Sun et al. BioData Mining (2024) 17:39
https://doi.org/10.1186/s13040-024-00392-y

BioData Mining

RESEARCH Open Access

MDVarP: modifier ~ disease-causing variant pairs predictor

Hong Sun[1*†], Yunqin Chen[2†] and Liangxiao Ma[3†]

†Hong Sun, Yunqin Chen and Liangxiao Ma equally contributed to this article as the co-first author.

*Correspondence: shpolor@163.com

[1] Shanghai Engineering Research Center for Big Data in Pediatric Precision Medicine, Center for Biomedical Informatics, School of Medicine, Shanghai Children's Hospital, Shanghai Jiao Tong University, Shanghai 200062, China
[2] Shanghai-MOST Key Laboratory of Health and Disease Genomics, Shanghai Institute for Biomedical and Pharmaceutical Technologies (SIBPT), Shanghai 200237, China
[3] Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Science, Shanghai 200031, China

Abstract

Background: Modifiers significantly impact disease phenotypes by modulating the effects of disease-causing variants, resulting in varying disease manifestations among individuals. However, identifying genetic interactions between modifier and disease-causing variants is challenging.

Results: We developed MDVarP, an ensemble model comprising 1000 random forest predictors, to identify modifier ~ disease-causing variant combinations. MDVarP achieves high accuracy and precision, as verified using an independent dataset with published evidence of genetic interactions. We identified 25 novel modifier ~ disease-causing variant combinations and obtained supporting evidence for these associations. MDVarP outputs a class label ("Associated-pair" or "Nonrelevant-pair") and two prediction scores indicating the probability of a true association.

Conclusions: MDVarP prioritizes variant pairs associated with phenotypic modulations, enabling more effective mapping of functional contributions from disease-causing and modifier variants. This framework interprets genetic interactions underlying phenotypic variations in human diseases, with potential applications in personalized medicine and disease prevention.

Keywords: Variants, Interacting unit, Prediction, Phenotypical expression

Introduction

Understanding the impact of genetic factors on human disease phenotypes is crucial for genetic diagnostics. Genetic factors contribute to disease etiology in varying degrees, from single mutation causing monogenicdiseases to complex interactions involving multiple genetic and environmental factors [1]. Understanding the forms of plasticity found in the human genetic architecture may assist the identification of genetic mechanisms responsible for the phenotypic variety of human diseases, and may be exploited as a preclinical knowledge base in precision medicine [2, 3].

Recent advancements in DNA sequencing and computational analysis have significantly advanced the field of human genetics [4, 5], providing insights into the genetic architecture of many diseases [6]. However, as genes do not act in isolation, interpreting genetic variants is complicated by interactions between genetic elements [7, 8]. Genetic modifiers play a crucial role in modulating the effects of deleterious variants, leading to phenotypic variability. Evidence for the action of modifier variants is extensive, both in humans and model organisms, and the effects of modifiers on the phenotypic expression of disease-causing variants can be subtle or profound [9–11]. Assessment of a broader spectrum of genetic contributions to disease risk, including both pathogenic variants and modifier variants, will be important to answer clinical questions about phenotypic variation among individuals [1, 12, 13].

While comprehensive assessment of genetic variations is becoming routine in clinical genetics, interpreting the effects of modifying factors beyond monogenic inheritance remains limited. Interactions between modifier variants and the targeted disease-causing variants is the key characteristic of modifier genetics [1, 12]. It is a significant challenge for experimental identification of genetic interactions that requires comparison of the phenotypic consequences of perturbing two genomic loci either singly or in combination. A high-quality experimental atlas of functional contributions including both disease-causing variants and modifier variants may still be decades into the future, therefore, there is a pressing need for in silico algorithms that predict modifying interactions between variants.

