's Hospital Medical Center, 3333 Burnet Avenue, MI 7032, Cincinnati, OH 45229 (US). HUSAMI, Ammar; c/o Children's Hospital Medical Center, 3333 Burnet Avenue,

MI 7032, Cincinnati, OH 45229 (US). MATHUR, Abhinav; c/o Children's Hospital Medical Center, 3333 Burnet Avenue, MI 7032, Cincinnati, OH 45229 (US).

(74) Agent: LIBERTO, Muriel et al.; Mintz Levin Cohn Ferris Glovsky And Popeo, P.C., One Financial Center, Boston, MA 02111 (US).

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

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK,

(54) Title: CLINICAL DECISION SUPPORT USING WHOLE EXOME ANALYSIS

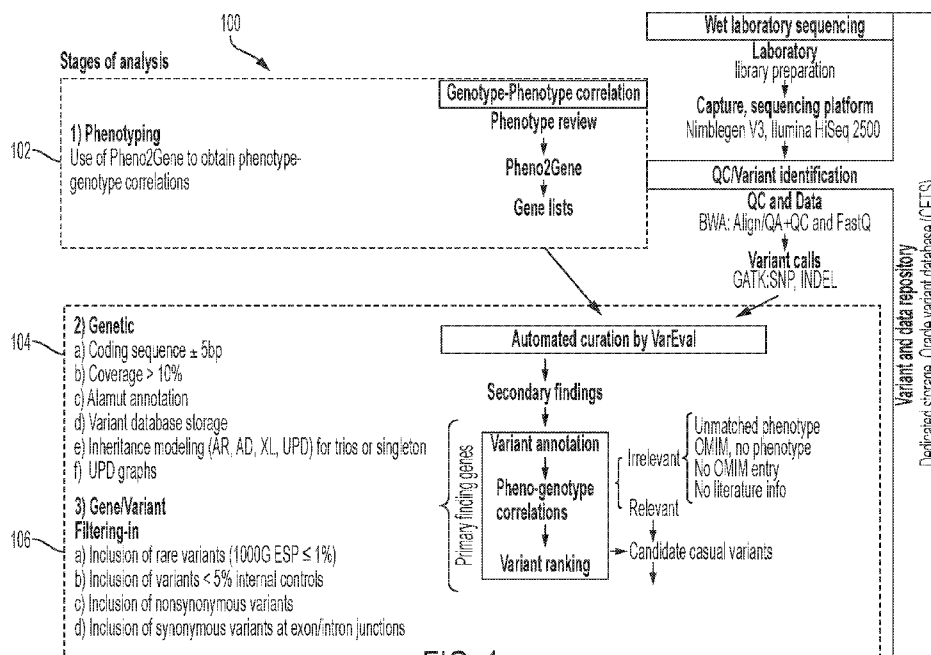


FIG. 1

(57) Abstract: The disclosure provides methods, including computer implemented methods, for clinical decision support comprising whole exome sequencing and the analysis of the resulting genetic variants of a human subject in order to identify the most likely causative variants underlying a patient's clinical phenotype.

EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV,
MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,
TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW,
KM, ML, MR, NE, SN, TD, TG).

Published:

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

CLINICAL DECISION SUPPORT USING WHOLE EXOME ANALYSIS

TECHNICAL FIELD

[0001] This disclosure relates generally to clinical decision support using whole exome sequencing to identify the most likely causative genetic variants to explain a patient's clinical phenotype, thereby improving clinical diagnosis of difficult to diagnose genetic diseases and disorders.

BACKGROUND

[0002] Recent advances in exome sequencing technologies have allowed discovery of genes that cause various disorders in a more comprehensive way by scanning the protein coding exome sequence space. See, for example, Choi M. *et al. Proc Natl Acad Sci USA*. 14 2009;106:19096–101. Whole exome sequencing (“WES”) technology has been applied to the task of identifying candidate genetic variants that are causative of Mendelian disorders including, for example, Miller syndrome, Fowler syndrome, and Perrault Syndrome. However, identifying the disease causing variant for a specific disorder from among the thousands of variants typically identified by WES is a difficult, laborious and time consuming task. See, for example, Majewski J. *et al., J Med Genet*. 2011;48:580–9.

[0003] Variant filtering strategies based on statistical genetics, predicted degree of deleteriousness and comprehensive annotation have been employed to narrow the list of candidate variants. For example, statistical genetics methods which prioritize genomic regions based on identity-by-descent polymorphisms and/or genetic linkage co-segregation include BEAGLE, GERMLINE, PLINK IBD, and MERLIN. See, for example, Browning BL and Browning SR. *Am J Hum Genet*. 2011;88:173–82; Rödelberger C. *et al. Bioinforma Oxf Engl*. 2011 27:829–36; and Abecasis GR *et al. Nat Genet*. 2002 30:97–101.

[0004] Alternative methods focus on deleterious predictions of a non-synonymous single nucleotide variant (“SNP”) in a protein-coding gene by using computational algorithms based on amino acid physicochemical properties, protein structure and cross-species conservation, namely, scale-invariant feature transform (SIFT), polymorphism phenotyping v2 (PolyPhen-2), likelihood ratio test (LRT), Gratham scale MutationTaster, and PhyloP. See, for example, Chun S, Fay JC. *Genome Res.* 2009 19:1553–61; Schwarz JM *et al. Nat Methods.* 2010 7:575–76; and Liu X, Jian X, Boerwinkle E. *Hum Mutat.* 2011 32:894–9.

[0005] The third type of analysis approach, utilized by SeattleSeq, annotate variation (ANNOVAR) and Alamut, comprehensively annotates variants using information from bioinformatics resources which are then used to bioinformatics resources which are then used to prioritize variants. See, for example, ANNOVAR reported by Wang K, Li M, Hakonarson H. *Nucleic Acids Res.* 2010 38:e164. Typically, functional information is found scattered across various tools and resources and may include inconsistent functional site predictions, making it challenging to get a list candidate genes and their respective variants for follow-up experimental validation. Furthermore, other important resources have yet to be incorporated into existing methods such as biological pathways and peer-reviewed literature.

[0006] Currently, clinical exome analysis groups have reported various filtering strategies. To identify the causative gene variant(s), the filtering algorithms for exome analysis pipelines are based on population and molecular genetic principles, namely, minor allele frequency (MAF) from public databases, examination of coding regions ± 2 base pair (“bp”), alterations present in the Online Mendelian Inheritance in Man (OMIM) and/or Human Gene Mutation Database (HGMD), inheritance modeling or co-segregation, *in silico* predictions, and phenotypic overlap among the proband and reported patients. See, for

example, Farwell KD *et al. Genet Med Off J Am Coll Med Genet.* 2015 17:578–86; and Yang Y, Muzny DM *et al. N Engl J Med.* 2013 369:1502–7 11.

[0007] After the application of these filters, there may be in the range of 300-700 variants that need to be examined, which is typically too many for efficient and cost-effective analysis. Accordingly, a ranking system that allows for the most likely causative gene variant to appear at the top of the list is needed. Furthermore, the identification of genes that overlap with clinical features is essential for exome analysis. Clinical laboratories have reported the utilization of OMIM, HGMD and Human Phenotype Ontology (HPO) to provide phenotype to genotype associations. But these search tools may miss certain associations due to the lack of frequent updates, synonymous word challenges and the lack of comprehensive search engines. While new computational algorithms are being developed to address these challenges, such tools have not been validated with clinical exome samples. See, for example, exomiser, a candidate variant ranking tool, based on disease gene phenotypes, model organism phenotypes, and protein-protein association neighbors, from simulated exomes. Bone WP *et al. Genet Med Off J Am Coll Med Genet.* 2016 18:608–17.

[0008] Thus, there is a need for improved methods to identify causative genetic variants in Mendelian diseases and disorders, and in particular methods to identify the most likely causative genetic variants from among the several hundred typically remaining after applying standard filters to clinical exome data. The present invention address this need.

SUMMARY

[0009] The present disclosure relates to clinical decision support using whole exome sequencing to identify the most likely causative genetic variants to explain a patient's clinical phenotype, thereby improving clinical diagnosis of difficult to diagnose genetic diseases and disorders. In some implementations, the clinical decision support tool is a computer implemented diagnostic method for diagnosing a genetic disorder in a human subject. The

term “patient” as used herein refers to a human subject. In some implementations, the method is a computer implemented diagnostic method for analysis of one or more genetic variants of a human subject in need of diagnosis for a genetic disorder, the method comprising steps of executing a query using at least one parameter identifying at least one or a plurality of genetic variants associated with the human subject; accessing, based on the executed query, at least one database communicatively coupled to a plurality of data sources, the database containing a plurality of phenotype data and a plurality genetic variants data transmitted from the plurality of data sources; generating, based on the at least one query parameter, at least one correlation between the at least one or a plurality of genetic variants associated with the human subject and one or more genetic variants contained in the at least one database, and outputting, using the at least one database, the one or more genetic variants and associated phenotypes for presentation on a user interface communicatively coupled to the at least one database; correlating the one or more genetic variants and associated phenotypes with a set of clinical features of the human subject; applying a filtering and weighting algorithm to each of the one or more genetic variants which incorporates one or more features selected from the population frequency of each genetic variant in a human population, preferably a human population corresponding to that of the human subject, the type of genetic variant, for example whether it is a synonymous or nonsynonymous nucleotide substitution, a frameshift mutation, a nonsense mutation, or a mutation that alters a start or stop codon, the presence of each genetic variant in at least one database of human mutations, such as the Human Gene Mutation Database, the presence of each genetic variant in at least one database connecting gene-level datasets to biological processes and disease, such as the GenMAPP database, and the strength of the correlation between the associated phenotype of each variant and one or more of the set of clinical features of the human subject; outputting a ranked list of the one or more genetic variants; and displaying, using the user interface, the ranked list of the one or

more genetic variants. The ranked list provides at the top of the list the most likely causative genetic variants for the subject's clinical features, thereby providing clinical decision support to aid the clinician in arriving at the most likely diagnosis for the subject.

[0010] In some implementations, the method further comprises a step of classifying the one or more genetic variants by inheritance mode using an inheritance model. In some implementations, the inheritance model is selected from an autosomal dominant model, an autosomal recessive model, an X-linked model, and a uniparental disomy model.

[0011] In some implementations, the human subject in need of diagnosis has undergone one or more diagnostic evaluations prior to implementation of the method, for example, one or more of a comparative genomic hybridization assay to analyze copy number variations, metabolic screening, and microarray based gene expression profiling. In some implementations, the method further comprises obtaining a biological sample from the human subject, extracting genomic DNA from the sample, and subjecting the DNA to whole exome sequencing. In some implementations, the step of subjecting the DNA to whole exome sequencing comprises one or more steps of sonication, ligation of the DNA to an adapter molecule, amplification of the DNA by a polymerase chain reaction, hybridization, and sequencing.

[0012] The disclosure provides computer implemented methods and tools directed to the problem of identifying one or more causative genetic variants of a Mendelian disease or disorder and thereby providing or informing a clinical diagnosis for the disease or disorder for a patient. The methods may also comprise a set of laboratory and computer implemented steps directed to whole exome sequencing (WES) of a sample of genomic DNA from the patient in need of diagnosis. In some implementations, the current subject matter's methods and tools can be applicable to whole exome sequencing as well as whole genome sequencing (WGS) processes. Generally, the output of WES is the identification of a set of genetic

variants carried by the patient. According to the methods described here, two computer-implemented processes are used to identify the causative variant(s) for the patient's disease or disorder from among the thousands of variants that may be identified by WES. The first receives as input clinical text describing the patient's phenotype and queries one or more databases to identify a list of genes correlating with the features of the patient's clinical phenotype, thereby generating a set of phenotype to genotype correlations that may underlie the patient's clinical features. The second is a weighting, filtering, and ranking tool that utilizes variant annotation, the phenotype to genotype correlations, and a ranking algorithm to place the most likely causative genetic variants at the top of the list, thereby assisting in the diagnosis of the disease or disorder.

