A Bayesian framework for efficient and accurate variant prediction

Dajun Qian, Shuwei Li, Yuan Tian, Jacob W. Clifford, Brice A. J. Sarver, Tina Pesaran, Chia-Ling Gau, Aaron M. Elliott, Hsiao-Mei Lu, Mary Helen Black*

Ambry Genetics, Aliso Viejo, California, United States of America

* mblack@ambrygen.com

Abstract

There is a growing need to develop variant prediction tools capable of assessing a wide spectrum of evidence. We present a Bayesian framework that involves aggregating pathogenicity data across multiple *in silico* scores on a gene-by-gene basis and multiple evidence statistics in both quantitative and qualitative forms, and performs 5-tiered variant classification based on the resulting probability credible interval. When evaluated in 1,161 missense variants, our gene-specific *in silico* model-based meta-predictor yielded an area under the curve (AUC) of 96.0% and outperformed all other *in silico* predictors. Multifactorial model analysis incorporating all available evidence yielded 99.7% AUC, with 22.8% predicted as variants of uncertain significance (VUS). Use of only 3 auto-computed evidence statistics yielded 98.6% AUC with 56.0% predicted as VUS, which represented sufficient accuracy to rapidly assign a significant portion of VUS to clinically meaningful classifications. Collectively, our findings support the use of this framework to conduct large-scale variant prioritization using *in silico* predictors followed by variant prediction and classification with a high degree of predictive accuracy.

Introduction

The recent surge of sequencing-based clinical genetic testing has put a spotlight on associated challenges in data interpretation. While advances in genomics allow for the development of new genetic tests at an unprecedented pace and complexity, the interpretation of results has remained a largely manual and time-consuming process that is not scalable to the volume and diversity of available data [1]. In particular, the rich evidence in well classified variants is not effectively incorporated in classification schemes relying on manual processing of large-scale information.

The pathogenicity of a genetic variant can be assessed by various evolutionary, functional and structural *in silico* scores and a range of evidence from clinical, family history, co-occurrence and co-segregation data, as well as the published findings of case-control, cohort or family-based studies [2, 3]. One approach to variant classification is to follow a rule-based scoring system, in which each line of evidence is converted into a score and the summary score from all available evidence is used to determine the classification [4–6]. This rule-based approach

assumes variants have strong and consistent evidence that is equally applicable across the genome. It also makes use of only qualitative evidence and does not leverage existing knowledge pertaining to other well classified variants. Another approach to variant classification is the multifactorial likelihood method, in which the prior probability of pathogenicity is obtained by using a genome-wide *in silico* (S1 Table) or ensemble (S2 Table) prediction method and the posterior probability is derived by aggregating the prior probability over multiple quantitative evidence [7, 8]. While this multifactorial method is computationally simple, it does not account for gene-specific differences and variability of the estimated prior probability. A third approach uses allele frequencies and multiple *in silico* predictors in a Bayesian logistic regression model that includes priors based on case-control proportions of carriers reported in the literature or public databases, a categorical covariate for gene-specific effects, and fixed model terms for each gene tested [9]. Such models improved prediction accuracy over alternative approaches, but did not allow for all parameters to be freely estimated for each gene or make use of the wide range of available evidence.

To efficiently and accurately determine the pathogenicity of genomic variants, there is a growing need to develop data-driven tools that are capable of assessing a wide array of evidence associated with each variant while leveraging information that is readily available for well classified variants. To address this need, we developed a Bayesian framework for variant prediction that aggregates multiple *in silico* scores and evidence statistics in both quantitative and qualitative forms, and validated the models in genes associated with hereditary cancer syndromes. Our approach improves upon existing methods by leveraging the vast information available for classified variants, quantifying gene-specific *in silico* effects while incorporating both quantitative and qualitative evidence, and predicting the pathogenicity of each variant using a probability distribution to account for uncertainty.

Results

Analytical framework

Our Bayesian framework is a data-driven model-based tool for variant prediction and classification analysis (Fig 1). Initially, a set of gene-specific *in silico* predictors is selected from publically available *in silico* scores. An *in silico* variant prediction (IVP) model is used to predict a preliminary level of variant pathogenicity based on a training dataset that contains variants with known classification and their accompanying 16 *in silico* predictors. Next, the prior probability distribution of pathogenicity for each variant is estimated from a corresponding IVP model followed by a rescaled transformation, and the posterior probability distribution is estimated from a multifactorial variant prediction (MVP) model that aggregates the prior distribution over multiple evidence predictors (Fig 1A). Finally, variant classification is assigned to 5-tiered classes of benign, variant of likely benign (VLB), variant of uncertain significance (VUS), variant of likely pathogenic (VLP) and pathogenic based on the probability distribution of pathogenicity (Fig 1B).

In silico model analysis

For each of 10 genes that collectively constitute the multigene panel test (MGPT) data containing missense variants with known ClinVar consensus classification outcomes of benign, VLB, VLP and pathogenic, we constructed a training dataset and derived an IVP model that retained 1 to 4 *in silico* predictors from the 16 candidate standalone scores tested (S3 Table). When these gene-specific IVP models were evaluated in all 1,161 class-known variants using a leave-one-out cross validation (LOOCV) method and the 5-tiered classification scheme, 360 (31.0%) were predicted to be concordant with their known classes, 277 (23.9%) were
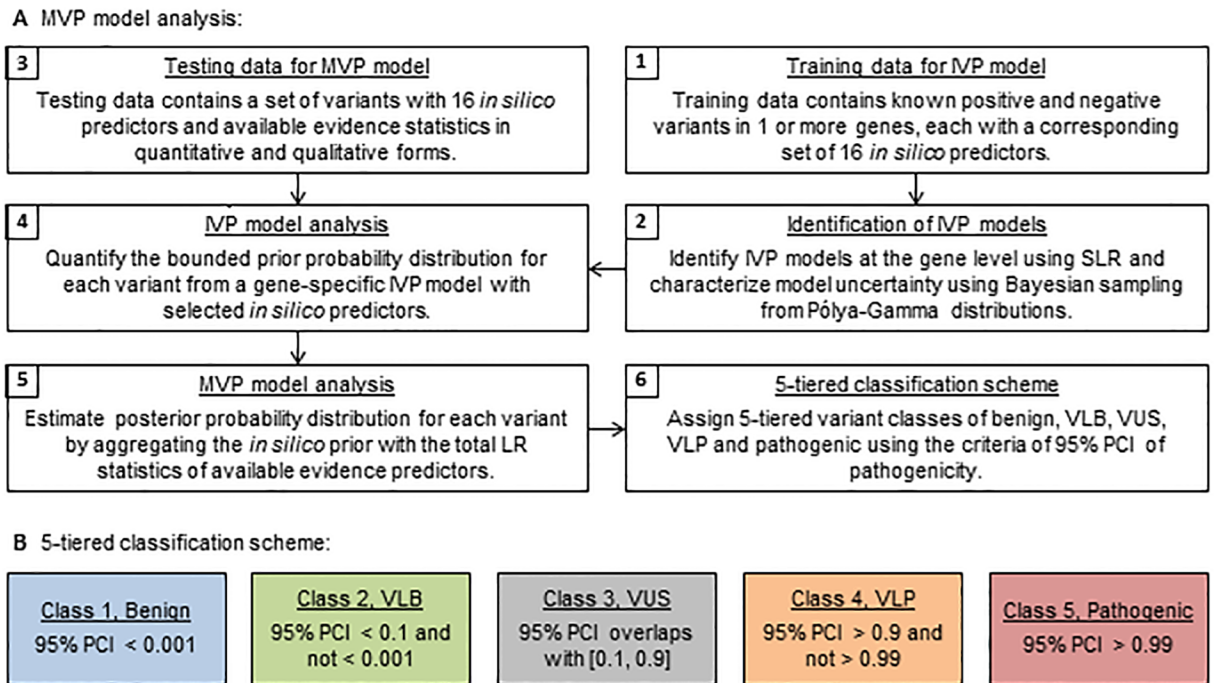
Fig 1. Algorithm modules of multifactorial model analysis for variant prediction and classification. (**A**) Modules of Bayesian multifactorial model analysis for variant prediction and classification. SLR = stepwise logistic regression. (**B**) 5-tiered variant classification scheme based on the estimated 95% probability credible interval (PCI) of variant pathogenicity.