Most computational methods focus on predicting pathogenicity of single variants [14, 15]. Methods developed to predict multilocus genetic patterns have mainly been focused on pathogenic variant combinations [16, 17]. Co-inherited genetic modifiers differ from digenic inheritance in that biallelic mutations are necessary and sufficient to cause the pathology, whereas modifiers contribute to the phenotypic variability of a disease. Current methods rarely address predictions of interactions between modifier variants and disease-causing variants. To address this, we developed the Modifier ~ Disease-causing Variant Pairs predictor (MDVarP), a novel bioinformatics tool that accurately identifies modifier ~ disease-causing variant combinations.

A major limitation of in-depth studies is that the current knowledge on genetic modifiers is much scarcer than that of disease-causing variants. We previously constructed a manually curated genetic modifier database, PhenoModifier [18] and performed systematic analyses on the differences and relationships between modifier variants and disease-causing variants [19]. We called a pair of a modifier locus variant and a disease causing locus variant, that may or may not have a specific phenotypical expression, a modifier ~ disease-causing variant combination. Building on our previous works, we developed MDVarP, which employs variant, gene, and gene pair information to predict modifier ~ disease-causing variant combinations. We validated MDVarP using an independent dataset consisting of experimentally identified modifier ~ disease-causing variant combinations, demonstrating its effectiveness in terms of accuracy and sensitivity.

Methods

Data and annotations

We annotated datasets using variant, gene and gene pair information (Table S1). We calculated population genetical statistics for each variant using the 1000 Genomes Project data (1KGP) [4], including nucleotide diversity (pi), population differentiation (F_{ST}) [20] among the five super-populations (Africans, Admixed Americans, East Asians, Europeans and South Asians), Hardy–Weinberg equilibrium [21] and derived allele frequency

(Daf). A derived allele of an SNP site was defined relative to an ancestral allele which was inferred from an alignment of multiple genomes [22].

The combined annotation dependent depletion (CADD) is an integrative annotation built from diverse genomic features derived from surrounding sequence context, gene model annotations, evolutionary constraint, epigenetic measurements and functional predictions [23], and this gives CADD the capacity to evaluate deleteriousness for both coding and non-coding variants [24]. We used scaled CADD scores (Phred scores, version 1.6) for comparability and used the highest score per variant locus.

Additionally, we used Phylop scores [25] to evaluate conserved or accelerated evolution, information from Pfam [26] to predict variant impact on conserved protein domains. For the gene features, we used the gene haploinsufficiency data from Huang et al. [27, 28] and recessiveness probabilities from MacArthur et al. [29]. For the gene pair features, we exploited the biological distance obtained from Itan et al. [30], protein–protein interaction data from BioGRID [31], and tissue-specific gene interaction data from GIANT [32].

We represented variant combinations as vectors of categorical and numerical features (Table S1), handling missing values as described in Table S2. After annotating both the training and control datasets, we initially had 65 features per entry, which were reduced to 57 after feature selection. All input features were mapped to the human reference build hg19.

Training

We trained MDVarP using 3351 modifier ~ disease-causing variant combinations, randomly split into training and test sets (8:2 ratio). We created 1000 balanced training sets, each containing 2680 modifier ~ disease-causing variant combinations and 2680 randomly selected control pairs.

We used the Random Forest (RF) algorithm [33] implemented in R package randomForest (version 4.7–1.1) [34] as the classifier. Each RF consisted of 500 decision trees using bootstrapping. We optimized the RF classifier using the tuneRF function, with parameters stepFactor = 1.5 and improve = 0.01, to determine the optimal number of variables (mtry) sampled at each split.

Construction of an independent testing dataset

To evaluate the predictor's accuracy and sensitivity, we manually collected a testing dataset from published papers with clear evidence of genetic interactions. We searched PubMed using keywords like 'genetic interaction', 'epistasis', words that describe interactive effects, and their synonyms, and extracted variant pairs meeting the following criteria: one variant is disease-causing or associated while the other is not in the HGMD database. The independent testing dataset consists of 108 modifier ~ disease-causing variant combinations, involving 75 modifier variants, 55 disease-causing variants, 43 modifier genes and 40 disease-causing genes.