[0013] Accordingly, in some implementations, the method can include applying a unique and comprehensive phenotype to genotype association tool in conjunction with a unique weighting, filtering, and ranking algorithm that can rapidly place the most likely causative genetic variants within a "short list", for example of the top 20 or 50 variants. In some implementations, various additional reviews of the obtained results can be performed. By way of non-limiting examples, one or more of the following reviews can be performed: OMIM review based on a phenotype, in-depth review of top candidates by literature review, if available, (e.g., OMIM, PUBMED, pathway analysis, HGMD, ClinVar, Google, etc.), variant classification using the ACMG guidelines (for example, to classify the variants in various categories: pathogenic, likely pathogenic, variant of unknown clinical significance, likely benign, benign, etc.), where the guidelines can rely on at least one of the following: population data, computational and predictive data, functional data (literature), segregation data, de novo data, allelic data, and/or any other database(s), and/or any other data, and/or any combination thereof.

[0014] Thus, in some implementations, the current subject matter relates to a computer-implemented method for identifying the most likely causative genetic variants from among a set of variants identified, for example, by WES.

[0015] The method can include receiving data (e.g., from a health care provider) containing information relating to one or more medical conditions (e.g., de-identified patient data, diagnosis data, etc.), translating the received data to generate at least one gene candidate list, generating a plurality of gene variants, filtering and prioritizing the plurality of gene variants, and displaying a list of filtered gene variants indicative of a diagnosis of the one or more medical conditions. In some implementations, the current subject matter can process data that is received in a variant call format (VCF), which specifies the format of a text file used in bioinformatics for storing gene sequence variations. As can be understood, other data formats (e.g., fastQ format, etc.) can be used. Further, gene candidate lists can be generated in one or more text files for the purposes of further filtration, weighing, ranking, etc.

[0016] In some implementations, the current subject matter can include one or more of the following optional features. The translation of the received data can include conversion of the received data into a predetermined format, mapping converted data to human-genome sequences, and identifying reference sequences. Based on the identified reference sequences, at least one variant is identified. The identified variants can then be annotated (i.e., filtered, ranked, etc.) and variant filtering and prioritization can be performed to generate the list of filtered variants.

[0017] In some implementations, the generation of filtered gene variants can be performed using one or more processors of at least one computing systems. The processors can perform obtaining at least one phenotype-genotype correlation from the received data, genetic inheritance modeling and/or disease segregation (using one or more models) on the obtained phenotype-genotype correlations, determine at least one functional effect of one or

more gene(s) and/or gene variant(s), and apply database knowledge to the determined gene variants.

[0018] In some implementations, the phenotype-genotype correlation can be obtained from one or more databases (e.g., data in the databases can be obtained from various data sources, such as, HPO, OMIM, GTR, Orphanet, in-house curations, etc.) that can be configured to convert clinical keywords to gene lists (e.g., “short stature”, etc.). The converted clinical keywords can be used to search the database(s) to generate a gene list to generate one or more phenotype to genotype correlation(s) for each clinical feature.

[0019] In some implementations, the current subject matter can also process the received data (e.g., in VCF format, fastQ format, etc.) to determine gene variants. This process can be performed using a genetic analysis stage and a gene/variant analysis stage, where the gene/variant analysis stage includes a filtering and a variant/gene ranking sub-stages. During the genetic analysis stage, the current subject matter can perform coding of sequences ± 5 base pairs (bp), set a threshold coverage of at least 10x, perform annotation (e.g., Alamut annotation), store variants in the database, perform inheritance modeling (e.g., AR, AD, XL, UPD) for trios/singleton, and generate uniparental disomy (UPD) graphs. For the filtering-in sub-stage of the gene/variant analysis stage, the current subject matter can include rare variants (e.g., 1000G and $ESP \leq 1\%$), variants $<5\%$ internal controls, nonsynonymous variants, synonymous variants at exon/intron junctions, and/or other parameters. Based on these parameters, the current subject matter can perform variant/gene ranking by weighing a gene/variant on the list based on frequency (e.g., $\leq 1\%$ in public databases), weighing a gene/variant on the list based on coding effect (e.g., frameshift, nonsense, start/stop loss, etc.), weighing a gene/variant on the list based on pathogenicity predictions (e.g., Grantham scale, SIFT, MAPP, etc.), weighing a gene/variant on the list based on presence in protein domain, weighing a gene/variant on the list based on existence

in mutation (HGMD) and clinical (ClinVar) databases, and weighing a gene/variant on the list based gene overlap with phenotype-genotype list. Based on the weighing, the current subject matter can generate a list of top variants. In some implementations, if there are variants found in the phenotype-genotype generated gene lists, then these variants can be ranked by assigning them a higher weight because they can correspond to an important match in a phenotype. Thus, variants in an important phenotype related genes can be determined and top variants identified.

[0020] Non-transitory computer program products (i.e., physically embodied computer program products) are also described that store instructions, which when executed by one or more data processors of one or more computing systems, causes at least one data processor to perform operations herein. Similarly, computer systems are also described that may include one or more data processors and memory coupled to the one or more data processors. The memory may temporarily or permanently store instructions that cause at least one processor to perform one or more of the operations described herein. In addition, methods can be implemented by one or more data processors either within a single computing system or distributed among two or more computing systems. Such computing systems can be connected and can exchange data and/or commands or other instructions or the like via one or more connections, including but not limited to a connection over a network (e.g., the Internet, a wireless wide area network, a local area network, a wide area network, a wired network, or the like), via a direct connection between one or more of the multiple computing systems, etc. As stated above, to identify phenotype-genotype correlations, clinical terms describing the phenotype of a patient can be received/used and to determine top gene variants, VCF, FastQ, etc. data can be used. The clinical terms can be manually entered (e.g., by a clinician, doctor, researcher, etc.) and/or automatically generated.

[0021] In some implementations, the methods described here may also include one or more steps of obtaining a genomic DNA sample from a patient, fragmenting the DNA, for example by sonication, ligating the DNA to an adaptor molecule, amplification of the DNA by a polymerase chain reaction, hybridization or capture based target enrichment, and sequencing of the amplified DNA, for example paired-end sequencing and/or Sanger sequencing.

[0022] The details of one or more variations of the subject matter described herein are set forth in the accompanying drawings and the description below. Other features and advantages of the subject matter described herein will be apparent from the description and drawings, and from the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0023] The accompanying drawings, which are incorporated in and constitute a part of this specification, show certain aspects of the subject matter disclosed herein and, together with the description, help explain some of the principles associated with the disclosed implementations. In the drawings,

[0024] FIG. 1 illustrates an exemplary process for identification of gene variants, according to some implementations of the current subject matter;

[0025] FIG. 2 illustrates exemplary sources used in identification of gene variants, according to some implementations of the current subject matter;

[0026] FIG. 3 illustrates an exemplary experimental summary of a gene identification process, according to some implementations of the current subject matter;

[0027] FIG. 4 illustrates an exemplary process for identification of gene variants along with corresponding plots illustrating filtering of gene variants, according to some implementations of the current subject matter;

[0028] FIG. 5 is an exemplary plot illustrating identified gene variants, according to some implementations of the current subject matter;

[0029] FIG. 6 is an exemplary table summarizing experimental results of the gene variant identification process, according to some implementations of the current subject matter;

[0030] FIG. 7 illustrates an exemplary system, according to some implementations of the current subject matter; and

[0031] FIG. 8 illustrates an exemplary method, according to some implementations of the current subject matter.

DETAILED DESCRIPTION

[0032] In some implementations, the current subject matter relates to a method, a system, and a computer program product for effectively identifying gene variants that are causative of a Mendelian genetic disorder or disease.

[0033] In some implementations, the current subject matter provides for effectively identifying gene variants based on received data for the purposes of providing diagnosis of various diseases and/or disorders. The current subject matter can be implemented using hardware, software, and/or any combination thereof. One or more processors of one or more computing systems can be involved in various operations associated with the identification of gene variants, whereby data can be obtained from one or more databases, servers, etc. and can be transmitted over one or more networks (e.g., wireless, wired, and/or any combination thereof). The current subject matter processes can be implemented using at least one of the following: a telephone, a smartphone, a personal computer, a laptop, a tablet computer, a personal digital assistant, and/or any other computing device and/or any combination thereof.

[0034] In some implementations, the current subject matter can receive various data (e.g., medical provider data), identify gene variants, filter and prioritize identified gene

variants, and produce a “short list” of gene variants that can be indicative of a particular disorder.

[0035] In some implementations, the identification of a “short list” of gene variants can be performed using a clinical exome analysis that can utilize a four-level framework (as for example is shown in FIG. 1) which can apply a phenotype-genotype association tool in conjunction with a unique weighing, filtering and ranking algorithm into a single analysis procedure that can identify the most likely causative gene variant of a disorder at the top of the variant “short list”. In particular, to generate such “short list” of variants, clinical case data can be received from one or more healthcare providers, phenotype information contained in the received data can be translated into candidate gene lists, and, finally, filtering/prioritization of the gene lists can be performed. An experimental clinical exome analysis performed in accordance with the exemplary implementations of the current subject matter (as discussed below in the Exemplary Experiments section) is referred to as the Cincinnati Clinical Exome Pipeline Analysis Suite (“CCEPAS”). The CCEPAS was developed and validated using 100 clinical exome cases.

[0036] FIG. 1 illustrates an exemplary four-level framework 100 that can perform the identification of the gene variant list, according to some implementations of the current subject matter. The framework 100 can include the following stages (as shown in FIG. 1):

- 1) phenotype-genotype correlation stage 102;
- 2) genetic inheritance modeling and disease segregation stage 104;
- 3) determination of gene/variant functional effects (e.g., deleteriousness) stage 106;
- 4) application of database knowledge-based evidence stage 108.

[0037] Each of these stages are further discussed below.

Phenotyping Stage 102

[0038] In some implementations, during phenotyping stage 102, phenotype-genotype correlations can be obtained from the received data using phenotype keywords, dynamically searching databases and displaying results that match those keywords (as, for example, shown in FIG 2). In some exemplary implementations, phenotype keywords can be used to identify one or more phenotypes/disorders and then, based on the identified phenotypes/disorders, select a specific phenotype and/or disorder, where a gene list can be downloaded (e.g., as a text file). The gene list can then be used for the purposes of gene and variant prioritization.

[0039] As shown in FIG. 2, users can be allowed to enter various gene associations in connection with the displayed results. In some implementations, the gene associations can be automatically generated. For example, a keyword can be entered using a user interface (e.g., using dropdown menus, prompts, etc.), a query can be generated to a database (e.g., Pheno2Gene database shown in FIG. 2) that can include phenotype-genotype correlations, the database can be searched, and in response to the query, a gene list can be generated based on the entered keyword (e.g., as a text file, and/or any other file). In some exemplary implementations, the Pheno2Gene database can include gene sources that can be obtained from various sources, such as in-house curations, and genetic testing registry (GTR). Genes that are obtained from Pheno2Gene database that overlap with gene variants determined during variant evaluation stage can be weighed and/or ranked in order to generate a list of top candidate genes for any further analysis. Contrary to some conventional systems, Pheno2Gene can provide specific genes lists of phenotype-to-gene correlations for each clinical term (e.g., keyword) and/or clinical phenotype feature.