categorized one level above or below their known classes (e.g., benign as VLB, pathogenic as VLP, etc.), 520 (44.8%) were classified as VUSs, and 4 (0.3%) were discordant (e.g., pathogenic/VLP classified as benign/VLB or vice versa) and therefore noted as false negatives or false positives (Fig 2A; S4 Table). Thus, while the IVP model yielded high positive predictive
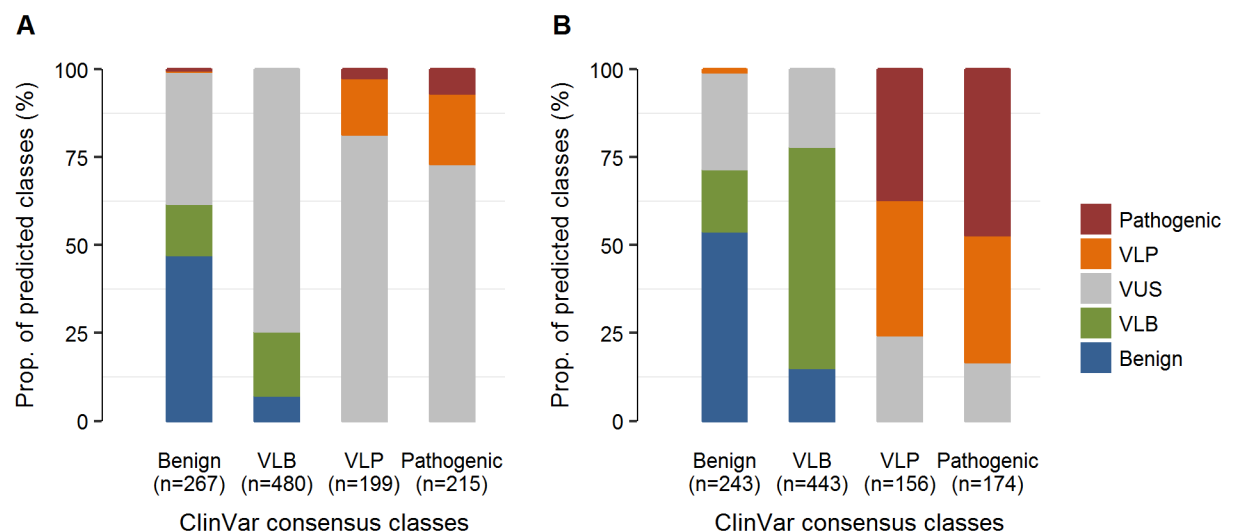
Fig 2. Outcome of 5-tiered predicted classes in MGPT data. (**A**) The proportions of predicted classes from gene-specific IVP model analysis in which each prediction was evaluated from a subset of 16 *in silico* predictors. The analysis was based on 1,161 class known missense variants in 10 genes using LOOCV. (**B**) The proportions of predicted classes from MVP model analysis in which each prediction aggregated a prior distribution from IVP model with the available evidence predictors. The analysis was based on 1,016 variants with any available evidence statistics.

Table 1. *In silico* variant prediction in MGPT data.

Method[a]	No. by Predicted Outcomes[b]					Performance Statistics[c]						
	TP	TN	FP	FN	VUS	Sen	Spe	PPV	NPV	Acc	AUC	P_VUS
Standalone predictor:												
MutPred	101	496	1	17	546	0.244	**0.664**	0.990	0.967	0.971	**0.939**	**0.470**
phyloP vertebrate	75	438	6	15	627	0.181	0.586	0.926	0.967	0.961	0.915	0.540
MutationAssessor	71	315	0	9	766	0.171	0.422	**1.000**	0.972	0.977	0.891	0.660
FATHMM	99	276	2	6	778	0.239	0.369	0.980	0.979	0.979	0.884	0.670
AGVGD	217	0	18	0	926	**0.524**	0.000	0.923	NA	0.923	0.878	0.798
Polyphen2 HVAR	0	414	0	21	726	0.000	0.554	NA	0.952	0.952	0.878	0.625
Siphy	0	277	0	5	879	0.000	0.371	NA	**0.982**	**0.982**	0.863	0.757
LRT	0	291	0	8	862	0.000	0.390	NA	0.973	0.973	0.862	0.742
GERP++	0	329	0	12	820	0.000	0.440	NA	0.965	0.965	0.860	0.706
Polyphen2 HDIV	0	404	0	26	731	0.000	0.541	NA	0.940	0.940	0.859	0.630
SIFT	0	315	0	12	834	0.000	0.422	NA	0.963	0.963	0.856	0.718
PROVEAN	50	102	1	2	1,006	0.121	0.137	0.980	0.981	0.981	0.855	0.866
phastCons mammalian	0	361	0	14	786	0.000	0.483	NA	0.963	0.963	0.838	0.677
phastCons vertebrate	0	455	0	10	696	0.000	0.609	NA	0.978	0.978	0.836	0.599
phyloP mammalian	0	213	0	5	943	0.000	0.285	NA	0.977	0.977	0.742	0.812
Grantham	0	0	0	0	1,161	0.000	0.000	NA	NA	NA	0.662	1.000
Meta-predictor:												
IVP	148	489	0	4	520	**0.357**	0.655	**1.000**	**0.992**	**0.994**	**0.960**	**0.448**
REVEL	90	475	4	20	572	0.217	0.636	0.957	0.960	0.959	0.942	0.493
MetaSVM	0	518	0	16	627	0.000	**0.693**	NA	0.970	0.970	0.940	0.540
Eigen	22	497	0	4	638	0.053	0.665	**1.000**	**0.992**	0.992	0.932	0.550
Eigen PC	12	506	0	5	638	0.029	0.677	**1.000**	0.990	0.990	0.927	0.550
CADD	61	330	17	3	750	0.147	0.442	0.782	0.991	0.951	0.906	0.646
MutationTaster	0	517	0	13	631	0.000	0.692	NA	0.975	0.975	0.894	0.543

[a]The *in silico* variant prediction analyses were evaluated from gene-specific prediction models from each of 16 standalone predictors and 7 meta-predictors, respectively, in which the IVP predictor was derived from the 16 standalone predictors. Each analysis was evaluated in MGPT data containing 1,161 missense variants (747 negatives, 414 positives) using LOOCV. Results are listed in descending order of AUC values among models using standalone predictors and meta-predictors, respectively.