Feature selection and interpretation

We applied five times repeated tenfold cross-validation for feature selection on a balanced set with median performance among all the training sets using the rfcv function in 'randomForest' package. Feature elimination order (starting from all features to 1) was ranked by feature importance. We applied log step, *i.e.* reducing a fixed proportion (1/3) of features at each run. The rfcv output is ranked by decreasing feature importance and the number of remaining features left was chosen so as to minimize the error rate.

Classification score and support score

We calculated two scores to indicate the strength of a prediction for two variant loci to be an "Associated-pair": a classification score (C-score) and a support score (S-score) [17]. The C-score is defined as the median of the "Associated-pair" class probabilities over all 1000 independent RFs for the two variant loci combination: $\underset{t=1,\ldots T}{\mathrm{Median}}\{P_t\}$, for each random forest tree t, T is the total number of random forests. The S-score is defined as the percentage of random forests that agree on the "Associated-pair" class for a two variants combination: $\sum_t^T d_t \times 100/T$, for each random forest tree t, T is total number of random forests and $d=1$ when it gives a "Associated-pair" decision, $d=0$ when it gives a "Nonrelevant-pair" decision.

Results and discussion

Collection of modifier ~ disease-causing variant combinations for training the model

The dataset of modifier variants was extracted from our PhenoModifier database [18] and from subsequent collection by manual collection, which contains 3770 records of modifier information, involving 303 disorders, 2183 genetic modifier variants and 943 modifier genes. The dataset of disease-causing variants were extracted from the HGMD database [6], variants being considered as disease-causing if the mutations were flagged as DM (disease-causing mutations).

A specific procedure was used for generating modifier ~ disease-causing variant combinations. Firstly, we got all combinations of modifier variants and disease-causing variants that are involved in the same disease. We collected all the associated diseases for each modifier variant, searched the HGMD database for a match and collected all the corresponding disease-causing variants. As many more tumor-associated variants were annotated both in HGMD and in phenModifier than variants related to other diseases, tumor diseases were excluded from the study to avoid bias. Secondly, we assigned a gene to each variant, meeting the criteria that the variant is located either in the genic region, or in the upstream/downstream regulatory region based on annotations from the Ensembl Variant Effect Predictor [35]. When more than one gene was assigned to a variant locus, gene pairs in which both genes are associated with the same disease based on the annotations in PhenoModifier, HGMD or DisGeNET database [36] were retained in the dataset. Thirdly, we introduced two conditions to decrease the probability of random associations: 1) the two variant loci are genotype dependent, tested by chi-squared statistics based on the 2504 genotypes derived from 1KGP [4] (p value < 0.05), and 2) the two genes are involved in the same pathway if no variation is observed among individuals in 1KGP in either the modifier gene or the disease-causing gene.

The filtered variant combinations were used as the training dataset on which we performed feature selection and parameter tuning for our model. The training dataset contained 3351 pairs of modifier ~ disease-causing variant combinations, involving 54 disorders, 530 modifier variants and 1432 disease-causing variants corresponding to 273