Genetic Stage 104

[0040] In some implementations, during the genetic stage 104, analysis can be restricted to coding sequences ± 5 bp of intron/exon boundaries and/or any other boundaries. In addition, inheritance modeling of the variants can be performed. These models serve as guides for the potential inheritance of discovered variants, and also allows the pipeline to filter those variants with non-informative inheritance. Inheritance modelling can also provide focus for analysis, such that variant zygosity can be taken into account when considering a disorder with known inheritance.

[0041] In some implementations, during stage 104, the current subject matter can process the received data (e.g., in VCF format, fastQ format, etc.) to determine gene variants. The gene/variant analysis stage 104 can include a filtering and a variant/gene ranking sub-stages. During the genetic analysis stage, the current subject matter can perform coding of sequences ± 5 base pairs (bp), set a threshold coverage of at least 10x, perform annotation (e.g., Alamut annotation), store variants in the database, perform inheritance modeling (e.g., AR, AD, XL, UPD) for trios/singleton, and generate uniparental disomy (UPD) graphs. For the filtering-in sub-stage of the gene/variant analysis stage, the current subject matter can include rare variants (e.g., 1000G and ESP $\leq 1\%$), variants $<5\%$ internal controls, nonsynonymous variants, synonymous variants at exon/intron junctions, and/or other parameters. Based on these parameters, the current subject matter can perform variant/gene ranking (as described below with regard to stage 106) by weighing a gene/variant on the list based on frequency (e.g., $\leq 1\%$ in public databases), weighing a gene/variant on the list based on coding effect (e.g., frameshift, nonsense, start/stop loss, etc.), weighing a gene/variant on the list based on pathogenicity predictions (e.g., Grantham scale, SIFT, MAPP, etc.), weighing a gene/variant on the list based on presence in protein domain, weighing a gene/variant on the list based on existence in mutation (HGMD) and clinical (ClinVar)

databases, and weighing a gene/variant on the list based gene overlap with phenotype-genotype list. Based on the weighing, the current subject matter can generate a list of top variants. In some implementations, if there are variants found in the phenotype-genotype generated gene lists, then these variants can be ranked by assigning them a higher weight because they can correspond to an important match in a phenotype. Thus, variants in an important phenotype related genes can be determined and top variants identified.

[0042] In some exemplary implementations, specifically related to inheritance modeling, the current subject matter can determine which parent (e.g., mother, father, etc.) each specific gene variant is originating from and can further ascertain if variants have inherited an autosomal recessive, autosomal dominant, X-Linked, or in *de novo* fashion. In this case, the inputs can include at least one of the following variants: (1) variants in coding sequences ± 5 bp, (b) variants with coverage $>10X$, and/or any other variants.

Gene/Variant Stage 106

[0043] In some implementations, during the gene/variant stage, filtering of variants can be performed. The filtering algorithm can be implemented using the following formula:

$$S_i = \sum_{j=1}^n w_j x_j \quad (1)$$

[0044] where S_i is a “combined score” of a variant i , w_j is a weight of a prediction algorithm j , and x_j is a score of the prediction algorithm j for variant i . The same equation can be applied to gene weights as well as variant weights independently. At the gene stage, weight can be applied to genes identified by the phenotype-genotype correlation categorizing them into phenotype overlapping and non-OMIM genes. The top variants can be examined to ensure that all criteria are met.

Knowledge Stage 108

[0045] In some implementations, during the knowledge stage 108, the top variants can become top candidates following an OMIM record review based on a phenotype. Then,

variants that made it through can be assessed in-depth by pathway analysis (e.g., HGMD, ClinVar, PUBMED and GOOGLE searches, and/or any others). Variants can be classified according to the American College of Medical genetics (ACMG) guidelines into five categories: pathogenic, likely pathogenic, variant of unknown clinical significance (VUCS), likely benign or benign. Secondary findings can be reported only if they met the criteria of being likely pathogenic or pathogenic variants and the proband/families opted to receive or reject the secondary findings.

EXEMPLARY EXPERIMENTS

[0046] The exemplary experiments performed in accordance with the current subject matter processes and systems, were conducted using output data from the Illumina HiSeq 2500 that was converted from binary base call (“bcl”) files to FastQ files using the Illumina Consensus Assessment of Sequence and Variation (CASAVA) software, version 1.8, and mapped to the reference haploid human-genome sequence (hg19) with the BWA program 0.5.9. Variant calls, which differed from the reference sequence, were obtained with the use of GATK 7.7.4. Finally, Alamut HT 1.1.8 was used for variant annotation.

[0047] In this exemplary experiment, one hundred pediatric patients referred for exome sequencing have had the analysis and results disclosure completed. The patients in this cohort had diverse clinical features. Before referral, all patients had undergone extensive diagnostic evaluations (e.g. aCGH microarray, targeted gene tests/panels, metabolic screening, clinical genetic evaluations, and other laboratory workup) that did not lead to a unifying diagnosis. Consent for clinical WES was obtained from the patients and/or their family. Internal review board (IRB) approval was obtained at Cincinnati Children's Hospital Medical Center (CCHMC) for this retrospective study.

[0048] WES and analysis protocols were developed and validated by the CCHMC molecular genetics laboratory of the Division of Human Genetics. Briefly, genomic DNA samples from patients were fragmented by sonication, ligated to Illumina multiplexing paired-end adapters, amplified by means of a polymerase-chain-reaction and hybridized to biotin-labeled NimbleGen V3 exome capture reagent (Roche NimbleGen). Hybridization was achieved at 47°C for 64 to 72 hours. After washing and re-amplification, paired-end sequencing (2x100 bp) was performed on the Illumina. HiSeq 2500 platforms to provide average sequence coverage of more than 100x, with more than 97% of the target bases having at least 10x coverage. Clinically relevant variants, from proband and parental samples (whenever available), were confirmed by Sanger sequencing.

[0049] As stated above, one hundred pediatric patients referred for exome sequencing were analyzed using the Cincinnati Clinical Exome Pipeline Analysis Suite discussed above (as well as shown in Figure 1). The patients in this cohort had diverse array of clinical features including immunodeficiency, neurological disorders and multiple congenital anomalies. There were 32 positive cases, including 9 singletons and 23 trios, from a total of 100 consecutive cases.

[0050] As quality control/quality assurance parameters, the average coverage, percentage at 10x and 20x per family were measured. The typically QC/QA acceptable average coverage was >100x and the percentage of >95% at 10x. The mean average coverage for the 100 case cohort was 114.60x. In addition, the mean percentage coverage was at 10x and 20x was 97.01 and 95.37%, respectively.

Phenotyping stage 102

[0051] A phenotyping process (e.g., Pheno2Gene tool) utilized The Human Phenotype Ontology (HPO), Online Mendelian Inheritance in Man (OMIM), Genetic Testing Registry (GTR), ORPHANET and in-house manual curations as the main sources of

information. A database update script was created to download these data from static URLs, and each disorder and phenotype was passed through a text collapsing algorithm to reduce the phenotype to its base elements, often stripping off unnecessary qualifiers ("Type I", "IA", "In males", and others). In this way the phenotypes were merged with their synonymous equivalents from other data sources. A second web-based system was built to allow users to add manual curations in the form of new gene to phenotype associations that were found in the primary literature and not in other databases. These associations went through a review and approval process before they were added to the Pheno2Gene database, thus, adding novel genes to the relevant phenotype list. In this way Pheno2Gene was not limited to external curated data but phenotype-genotype associations were gathered and curated, thereby growing the knowledge base.

[0052] Variants were analyzed and interpreted using CCEPAS implementing a weighing and ranking system based on phenotype-genotype correlations, genetic principles, gene/variant deleteriousness and database knowledge-based evidence (as shown in FIG. 1). Pheno2Gene tool was implemented HPO, OMIM, GTR, ORPHANET and in-house manual curations, as the sources of information phenotype-genotype correlations (as shown in FIG. 2). Phenotype keywords were entered and once a phenotype and/or disorder was selected the process allowed for the gene list to be generated and downloaded as a text file. This text file was used for gene prioritization by variant identification process (e.g., VarEval).

[0053] Before implementation, a phenotype-genotype validation was performed in two phases: 1) the accuracy, specificity and sensitivity of 10 known disorders, namely, Fanconi anemia, CHARGE syndrome, Sotos syndrome, Smith-Lemli-Opitz syndrome, Wilson disease, MCAD deficiency, Joubert syndrome, Osteogenesis imperfecta, Marfan syndrome and Rett syndrome; and 2) the concordance of 10 positive exome cases. To test Pheno2Gene, ten representative genetic syndromes with diverse clinical features and disorder

prevalence with well-defined causative genes were selected to determine whether the correct gene lists were provided. Specifically, the disorders were binned into 3 categories: rare ($<1/50000$), fairly common ($1/20000$ - $1/50000$) and common ($>1/20000$) (as shown in FIG. 3). Additionally, the features of the selected syndrome and included hematological (1 syndrome), connective tissue/skeletal (2 syndromes), neurological (2 syndromes), multiple congenital defects (2 syndromes) and metabolic clinical features (2 syndromes). For each syndrome (phenotype-genotype correlation), a gene list output of syndromes was compared to the performance to other popular phenotype-genotype correlational software options, namely, Phenomizer (<http://compbio.charite.de/phenomizer>) and Ingenuity (by Qiagen). True Positives were defined as genes that are the output by each method that were in the "Gene list" Gene # Column (i.e., 16 for Fanconi Anemia). False Positives were those genes output by each method that are not in the "Gene list." As for True Negatives, this was the entire set of protein coding genes for human, minus those genes output by each method. False Negatives were genes that are associated with the syndrome and in the "Gene list" but were missed by the method. Pheno2Gene tool accuracy was similar to that of Phenomizer and Ingenuity softwares (as shown in FIG. 3). However, the sensitivity of Pheno2Gene is significantly higher than Phenomizer, but the same as Ingenuity. Notably, Pheno2Gene outperformed Phenomizer and Ingenuity on specificity.

Genetic Stage 104 and Gene/Variant Stage 106

[0054] The inferred inheritance models for trio cases include autosomal dominant, autosomal recessive, X-Linked, as well as a mechanism for detecting uniparental disomy. However, for singleton cases, inheritance modelling was not performed, so the analysis is driven by phenotyping and variant prioritization alone.

[0055] Variants were filtered (using a VarEval tool) on the basis of low frequency found in public databases (dbSNP and ESP database frequencies of $< 1\%$ or frequency not

available) and internal normal control database (frequency of 5%) as well as variant type (inclusion of nonsynonymous and synonymous at exon junctions). VarEval weighs and ranks variants based on low frequency (<1% in public databases: dbSNP, ESP, 1000 genomes and ExAC databases), coding effect (frameshift, nonsense and start/stop loss), pathogenicity predictions (Grantham scale, SIFT and MAPP), presence in protein domain, existence in mutation (HGMD "DM" and "DM?") and clinical (ClinVar "pathogenic") variant databases.

[0056] At the genetic stage, VarEval tool filtered-in variants in coding sequences ± 5 bp of intron/exon boundaries (exome region of interest) that are greater than 10X coverage and fitted variants into inheritance models. VarEval tool then filtered-in non-synonymous and filtered out variants in pseudogenes, non-HGMD at minor allele frequency (MAF) >1% ESP and HGMD at MAF >5%. In addition to filtering, VarEval weighed genes, based on the match to phenotype keywords and variant pathogenicity parameters: <1% in public databases, namely, dbSNP, ESP, 1000 genomes and ExAC databases, coding effect (frameshift, nonsense and missense), pathogenicity predictions (Grantham scale, SIFT and MAPP), presence in protein domain, existence in mutation (HGMD "DM" and "DM?") and clinical (ClinVar "pathogenic") variant databases. For example, a mutation in *PAX1* was found rapidly in a clinical case by applying the VarEval algorithm in combination with the utilization of the Pheno2Gene tool (as shown in FIG. 4, part A). Specifically, prior to VarEval there were 153,376 variants, however, the variant number rapidly dropped to 1027 after performing VarEval filtering. This example demonstrated the utility of VarEval and Pheno2Gene to rapidly prioritize potential candidate gene variants. Generally, on average prior to VarEval clinical exome cases had approximately 150,000 variants and 500 variants post-VarEval filtering (as shown in FIG 4, parts B and C, respectively).