[b]Predicted outcomes were derived from the predicted positive/negative categories and the known ClinVar consensus classes. TN = true negative, TP = true positive, FN = false negative, FP = false positive.

[c]Performance statistics were reported as Sen = sensitivity, Spe = specificity, PPV, NPV, Acc = accuracy, AUC, and P_VUS = proportion of variants classified as VUS. NA = not able to calculate. The best performance statistics among comparison *in silico* prediction methods are highlighted in bold.

value (PPV), negative predictive value (NPV) and accuracy (100.0%, 99.2% and 99.4%, respectively), sensitivity and specificity were modest (35.7% and 65.5%, respectively) (Table 1, lower section).

We next compared the predictive performance of the IVP model to that of each individual *in silico* predictor. Using the same 1,161 class-known variants, the gene-specific IVP models had the highest sensitivity (35.7%), PPV (100.0%), NPV (99.2%), accuracy (99.4%), and AUC (96.0%), and lowest proportion of VUS (44.8%) compared to each of the 16 standalone and 6 meta-predictor models (Table 1). Among all 23 *in silico* models, the AUC statistics were highest for IVP model and followed by REVEL and MetaSVM (Fig 3A and 3B; Table 1: $AUC_{IVP} = 0.960$ vs. $AUC_{REVEL} = 0.942$, p = 0.01; $AUC_{IVP} = 0.960$ vs. $AUC_{MetaSVM} = 0.940$, p = 0.002).
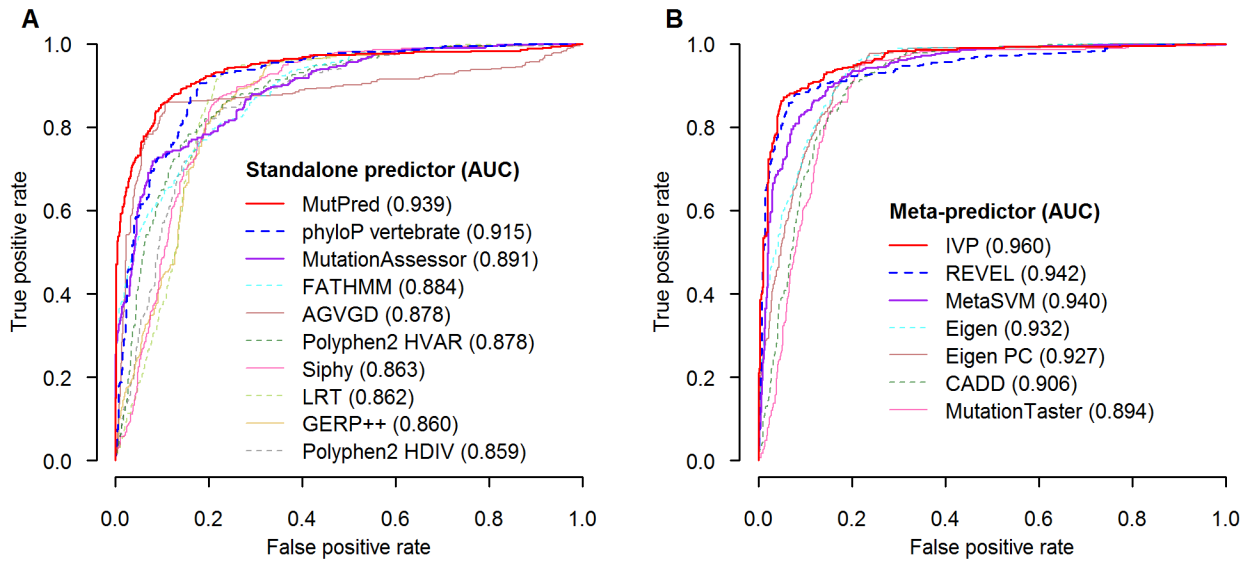
Fig 3. Comparison of AUC statistics of standalone and meta *in silico* predictors in MGPT data. (**A**) AUC statistics of top 10 standalone *in silico* predictors. (**B**) AUC statistics of 7 meta *in silico* predictors. The analysis models in legend were listed in descending order of AUC values. Abbreviations: AUC = area under the receiver operating characteristic curve; IVP = *in silico* variant prediction; MGPT = multigene panel test.

https://doi.org/10.1371/journal.pone.0203553.g003

Multifactorial model analysis

Among the 1,161 variants included in this analysis, 1,016 (87.5%) had data available for at least one evidence statistic from qualitative and/or quantitative sources. When applying our MVP model analysis to these variants, predictive performance improved as expected when compared to that of the IVP models (Table 2, upper section vs. lower section). The proportions of variants classified as concordant with their known classes, categorized one level above or

Table 2. Multifactorial variant prediction in MGPT data.

Method and Data (n-, n+)[a]	No. by Predicted Outcomes[a]					Performance Statistics[a]						
	TP	TN	FP	FN	VUS	Sen	Spe	PPV	NPV	Acc	AUC	P_VUS
Using all evidence data:												
Any evidence data (686, 330)	263	519	2	0	232	0.797	0.757	0.992	1.000	0.997	0.998	0.228
LR < 0.1 or > 10 (372, 226)	210	358	2	0	28	0.929	0.962	0.991	1.000	0.996	0.999	0.047
LR < 0.01 or > 100 (223, 74)	74	223	0	0	0	1.000	1.000	1.000	1.000	1.000	1.000	0.000
LR < 0.001 or > 1,000 (155, 23)	23	155	0	0	0	1.000	1.000	1.000	1.000	1.000	1.000	0.000
Using only auto-computed evidence:												
Any evidence data (618, 255)	95	287	2	0	489	0.373	0.464	0.979	1.000	0.995	0.986	0.560
LR < 0.1 or > 10 (225, 35)	31	212	2	0	15	0.886	0.942	0.939	1.000	0.992	0.996	0.058
LR < 0.01 or > 100 (174, 8)	8	174	0	0	0	1.000	1.000	1.000	1.000	1.000	1.000	0.000
LR < 0.001 or > 1,000 (107, 0)	0	107	0	0	0	NA	1.000	NA	1.000	1.000	NA	0.000

