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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
4 August 2016 (04.08.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/123322 A1

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

(21) International Application Number:
PCT/US2016/015333

(22) International Filing Date:
28 January 2016 (28.01.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/110,022 30 January 2015 (30.01.2015) US

(71) Applicant: **BAYLOR COLLEGE OF MEDICINE** [US/US]; One Baylor Plaza, BCMD 600D, Houston, TX 77030 (US).

(72) Inventors: **ZANEVELD, Jacques, Eric**; 2300 Old Spanish Trail #1125, Houston, TX 77054 (US). **SALVO, Jason, Scott**; 2111 Holly Hall Street, Apartment 2203, Houston, TX 77030 (US). **CHEN, Rui**; 2715 E. Bainbridge Circle, Pearland, TX 77584 (US).

(74) Agents: **WAKIMURA, Mary, Lou** et al.; Hamilton, Brook, Smith & Reynolds, P.C., 530 Virginia Rd, P.O. Box 9133, Concord, MA 01742-9133 (US).

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

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK,

[Continued on next page]

(54) Title: EXPERT PROGRAM AUTOMATES IDENTIFICATION OF HUMAN DISEASE CAUSING MUTATIONS LEVERAGING LARGE PATIENT DATA SETS

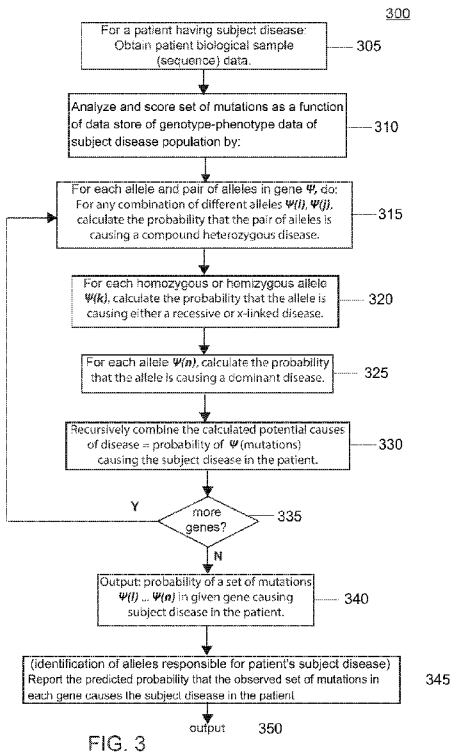

FIG. 3

(57) Abstract: Next-Generation Sequencing (NGS) based genetic diagnosis has immense potential to improve medical practice. Two major remaining obstacles to the widespread clinical use of NGS are the cost of supporting a team of geneticists and concerns over reproducibility of results. To address these concerns, a computer-based method and system is presented that automates the process of performing a genetic diagnosis by leveraging sequencing and phenotype data mined from thousands of patients. The resulting EXPERT system lists the probability that every mutant gene causes disease. Taking the highest scoring gene generates a genetic diagnosis identical to that of human experts 90% of the time, with differences occurring largely in unclear cases. In addition, the probability assigned to a gene by the EXPERT system is closely correlated with the subjective "confidence" of a prediction assigned by human experts examining the same patient. This system thus automates the genetic diagnosis process, and has the potential to allow for fast, simple, and reliable genetic diagnosis.

WO 2016/123322 A1

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:

— *of inventorship (Rule 4.17(iv))*

Published:

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

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

EXPERT PROGRAM AUTOMATES IDENTIFICATION OF HUMAN DISEASE CAUSING MUTATIONS LEVERAGING LARGE PATIENT DATA SETS

RELATED APPLICATION

[0001] This application claims the benefit of U.S. Provisional Application No. 62/110,022, filed on January 30, 2015. The entire teachings of the above application are incorporated herein by reference.

GOVERNMENT SUPPORT

[0002] This invention was made with government support under Grant Numbers: R01EY022356, R01EY018571 and EY07102 from National Institutes of Health (NIH) and Neuroscience Education Institute (NEI) respectively. The government has certain rights in the invention.

BACKGROUND OF THE INVENTION

[0003] Molecular diagnosis, or identification of the alleles responsible for human disease, is of growing clinical importance. Personalized medicine techniques, including gene therapy, require a molecular diagnosis of the patient before they can be used. It has been demonstrated in many studies that Next Generation Sequencing (NGS) based molecular diagnosis is much more accurate than other methods like genotyping arrays. However, despite the growing demand for NGS based molecular diagnosis, it remains expensive and is only performed in a small number of clinical labs. This is in part because identifying the mutations causing human disease is a time consuming task than can currently only be performed by an expert with extensive training. These experts rely on their experience to identify hypermutable genes, how frequently each gene causes the disease in question and how frequently each disease has any given inheritance pattern. In addition, these experts must know how to estimate the probability that a novel allele is pathogenic based on how severe the allele is, the gene in which it occurs, and how frequently alleles that severe occur by chance in that particular gene. Learning all of these factors for any given disease is a time intensive task that remains very costly.

[0004] If genetic diagnosis is to flourish in the clinic, it is important to minimize the cost of diagnosis while maximizing the accuracy and speed of obtaining results. Currently, one of the largest costs associated with molecular diagnosis is the support of a team of professionals with the ability to make molecular diagnosis decisions. This makes molecular diagnosis prohibitively expensive for small institutions and will likely be the dominant cost associated with molecular diagnosis as sequencing prices continue to fall. Further, routine molecular diagnosis takes time away from experts who could otherwise concentrate on complicated and interesting cases.

SUMMARY OF THE INVENTION

[0005] The present invention addresses the foregoing problems in the art. In particular, Applicants present a computer based method and apparatus (referred to herein as "EXPERT system") that automatically identifies disease causing mutations in a patient cohort, providing an estimate of the probability that any given set of mutations will cause disease. Applicants' method and apparatus take into account all of the parameters mentioned above leveraging data from a massive cohort of molecularly diagnosed human patients, allowing for fast, reliable, and inexpensive molecular diagnosis of human disease.

[0006] A molecular diagnosis method and apparatus embodying the present invention utilize a data store of genotype-phenotype data of the population with and without various diseases. A digital processor receives sequence data of a patient that has the subject disease. The processor determines the probability that the subject disease in the patient is caused by any combination of two recessive heterozygous alleles observed in the sequencing data in a given gene (the first potential cause). The processor determines the probability that the subject (patient) has an X-linked or recessive disease caused by a homozygous or hemizygous allele that was observed in the sequencing data (the second potential cause). The processor determines the probability that the subject has a dominant disease due to an allele observed in the sequencing data (the third potential cause). Next the processor recursively combines all instances of the determined first, second and third potential causes. In particular, this combining calculates the probability of a set of mutations in a given gene being the cause of the subject disease in the patient.

[0007] As used herein, a set of mutations can include no mutation, one mutation or plural mutations.

[0008] The processor, accessing a memory storing the data store, can use the data from the data store and analyzes and scores the set of mutations. The analyzing and scoring can include:

(i) calculating probability that the subject disease has a particular inheritance mode (is caused by a dominant gene or is caused by a recessive gene in the subject disease population);

(ii) for each gene Ψ with at least one mutation in the subject, calculating the probability that the gene can cause the subject disease with a given inheritance mode; and

(iii) considering gene-specific differences between genes represented in the data store.

[0009] In particular, use P(Ψ causes disease with given inheritance mode) = 1 if the gene is known from the data store data to cause the subject disease with that inheritance model (dominant or recessive). Otherwise assign P(Ψ causes dominant disease) = frequency of a dominant gene causing the subject disease in the data store data, and P(Ψ causes recessive disease) = frequency of a recessive gene causing the subject disease in the data store data.

[0010] From the analysis method and data from the data store, the processor outputs a determination of the probability that each mutant gene in the patient is responsible for the subject disease, i.e., a general identification of the alleles responsible for the subject disease in the patient.

[0011] In at least one embodiment, the subject disease is retinal disease. In other embodiments, the subject disease is a Mendelian disorder. Other diseases are suitable for diagnosing in embodiments of the present invention. Examples may include, but are not limited to, cancer, heart disease, cystic fibrosis, alcoholism, and others.

[0012] In embodiments, the analyzing and scoring considers gene-specific differences including computing probability that any given allele occurring in a gene severely disrupts the function of that gene. To compute this probability, the processor (a) uses the frequency of observing mutations in non-diseased individuals in the data store data to calculate probability of that gene being the cause of the subject disease, and (b) further determines probability of an allele in that gene affecting protein function of the gene.