[0057] In addition to filtering, VarEval weighed genes based upon their overlap with gene lists produced by Pheno2Gene, that represent the clinical features of the clinical case.

VarEval also weighed variants based on population frequency, computed pathogenicity characteristics, and presence in mutation databases such as ClinVar and HGMD. To demonstrate the effect of filtering and weighing, the causative variants of 100 clinical exome cases were ranked and divided by inheritance mode. In autosomal dominant cases, the causative variant was found in the top 1, top 10, top 20 and top 50 for 70%, 90%, 100% and 100% of the cases, respectively. Similarly, for autosomal recessive cases, the two causative variants were found in the top 1 and top 20 for 50% and 100% of the cases. Compared to this ranking, X-linked cases were always found to have the causative variants as the top hit. In contrast to trio exome cases, singleton (proband only) cases had a slight reduction of variants in the top 1 (50%), top 10 (70%) and top 20 (70%). For case 7 as an example, the filtering process narrowed down the variants to 1027 and when the gene and variant weights were accounted for, the causative *PAXI* variant moved to the 17th position (as shown in FIG. 4, part A). However, it occupied the 1st position of the homozygous inheritance model. Generally, the causative variants were found in the top 50 list for all cases. This example demonstrated the utility of VarEval and Pheno2Gene to rapidly prioritize potential candidate gene variants from thousands of variants to the causative one(s).

Knowledge Stage 108

[0058] At this stage, clinical exome sequencing data interpretation was performed by a team represented by both molecular and clinical geneticists, pediatric subspecialists and genetic counselors.

[0059] The filtered variants became top candidates following an OMIM record review based on phenotype. The variants that made it through were assessed in-depth by pathway analysis, HGMD, ClinVar, PUBMED and GOOGLE searches and were classified into five categories: pathogenic, likely pathogenic, VUCS, likely benign or benign. Approximately 50% of the likely pathogenic or pathogenic variants have not been previously reported. In

contrast, a significant number of reported variants in our exome cases were only recently known on disease – gene discoveries.

[0060] To assess the performance of Pheno2Gene tool, a validation comparison between Pheno2Gene, Ingenuity and Phenomizer tools utilizing keywords and genes from 10 genetic, rare to ultrarare, disorders with a wide range of clinical features was performed (as shown in FIG. 3, part A). Pheno2Gene's accuracy, 94.8%, was identical to Phenomizer and Ingenuity (as shown in FIG. 3, part A). However, Pheno2Gene outperformed Phenomizer and Ingenuity in terms of specificity and only Phenomizer in terms of sensitivity. Other reports on clinical exome analysis taken into account phenotype-based analysis, however, they only utilized OMIM and HGMD. OMIM clinical feature searches can be non-specific with a large number of pages as an output for each keyword, while HGMD phenotype searches miss a number of known and recent phenotype-gene associations. Another group reported using the Human Phenotype Ontology (HPO) and OMIM to make the phenotype to gene associations. Pheno2Gene is a comprehensive tool that has five resources (HPO, OMIM, Orphanet, in-house curations and Genetic Testing Registry) for obtaining gene lists from clinical features that is frequently updated from resource databases, thus, providing the latest phenotype to gene relationships.

[0061] The gene/variant weighing, filtering and ranking stages of the analysis process have been consolidated by one algorithm, named VarEval (Variant Evaluator; as shown in FIGS. 1, 4, 5). VarEval uses three strategies, namely, weighing (genes and variants), filtering (variants), and ranking (genes and variants), to identify causative gene variants. VarEval weighed variants by the pathogenicity assessment and genes by their association to phenotype. In addition, VarEval filtering in aforementioned parameters rapidly decreased the number of variants from ~150,000 to roughly ~500 (as shown in FIG. 4, parts A, B, C). The third component utilized a ranking approach, whereby, the gene ranking is performed first,

most phenotypes matching to the top, and within the gene, variants are sorted by decreasing order of pathogenicity. Thus, the outcome of this ranking approach permitted the phenotypically matching genes with deleterious variants to be, in general, within the top 50 variants for both trio and singleton cases (as shown in FIG. 4, part A, and FIG. 5). In fact, for autosomal dominant cases, 70% of gene variants that explained the phenotype of the case was, indeed, the top 1 variant, and 90% of cases had the causative gene variants in the top 10 variants. This approach, of utilizing Pheno2Gene and VarEval, was validated with 100 clinical exome cases (as shown in FIGS. 5 and 6). Due to the nature of the weighing, filtering and ranking, CCEPAS also has the potential of being used in a rapid exome situations whereby critical cases are involved that require fast turnaround times.

[0062] In some implementations, the current subject matter can be configured to be implemented in a system 700, as shown in FIG. 7. The system 700 can include a processor 710, a memory 720, a storage device 730, and an input/output device 740. Each of the components 710, 720, 730 and 740 can be interconnected using a system bus 750. The processor 710 can be configured to process instructions for execution within the system 700. In some implementations, the processor 710 can be a single-threaded processor. In alternate implementations, the processor 710 can be a multi-threaded processor. The processor 710 can be further configured to process instructions stored in the memory 720 or on the storage device 730, including receiving or sending information through the input/output device 740. The memory 720 can store information within the system 700. In some implementations, the memory 720 can be a computer-readable medium. In alternate implementations, the memory 720 can be a volatile memory unit. In yet some implementations, the memory 720 can be a non-volatile memory unit. The storage device 730 can be capable of providing mass storage for the system 700. In some implementations, the storage device 730 can be a computer-readable medium. In alternate implementations, the storage device 730 can be a floppy disk

device, a hard disk device, an optical disk device, a tape device, non-volatile solid state memory, or any other type of storage device. The input/output device 740 can be configured to provide input/output operations for the system 700. In some implementations, the input/output device 740 can include a keyboard and/or pointing device. In alternate implementations, the input/output device 740 can include a display unit for displaying graphical user interfaces.

[0063] FIG. 8 illustrates an exemplary method 800 for identifying gene variants, according to some implementations of the current subject matter. At 802, data (e.g., from a health care provider) containing information relating to one or more medical conditions (e.g., de-identified patient data, diagnosis data, etc.) can be received. At 804, the received data can be translated to generate at least one gene candidate list. At 806, a plurality of gene variants can be generated. At 808, the plurality of gene variants can be filtered and prioritized. A list of filtered gene variants indicative of a diagnosis of the one or more medical conditions can be generated and/or displayed, at 810.

[0064] In some implementations, the current subject matter relates to a computer implemented method for generating a list of gene variants. The method can include executing a query using at least one parameter identifying a phenotype associated with a patient; accessing, based on the executed query, at least one database communicatively coupled to a plurality of data sources, the database containing a plurality of phenotype data and a plurality genotype data transmitted from the plurality of data sources; generating, based on the query parameter(s), at least one correlation between the identified phenotype and one or more genotypes contained in the database, and outputting, using the database, one or more genotypes for presentation on a user interface communicatively coupled to the database; generating, based on the query parameter, one or more gene variants corresponding to the identified phenotype; selecting at least one gene variant from the generated one or more gene

variants and correlating the selected gene variant with the generated one or more genotypes; ranking one or more gene variants based on the correlation of the selected gene variants and the generated one or more genotypes; and displaying, using the user interface, a ranked list of the one or more gene variants. The ranking can also include weighing one or more gene variants.

[0065] In some implementations, the current subject matter can include one or more of the following optional features. The translation of the received data can include conversion of the received data into a predetermined format, mapping converted data to human-genome sequences, and identifying reference sequences. Based on the identified reference sequences, at least one variant is identified. The identified variants can then be annotated (i.e., filtered, ranked, etc.) and variant filtering and prioritization can be performed to generate the list of filtered variants.

[0066] In some implementations, the generation of filtered gene variants can be performed using one or more processors of at least one computing systems. The processors can perform obtaining at least one phenotype-genotype correlation from the received data, genetic inheritance modeling and/or disease segregation (using one or more models) on the obtained phenotype-genotype correlations, determine at least one functional effect of one or more gene(s) and/or gene variant(s), and apply database knowledge to the determined gene variants.

[0067] The systems and methods disclosed herein can be embodied in various forms including, for example, a data processor, such as a computer that also includes a database, digital electronic circuitry, firmware, software, or in combinations of them. Moreover, the above-noted features and other aspects and principles of the present disclosed implementations can be implemented in various environments. Such environments and related applications can be specially constructed for performing the various processes and

operations according to the disclosed implementations or they can include a general-purpose computer or computing platform selectively activated or reconfigured by code to provide the necessary functionality. The processes disclosed herein are not inherently related to any particular computer, network, architecture, environment, or other apparatus, and can be implemented by a suitable combination of hardware, software, and/or firmware. For example, various general-purpose machines can be used with programs written in accordance with teachings of the disclosed implementations, or it can be more convenient to construct a specialized apparatus or system to perform the required methods and techniques.

[0068] The systems and methods disclosed herein can be implemented as a computer program product, i.e., a computer program tangibly embodied in an information carrier, e.g., in a machine readable storage device or in a propagated signal, for execution by, or to control the operation of, data processing apparatus, e.g., a programmable processor, a computer, or multiple computers. A computer program can be written in any form of programming language, including compiled or interpreted languages, and it can be deployed in any form, including as a stand-alone program or as a module, component, subroutine, or other unit suitable for use in a computing environment. A computer program can be deployed to be executed on one computer or on multiple computers at one site or distributed across multiple sites and interconnected by a communication network.

[0069] As used herein, the term “user” can refer to any entity including a person or a computer.

[0070] Although ordinal numbers such as first, second, and the like can, in some situations, relate to an order; as used in this document ordinal numbers do not necessarily imply an order. For example, ordinal numbers can be merely used to distinguish one item from another. For example, to distinguish a first event from a second event, but need not imply any chronological ordering or a fixed reference system (such that a first event in one

paragraph of the description can be different from a first event in another paragraph of the description).

[0071] The foregoing description is intended to illustrate but not to limit the scope of the invention, which is defined by the scope of the appended claims. Other implementations are within the scope of the following claims.

[0072] These computer programs, which can also be referred to programs, software, software applications, applications, components, or code, include machine instructions for a programmable processor, and can be implemented in a high-level procedural and/or object-oriented programming language, and/or in assembly/machine language. As used herein, the term “machine-readable medium” refers to any computer program product, apparatus and/or device, such as for example magnetic discs, optical disks, memory, and Programmable Logic Devices (PLDs), used to provide machine instructions and/or data to a programmable processor, including a machine-readable medium that receives machine instructions as a machine-readable signal. The term “machine-readable signal” refers to any signal used to provide machine instructions and/or data to a programmable processor. The machine-readable medium can store such machine instructions non-transitorily, such as for example as would a non-transient solid state memory or a magnetic hard drive or any equivalent storage medium. The machine-readable medium can alternatively or additionally store such machine instructions in a transient manner, such as for example as would a processor cache or other random access memory associated with one or more physical processor cores.

[0073] To provide for interaction with a user, the subject matter described herein can be implemented on a computer having a display device, such as for example a cathode ray tube (CRT) or a liquid crystal display (LCD) monitor for displaying information to the user and a keyboard and a pointing device, such as for example a mouse or a trackball, by which the user can provide input to the computer. Other kinds of devices can be used to provide for

interaction with a user as well. For example, feedback provided to the user can be any form of sensory feedback, such as for example visual feedback, auditory feedback, or tactile feedback; and input from the user can be received in any form, including, but not limited to, acoustic, speech, or tactile input.