[a]Multifactorial variant predictions by MVP model analyses were conducted using either all available evidence data, which includes all quantitative and qualitative evidence predictors (total variants = 1,016), or only the 3 auto-computed predictors from readily available databases: family history, co-occurrence and mutation hotspot indicator (total variants = 873), respectively. The prior model for the MVP analysis was constructed for each variant using LOOCV. The n- and n+ values refer to the numbers of negative and positive variants, respectively, in the analytical dataset. LR = total LR from all evidence statistics or auto-computed ones. The abbreviation terms of predicted outcomes and performance statistics are same as those in footnotes b and c of Table 1.

https://doi.org/10.1371/journal.pone.0203553.t002

below their known classes, as VUSs, and discordant between benign/VLB and pathogenic/ VLP were 551 (54.2%), 231 (22.7%), 232 (22.8%) and 2 (0.2%), respectively (Fig 2B and S4 Table). With 2 false classifications and 782 appropriately categorized as benign/VLB or pathogenic/VLP variants, the MVP model analysis achieved 99.2% PPV (95% CI: 97.1% or above) and 100.0% NPV (95% CI: 99.1% or above). Moreover, in subsets of variants with moderate evidence (total likelihood ratio (LR) statistic <0.1 or >10) or strong evidence (total LR statistic <0.01 or >100) evidence for benign or pathogenic, 95.0% (568/598) and 100% (297/297) were correctly classified into the clinically informative classes of benign/VLB and pathogenic/VLP, respectively (Table 2, upper section). In particular, for the subset of variants with total LR statistic <0.01 or >100, we observed ideal performance statistics: 100.0% PPV, 100% NPV, 100% AUC, and 0% predicted VUS (Table 2, upper section).

Given that some evidence may lend itself to automated computation, while others may require manual examination or adjustment [10], we also applied MVP model analysis using a subset of auto-computed predictors for which data were readily available: co-occurrence, family history and mutation hotspot. Among the 1,161 variants initially included, 873 (75.2%) had data for at least one of these 3 predictors. Results of the constrained MVP model for these variants were highly accurate, with 98.6% AUC (Table 2, lower section); the proportions of variants classified concordant with their known classes, categorized one level above or below their known classes, as VUSs, and discordant between benign/VLB and pathogenic/VLP were 260 (29.8%), 122 (14.0%), 489 (56.0%) and 2 (0.2%), respectively (Table 2, lower section and S4 Table). Overall, 32.9% (382/1,161) of variants were appropriately categorized as benign/VLB or pathogenic/VLP. Among variants with strong evidence for a benign or pathogenic classification (total LR statistic <0.01 or >100, respectively), 100.0% (182/182) of variants were correctly classified as benign/VLB and pathogenic/VLP (Table 2, lower section).

Discussion

We present a Bayesian framework consisting of IVP models that assess variant pathogenicity using a subset of gene-specific *in silico* predictors, and an MVP model that aggregates this result with information from a variety of qualitative and quantitative evidence sources to accurately and robustly predict the pathogenicity classes of missense variants. The performance of MVP model analysis demonstrates that this approach is capable of leveraging a vast constellation of pathogenicity information available in large-volume data, and has important implications for increased clinical utility given its high predictive accuracy, which was cross validated in over 1,000 classified missense variants.

The IVP and MVP model analyses have several distinct features that allow for improved prediction accuracy and practical utility. First, IVP model prediction is conducted with data-driven model terms derived from gene-specific training data, and is supported by a data expansion procedure to accommodate a sparse data scenario in which the training data contains an insufficient number of class-known variants. Notably, our IVP model quantifies the pathogenicity of each variant with a probability distribution, instead of a probability value, which allows for improved prediction robustness and accuracy by accounting for estimation uncertainty. Second, an exact and fast Bayesian sampling procedure using data-independent Pólya-Gamma distributions is adapted for model estimation [11, 12], which facilitates analysis without the estimation of data-dependent priors on a gene-by-gene basis. Third, the incorporation of evidence statistics in qualitative forms, which are typically collected for rule-based classification schemes, makes the MVP model analysis capable of incorporating broader types of evidence statistics for improved practical utility, although the MVP model is not limited by the use of these data. We demonstrated that MVP model performance is highly accurate even

when only three auto-computed quantitative evidence predictors are included in the model. Lastly, the use of 95% PCI is designed to reduce the potential misclassification between benign/VLB and pathogenic/VLP by accounting for the variability in the predicted probability of variant classification.

When applied to the 10 genes with varying numbers of class-known missense variants, the gene-specific IVP models outperformed 16 standalone and 6 meta-predictors each based on a genome-wide *in silico* score. These results support the notion that universal prediction models tend to have varied performance across different genes, and gene-specific classifiers incorporating phenotype data and established disease-causing evidence can improve prediction accuracy [13–15]. For multifactorial variant classification, the MVP model correctly assigned 77% of variants to its precise or closely matched class and misclassified only 0.2% of variants between benign/VLB and pathogenic/VLP. In addition, we show that using auto-computed evidence statistics derived from commonly collected and readily available phenotype and genotype data, 33% of evaluated variants can be correctly classified to their precise or closely matched classes. These results highlight the practical utility of applying IVP model analysis for large-scale variant prioritization and MVP model analysis for variant classification.

Despite the fact that gene-specific models outperform those based on genome-wide information, IVP models can be unstable when the gene-specific training data is of small sample size and/or contains many ambiguously classified variants. Other approaches, such as a weighted average of gene-specific and panel-specific models might improve model robustness and prediction accuracy, and remain to be investigated. The IVP model presented here is limited to continuous *in silico* predictors, whereas categorical features such as domain effects and variant types may also be informative [9]. Availability of evidence statistics is also a limiting aspect of MVP model analysis; although the Bayesian MVP model we presented classified a large proportion of variants with a high degree of accuracy, the remaining VUS were due to either no (145/1,161 = 12.5%) or insufficient (232/1,161 = 20.0%) evidence statistics. In addition, models that account for correlations among evidence statistics [16], incorporate interaction and non-linear effects [9], and/or integrate LR statistics using distribution-based sampling approaches have the potential for improved variant prediction and classification.

Our Bayesian IVP and MVP model analyses form a data-driven framework for variant prediction and classification in aggregating a broad spectrum of pathogenic information. These model-based approaches are adaptive to the complexity of large-scale data, and are applicable to a wide variety of genes and phenotype conditions, provided that suitable training data are available. Importantly, these models afford an opportunity to accurately and efficiently reclassify VUS, and as such have the potential to improve the information on which clinical decisions are based.