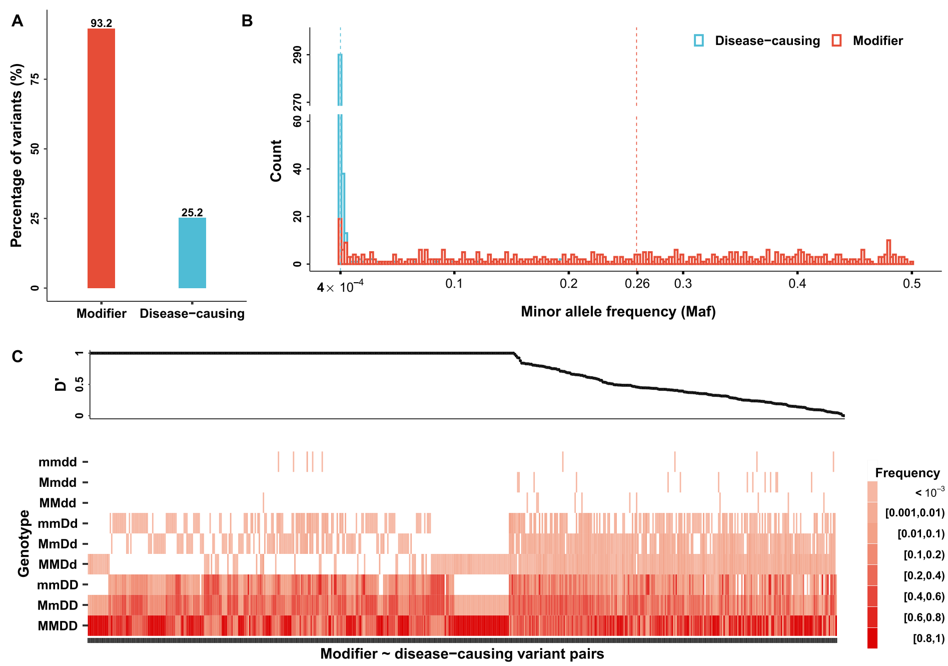

Fig. 1 Overlapping variants and modifier ~ disease-causing variant combinations between the training dataset and 1KGP. **A** Percentage of variants carried by individuals in 1KGP. **B** Histogram of minor allele frequency of variants in 1KGP. The dashed lines indicate median values. **C** Genotype frequencies of the 548 modifier ~ disease-causing variant combinations, for which the mutant alleles were designated in the HGMD database. For each modifier ~ disease-causing variant combinations, M/m designate the allele at the modifier locus and D/d the allele at the disease-causing variant locus; uppercase stands for the wide-type allele and lowercase stands for the mutant allele which is recorded as the allele responsible for the reported disease in the HGMD database. The upper panel shows the linkage disequilibrium statistics D′ for each modifier ~ disease-causing variant combination

modifier genes and 168 disease-causing genes. The dataset contained 3213 pairs in which both the modifier and disease-causing variants are located in genic regions and 138 pairs in which the disease-causing locus is genic while the modifier locus is non-genic.

Pairs of variants in the training data show non-random associations among individuals of 1KGP

Because of experimental challenges, interactions between modifier variants and disease-causing variants are difficult to identify, leading to limited empirical evidence for interacting genes, especially for interacting variants [12]. The ultimate phenotypic manifestation of modifier variants depends on interactions between the specific modifier allele and target disease-causing allele, usually in the context of functional networks [7, 8, 37] and often shows genetic linkage associations [12, 38]. To get as much training data as possible, we focused on all the manually collected modifier variants, introducing restrictions of genetic linkage and functional relevance, and screened for all possible modifier ~ disease-causing variant combinations.

To assess evidence of non-random associations in the training dataset, we further examined variant combinations in 1KGP. For variants included in the training dataset, 93% of the modifier locus variants were found among individuals in 1KGP; for the disease-causing loci, this figure was 25% (Fig. 1A). The median frequency of minor alleles was 0.26 for modifier variants, compared to 4×10^{-4} for disease-causing variants (Fig. 1B). We next examined the frequency in 1KGP of the mutant allele which is recorded as the allele associated with the reported disease in HGMD. In the training dataset, we found 548 modifier ~ disease-causing variant combinations in which the mutant allele was seen in at least one individual in 1KGB at both the modifier locus and the disease-causing locus. The disease-associated mutants were less common than the modifier mutants, and homozygous mutant alleles are even rarer (Fig. 1C). We computed the normalized linkage disequilibrium value D' for all 2-locus haplotypes of modifier ~ disease-causing variant combinations, and found that most allele combinations co-occur in high linkage disequilibrium (Fig. 1C). The linkage disequilibrium analysis thus suggests that there are a number of nonrandom associations of modifier ~ disease-causing variant combinations in the training dataset.