[0074] The subject matter described herein can be implemented in a computing system that includes a back-end component, such as for example one or more data servers, or that includes a middleware component, such as for example one or more application servers, or that includes a front-end component, such as for example one or more client computers having a graphical user interface or a Web browser through which a user can interact with an implementation of the subject matter described herein, or any combination of such back-end, middleware, or front-end components. The components of the system can be interconnected by any form or medium of digital data communication, such as for example a communication network. Examples of communication networks include, but are not limited to, a local area network (“LAN”), a wide area network (“WAN”), and the Internet.

[0075] The computing system can include clients and servers. A client and server are generally, but not exclusively, remote from each other and typically interact through a communication network. The relationship of client and server arises by virtue of computer programs running on the respective computers and having a client-server relationship to each other.

[0076] The implementations set forth in the foregoing description do not represent all implementations consistent with the subject matter described herein. Instead, they are merely some examples consistent with aspects related to the described subject matter. Although a few variations have been described in detail above, other modifications or additions are possible. In particular, further features and/or variations can be provided in addition to those set forth herein. For example, the implementations described above can be directed to various

combinations and sub-combinations of the disclosed features and/or combinations and sub-combinations of several further features disclosed above. In addition, the logic flows depicted in the accompanying figures and/or described herein do not necessarily require the particular order shown, or sequential order, to achieve desirable results. Other implementations can be within the scope of the following claims.

What is claimed is:

1. A computer implemented method for clinical decision support, the method comprising the analysis of one or more genetic variants of a human subject in need of diagnosis for a genetic disorder, the method comprising steps of
 - executing a query using at least one parameter identifying at least one or a plurality of genetic variants associated with the human subject;
 - accessing, based on the executed query, at least one database communicatively coupled to a plurality of data sources, the database containing a plurality of phenotype data and a plurality genetic variants data transmitted from the plurality of data sources;
 - generating, based on the at least one query parameter, at least one correlation between the at least one or a plurality of genetic variants associated with the human subject and one or more genetic variants contained in the at least one database, and outputting, using the at least one database, the one or more genetic variants and associated phenotypes for presentation on a user interface communicatively coupled to the at least one database;
 - correlating the one or more genetic variants and associated phenotypes with a set of clinical features of the human subject;
 - applying a filtering and weighing algorithm to each of the one or more genetic variants which incorporates one or more features selected from the population frequency of each genetic variant in a human population, preferably a human population corresponding to that of the human subject, the type of genetic variant, for example whether it is a synonymous or nonsynonymous nucleotide substitution, a frameshift mutation, a nonsense mutation, or a mutation that alters a start or stop codon, the presence of each genetic variant in at least one database of human mutations, such as the Human Gene Mutation Database, the presence of each genetic variant in at least one database connecting gene-level datasets to biological processes and disease, such as the GenMAPP database, and the strength of the correlation

between the associated phenotype of each variant and one or more of the set of clinical features of the human subject;

outputting a ranked list of the one or more genetic variants; and

displaying, using the user interface, the ranked list of the one or more genetic variants, thereby providing clinical decision support in the form of the most likely genetic variant(s) underlying the subject's clinical features.

2. The method according to claim 1, further comprising a step of classifying the one or more genetic variants by inheritance mode using an inheritance model.

3. The method of claim 2, wherein the inheritance model is selected from an autosomal dominant model, an autosomal recessive model, an X-linked model, and a uniparental disomy model.

4. The method of any one of claims 1-3, wherein the human subject in need of diagnosis has undergone one or more diagnostic evaluations prior to implementation of the method.

5. The method of claim 4, wherein the one or more diagnostic evaluations include one or more of a comparative genomic hybridization assay to analyze copy number variations, metabolic screening, and microarray based gene expression profiling.

6. The method of any one of claims 1-5, further comprising obtaining a biological sample from the human subject, extracting genomic DNA from the sample, and subjecting the DNA to whole exome sequencing.

7. The method of claim 6, wherein the step of subjecting the DNA to whole exome sequencing comprises one or more steps of sonication, ligation of the DNA to an adapter molecule, amplification of the DNA by a polymerase chain reaction, hybridization, and sequencing.

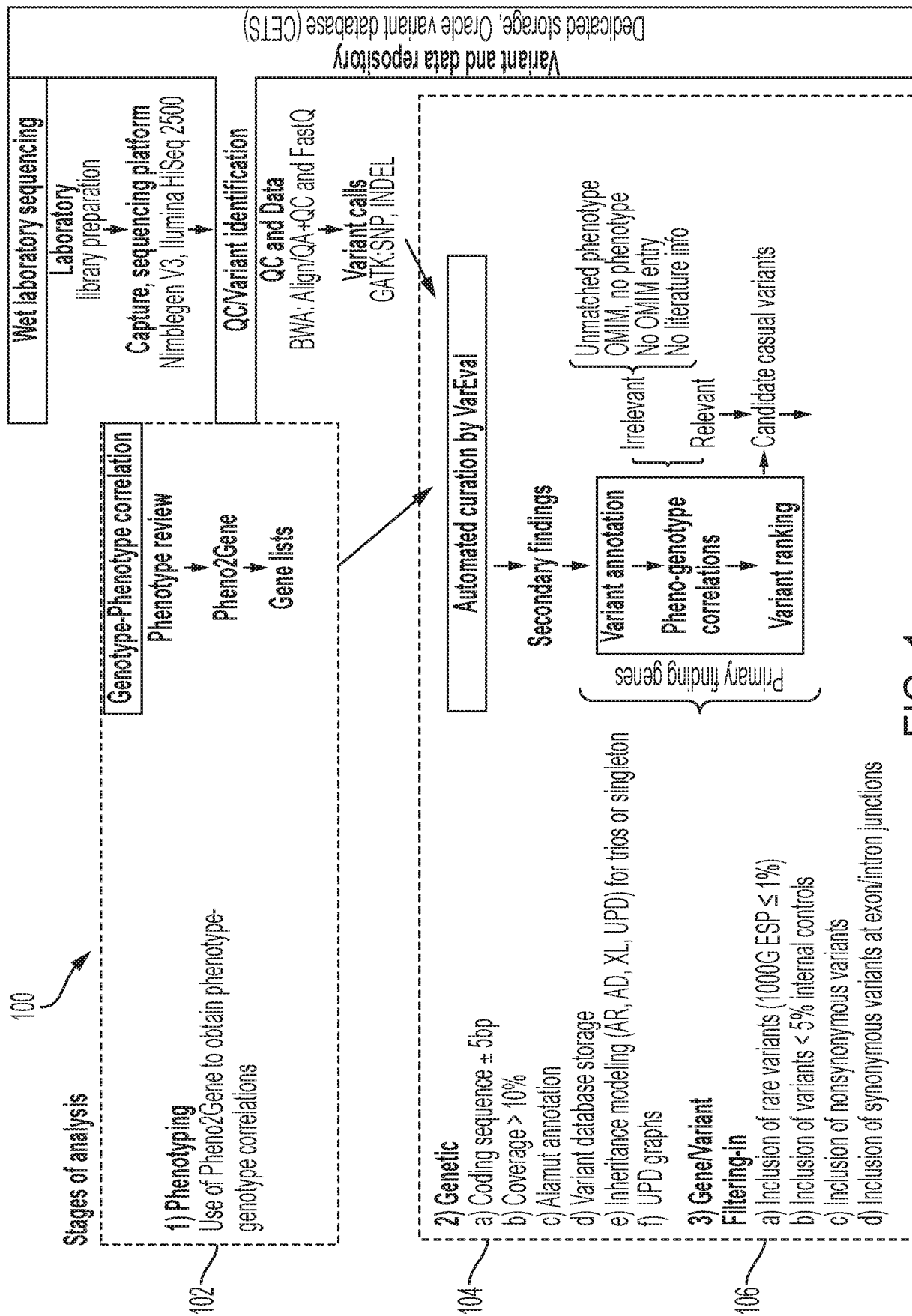


FIG. 1

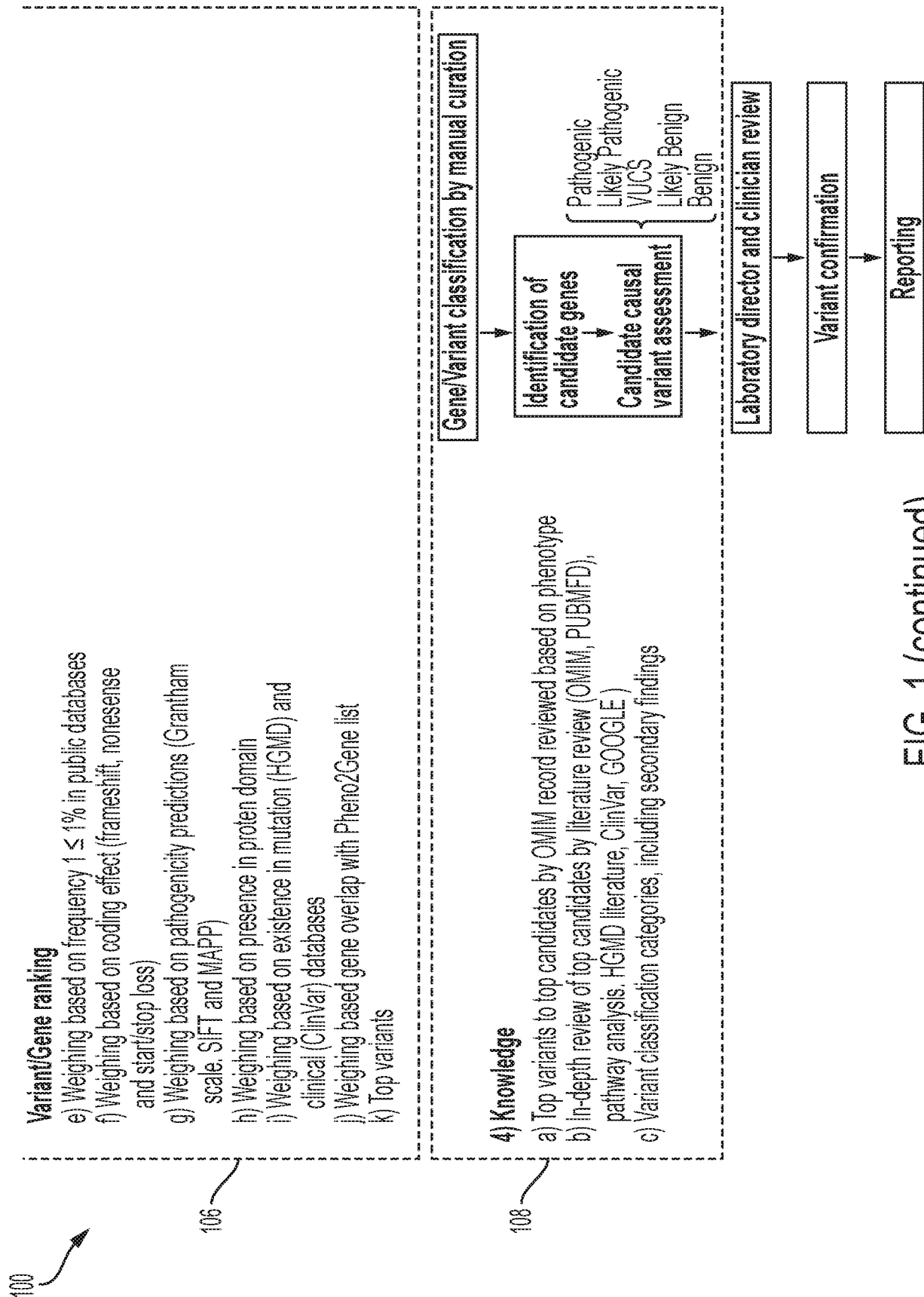


FIG. 1 (continued)