Methods

Data

To assess prediction performance, we obtained 1,161 classified missense variants identified in 10 MGPT genes (*BRCA1*, *BRCA2*, *CDH1*, *PALB2*, *PTEN*, *TP53*, *MLH1*, *MSH2*, *MSH6* and *PMS2*) from ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/, downloaded in April 2017), compiled *in silico* predictor information from dbNSFP (https://sites.google.com/site/jpopgen/dbNSFP, downloaded in June 2017) and AGVGD (http://agvgd.hci.utah.edu/, released Sept 2014) databases, and collected variant-specific evidence statistics from Ambry Genetics databases, respectively (S1 Data: MGPT data with all relevant variables). Variants obtained from ClinVar were those with classifications established by expert panel review or deposited by any of the 6 submitters that consistently provide assertion criteria: Ambry Genetics, Emory,

GeneDx, InSiGHT, InVitae and SCRP. For each variant, we defined its consensus class per the following hierarchy: 1) we selected the most supported category among negative, positive and VUS, where negative and positive are designated as benign/VLB and pathogenic/VLP, respectively (i.e., positive if $N_{positive} > \max(N_{negative}, N_{VUS})$ or negative if $N_{negative} > \max(N_{positive}, N_{VUS})$); and 2) if the number of submitters is tied, and there was no conflict between positive and negative classes, we assigned positive or negative as appropriate (i.e., positive if $N_{positive} = N_{VUS}$ and $N_{negative} = 0$ or negative if $N_{negative} = N_{VUS}$ and $N_{positive} = 0$). Variants with conflicting classes (i.e., $N_{positive} = N_{negative}$), or classified as VUS by all submitters, were excluded from analysis. The number of variants with ClinVar consensus classes varied from 22 to 385 per gene (S5 Table, upper section).

Predictors and their derivation

The dataset used for IVP model training contained a set of class-known variants which are designated benign, VLB, VLP and pathogenic. The response variable *y* is a binary outcome for pathogenicity, with 1 for pathogenic or VLP and 0 for benign or VLB.

The predictor variables for IVP model analysis included 16 individual *in silico* scores from publically available variant prediction tools (i.e., Grantham [17], GERP++ [18], phastCons (for vertebrate and mammalian genomes) [19], AGVGD [20], SIFT [21], MutPred [22], SiPhy [23], LRT [24], phyloP (for vertebrate and mammalian genomes) [25], Polyphen2 (built on HumVar and HumDiv datasets) [26], MutationAssessor [27], PROVEAN [28] and FATHMM [29]; description in S1 Table). Six meta *in silico* scores were collected for comparisons of prediction performance (i.e., MutationTaster [30], CADD [31], REVEL [32], Eigen (for overall and eigendecomposition) [33] and MetaSVM [34]; description in S2 Table). Missing values of *in silico* predictors were imputed with the *k*-nearest neighbors method implemented in R package DMwR2 version 0.02. The missing values for a given variant were assigned the average value of the non-missing values for that predictor from its $k = 40$ nearest neighboring variants. *In silico* predictors with non-unit scale values (i.e., Grantham, GERP++, AGVGD, SiPhy, phyloP, MutationAssessor, PROVEAN, FATHMM, CADD, Eigen and MetaSVM) were transformed to unit scale of 0 to 1 by $(x_{raw} - x_{min})/(x_{max} - x_{min})$, where x_{raw} is the *in silico* score in its original scale and x_{min} and x_{max} are the minimum and maximum score values, respectively.

The predictor variables for MVP model analysis included 7 evidence predictors quantified as LR statistics for a) frequency and association, b) co-occurrence, c) co-segregation, d) family history, e) functional evidence, f) structural evidence, and g) other supporting data, respectively (description in S6 Table). For each of the 7 evidence predictors, we extracted qualitative evidence from variant-level classification records characterized by two parameters of p_{cut} and p_{frac}. Here p_{cut} is a threshold probability of 0.001, 0.1, 0.9 or 0.99 for assigning a variant to a targeted class of benign, VLB, VLP or pathogenic, respectively, as recommended in the American College of Medical Genetics and Genomics (ACMG) guidelines [2]. Also, p_{frac} is a fraction of evidence needed to reach a targeted class (e.g., 1, 0.5 and 0.25 are example values for evidence predictors in qualitative form; see S6 Table for numerical illustration). Applying the Bayes rule equation under a null probability of $p_{null} = 0.5$, the LR statistic for qualitative pathogenicity evidence is

$$LR(p_{cut}, p_{frac}) = (p_{cut}/(1 - p_{cut}))^{p_{frac}}. \tag{1}$$

For auto-computed predictors, we derived LR statistics directly from real-time in-house subject-level genotype and phenotype data. The LR statistic of co-occurrence evidence was estimated from a binomial likelihood model [35] under the rationale that a pathogenic variant is less likely to co-occur with a known pathogenic mutation in trans. The LR statistic of family

history evidence was derived from a history weighting score model [36] based on the premise that pathogenic mutations are more likely to occur in high-risk individuals while the presence of benign variants is unrelated to personal and family history. The mutation hotspot evidence was filtered out for the existence of at least one pathogenic variant at the same amino acid residue. For evidence statistics derived from two sources, denoted LR_1 and LR_2, the combined evidence of two correlated LR statistics is quantified as

$$LR(LR_1, LR_2) = \begin{cases} \max(LR_1, LR_2) & \textit{if both} > 1, \\ \min(LR_1, LR_2) & \textit{if both} < 1, \\ LR_1 \times LR_2 & \textit{otherwise}. \end{cases} \tag{2}$$

Thus, the available pathogenicity evidence for each variant was summarized into 7 LR statistics, where LR statistic was set to1 for each missing evidence predictor.

Training data for IVP model

To build a gene-specific IVP model for gene G, we constructed a training data $\mathbb{A}$ containing all class-known variants in that gene under a minimal sample size requirement of $n_{negative} \geq 5$ and $n_{positive} \geq 5$. Here we define $n_{negative} = n_{benign} + 0.5 \times n_{VLB}$ and $n_{positive} = n_{pathogenic} + 0.5 \times n_{VLP}$, where each n_* represents the number of class * variants in training data $\mathbb{A}$. The variant classes of benign, VLB, VLP and pathogenic are based on ClinVar consensus classifications, as previously described. For a sparse data scenario in which the variants in training data $\mathbb{A}$ satisfy $n_{negative} + n_{positive} \geq 5$ and either $n_{negative} < 5$ or $n_{positive} < 5$, we implemented a data expansion procedure based on the assumption that variants with similar *in silico* scores residing in two genes known to influence the same phenotype should have a more similar degree of pathogenicity than those from two randomly selected genes. This procedure resulted in a training data $\mathbb{A}$ for gene G that includes additional variants from other similar genes, and was implemented as follows:

1. Computed the mean distance for all variant pairs between gene G and each other gene $H \subset G^*$, where G^* is a gene set for all genes from the same panel or pathway except gene G. The distance between variants $g \in G$ and $h \in H$ was defined as the Euclidean distance of $d_{gh} = sqrt(\Sigma_{i=1,\cdots,I}(x_{gi} - x_{hi})^2)$, where vectors $(x_{g1}, \cdots, x_{gI})$ and $(x_{h1}, \cdots, x_{hI})$ are the $I = 16$ unit scaled individual *in silico* predictor values of variants g and h, respectively.