[0013] A computer-based tool carrying out molecular diagnosis includes, in a digital processor, an input configured to obtain biological sample sequencing data of a patient having a subject disease and a driver configured access a memory storing a data store of

genotype-phenotype data of a population having the subject disease. The computer-based tool further includes a detector configured to use the data from the data store to identify the probability of each set of alleles in a given gene being responsible for the subject disease in the patient by:

[0014] determining a first potential cause by calculating the probability that the subject disease in the patient is caused by any combination of two recessive heterozygous alleles observed in the sequencing data in a given gene;

[0015] determining a second potential cause by calculating the probability that the patient has an X-linked or recessive disease caused by a homozygous or hemizygous allele that was observed in the sequencing data;

[0016] determining a third potential cause by calculating the probability that the patient has a dominant disease due to an allele observed in the sequencing data; and

[0017] recursively combining all instances of the determined first, second and third potential causes for each gene, and outputting the results of said combining, the results indicating the probability a set of mutations in a given gene causes the subject disease in the patient.

BRIEF DESCRIPTION OF THE DRAWINGS

[0018] The foregoing will be apparent from the following more particular description of example embodiments of the invention, as illustrated in the accompanying drawings in which like reference characters refer to the same parts throughout the different views. The drawings are not necessarily to scale, emphasis instead being placed upon illustrating embodiments of the present invention.

[0019] Fig. 1 is a graph illustrating EXPERT system P-values categorized by human diagnosis confidence levels.

[0020] Fig. 2 is a graph illustrating EXPERT score distributions for RP (Retinitis Pigmentosa) patients and controls.

[0021] Fig. 3 is a flow diagram of a computer tool and/or computer process 300 embodying the present invention.

[0022] Fig. 4 is a schematic view of a computer network environment in which embodiments of the present invention are deployed.

[0023] Fig. 5 is a block diagram of a computer node in the network of Fig. 4 embodying the present invention.

DETAILED DESCRIPTION OF THE INVENTION

[0024] A description of example embodiments of the invention follows.

[0025] Embodiments related to a computer based method and apparatus (referred to herein as "EXPERT system") that automatically identifies disease causing mutations in a patient cohort, providing an estimate of the probability that any given set of mutations will cause disease.

EXPERT system process and computer program

[0026] Consider a group of n alleles in gene Ψ that may or may not cause disease y. Let $\Psi(i)$ be the ith allele in gene Ψ in the patient. There are three possible causes of disease considered by embodiments of the present approach. The first is that a combination of two recessive heterozygous alleles in gene Ψ causes recessive disease in the patient. For any combination of different alleles $\Psi(i)$, $\Psi(j) \mid i \neq j$; calculate the probability that the set of alleles is causing compound heterozygous disease using the following formula:

$$P[\Psi(i), \Psi(j) \text{ causes compound heterozygous disease}] =$$
$$P(\text{Disease is recessive}) * P(\text{gene } \Psi \text{ causes disease } y \text{ with recessive inheritance}) * P[\Psi(i) \text{ disrupts gene } \Psi\text{'s function}] * P[\Psi(j) \text{ disrupts gene } \Psi\text{'s function}]$$

[0027] Simply stated, the probability that the two alleles cause disease equals the probability that the patient's disease has recessive inheritance multiplied by the probability that the gene in which the alleles occur causes the recessive patient disease multiplied by the probability that the alleles are damaging to the protein in which they occur. The second possible cause of disease considered by embodiments is that a homozygous or hemizygous allele causes either recessive or X-linked disease. For each homozygous or hemizygous allele $\Psi(k)$, calculate the probability of this possible cause as follows:

$$P(\text{homozygous allele } \Psi(k)) \text{ causes recessive disease} =$$
$$P(\text{Disease is recessive}) * P(\text{gene } \Psi \text{ causes disease } y \text{ with recessive inheritance}) * P[\Psi(k) \text{ disrupts gene } \Psi\text{'s function}]$$

[0028] The third possible cause of disease considered by embodiments is that an allele causes dominant disease. For each allele, the probability of this value is calculated as follows:

$$P[\text{allele } \Psi(l) \text{ causes dominant disease}] =$$
$$P(\text{Disease is dominant}) * P(\text{gene } \Psi \text{ causes disease } y \text{ with dominant inheritance}) * P[\Psi(l) \text{ disrupts gene } \Psi\text{'s function}]$$

[0029] This set of formulas can provide the probability that each potential cause in isolation is responsible for disease. However, if there is more than one potential cause (i.e., there are two potential dominant alleles, or three alleles for a recessive gene giving three potential compound heterozygous combinations of alleles possibly leading to disease), they cannot be simply added. For example, consider two dominant alleles each with a probability to cause disease of 0.6. Summing the two values yields a probability of 1.2, which is clearly false. Instead, consider the fact that the second potential cause can only be correct if the previous potential cause is NOT true. Thus, the correct probability that Ψ causes disease in this case is 0.6+(1-0.6)*(0.6) =0.84.

[0030] To generalize this concept, let $P(\Psi$ causes disease $|\ y$ potential causes) be the probability gene Ψ causes disease after considering the first y potential causes. Then one can combine all of the possible causes recursively as follows:

Initiation step: $P(\Psi$ causes disease$|1^{st}$ possible cause) = $P(1^{st}$ possible cause)

Recursive step: $P(\Psi$ causes disease$|\ x$ candidates)= $P(\Psi$ causes disease$|\ x$-1 candidates) + {[1-$P(\Psi$ causes disease$|\ x$-1 candidates)] *$P(x^{th}$ candidate)}

[0031] By enumerating all alleles and pairs of alleles and applying these formulas, we can thus calculate the probability a set of mutations $\Psi(1)$ ... $\Psi(n)$ in a particular gene Ψ causes disease.

[0032] For this approach to work, there are three values to obtain. First, obtain the probability that the patient's disease has a particular inheritance mode. Second, obtain the probability that any given gene can cause the patient's disease with a given inheritance mode. Third, obtain the probability that any given allele severely disrupts the function of the gene in which it occurs. There are many possible ways to estimate these three values (probabilities) that may be considered. Embodiments use a data driven approach, deriving estimates of all three values using genotype-phenotype data from a population of >2,000 retinal disease patients and >2,000 controls.

[0033] Specifically, if the user specifies that the inheritance mode is unknown, embodiments can calculate the probability of a patient having each inheritance model (e.g., dominant or recessive) based off of the observed frequency of each inheritance mode in a large patient cohort. When estimating the probability that a gene can cause a disease with a particular inheritance model, embodiments can set the value to 1 if the gene is known to cause the disease with that inheritance model. However, due to the genetic and phenotypic

heterogeneity of disease, patients frequently are found to have a molecular cause of disease only previously associated with a related disorder. For example, retinal disease patients with a clinical diagnosis of recessive Leber's Congenital Amaurosis (LCA) are often found to have disease causing mutations in a gene only previously reported to cause recessive RP. Embodiments can use the observed frequency of these discrepancies to calculate the probability that a gene not reported to cause the disease could in fact be the cause. In this case, 4% of LCA patients in an example training cohort were found to have disease caused by mutations in a gene only previously reported to cause recessive RP, thus an example embodiment can set P(any known recessive RP gene causes LCA)=0.04.

[0034] The final remaining value, the probability that an allele affects protein function, has been subject of extensive research with many programs (including Sift, Polyphen, and Evolutionary Trace) already existing that give sufficient predictions. Embodiments utilize one such program, CADD, as the backbone for the calculation. One difficulty with these programs, however, is that certain genes frequently have mutations that are ranked as extremely detrimental, even in controls. In other genes, even a mild change may be significant as variants are almost never observed in controls.

[0035] These gene-specific differences can be due to different mutation rates and selective pressures acting on genes, as well as the presence of pseudo genes and technical errors. To capture these gene-specific differences, embodiments according to the present approach compare the score of each tested mutation with the scores of variants from the same gene found in healthy individuals from the thousand genomes project. This data can be used to calculate the probability that a prediction as severe as the one in question would be observed by chance in a control individual. Embodiments then multiply the probability that the allele in question would not be found by chance with the percentile score of the allele from CADD to derive the probability that the allele is severely affecting protein function in a particular gene. This can be expressed as follows:

$$P[\Psi(m) \text{ affects gene function}] = [\Psi(m) \text{ CADD score}] * [1 - (\text{fraction of controls with a mutation in } \Psi \text{ scoring} \geq \Psi(m) \text{ CADD score})]$$

[0036] For example, if a mutation is in the top 5% of CADD scores and an allele as severe is only observed in 15% of controls then one would obtain: P(Allele affects gene function)=(1-0.15)*(0.95)= 0.8075. By contrast, even if a mutation is in the top 5% of CADD scores, if mutations as severe as it are observed in 90% of controls it is given a

probability score of 0.095. This step accounts for hypermutable genes, allowing only mutations more severe than those regularly observed in controls to be given high scores. As a final filtering step, variants may not considered as candidate dominant mutations if they have a frequency in any control cohort of >1/1,000. This threshold was used in embodiments of the present approach because dominant eye disease alleles are very rare in the population due to high selective pressure against them in human history. The threshold can easily be set as desired by the user for a particular disease of interest.