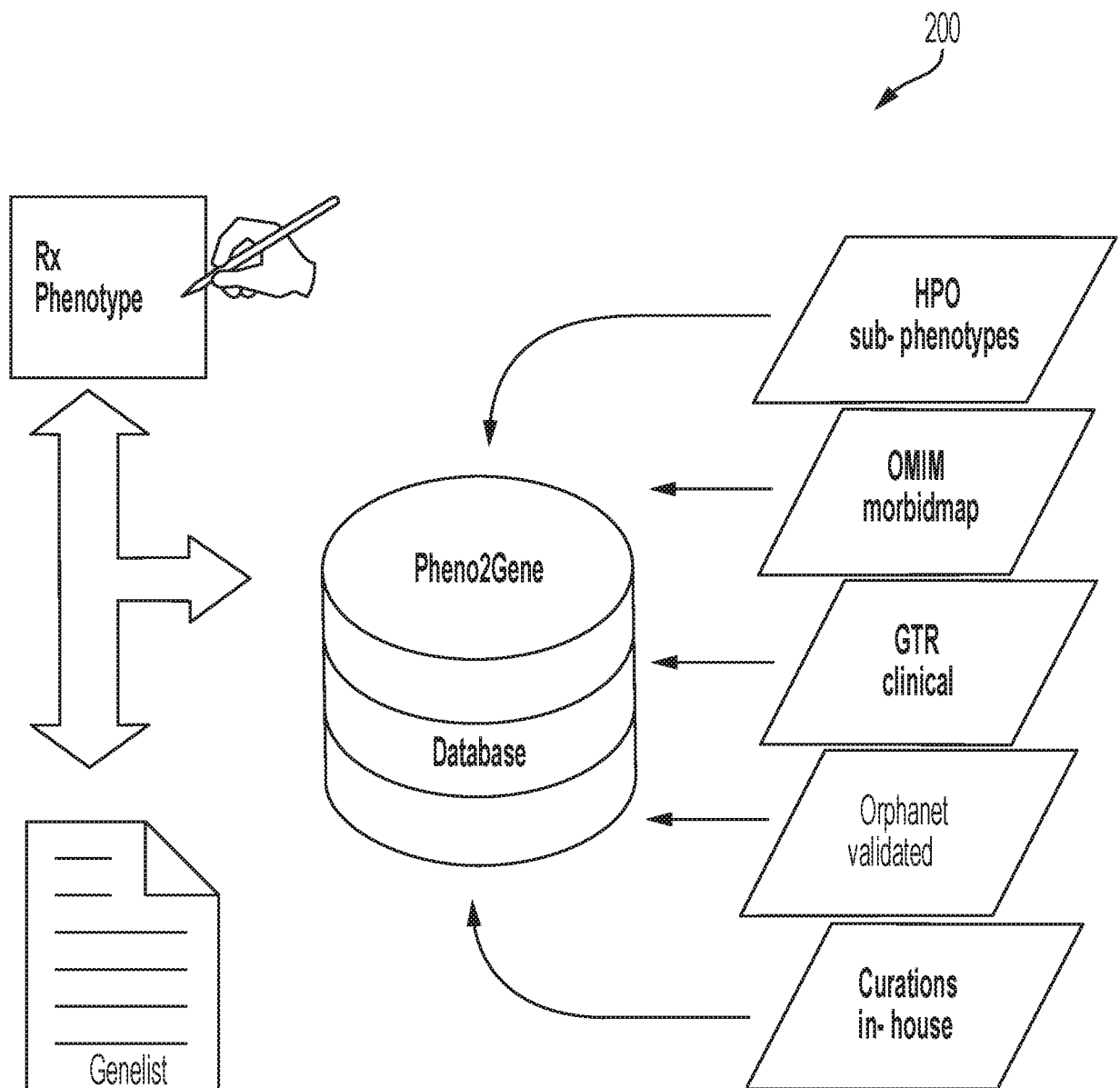


FIG. 2

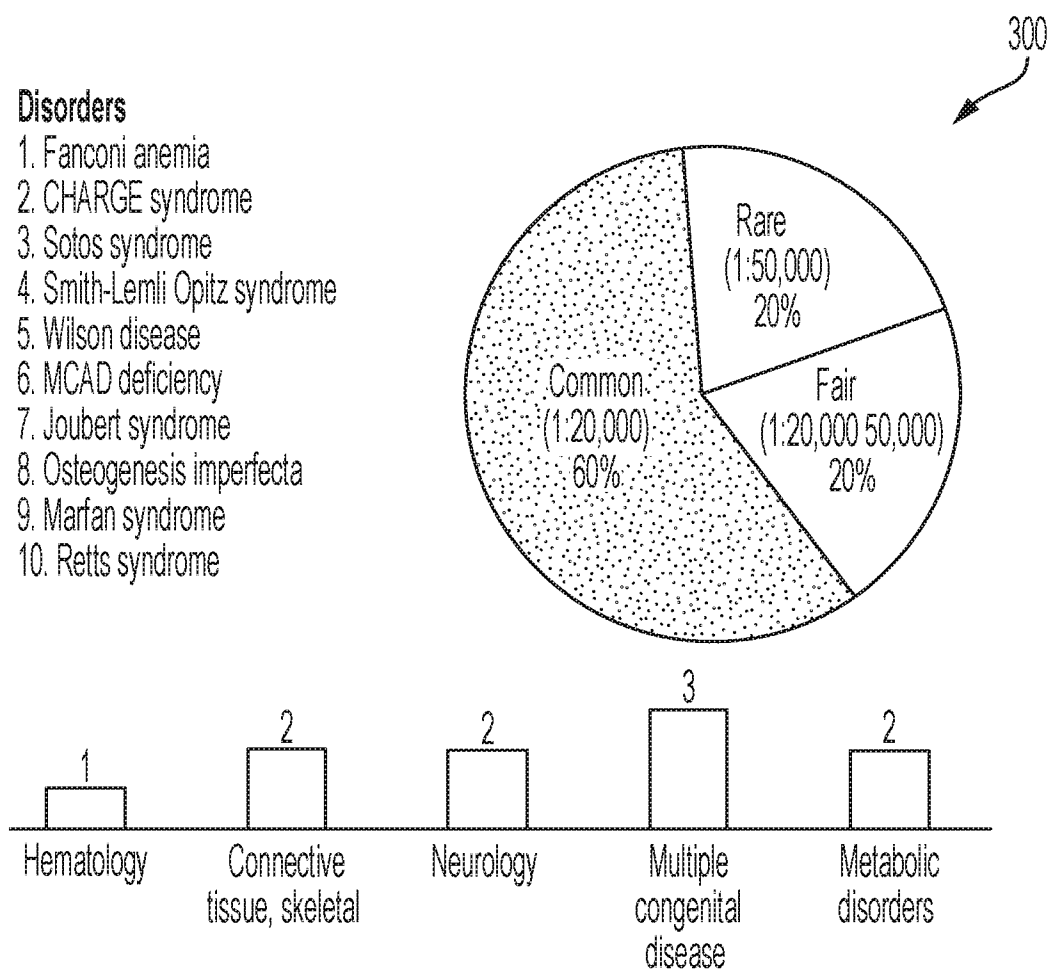


FIG. 3A

Panel	Summary (%)		
Method	Phenomizer	Ingenuity	Pheno2Gene
Accuracy	94.8	94.8	94.8
Sensitivity	88.3	100.0	100.0
Specificity	94.8	99.9	100.0