2. Chose a most similar gene $G' \subset G^*$ that showed the minimal mean distance with analysis gene G and merged the negative and/or positive variants one by one from G' to $\mathbb{A}$ in descending order of between-variant distance until both $n_{negative} \geq 5$ and $n_{positive} \geq 5$ were met or all variants in G' were merged into the training data $\mathbb{A}$.

3. If $n_{negative} < 5$ or $n_{positive} < 5$ remains true, repeated step (2) by choosing another gene $G'' \subset (G^* - G')$ and merging variants from G'' to $\mathbb{A}$ until $n_{negative} \geq 5$ and $n_{positive} \geq 5$ were met.

Thus, this data expansion procedure provided a means by which gene-specific training data may be constructed for genes that contain only a few classified variants.

Identification of IVP model

We derived a gene-specific IVP model by selecting a subset of *in silico* predictors using stepwise logistic regression (SLR) and quantified the probability distribution of pathogenicity using Bayesian logistic regression with coefficients sampled from data-independent Pólya-

Gamma distributions. The set of predictors yielding the minimal cross validation error rate among penalty coefficients (ranging from 2 to 8) were retained. The derived IVP model took the form

$$logit(y) = b_0 + b_1 z_1 + \cdots + b_K z_K, \tag{3}$$

where $Z = (z_1, \ldots, z_K)$ is a subset of gene-specific *in silico* predictors retained from SLR analysis, $B = (b_0, b_1, \ldots, b_K)$ are regression coefficients for intercept and slopes, and $logit(y)$ is the logit transformation of response variable y for negative and positive variants. The distributions of regression coefficients B of an IVP model were estimated by Bayesian Markov chain Monte Carlo updated from Pólya-Gamma distributions [11], as implemented by the logit function in R package BayesLogit version 0.5.1. For variant prediction purposes, the distributions of $K + 1$ regression coefficients, denoted $\mathbb{B} = (b_{kn}, 0 \leq k \leq K, 1 \leq n \leq N)$, are estimated jointly using $N = 1,000$ Gibbs samples after a burn-in period of 20,000 samples. Thus, the IVP probability distribution of each variant was estimated by its logistic transformation, denoted $\dot{\mathbb{P}} = (\dot{p}_n, 1 \leq n \leq N | \dot{p}_n = logit^{-1}(b_{0n} + \sum_k b_{kn} x_{kn}))$.

Estimation of IVP prior probability distribution

As a principle of multifactorial variant classification analysis, at least two lines of evidence are required for assigning a variant class to benign, VLB, VLP or pathogenic (i.e., classification based on the probability distribution from *in silico* analysis alone is not possible) [2]. To accommodate this framework, the IVP prior distribution of each variant was derived from a rescaled IVP probability distribution in the range of 0.1 to 0.9 with standard deviation proportional to the expected value. Specifically, the IVP prior distribution of a variant, denoted $\mathbb{P} = (p_n, 1 \leq n \leq N)$, was quantified by a 2-sided shifted probability function

$$p_n = \ddot{p}_{med} + (\ddot{p}_n - \ddot{p}_{med}) \times sd(\ddot{p}_{med})/sd(\dot{p}_{med}), \tag{4}$$

where $\ddot{\mathbb{P}} = (\ddot{p}_n, 1 \leq n \leq N | \ddot{p}_n = 0.8 \times \dot{p}_n + 0.1)$ is a linearly shifted probability distribution in range of 0.1 to 0.9, $\dot{p}_{med}$ and $\ddot{p}_{med}$ are the medians of IVP distribution $\dot{\mathbb{P}}$ and its shifted distribution $\ddot{\mathbb{P}}$, respectively, and $sd(\dot{p}_{med}) = sqrt(\dot{p}_{med} \times (1 - \dot{p}_{med}))$ and $sd(\ddot{p}_{med}) = sqrt(\ddot{p}_{med} \times (1 - \ddot{p}_{med}))$ are the expected standard deviations of $\dot{p}_{med}$ and $\ddot{p}_{med}$, respectively. The use of a rescaled prior distribution, as quantified by Eq (4), ensures the variability of IVP distribution is maintained in the prior distribution.

MVP model analysis

For each variant, the Bayesian MVP model analysis employed a distribution-based formation of Bayes rule to quantify the posterior probability distribution of pathogenicity. Such a posterior distribution, denoted $\mathbb{Q} = \{q_n, 1 \leq n \leq N\}$, was computed by aggregating the prior probability distribution $\mathbb{P}$ from an IVP model analysis and the available LR statistics from 7 evidence predictors through the equation:

$$q_n = p_n LR_{total}/(1 - p_n + p_n LR_{total}). \tag{5}$$

where p_n is a sample value from the prior distribution, and $LR_{total} = LR_{FAA} \times LR_{COC} \times LR_{CSG} \times LR_{FHX} \times LR_{FUN} \times LR_{STR} \times LR_{OTH}$ is the total LR statistic calculated under the assumption that the 7 evidence predictors are statistically independent (see S6 Table for definition of each LR statistic).

5-tiered variant classification scheme

We employed a 5-tiered variant classification scheme to assign the predicted classes of benign, VLB, VUS, VLP and pathogenic at the targeted probability thresholds of 0.001, 0.1, 0.9 and 0.99, respectively, following the ACMG guideline (Fig 1A) [2, 37]. The predicted class of each variant was assigned based on the 95% probability credible interval (PCI) of pathogenicity obtained from an IVP model or MVP model (Fig 1B). Here the 95% PCI is defined as the 2-sided 95% range of a probability distribution estimated from a corresponding prediction model using Bayesian sampling from Pólya-Gamma distributions [11]. The use of 95% PCI, instead of a point estimate of probability value, was designed to control the occurrence of false events by accounting for the variability of probability distribution.

Performance evaluation

To evaluate the performance of variant prediction, we assessed predicted outcomes including true positives (TP; predicted pathogenic/VLP concordant with known class), true negatives (TN; predicted benign/VLB concordant with known class), false positives (FP; benign/VLB predicted to be pathogenic/VLP), and false negatives (FN; pathogenic/VLP predicted to be benign/VLB). Performance statistics such as sensitivity, specificity, PPV, NPV, accuracy, AUC and proportion of predicted variants of uncertain significance (P_{VUS}) were assessed [38] (S7 Table). All performance statistics of IVP and MVP model analyses were evaluated using LOOCV to control model overfitting. The 95% confidence interval (CI) of a proportion was estimated from binomial distribution. Comparison of AUC estimates for different methods was performed using Delong's test [39]. All analyses were conducted with R for Statistical Computing version 3.3.3.