Sequencing

[0037] In an embodiment, the EXPERT system was run on sequencing data generated in the Human Genome Sequencing Center (HGSC) at Baylor College of Medicine. The sequencing approach has been described previously. Briefly, samples were prepared using standard techniques and captured using a custom designed Agilent Sureselect array. Reads were aligned using BWA. Recalibration and realignment was performed via GATK. Atlas SNP and Atlas indel were used to call variants. Variants were annotated using ANNOVAR. Variants were annotated with control frequency scores using the 1,000 genomes, CHARGE, ESP6500, and an internal database in the HGSC.

Results

[0038] Patients with ocular disease were diagnosed in parallel by a human geneticist specializing in the genetic diagnosis of ocular disease (hereafter "human") and the EXPERT system, with the goal of identifying the genes and alleles causing disease. The vast majority (71/79 or 90%) of patients received the same or a similar call (defined as only the EXPERT system calling variants given a probability <0.05 when the human called that there were no potentially pathogenic variants, or the EXPERT system recognizing the human top call as the 2nd place most likely cause of disease with a score within 0.1 of the 1st place score) from the EXPERT system and a human. The differences between the calls made by the EXPERT system and a human are often due to difficult or unclear cases rather than a clear mistake on the part of either party. For example, one can consider results from one group of eye disease patients with Retinitis Pigmentosa (RP).

[0039] Table 1 shows data for 24 RP patients. These patients all had either autosomal recessive (ar), autosomal dominant (ad) or X-linked (X) RP. At the time each patient was first seen, the particular variant of RP may not have been known. Thus, some patients are identified as "RP" in Table 1. The "Top call" is the highest scoring gene, while the 2nd place

call is the second highest scoring gene for a patient. Patients receiving different top calls from the human are displayed in bold. Each potential cause of disease is matched with a score, or in the case of the human, a subjective confidence level.

[0040] As shown in Table 1, in this cohort of 24 RP patients with varying inheritance patterns, the top call made by the EXPERT system and the top call made by a human were identical in 20 cases. In four cases, discrepancies were seen between the EXPERT system diagnosis and the human diagnosis (see rows in bold text in Table 1). One may consider each discrepancy individually. The first discrepancy was observed in a patient with adRP. In this case, the EXPERT system gave a similar score to two genes, *PRPF31* and *IMPDH1* indicating that both are potential candidates. The human correctly identified that the gene *IMPDH1* is caused by dominant negative rather than haploinsufficient mutations, and is thus unlikely to be caused by the patient's nonsense mutation. This information was not considered by the EXPERT system causing it to inaccurately give the *IMPDH1* mutation a high score. However, because the EXPERT system also gave the *PRPF31* mutation a high score, further testing could have quickly caused the correct diagnosis to be made.

[0041] In the second case, the EXPERT system identified *PRPF6* as a potential cause of disease in an RP patient while the human listed that there was no potential cause of disease observed. Upon additional consultation, it appeared that the human considers this mutation possible but very unlikely due to the relatively high frequency of the mutation in control cohorts. The third difference occurred when the EXPERT system called a homozygous *USH2A* mutation as the likely cause of disease in another RP patient. The human had overlooked the fact that this mutation was homozygous and failed to call it as a candidate. After review, the human agreed with the system's prediction and has now updated the likely cause of disease in this patient to be *USH2A* in agreement with the EXPERT system prediction. This demonstrates the utility of the EXPERT system in reviewing and assisting in patient diagnosis. In the final difference, the EXPERT system identified compound heterozygous *RP1L1* mutations as a possible cause of disease in an RP patient, although the probability it gave this gene of causing disease was extremely low (less than a 2%). The human originally did not consider the mutation a possible cause of disease, but upon consultation could not rule it out, stating it was a very unlikely cause of disease. This agrees with the prediction made by the EXPERT system.

[0042]

Patient disease	EXPERT top call	Top call score	EXPERT 2nd place call	2nd place call score	Human call	Confidence
RP	*USH2A*	0.969	*RPGR*	0.470	*USH2A*	High
RP	*CNGB1*	0.759	*RIMS1*	0.061	*CNGB1*	High
XRP	*RPGR*	0.241	None		*RPGR*	High
adRP	*EYS*	0.190	*HMNC1*	0.020	*EYS*	High
adRP	*PRPF31*	0.980	*RP1L1*	0.018	*PRPF31*	High
adRP	*PRPF3*	0.978	*HCMN1*	0.020	*PRPF3*	High
RP	*ABCA4*	0.366	None		*ABCA4*	High
arRP	*RP1*	0.020	None		*RP1*	High
adRP	***IMPDH1***	**0.965**	***PRPF31***	**0.882**	***PRPF31***	**High**
XRP	*PRPF31*	0.020	None		*PRPF31*	High
arRP	*WDR19*	0.946	None		*WDR19*	High
arRP	*RP1*	0.978	None		*RP1*	High
arRP	*CDHR1*	0.978	None		*CDHR1*	High
arRP	*EYS*	0.980	*CNGA1*	0.909	*EYS*	High
adRP	*IMPG1*	0.019	None		*IMPG1*	Low
RP	***PRPF6***	**0.404**	***SNRNP200***	**0.240**	**None**	
RP	None		None		None	
RP	***USH2A***	**0.295**	**None**		**None**	
RP	None		None		None	
RP	None		None		None	
adRP	None		None		None	
arRP	***RP1L1***	**0.018**	**None**		**None**	
arRP	None		None		None	
arRP	None		None		None	

Table 1: Comparison of human and EXPERT system molecular diagnosis

[0043] It is also interesting to consider whether the confidence levels of the human are related to the probability scores obtained from the EXPERT system. To test this hypothesis, the P-values of the top EXPERT system calls were placed into three categories based on the human confidence level, which was either high, low, or no mutation found (Fig. 1). Strikingly, the EXPERT system assigned probability for human "high confidence" calls was on average three fold higher than that of the human "low confidence" calls, and low confidence calls had more than double the average score of patients where the human could not identify a causative mutation. The number of patients receiving a "low confidence" call in this study was low (n=6) so the differences between the "low confidence" and "no mutation found" categories was not significant, though all other differences were highly significant ($P<10^{-4}$). As shown in Fig. 1, the average EXPERT system P-values of the top gene in patients receiving high confidence, low confidence, or no call from a human were 0.60, 0.16,

and 0.07, respectively. The error bars represent 1 standard deviation. The figure illustrates that EXPERT system P-values correlate with human diagnosis confidence levels.

[0044] One possible critique of this approach would be if control individuals also received a molecular diagnosis, as this would indicate the presence of false positive calls. To determine if the pattern of scores in patients were different from controls, the EXPERT system was run on 500 control individuals from the CHARGE cohort, asking the program to treat them as RP patients and come to a diagnosis. These results were then compared with those from 63 RP patients participating in this study. For each patient and control, the top scoring gene was placed in a score bin, with increments of 0.1.

[0045] A plot of the distribution of scores for RP patients and controls is shown in Fig. 2. The percentage of the cohort in each bin is displayed using a logarithmic scale on the Y axis. As expected, controls are highly enriched for low scores while patients are highly enriched for high scores. This indicates that the EXPERT system can correctly differentiate true positive signals from background mutations. The differences between the cohorts were statistically significant in a student's T-test ($P<10^{-7}$). As expected, the majority of controls (374, or 75%) had a score in the bottom bin (0-0.1) while only 17 (3%) had a score >0.5. By contrast, one third of RP patients had a score in the bottom bin while over 29% had a score >0.5. This indicates a clear separation of scores, with RP patients almost exclusively containing high scores and controls being enriched for low scores. Further, the average P-value for a gene in the control cohort was 0.08, while the average P-value for a gene in RP patients was more than fourfold higher at 0.32. This demonstrates that the majority of calls made by the EXPERT system are correct and that the approach described here has minimal false positives.

[0046] To further compare the EXPERT system with the results obtained by experts, the sequencing data for all controls with a P-value >0.5 on the EXPERT system were presented to a geneticist who specializes in molecularly diagnosing RP patients. When lied to, and told these controls actually had RP, the geneticist made the same diagnosis as the EXPERT system in 12/17 cases (Table 2), and had high confidence for 11/12 of these predictions. Thus, the majority of false positive calls made by the EXPERT system are due to the limitations of current knowledge, rather than a mistake made by the EXPERT system.

[0047]

Phenotype	EXPERT Call	Probability	Human Call	Confidence
Control	*PROM1*	0.975	*PROM1*	High confidence
Control	*RPGR*	0.880	*RPGR*	High confidence
Control	*NRL*	0.815	*NRL*	High confidence
Control	*MERTK*	0.731	*MERTK*	High confidence
Control	*LRAT*	0.726	No candidate	
Control	*WDR19*	0.690	*WDR19*	High confidence
Control	*TULP1*	0.690	*TULP1*	Low confidence
Control	*USH2A*	0.689	*USH2A*	High confidence
Control	*RPGR*	0.671	No candidate	
Control	*GPR15*	0.660	*GPR15*	High confidence
Control	*ABCA4*	0.625	*ABCA4*	High confidence
Control	*C8orf37*	0.569	No candidate	
Control	*C2orf71*	0.539	*C2orf71*	High confidence
Control	*C2orf71*	0.539	*C2orf71*	High confidence
Control	*RPGR*	0.514	No candidate	
Control	*RPGR*	0.514	*RPGR*	High confidence
Control	*RPGR*	0.514	No candidate	

Table 2: Comparison of controls analyzed by the EXPERT system and a geneticist.