FIG. 3B

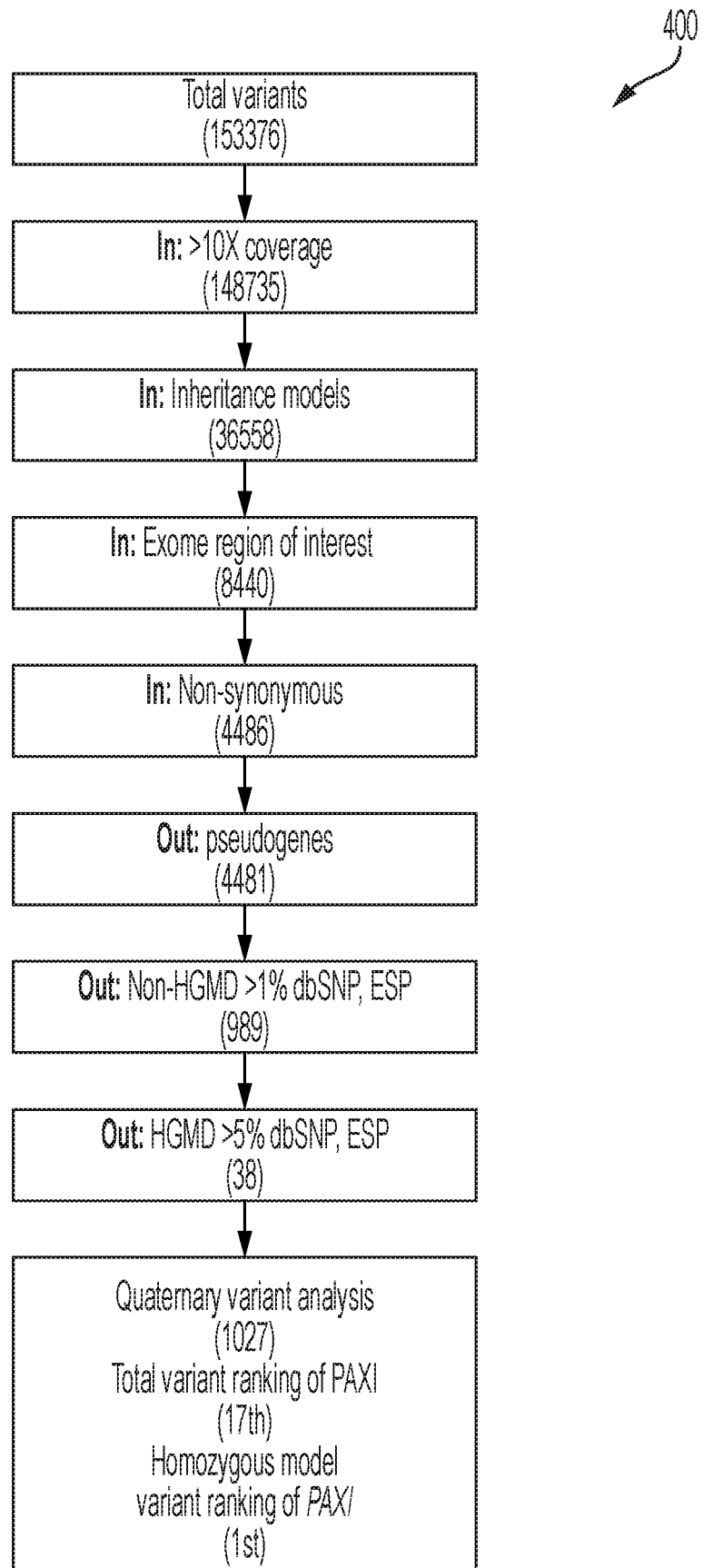


FIG. 4A

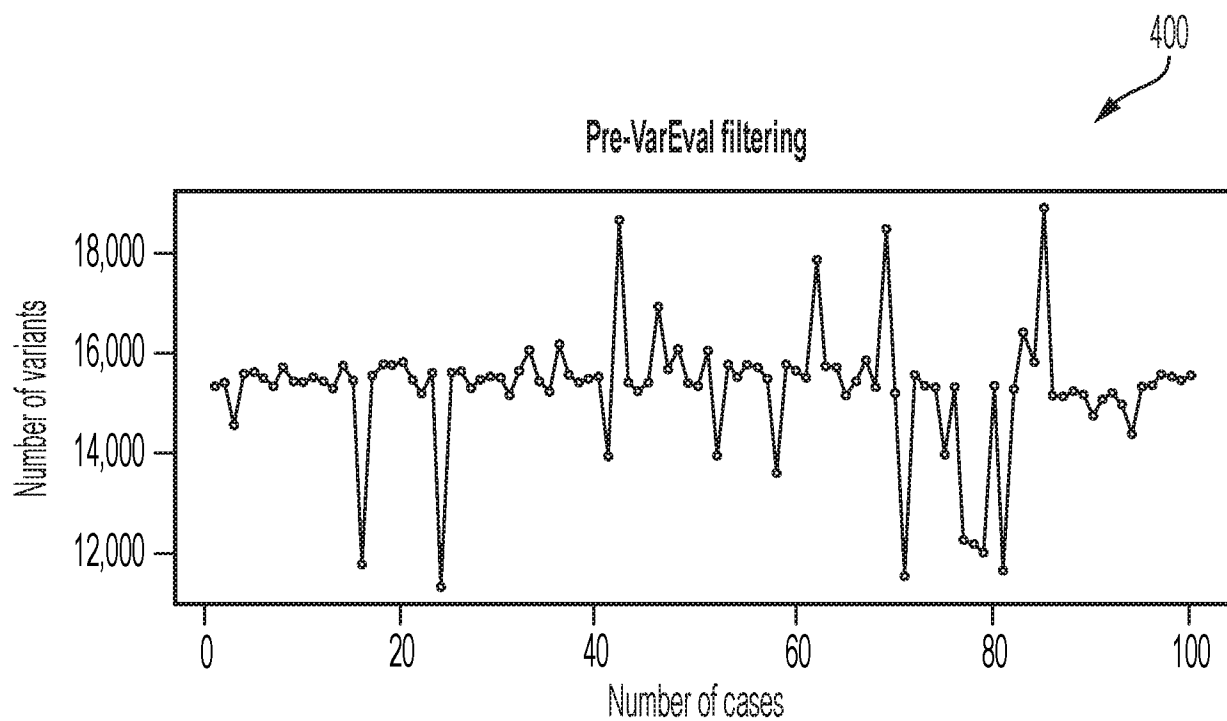


FIG. 4B

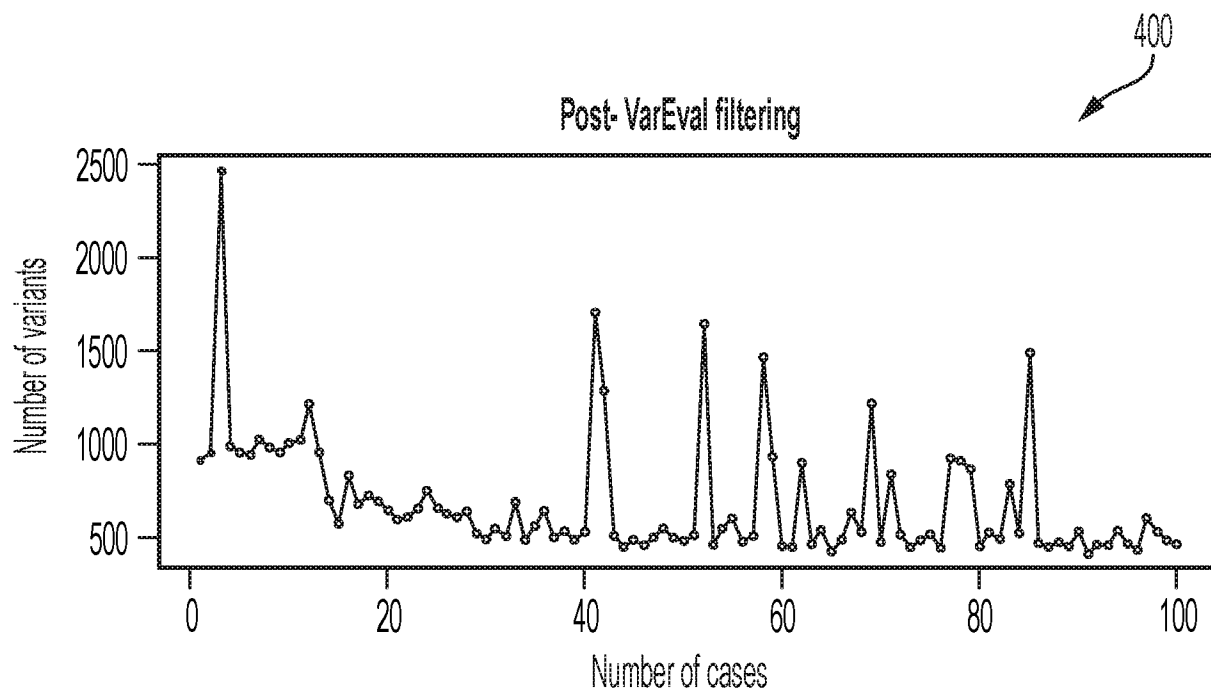


FIG. 4C

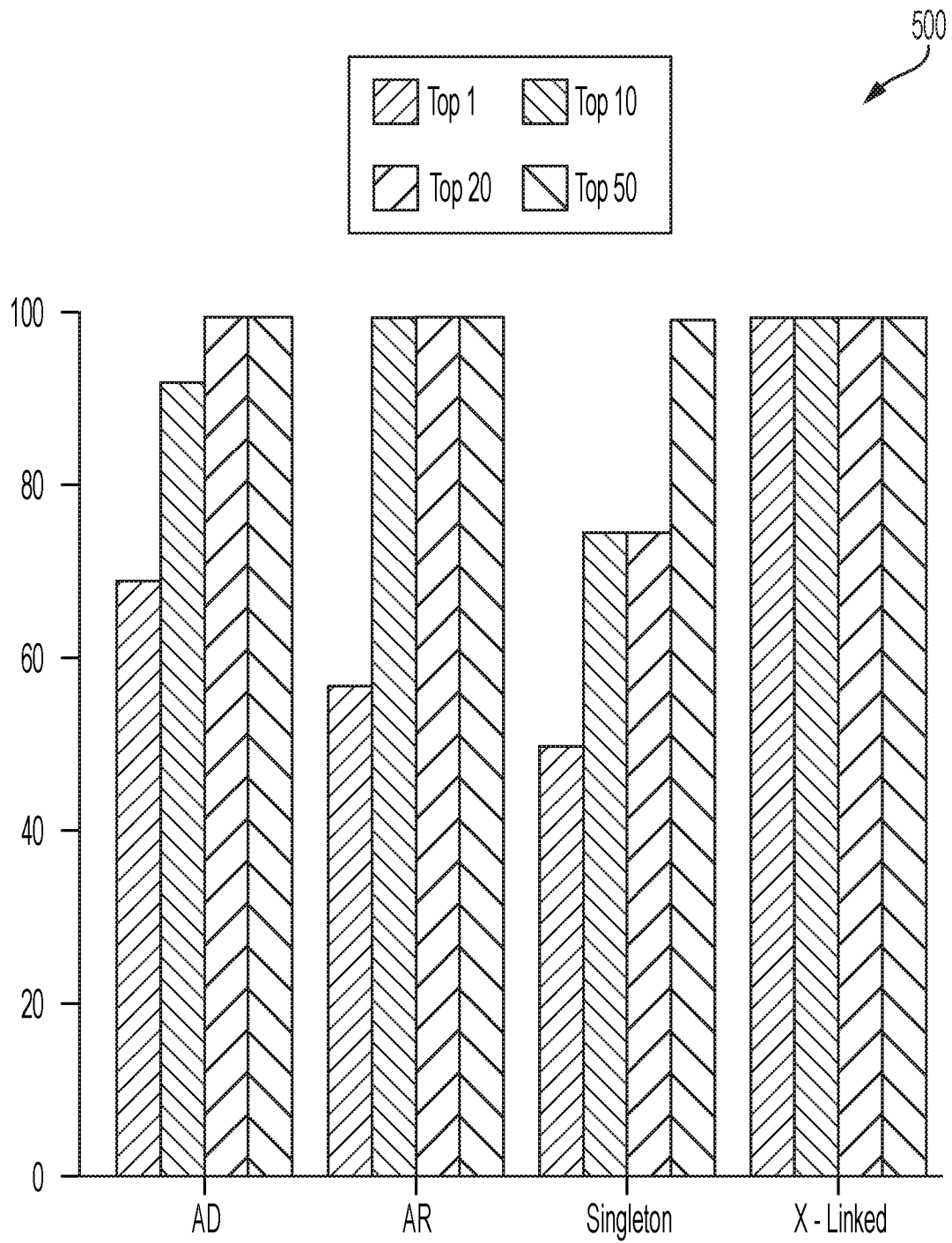


FIG. 5

600

Family ID	Clinically Reported likely pathogenic/ pathogenic Variants				Inheritance Model	Model Ranking	VarEval 2.0 Output			
	Gene	Base Change	Zygosity	Inheritance			Top 50	Top 20	Top 10	Top 1
4	NFKB2	c.2598_2599insT	Het	AD	De novo	1	Y	Y	Y	Y
	TNFRSF13B	c.706G>T	Het	AD	AD	1	Y	Y	Y	Y
7	COL5A1	c.1588G>A	Het	AD	AD	4	Y	Y	Y	N
11	DYRK1A	c.691C>T	Het	AD	De novo	1	Y	Y	Y	Y
	STXBP1	c.1217G>A	Het	AD	De novo	1	Y	Y	Y	Y
16	GJB2	c.416G>A	Het	AR	Unknown	5	Y	Y	Y	N
19	ASXL3	1383_1386dupTA	Het	AD	De novo	4	Y	Y	Y	N
20	INS	c.94G>A	Het	AD	De novo	1	Y	Y	Y	Y
22	DCTN1	c.3146G>A	Het	AD	Dom	11	Y	Y	N	N
25	CPA6	c.799G>A	Het	AD	Dom	7	Y	Y	Y	N
26	GABRB3	c.905A>C	Het	AD	De novo	1	Y	Y	Y	Y
29	PIK3CD	c.1002C>G	Het	AD	De novo	1	Y	Y	Y	Y
38	TUBB4A	c.1172G>A	Het	AD	De novo	1	Y	Y	Y	Y
41	SYNE1	c.6929A>T	Het	AD	Unknown	25	Y	N	N	N
	SYNE1	c.9655G>A	Het	AD	Unknown	26	Y	N	N	N
47	RYR1	c.3389G>A	Het	AD	Unknown	1	Y	Y	Y	Y
	DYNC1H1	c.1013A>T	Het	AD	De novo	1	Y	Y	Y	Y
48	PAX1	c.463_465del	Homo	AR	Hom	1	Y	Y	Y	Y
50	NPC1	c.3019C>G	Het	AR	CoH	1	Y	Y	Y	Y
60	CACNA1C	c.3577G>T	Het	AD	Unknown	46	Y	N	N	N