Supporting information

S1 Table. Standalone *in silico* predictors. [a] The 16 standalone predictors in this table will be used for building the *in silico* variant prediction (IVP) models.
(DOCX)

S2 Table. Meta *in silico* predictors. [a] The 6 meta predictors in this table will be used to evaluate the prediction performance.
(DOCX)

S3 Table. *In silico* predictors retained in gene-specific IVP models. [a] No. of variants in gene-specific training data included the classified variants in MGPT data from the same gene and those from other genes with minimal average distance to meet the sample size criterion of $n_{neg} \geq 5$ and $n_{pos} \geq 5$. [b] The IVP models were built on a subset of 16 standardized scores (gra, . . ., mua) in unit range of 0 to 1 from their original scores (Grantham, . . ., FATHMM), as follows: gra = (Grantham− 5)/215, ger = (GERP++ + 12.3)/18.5, pcv = phastCons_vertebrate, pcm = phastCons_mannalian, agv = AGVGD/65, sif = 1 −SIFT, mup = MutPred, sip = Siphy/ 38, lrt = LRT, ppv = (phyloP_vertebrate + 20)/30, ppm = (phyloP_mammalian + 13.3)/14.5, pov = Polyphen2_HVAR, pod = Polyphen2_HDIV, mua = (MutationAssessor + 5.2)/11.7, pro = (PROVEAN + 14)/28, and fat = (FATHMM + 16.2)/26.9.
(DOCX)

S4 Table. Cross validation for variant classification in MGPT data. [a] IVP model analysis was conducted with 1,161 missense variants. MVP model analyses were evaluated in 1,016 variants with any available evidence and 873 variants with only auto-computed evidence. Total numbers of variants for MVP model analysis were reduced due to some variants did not have

the required evidence statistics. Known classes were based on ClinVar consensus classification outcomes, and the predicted classes were evaluated by IVP and MVP models, respectively, using LOOCV.
(DOCX)

S5 Table. Number of variants according to ClinVar consensus classification.
(DOCX)

S6 Table. Evidence items and their likelihood ratio statistics. [a] The description of each evidence item follows the Ambry's Variant Classification Scheme. [b] All evidence items are summarized into 7 categories for MVP model analysis: FAA: Frequency and association evidence from control populations and case-control studies; COC: Co-occurrence evidence with another mutation; CSG: Co-segregation evidence with disease; FHX: Evidence of personal and family history, *de novo* alternation in family and established diagnosis without other mutation; FUN: Evidence of functional validation or genomic features (includes mutational hotspot); STR: Structural evidence; OTH: Other supporting evidence not included in other evidence groups. [c] The effect of each evidence item is graded into 7 levels of P-1, P-4, LP-1, LP-4, LB-1, LB-2 and B-1, respectively. These effect levels are quantified using following criteria under the assumption of no prior knowledge of variant pathogenicity: One evidence item of effect level P-1, or 4 evidence items of effect level P-4, are required to classify a variant to pathogenic; One evidence item of effect level LP-1, or 4 evidence items of effect level LP-4, are required to classify a variant to VLP; One evidence item of effect level LB-1, or 2 evidence items of effect level LB-2, are required to classify a variant to VLB; One evidence item of effect level B-1 is required to classify a variant to benign. [d] The effect of each qualitative evidence is transferred to a LR statistic under a null probability of 0.5 (see Eq (1) in Methods section). Numerically, the LR statistics for different evidence effects are calculated as follows: $LR(P\text{-}1) = 0.99/(1-0.99) = 99$; $LR(P\text{-}4) = 99^{0.25} = 3.1543$; $LR(LP\text{-}1) = 0.95/(1-0.95) = 19$; $LR(LP\text{-}4) = 19^{0.25} = 2.0878$; $LR(LB\text{-}1) = 0.05/(1-0.05) = 0.0526$; $LR(LB\text{-}2) = (1/19)^{0.5} = 0.2294$; $LR(B\text{-}1) = 0.001/(1-0.001) = 0.0010$.
(DOCX)

S7 Table. Performance statistics for variant classification. [a] The number of "total positive variants evaluated" includes true positives (TP), false negatives (FN) and those positives predicted as variants of uncertain significance (VUS). The number of "total negative variants evaluated" includes true negatives (TN), false positives (FP) and those negatives predicted as VUS.
(DOCX)

S1 Data. MGPT dataset. The 1,161 missense variants in 10 genes for performance evaluation of *in silico* and multifactorial model analyses reported in this article.
(XLSX)

Acknowledgments

Author Contributions

Conceptualization: Dajun Qian, Aaron M. Elliott, Hsiao-Mei Lu.

Data curation: Shuwei Li, Yuan Tian, Jacob W. Clifford, Brice A. J. Sarver.

Formal analysis: Dajun Qian, Yuan Tian, Jacob W. Clifford.

Investigation: Dajun Qian, Shuwei Li, Yuan Tian, Jacob W. Clifford, Brice A. J. Sarver, Tina Pesaran, Chia-Ling Gau, Aaron M. Elliott, Hsiao-Mei Lu.

Methodology: Dajun Qian, Shuwei Li, Yuan Tian, Mary Helen Black.

Project administration: Brice A. J. Sarver.

Resources: Tina Pesaran, Chia-Ling Gau, Aaron M. Elliott, Hsiao-Mei Lu.

Supervision: Mary Helen Black.

Writing – original draft: Dajun Qian, Mary Helen Black.

Writing – review & editing: Dajun Qian, Shuwei Li, Yuan Tian, Jacob W. Clifford, Brice A. J. Sarver, Tina Pesaran, Chia-Ling Gau, Aaron M. Elliott, Hsiao-Mei Lu, Mary Helen Black.

References

1. Quintans B, Ordonez-Ugalde A, Cacheiro P, Carracedo A, Sobrido MJ. Medical genomics: The intricate path from genetic variant identification to clinical interpretation. Appl Transl Genom. 2014; 3(3):60–7. https://doi.org/10.1016/j.atg.2014.06.001 PMID: 27284505.

2. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015; 17(5):405–24. https://doi.org/10.1038/gim.2015.30 PMID: 25741868.

3. Maxwell KN, Hart SN, Vijai J, Schrader KA, Slavin TP, Thomas T, et al. Evaluation of ACMG-guideline-based variant classification of cancer susceptibility and non-cancer-associated genes in families affected by breast cancer. Am J Hum Genet. 2016; 98(5):801–17. https://doi.org/10.1016/j.ajhg.2016.02.024 PMID: 27153395.

4. Campuzano O, Allegue C, Fernandez A, Iglesias A, Brugada R. Determining the pathogenicity of genetic variants associated with cardiac channelopathies. Sci Rep. 2015; 5:7953. https://doi.org/10.1038/srep07953 PMID: 25608792.