[0048] A small (3%) number of controls were found to have mutations that could probably cause RP when the EXPERT system was instructed to consider the controls to be RP patients. In Table 2, the second and third columns ("EXPERT Call" and "Probability") give the top ranked gene and score, respectively, when analyzed by the EXPERT system. The last two columns ("Human Call" and "Confidence") give the top ranked gene and score, respectively, when analyzed by a geneticist specializing in RP who was incorrectly told these were RP patients. A majority of the calls made by the EXPERT system were also made by the geneticist, especially for genes given a very high score by the EXPERT system.

Discussion

[0049] Two of the main barriers to the adaptation of widespread, NGS based clinical genetic testing are the costs associated with professionals that perform diagnosis and concerns about the reliability and reproducibility of results. An automated program, such as the EXPERT system presented here, has the potential to address both of these problems by providing cheap, accessible and reliable genetic diagnoses. When tested, the EXPERT system appears to make identical calls to those of a human at least 90% of the time. Most discrepancies appear to occur in borderline cases, and on occasion the EXPERT system can even make correct predictions missed by a human.

[0050] Predictions made by humans are often subjective judgment calls. This makes generating standardized results difficult, as what one expert may call "high confidence" another may call "low confidence." Further, while humans can have incredible insights and typically perform very well, they can also be prone to missing details and lapses in judgment, especially when overworked. When health is on the line, it is important to minimize or prevent these types of mistakes. By contrast, the automated EXPERT system is not subject to these types of human errors, and will provide consistent results. In addition, the P-values associated with EXPERT system predictions are based on a well-defined underlying theory.

[0051] In some embodiments, as additional data becomes available, the EXPERT system improves in accuracy. For example, one implementation of the present approach only uses information about what diseases a gene is known to cause when calculating the probability that that gene could cause a different disease. This approach is not ideal because each gene may have a different probability of causing different diseases. As more data becomes available, more reliable frequencies for every individual gene can be had. In addition, as more data becomes available each gene can be considered separately.

[0052] Extensive testing of embodiments of the present approach shows that the P-values the EXPERT system generates closely correlate with the subjective confidence levels listed by humans. Further, it shows the clear ability to differentiate patient and control cohorts. The present approach is applicable to any Mendelian disease if sufficient datasets are available. The accuracy of this approach will likely improve as the size of data sets increase. Indeed, it seems possible in the near term to make a program with better performance than human experts because the program can consider more data than a human can easily review for every case. Regardless, further development of automated diagnostic programs will help facilitate the spread of clinical NGS-based diagnosis.

[0053] Fig. 3 is a flow diagram of a computer tool and/or computer process 300 embodying the present invention. At 305, a digital processor obtains biological sequence data of a patient that has the subject disease. At 310, a set of mutations is analyzed and scored as a function of a data store of genotype-phenotype data of the population with and without various diseases, including the subject disease, by performing subsequent processing blocks.

[0054] At 315, the processor determines the first potential cause by calculating the probability that the subject disease in the patient is caused by any combination of two

recessive heterozygous alleles observed in the sequencing data in a given gene. At 320, the processor determines the second potential cause by calculating the probability that the subject (patient) has an X-linked or recessive disease caused by a homozygous or hemizygous allele that was observed in the sequencing data. At 325, the processor determines the third potential cause by calculating the probability that the subject has a dominant disease due to an allele observed in the sequencing data. Next, at 330, the processor recursively combines all instances of the determined first, second and third potential causes of modules/steps 315, 320, 325. In particular, this combining calculates the probability of a set of mutations in a given gene being the cause of the subject disease in the patient. At 335, a decision point is reached. If there are more genes, the process returns to block 315, and steps/modules 315, 320, 325 and 330 can be repeated for each allele and pair of alleles in a given gene Ψ. Otherwise, if there are no more genes, the processor outputs, at 340, the probability of a set of mutations in a given gene causing the subject disease in the patient. The processor, accessing a memory storing the data store, can use the data from the data store and analyzes and scores the set of mutations.

[0055] From the analysis method and data from the data store, at 345, the processor outputs a determination of the probability that each mutant gene in the patient is responsible for the subject disease, i.e., a general identification of the alleles responsible for the subject disease in the patient. As shown in the example process 300, the processor can report, as output 350, the predicted probability that observed set of mutations in each gene causes the subject disease in the patient.

[0056] In embodiments, the process, e.g., the analyzing and scoring at 310, can include any combination of: (i) calculating the probability that the subject disease has a particular inheritance mode (e.g., is caused by a dominant gene or is caused by a recessive gene in the subject disease population); (ii) for each gene Ψ with at least one mutation in the subject, calculating the probability that the gene can cause the subject disease with a given inheritance mode; and (iii) considering gene-specific differences between genes represented in the data store.

[0057] It will be understood that various embodiments may skip certain steps. For example, an embodiment may ignore the inheritance mode, step (i) described above, and still produce a useful output. For example, skipping step (i) is believed to work as well as including that step in a case where all patients have the same inheritance mode. In another

example, skipping step (ii) described above may be applicable if one is only considering genes that are 100% known to cause the patient's disease given their inheritance model. Skipping step (iii) can work, but it would likely over-estimate the importance of hyper-mutable genes.

[0058] Fig. 4 illustrates a computer network or similar digital processing environment in which the present invention may be implemented. Client computer(s)/devices 50 and server computer(s) 60 provide processing, storage, and input/output devices executing application programs and the like. Client computer(s)/devices 50 can also be linked through communications network 70 to other computing devices, including other client devices/processes 50 and server computer(s) 60. Communications network 70 can be part of a remote access network, a global network (e.g., the Internet), a worldwide collection of computers, Local area or Wide area networks, and gateways that currently use respective protocols (TCP/IP, Bluetooth, etc.) to communicate with one another. Other electronic device/computer network architectures are suitable.

[0059] Fig. 5 is a diagram of the internal structure of a computer (e.g., client processor/device 50 or server computers 60) in the computer system of Fig. 4. Each computer 50, 60 contains system bus 79, where a bus is a set of hardware lines used for data transfer among the components of a computer or processing system. Bus 79 is essentially a shared conduit that connects different elements of a computer system (e.g., processor, disk storage, memory, input/output ports, network ports, etc.) that enables the transfer of information between the elements. Attached to system bus 79 is I/O device interface 82 for connecting various input and output devices (e.g., keyboard, mouse, displays, printers, speakers, etc.) to the computer 50, 60. Network interface 86 allows the computer to connect to various other devices attached to a network (e.g., network 70 of Fig. 4). Memory 90 provides volatile storage for computer software instructions 92 and data 94 used to implement an embodiment of the present invention (e.g., code detailed above and in Fig. 3 as process 300). Disk storage 95 provides non-volatile storage for computer software instructions 92 and data 94 used to implement an embodiment 300 of the present invention. Data 94 includes a working copy of the data store of genotype-phenotype data of the population with and without various diseases including the subject disease population. Central processor unit 84 is also attached to system bus 79 and provides for the execution of computer instructions. Central processor unit 84 is configured to implement process 300 by

executing corresponding instructions 92 with supporting data 94 and input patient sequence data.

[0060] In one embodiment, the processor routines 92 and data 94 are a computer program product (generally referenced 92), including a computer readable medium (e.g., a removable storage medium such as one or more DVD-ROM's, CD-ROM's, diskettes, tapes, etc.) that provides at least a portion of the software instructions for the invention system. Computer program product 92 can be installed by any suitable software installation procedure, as is well known in the art. In another embodiment, at least a portion of the software instructions may also be downloaded over a cable, communication and/or wireless connection. In other embodiments, the invention programs are a computer program propagated signal product 107 embodied on a propagated signal on a propagation medium (e.g., a radio wave, an infrared wave, a laser wave, a sound wave, or an electrical wave propagated over a global network such as the Internet, or other network(s)). Such carrier medium or signals provide at least a portion of the software instructions for the present invention routines/program 92.

[0061] In alternate embodiments, the propagated signal is an analog carrier wave or digital signal carried on the propagated medium. For example, the propagated signal may be a digitized signal propagated over a global network (e.g., the Internet), a telecommunications network, or other network. In one embodiment, the propagated signal is a signal that is transmitted over the propagation medium over a period of time, such as the instructions for a software application sent in packets over a network over a period of milliseconds, seconds, minutes, or longer. In another embodiment, the computer readable medium of computer program product 92 is a propagation medium that the computer system 50 may receive and read, such as by receiving the propagation medium and identifying a propagated signal embodied in the propagation medium, as described above for computer program propagated signal product.