FIG. 6

600

65	NPC1	c.3450C>A	Het	AR	CoH	2	Y	Y	Y	N
	CTPS1	c.1692-1G>C	Homo	AR	Hom	1	Y	Y	Y	Y
66	NPR2	c.263A>C	Homo	AR	UPD	1	Y	Y	Y	Y
	C2CD3	c.1736T>G	Het	AR	CoH	7	Y	Y	Y	N
68	C2CD3	593_4596delAA	Het	AR	CoH	6	Y	Y	Y	N
	TTC7A	c.211G>A	Het	AR	CoH	1	Y	Y	Y	Y
70	TTC7A	1013_1014delC	Het	AR	CoH	2	Y	Y	Y	N
	PLA2G6	c.278C>A	Het	AR	CoH	1	Y	Y	Y	Y
74	VHL	c.340+5G>C	Het	AD	Unknown	1	Y	Y	Y	Y
	HBB	c.424C>G	Het	AD	Unknown	2	Y	Y	Y	N
78	CTLA4	c.90_104del115	Het	AD	Unknown	1	Y	Y	Y	Y
	CTLA4	c.151C>T	Het	AD	Unknown	1	Y	Y	Y	Y
81	CTLA4	c.151C>T	Het	AD	Unknown	1	Y	Y	Y	Y
	CTLA4	c.151C>T	Het	AD	Unknown	1	Y	Y	Y	Y
84	STAT1	c.1154C>T	Het	AD	Unknown	3	Y	Y	Y	N
	PLA2G6	c.2070_2072del	Het	AR	CoH	2	Y	Y	Y	N
93	ECHS1	c.266G>T	Het	AR	CoH	1	Y	Y	Y	Y
	ECHS1	c.518C>T	Het	AR	CoH	2	Y	Y	Y	N
95	MEFV	c.586G>T	Homo	AR	Hom	1	Y	Y	Y	Y
	PCDH19	c.3355G>A	Het	XL	XLD	1	Y	Y	Y	Y
105	XIAP	c.664C>T	Het	XL-R	DeN	1	Y	Y	Y	Y

FIG. 6 (continued)

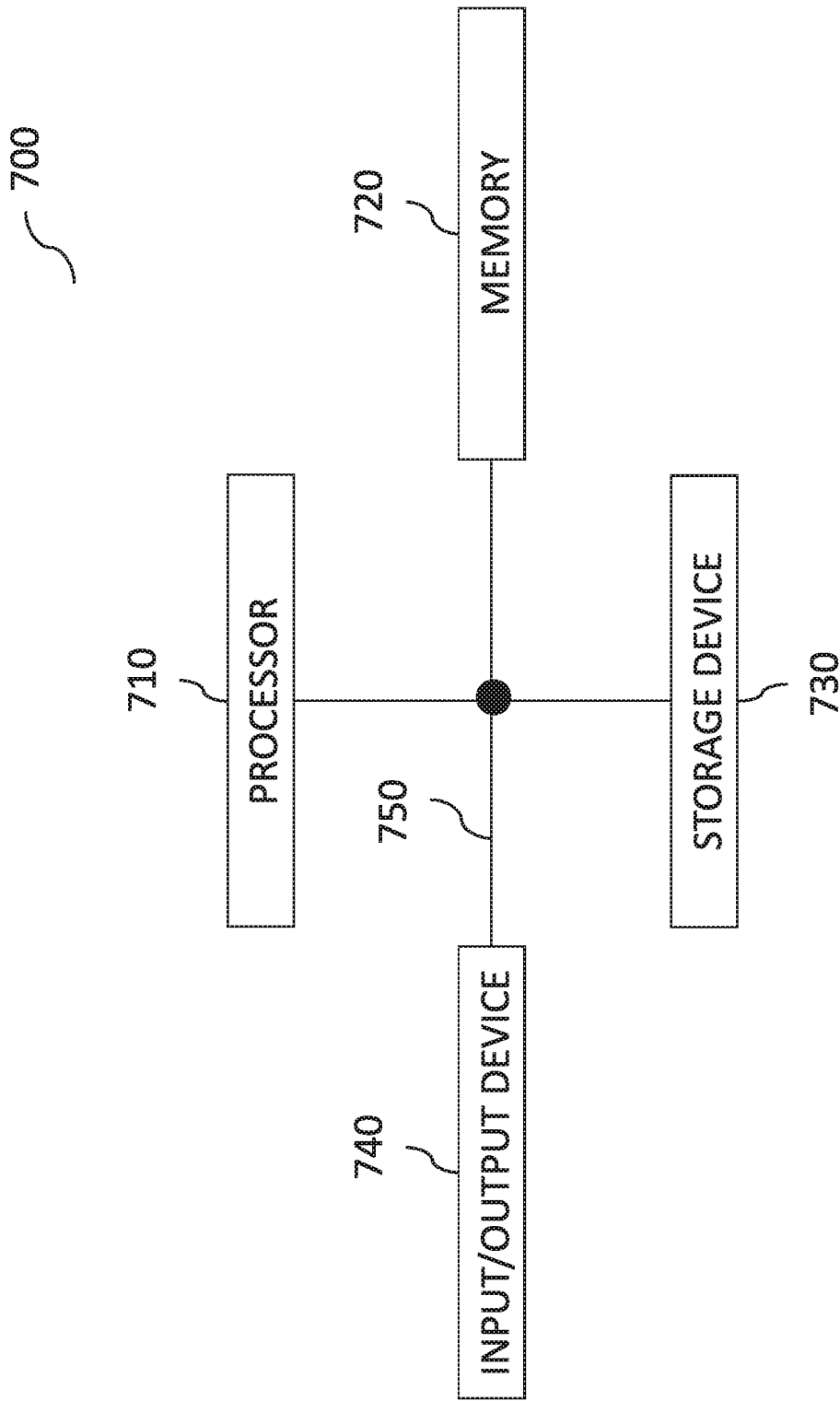


FIG. 7

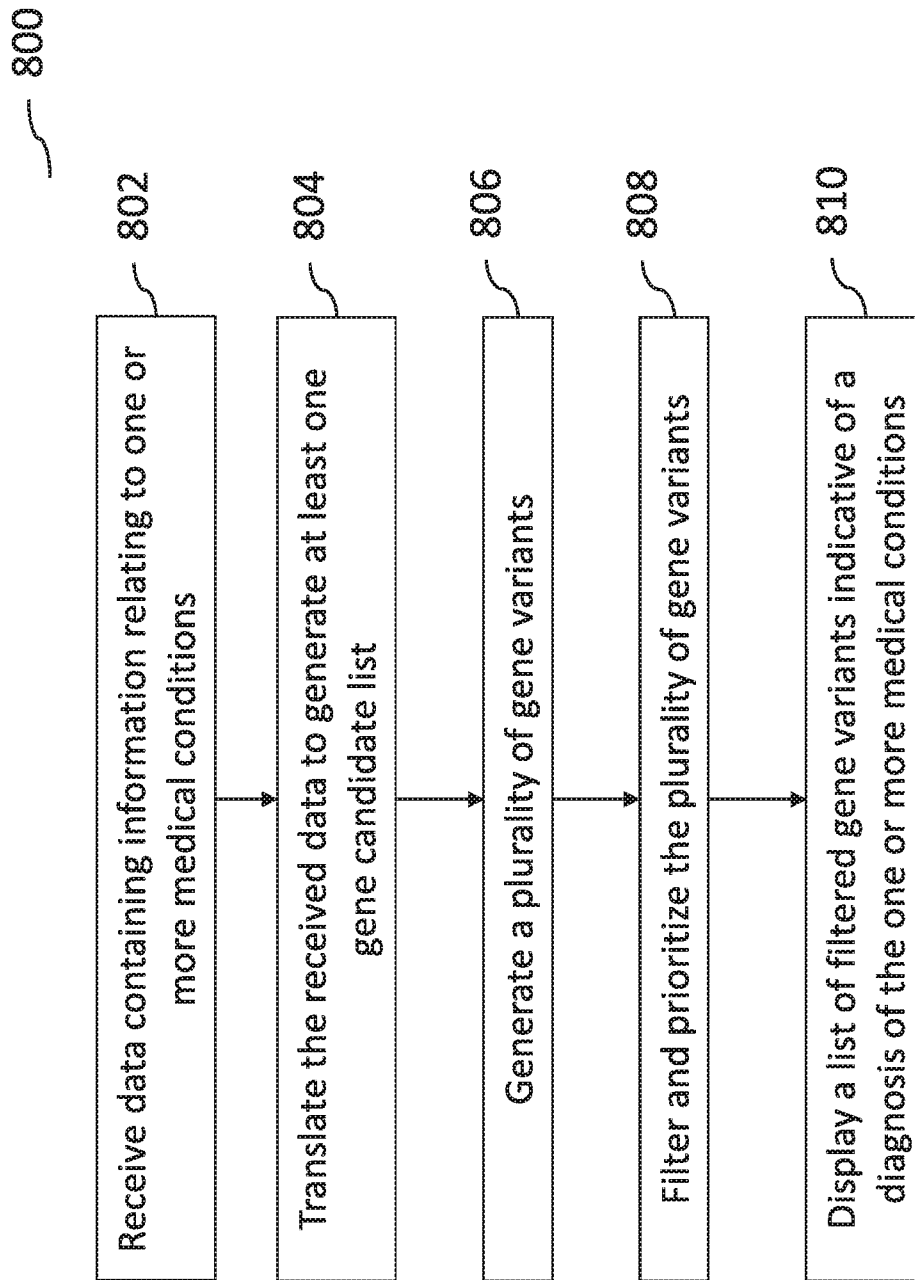


FIG. 8

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2018/066538

A. CLASSIFICATION OF SUBJECT MATTER
INV. G16B20/00
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G16B

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

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

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2015/310163 A1 (KINGSMORE STEPHEN [US] ET AL) 29 October 2015 (2015-10-29) paragraphs [0033] - [0212] -----	1-7
X	US 2016/092631 A1 (YANDELL MARK [US] ET AL) 31 March 2016 (2016-03-31) paragraphs [0079] - [0211] -----	1-7
X	US 2016/283484 A1 (CHANDRATILLAKE GEMMA L [GB] ET AL) 29 September 2016 (2016-09-29) paragraphs [0034] - [0095] -----	1-7
X	US 2016/048634 A1 (TORKAMANI ALI [US]) 18 February 2016 (2016-02-18) paragraphs [0057] - [0279] ----- -/-	1-7



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

29 March 2019

Date of mailing of the international search report

08/04/2019

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Díaz de Lezana, C

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2018/066538

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Regis A James ET AL: "A visual and curatorial approach to clinical variant prioritization and disease gene discovery in genome-wide diagnostics", 2 February 2016 (2016-02-02), XP055428373, DOI: 10.1186/s13073-016-0261-8 Retrieved from the Internet: URL:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4736244/ [retrieved on 2017-11-23] the whole document -----	1-7
X	C. ALEXANDER VALENCIA ET AL: "Clinical Impact and Cost-Effectiveness of Whole Exome Sequencing as a Diagnostic Tool: A Pediatric Center's Experience", FRONTIERS IN PEDIATRICS, vol. 3, 3 August 2015 (2015-08-03), XP055574942, DOI: 10.3389/fped.2015.00067 the whole document -----	1-7
A	KAREN EILBECK ET AL: "Settling the score: variant prioritization and Mendelian disease", NATURE REVIEWS GENETICS, vol. 18, no. 10, 14 August 2017 (2017-08-14), pages 599-612, XP055575548, GB ISSN: 1471-0056, DOI: 10.1038/nrg.2017.52 the whole document -----	1-7

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2018/066538

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2015310163 A1	29-10-2015	CA 2885834 A1 EP 2901345 A2 US 2015310163 A1 WO 2014052909 A2	03-04-2014 05-08-2015 29-10-2015 03-04-2014
US 2016092631 A1	31-03-2016	AU 2015206538 A1 CA 2936107 A1 CN 106575321 A EP 3095054 A1 US 2016092631 A1 US 2019065671 A1 WO 2015109021 A1	14-07-2016 23-07-2015 19-04-2017 23-11-2016 31-03-2016 28-02-2019 23-07-2015
US 2016283484 A1	29-09-2016	GB 2535066 A US 2016283484 A1 WO 2015051275 A1	10-08-2016 29-09-2016 09-04-2015
US 2016048634 A1	18-02-2016	CA 2942811 A1 US 2016048634 A1 WO 2014149972 A1	25-09-2014 18-02-2016 25-09-2014