5. Karbassi I, Maston GA, Love A, DiVincenzo C, Braastad CD, Elzinga CD, et al. A standardized DNA variant scoring system for pathogenicity assessments in Mendelian disorders. Hum Mutat. 2016; 37(1):127–34. https://doi.org/10.1002/humu.22918 PMID: 26467025.

6. Pesaran T, Karam R, Huether R, Li S, Farber-Katz S, Chamberlin A, et al. Beyond DNA: An integrated and functional approach for classifying germline variants in breast cancer genes. Int J Breast Cancer. 2016; 2016:2469523. https://doi.org/10.1155/2016/2469523 PMID: 27822389.

7. Lindor NM, Guidugli L, Wang X, Vallee MP, Monteiro AN, Tavtigian S, et al. A review of a multifactorial probability-based model for classification of BRCA1 and BRCA2 variants of uncertain significance (VUS). Hum Mutat. 2012; 33(1):8–21. https://doi.org/10.1002/humu.21627 PMID: 21990134.

8. Thompson BA, Goldgar DE, Paterson C, Clendenning M, Walters R, Arnold S, et al. A multifactorial likelihood model for MMR gene variant classification incorporating probabilities based on sequence bioinformatics and tumor characteristics: a report from the Colon Cancer Family Registry. Hum Mutat. 2013; 34(1):200–9. https://doi.org/10.1002/humu.22213 PMID: 22949379.

9. Ruklisa D, Ware JS, Walsh R, Balding DJ, Cook SA. Bayesian models for syndrome- and gene-specific probabilities of novel variant pathogenicity. Genome Med. 2015; 7(1):1–16.

10. Li Q, Wang K. InterVar: Clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. Am J Hum Genet. 2017; 100(2):267–80. https://doi.org/10.1016/j.ajhg.2017.01.004 PMID: 28132688.

11. Polson NG, Scott JG, Windle J. Bayesian inference for logistic models using Pólya–Gamma latent variables. J Am Stat Assoc. 2013; 108(504):1339–49. https://doi.org/10.1080/01621459.2013.829001

12. Choi HM, Hobert JP. The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. Electron J Statist. 2013; 7:2054–64. https://doi.org/10.1214/13-EJS837

13. Crockett DK, Lyon E, Williams MS, Narus SP, Facelli JC, Mitchell JA. Utility of gene-specific algorithms for predicting pathogenicity of uncertain gene variants. J Am Med Inform Assoc. 2012; 19(2):207–11. https://doi.org/10.1136/amiajnl-2011-000309 PMID: 22037892.

14. Li Q, Liu X, Gibbs RA, Boerwinkle E, Polychronakos C, Qu H-Q. Gene-specific function prediction for non-synonymous mutations in monogenic diabetes genes. PLoS One. 2014; 9(8):e104452. https://doi.org/10.1371/journal.pone.0104452 PMID: 25136813.

15. Wang M, Wei L. iFish: predicting the pathogenicity of human nonsynonymous variants using gene-specific/family-specific attributes and classifiers. Sci Rep. 2016; 6:31321. https://doi.org/10.1038/srep31321 PMID: 27527004.

16. Feng BJ. PERCH: A unified framework for disease gene prioritization. Hum Mutat. 2017; 38(3):243–51. https://doi.org/10.1002/humu.23158 PMID: 27995669.

17. Grantham R. Amino acid difference formula to help explain protein evolution. Science. 1974; 185 (4154):862–4. https://doi.org/10.1126/science.185.4154.862 PMID: 4843792

18. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005; 15(7):901–13. Epub 2005/06/21. https://doi.org/10.1101/gr.3577405 PMID: 15965027.

19. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005; 15(8):1034–50. https://doi.org/10.1101/gr.3715005 PMID: 16024819

20. Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, et al. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J Med Genet. 2006; 43(4):295–305. https://doi.org/10.1136/jmg.2005.033878 PMID: 16014699.

21. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protocols. 2009; 4(8):1073–81. https://doi.org/10.1038/nprot.2009.86 PMID: 19561590

22. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009; 25(21):2744–50. https://doi.org/10.1093/bioinformatics/btp528 PMID: 19734154.

23. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009; 25(12):i54–62. https://doi.org/10.1093/bioinformatics/btp190 PMID: 19478016.

24. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009; 19(9):1553–61. https://doi.org/10.1101/gr.092619.109 PMID: 19602639.

25. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010; 20(1):110–21. https://doi.org/10.1101/gr.097857.109 PMID: 19858363.

26. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7(4):248–9. https://doi.org/10.1038/nmeth0410-248 PMID: 20354512.

27. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011; 39(17):e118. https://doi.org/10.1093/nar/gkr407 PMID: 21727090.

28. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012; 7(10):e46688. https://doi.org/10.1371/journal.pone.0046688 PMID: 23056405.

29. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013; 34(1):57–65. https://doi.org/10.1002/humu.22225 PMID: 23033316.

30. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014; 11(4):361–2. https://doi.org/10.1038/nmeth.2890 PMID: 24681721.

31. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014; 46(3):310–5. https://doi.org/10.1038/ng.2892 PMID: 24487276.

32. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet. 2016; 99(4):877–85. https://doi.org/10.1016/j.ajhg.2016.08.016 PMID: 27666373.

33. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet. 2016; 48(2):214–20. https://doi.org/10.1038/ng.3477 PMID: 26727659.

34. Kim S, Jhong JH, Lee J, Koo JY. Meta-analytic support vector machine for integrating multiple omics data. BioData Min. 2017; 10:2. https://doi.org/10.1186/s13040-017-0126-8 PMID: 28149325.

35. Goldgar DE, Easton DF, Deffenbaugh AM, Monteiro ANA, Tavtigian SV, Couch FJ. Integrated evaluation of DNA sequence variants of unknown clinical significance: application to BRCA1 and BRCA2. Am J Hum Genet. 2004; 75(4):535–44. https://doi.org/10.1086/424388 PMID: 15290653.

36. Pruss D, Morris B, Hughes E, Eggington JM, Esterling L, Robinson BS, et al. Development and validation of a new algorithm for the reclassification of genetic variants identified in the BRCA1 and BRCA2 genes. Breast Cancer Res Treat. 2014; 147(1):119–32. https://doi.org/10.1007/s10549-014-3065-9 PMID: 25085752.

37. Thompson BA, Spurdle AB, Plazzer JP, Greenblatt MS, Akagi K, Al-Mulla F, et al. Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. Nat Genet. 2014; 46(2):107–15. https://doi.org/10.1038/ng.2854 PMID: 24362816.

38. Vihinen M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics. 2012; 13(Suppl 4):S2. https://doi.org/10.1186/1471-2164-13-S4-S2 PMID: 22759650

39. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988; 44(3):837–45. Epub 1988/09/01. https://doi.org/10.2307/2531595 PMID: 3203132.