[0062] Generally speaking, the term "carrier medium" or transient carrier encompasses the foregoing transient signals, propagated signals, propagated medium, storage medium and the like.

[0063] Widespread use of systems embodying the present invention has the potential to address the shortcomings and problems of the art. Currently, embodiments have demonstrated the ability to consistently produce high quality genetic diagnosis comparable to those of a highly trained expert for the genetic diagnosis of retinal disease. Diagnosis of a

hundred samples can be achieved daily by a single individual with moderate computer experience, massively reducing the overhead required for a molecular diagnosis facility. Advantageously, the results of a system embodying the present invention are reliable and consistent, and allow for precise comparison between institutions to ensure reliable results which are of the upmost importance in healthcare.

[0064] Certain embodiments have been tested for retinal disease for which there is sufficient disease-specific data. In order to test whether different embodiments work for a variety of diseases, access to phenotype-genotype data is required, as well as a comparison of the results of the EXPERT system/embodiments to the results derived by other labs. Further developments will allow Applicant to assess if principles of the present invention work for other diseases. Applicant's method and system are shown to be widely applicable. The present invention could potentially generate a robust and easily used pipeline for widespread medical use, and it currently has utility for the genetic diagnosis of retinal disease.

EXEMPLIFICATION

EXAMPLE 1 - RETINAL EXPERT SYSTEM (RES): AN INTEGRATED APPROACH TO THE DETERMINATION OF CAUSATIVE VARIANTS IN RETINAL EYE DISEASE

[0065] **Purpose:** With the decreasing cost of genomic sequencing, data driven analysis has become an increasingly important aspect of molecular diagnosis. Larger data sources have vastly improved post-sequencing processes such as annotation and filtering, yet ultimate determination of molecular causality is still left to experts of the field. Described is a process and system whereby a list of patient variants can be scored and ranked according to likeliness of disease causality using information gleaned from large stores of genomic data. Such a system preferably learns from ever growing datasets, continuously evolving and optimizing.

[0066] **Methods:** A statistical framework is described that integrates gene-disease association, inheritance pattern, and functional prediction, to rank mutant genes in a patient. Taking advantage of previously published mutant alleles and an internal database, the prior probabilities of each gene associated with a specific disease were estimated. This study focused on common retinal diseases of which many genes were associated with multiple diseases. An embodiment of the present approach scores genes as likely to be causative based on known disease associations, and ability to fit a given inheritance pattern. Functional prediction scores were integrated as a means to define variant potency, with normalization based on genic level analysis of common variants.

[0067] **Results:** A system embodying the present invention was trained with gene-disease association data from more than 1000 patients with various retinal diseases. The system was tested on a separate cohort of 30 patients and the results compared to molecular diagnoses determined by human experts. Strikingly, results showed over 85% correlation between the system and human analysis. Of 20 high confidence human calls, 18 system calls were correlated. Of 4 samples with no human determination, the system was similarly unable to call 3 samples while discovering in the fourth a possibly overlooked causative gene of RP1L1.

[0068] **Conclusions:** The Retinal EXPERT system (RES) represents a significant step toward streamlining NGS based molecular diagnosis as it prioritizes the most likely disease causing genes through quantification, allowing researchers and diagnosticians alike to work efficiently and effectively. More importantly, this tool will continue to improve with the accumulation of larger data sets. Similar methods can be implemented for other human diseases as well.

[0069] **Layman Abstract:** With the decreasing cost of genomic sequencing, it is important to be able to analyze samples accurately when finding underlying genetic causes of disease. Embodiments of the present invention, when given a list of a patient's genetic mutations, can determine the gene that most likely contains a mutation causing that patient's disease. Embodiments are based on a combination of several types of data: gene-disease association, inheritance pattern, population frequency, and the potential that a mutation is damaging. This scoring system and method was trained with data from over 1000 patients with likely causative variants previously determined by human experts, and tested on a separate cohort of 30. Strikingly, the results showed over 85% correlation between the top scoring gene determined by the algorithm and that determined by human expert analysis. Embodiments allow for a more efficient work flow, with less likelihood of major human error. Bringing the most likely disease causing genes to the front allows researchers and diagnosticians alike to work efficiently and effectively. More importantly, this tool will continue to improve with the accumulation of larger data sets. Similar methods can be implemented for other human diseases as well.

EXAMPLE 2 - USHER SYNDROME DATA

[0070] In an effort to explore the implications of the EXPERT System on multiple disease type an embodiment was used to examine a cohort of 32 molecularly diagnosed

Chinese Usher Syndrome patients. Unlike pure Retinitis Pigmentosa (RP), Usher Syndrome characterization includes hearing loss. Most patients with advanced Usher Syndrome are both blind and deaf. Additionally, Usher Syndrome differs on a molecular level, with several uniquely associated genes with the potential to harbor disease specific mutations including the USH family of genes. A comparison of human molecular diagnosis and EXPERT System molecular diagnosis is presented in Table 3.

[0071]

Sample ID	Human Call	EXPERT System Call	EXPERT System Score
SRF_1067	PCDH15	PCDH15	0.98
SRF_1076	MYO7A	MYO7A	0.98
SRF_1086	USH2A	USH2A	0.98
SRF_1120	USH2A	USH2A	0.98
SRF_1350	USH2A	USH2A	0.98
SRF_1362	USH2A	USH2A	0.27
SRF_1363	USH2A	USH2A	0.075
SRF_1369	**EYS**	**USH1G**	**0.98**
SRF_1402	RDH12	RDH12	0.02
SRF_1405	MYO7A	MYO7A	0.98
SRF_1425	MYO7A	MYO7A	0.269
SRF_1434	USH2A	USH2A	0.989
SRF_1442	USH2A	USH2A	0.979
SRF_1444	USH2A	USH2A	0.98
SRF_1446	USH2A	USH2A	0.273
SRF_1454	USH2A	USH2A	0.979
SRF_198	USH2A	USH2A	0.27
SRF_207	USH2A	USH2A	0.27
SRF_277	USH2A	USH2A	0.277
SRF_278	USH2A	USH2A	0.98
SRF_375	USH2A	USH2A	0.526
SRF_376	USH2A	USH2A	0.453
SRF_554	MYO7A	MYO7A	0.989
SRF_555	GPR98	GPR98	1
SRF_556	GPR98	GPR98	0.98
SRF_557	MYO7A	MYO7A	0.278
SRF_797	USH2A	USH2A	0.979
SRF_818	GPR98	GPR98	0.072
SRF_822	USH2A	USH2A	0.988
SRF_834	USH2A	USH2A	0.979
SRF_846	MYO7A	MYO7A	0.786
SRF_849	USH2A	USH2A	0.27

Table 3: Human and EXPERT System molecular diagnosis of Usher Syndrome Patients

[0072] Table 3 shows results for 32 patients. The first row ("Sample ID") list a randomized identifier, the second row ("Human Call") lists the human implicated gene, the third row ("EXPERT System Call") lists the EXPERT System implicated gene (i.e., the gene having highest probability of containing causative variant), and the fourth column ("EXPERT System Score") list the probability (0 to 1 probability score) determined by the EXPERT System for the implicated gene. Overall there was a 97% similarity between the human and EXPERT System molecular diagnoses. The only case in which there was inconsistency, SRF_1369, was due human error. For this patient, the human diagnostician improperly suggested that a homozygous mutation in EYS was the cause of the disease. Upon validation through Sanger sequencing, the EYS variant was found to be heterozygous and no additional follow-up was performed. On the other hand, the EXPERT System was not only able to discount the erroneous EYS variant (due to EYS not being associated with Usher Syndrome), but it also was able to determine an alternative molecular diagnosis (USH1G) that the human diagnostician had not considered.

EXAMPLE 3 - TESTED DISEASES AND THEIR DESCRIPTIONS

[0073] In addition to Retinitis Pigmentosa and Usher Syndrome, the EXPERT System has been tested on Leber Congenital Amaurosis, Stargardt disease, Familial Exudative Vitreoretinopathy, Cone Rod Dystrophy, and Bardet-Biedl syndrome. A description of these diseases is below:

[0074] **Retinitis Pigmentosa:** A retinal degenerative disorder characterized by pigmentation of the retina. The onset of this disease is variable but typically occurs in adulthood. Degeneration starts with peripheral cone cells and will typically progress inward, eventually causing blindness. This disease can either be recessive or dominant.

[0075] **Leber Congenital Amaurosis:** This retinal degenerative disorder is similar to Retinitis Pigmentosa, however onset is at birth. This disease is the leading cause of genetically linked blindness in school aged children. Although dominant cases have been found, LCA is mainly observed as a recessive disease.

[0076] **Stargardt disease:** This is a retinal degenerative disease that typically presents and progresses to complete blindness within the first 20 years of life. Unlike Retinitis Pigmentosa, degeneration with this disease begins with the macula and progresses outwards. The majority of Stargardt patients have a recessive form of the disease with few cases of the dominant disease observed.

[0077] **Familial Exudative Vitreoretinopathy:** Unlike the above-described diseases, this is not a retinal degenerative disease. Similar to wet Age-Related Macular Dystrophy, this disease is caused by an overabundance of blood vessel growth in the retina. This growth initially can cause hemorrhaging into the vitreous space, followed by folding and tearing of the retina. Ultimately, if left untreated, this disease can cause retinal detachment and blindness. Both recessive and dominate forms of this disease exist.

[0078] **Cone Rod Dystrophy:** This is another retinal degenerative disease, but unlike the previous diseases mentioned, it involves the simultaneous degeneration of both cone and rod cells within the retina. This disease can be recessive or dominant.

[0079] **Bardet-Biedl syndrome:** Similar to Usher Syndrome, this disease is characterized by several systematic phenotypes including Retinitis Pigmentosa, renal dysfunction, obesity, hypogonadism, and bone developmental problems like polydactyly. The systemic nature of this disease is due to the disruption of general cilia function, an important component of many cell and tissue types throughout the human body. This disease is an autosomal recessive disorder.

[0080] The retinal diseases on which the EXPERT System was originally developed have many diverse phenotypes and disease specific gene associations. As such, with the availability of similar data for additional diseases (gene/disease associations and inheritance pattern probability), the EXPERT System statistical framework can be useful for all diseases.

[0081] The teachings of all patents, published applications and references cited herein are incorporated by reference in their entirety.

[0082] While this invention has been particularly shown and described with references to example embodiments thereof, it will be understood by those skilled in the art that various changes in form and details may be made therein without departing from the scope of the invention encompassed by the appended claims.

[0083] The foregoing discussions reference by way of non-limiting example retinal disease as a subject disease of an embodiment of the present invention. Other diseases are suitable subjects of diagnosis by various embodiments of the present invention. For example, embodiments provide molecular diagnosis of Mendelian disorders, such as but not limited to, those listed in the Online Mendelian Inheritance in Man (OMIM) database. See the OMIM website (www.omim.org). OMIM® and Online Mendelian Inheritance in Man® are registered trademarks of the Johns Hopkins University.

CLAIMS

What is claimed is:

1. A molecular diagnosis method, comprising computer-implemented steps of:

in a digital processor:

obtaining biological sample sequencing data of a patient having a subject disease;

accessing a memory storing a data store of genotype-phenotype data of a population having the subject disease;

using the data from the data store, identifying the probability of each set of alleles in a given gene being responsible for the subject disease in the patient by:

determining a first potential cause by calculating the probability that the subject disease in the patient is caused by any combination of two recessive heterozygous alleles observed in the sequencing data in a given gene;

determining a second potential cause by calculating the probability that the patient has an X-linked or recessive disease caused by a homozygous or hemizygous allele that was observed in the sequencing data;

determining a third potential cause by calculating the probability that the patient has a dominant disease due to an allele observed in the sequencing data; and

recursively combining all instances of the determined first, second and third potential causes for each gene, and outputting the results of said combining, the results indicating the probability a set of mutations in a given gene causes the subject disease in the patient.

2. A method as claimed in Claim 1, further comprising:

analyzing and scoring as follows:

i) calculating the probability that the subject disease has a particular inheritance mode, the probability being indicative of the subject disease being caused by a dominant gene or being caused by a recessive gene;

ii) for each gene with a set of mutations, calculating the probability that the gene can cause the subject disease with a given inheritance mode; and

iii) considering gene-specific differences between genes represented in the data store; and

wherein the probability of each set of alleles is identified from results of the analyzing and scoring.

3. A method as claimed in Claim 2 wherein the analyzing and scoring considers gene-specific differences including computing probability that any given allele occurring in a gene severely disrupts the function of that gene, said computing using percent frequency of these discrepancies in the data store to calculate probability of that gene being the cause of the subject disease.

4. A method as claimed in Claim 3 wherein the computing further determines probability of an allele in that gene affects protein function of said gene.

5. A method as claimed in any one of Claims 1 to 4 wherein the data store includes genotype-phenotype data of a population with and without various diseases including the subject disease.

6. A method as claimed in any one of Claims 1 to 4 wherein the subject disease is a retinal disease, a Mendelian disorder, or other.

7. A computer-based tool carrying out molecular diagnosis comprising:

in a digital processor:

an input configured to obtain biological sample sequencing data of a patient having a subject disease;

a driver configured to access a memory storing a data store of genotype-phenotype data of a population having the subject disease;

a detector configured to use the data from the data store to identify the probability of each set of alleles in a given gene being responsible for the subject disease in the patient by:

determining a first potential cause by calculating the probability that the subject disease in the patient is caused by any combination of two recessive heterozygous alleles observed in the sequencing data in a given gene;

determining a second potential cause by calculating the probability that the patient has an X-linked or recessive disease caused by a homozygous or hemizygous allele that was observed in the sequencing data;

determining a third potential cause by calculating the probability that the patient has a dominant disease due to an allele observed in the sequencing data; and

recursively combining all instances of the determined first, second and third potential causes for each gene, and outputting the results of said combining, the results indicating the probability a set of mutations in a given gene causes the subject disease in the patient.

8. A computer-based tool as claimed in Claim 7, further comprising:

an analyzer to analyze and score by:

i) calculating the probability that the subject disease has a particular inheritance mode, the probability being indicative of the subject disease being caused by a dominant gene or being caused by a recessive gene;

ii) for each gene with a set of mutations, calculating the probability that the gene can cause the subject disease with a given inheritance mode; and

iii) considering gene-specific differences between genes represented in the data store; and

wherein the probability of each set of alleles is identified from results of the analyzing and scoring.

9. A computer-based tool as claimed in Claim 8 wherein the analyzer considers gene-specific differences by computing probability that any given allele occurring in a gene severely disrupts the function of that gene, said computing using percent frequency of these discrepancies in the data store to calculate probability of that gene being the cause of the subject disease.

10. A computer-based tool as claimed in Claim 9 wherein the analyzer further determines probability of an allele in that gene affects protein function of said gene.

11. A computer-based tool as claimed in any one of Claims 7 to 10 wherein the data store includes genotype-phenotype data of a population with and without various diseases including the subject disease.

12. A computer-based tool as claimed in any one of Claims 7 to 10 wherein the subject disease is a retinal disease, a Mendelian disorder, or other.

13. A molecular diagnosis system comprising:

a computer memory storing a data store of genotype-phenotype data of a population with and without various diseases including a subject disease of a patient; and

a processor communicatively coupled to the memory in a manner enabling access to the data store, the processor configured to be responsive to sequencing data of the patient and in response:

uses the data from the data store to identify the probability of each set of alleles in a given gene being responsible for the subject disease in the patient by:

determining a first potential cause by calculating the probability that the subject disease in the patient is caused by any combination of two recessive heterozygous alleles observed in the sequencing data in a given gene;

determining a second potential cause by calculating the probability that the patient has an X-linked or recessive disease caused by a homozygous or hemizygous allele that was observed in the sequencing data;

determining a third potential cause by computing the probability that the patient has a dominant disease due to an allele observed in the sequencing data; and

recursively combining all instances of the determined first, second and third potential causes for each gene, and outputting the results of said combining as indications of the probability that a set of mutations in a given gene causes the subject disease in the patient.

14. A system as claimed in Claim 13, wherein the processor further uses the data from the data store to analyze and score by:

i) calculating the probability that the subject disease has a particular inheritance mode, the probability being indicative of the subject

disease being caused by a dominant gene or being caused by a recessive gene;

ii) for each gene with a set of mutations, calculating the probability that the gene can cause the subject disease with a given inheritance mode; and

iii) considering gene-specific differences between genes represented in the data store; and

wherein the probability of each set of alleles is identified from results of the analyzing and scoring.

15. A system as claimed in Claim 14 wherein the processor analyzing and scoring considers gene-specific differences including computing probability that any given allele occurring in a gene severely disrupts the function of that gene, said computing using percent frequency of these discrepancies in the data store to calculate probability of that gene being the cause of the subject disease.

16. A system as claimed in Claim 15 wherein the processor computing further determines probability of an allele in that gene affects protein function of said gene.

17. A system as claimed in any one of Claims 13 to 16 wherein the subject disease is a retinal disease, a Mendelian disorder, or other.

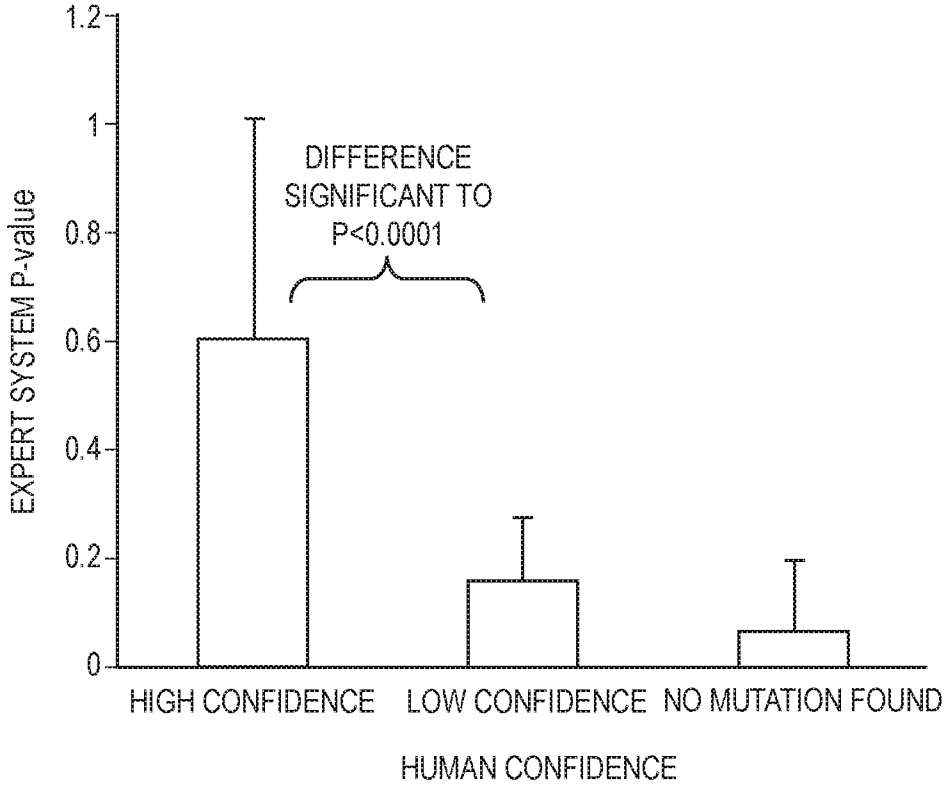
1.2
1
0.8
0.6
0.4
0.2
0
EXPERT SYSTEM P-value
DIFFERENCE
SIGNIFICANT TO
P<0.0001
HIGH CONFIDENCE
LOW CONFIDENCE
NO MUTATION FOUND
HUMAN CONFIDENCE

FIG. 1

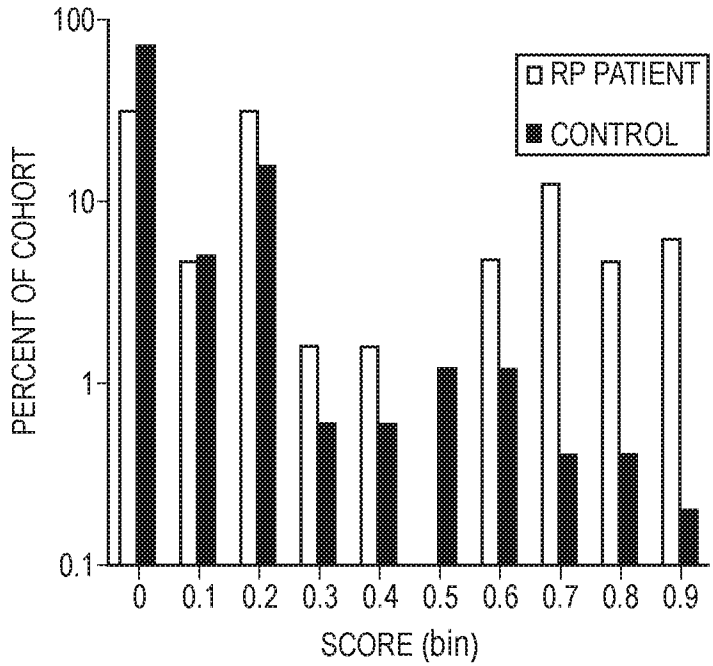
100
10
1
0.1
PERCENT OF COHORT
RP PATIENT
CONTROL
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
SCORE (bin)

FIG. 2

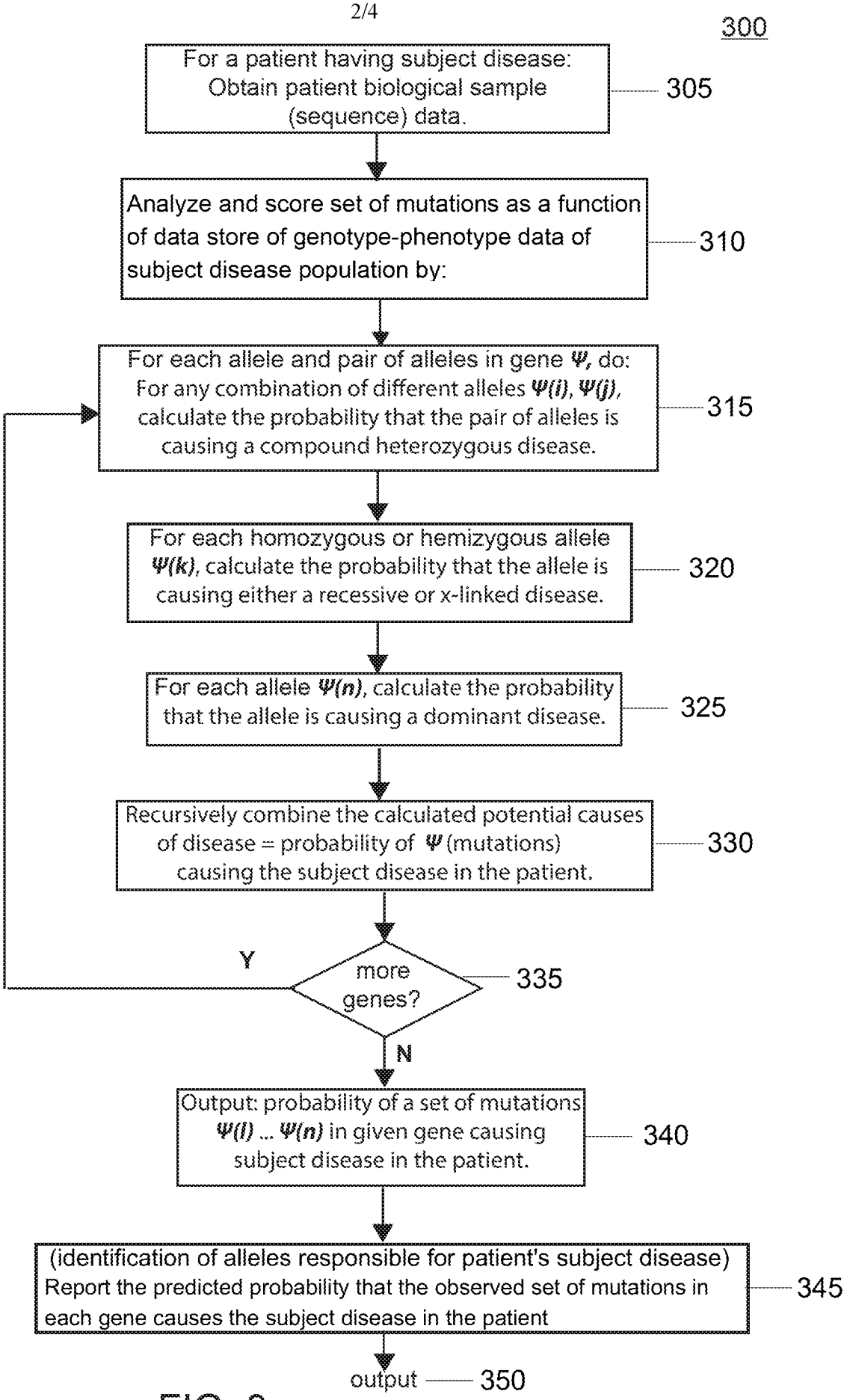
300
For a patient having subject disease: Obtain patient biological sample (sequence) data.
305
Analyze and score set of mutations as a function of data store of genotype-phenotype data of subject disease population by:
310
For each allele and pair of alleles in gene Ψ, do: For any combination of different alleles Ψ(i), Ψ(j), calculate the probability that the pair of alleles is causing a compound heterozygous disease.
315
For each homozygous or hemizygous allele Ψ(k), calculate the probability that the allele is causing either a recessive or x-linked disease.
320
For each allele Ψ(n), calculate the probability that the allele is causing a dominant disease.
325
Recursively combine the calculated potential causes of disease = probability of Ψ (mutations) causing the subject disease in the patient.
330
Y
more genes?
335
N
Output: probability of a set of mutations Ψ(l) ... Ψ(n) in given gene causing subject disease in the patient.
340
(identification of alleles responsible for patient's subject disease) Report the predicted probability that the observed set of mutations in each gene causes the subject disease in the patient
345
output
350

FIG. 3

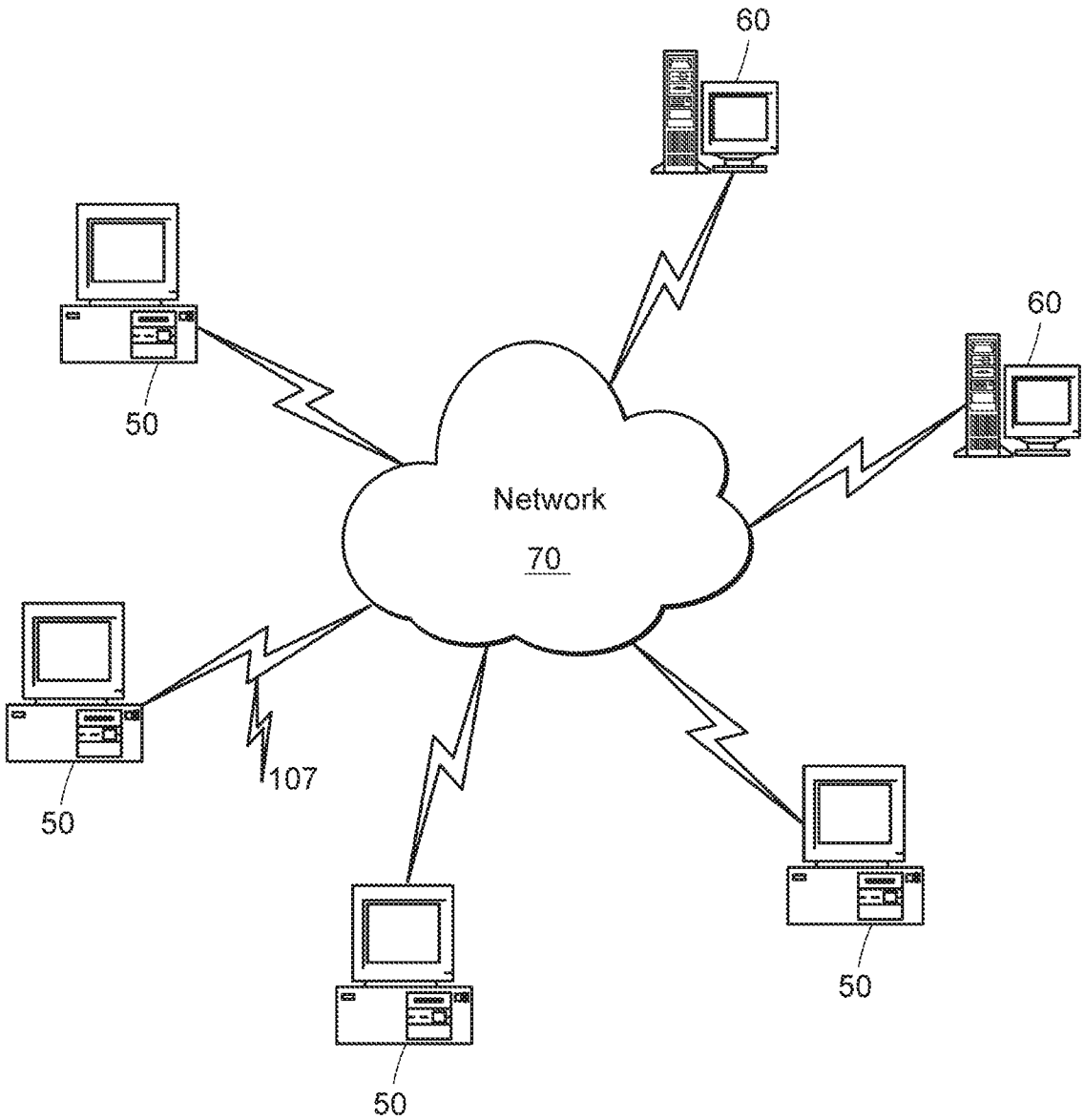

FIG. 4

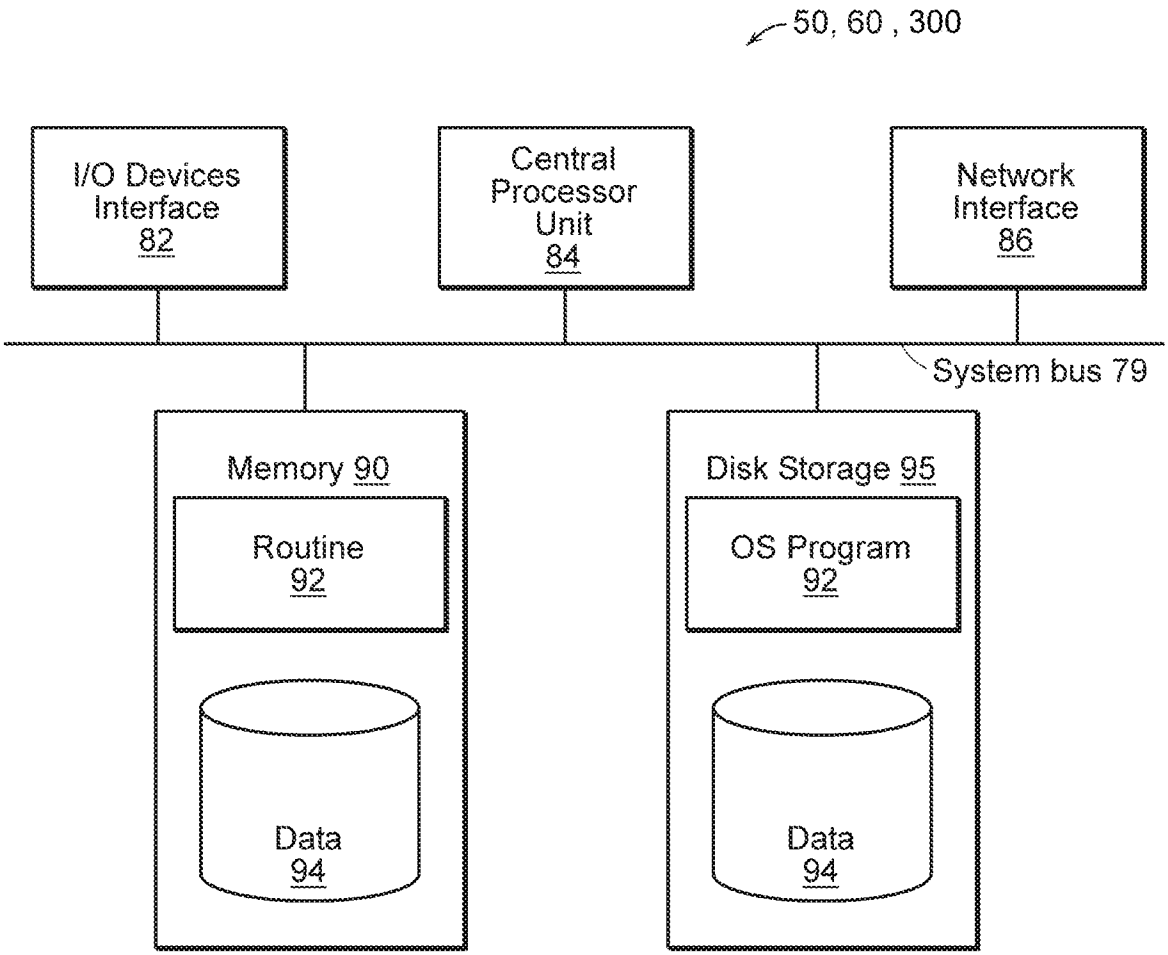
50, 60 , 300
I/O Devices Interface 82
Central Processor Unit 84
Network Interface 86
System bus 79
Memory 90
Routine 92
Data 94
Disk Storage 95
OS Program 92
Data 94

FIG. 5

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2016/015333

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F19/18 G06F19/24
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)
G06F

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, 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 2013/231404 A1 (SEGAL MICHAEL M [US]) 5 September 2013 (2013-09-05) abstract paragraph [0004] - paragraph [0014] paragraph [0088] paragraph [0124] - paragraph [0187] -----	1-17
X	US 2013/332081 A1 (REESE MARTIN G [US] ET AL) 12 December 2013 (2013-12-12) abstract paragraph [0005] - paragraph [0012] paragraph [0044] - paragraph [0064] paragraph [0097] - paragraph [0098] paragraph [0118] - paragraph [0149] -----	1-17
A	US 2014/359422 A1 (BASSETT JR DOUGLAS E [US] ET AL) 4 December 2014 (2014-12-04) paragraph [0172] paragraph [0238] - paragraph [0269] -----	1-17

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
13 May 2016	23/05/2016

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

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2016/015333

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2013231404 A1	05-09-2013	NONE	
US 2013332081 A1	12-12-2013	EP 2614161 A1	17-07-2013
		US 2013332081 A1	12-12-2013
		WO 2012034030 A1	15-03-2012
US 2014359422 A1	04-12-2014	NONE	