Feature evaluation

Our initial study on the data in the PhenoModifier database[18] revealed that biological features defined at the variant and gene level distinguish very well modifier variations and disease-causing variations [19]. Modifier variations differ from disease-causing variations in many aspects, including population genetics statistics, epigenetic features, evolutionary characteristics and functional properties of the variations. Genes containing modifier variation(s) exhibit a higher probability of being haploinsufficient and have a higher probability of recessive disease causation. The study further suggested that co-expression analysis is an effective methodology to predict functional associations between modifier genes and their potential target genes. It has been reported that gene interaction networks can be exploited as effective ways to identify specific target alleles of modifiers [11, 12].

Based on the research outcome so far, we used 34 types of biological information for classification (Table S1): 1) 28 values coding for variants, including 14 values representing variant consequences derived from the Ensembl Variant Effect Predictor [35], 6 values representing protein domain annotations inferred from the Ensembl Variant Effect Predictor [35], 3 values representing evolutionary conservation [25], CADD Phred score representing the degree of deleteriousness of a variation, and 4 variant population genetics statistics calculated based on 1KGP; 2) 3 gene values measuring the degree of gene haploinsufficiency [27, 28] and recessiveness [29]; 3) 3 gene pair values, including the biological distance obtained from Itan et al. [32], a metric of protein–protein interactions from BioGRID [33], and tissue-specific functional interactions from GIANT [34]. We then annotated our data with this information at the variant, gene, and gene-pair levels, leading to 65 features in total per variant pair.

MDVarP accurately identifies modifier ~ disease-causing variant combinations

To better understand the affecting features and identify more modifier ~ disease-causing variant combinations, we developed the Modifier ~ Disease-causing Variant Pairs predictor (MDVarP), a predictor of modifier ~ disease-causing variant combinations.

To create control datasets of variants, we used variant data collected from four databases: phenoModifier [18], HGMD [6], 1KGP of Phase 3 [4], and DisGeNET [36]. Our previous study found that the genomic loci of modifier variations differ from pathogenic loci in population genetics statistics [19]. We assigned each variant population genetics statistics calculated on the basis of 2504 genotypes from 1KGP [4]. For the variants annotated in the more up-to-date database, *e.g.* gnomAD [39], but not in 1KGP, data on population genetics statistics will be missing. To reduce bias due to the missing data, variants annotated by gnomAD were thus not considered in the construction of the control variant pool. We randomly selected two-variant combinations, and removed the modifier ~ disease-causing variant combinations. The randomly selected variant pairs were merged to form a neutral dataset, with an equal amount of two-variant combinations as well as an equal distribution of genic/non-genic variant location as in the training data. In this way, we extracted 1000 random sets of two-variant combinations as the control training sets.

For each entry in the training dataset, the modifier locus was ranked ahead of the disease-causing locus. To ensure a fair prediction process, we determined the order of variants and genes inside each randomly selected two-variant combination based on their CADD Phred score, with the variant present in the second gene always having the highest CADD Phred score. For the 138 pairs in which one locus is genic and the other non-genic, the non-genic locus was ranked ahead of the genic locus. For the 3213 pairs in which both loci were genic, we used the CADD Phred score to determine the order of the two loci, so that the second locus is the variant with the highest CADD Phred score, that is, with the higher probability of being disease-causing.

We found that there were no significant differences between the training and control datasets for eight of the 65 features (p value > 0.05, Wilcoxon Signed Rank test), including three features encoding consequences of the modifier variant, namely, whether the modifier variant is a non-coding change or located at a canonical splice site or in an intergenic region, and five features encoding consequences of the disease-causing variant, namely, whether the disease-causing variant is a non-coding change or occurred in a 5’UTR, in stop codon, or in a non-coding region or an intergenic region. The feature data was thus reduced to 57 dimensions.

We next developed a random forest algorithm (MDVarP), built on 1000 random forests, to predict whether two variant loci constitute a functional interacting unit that has a specific phenotypical expression. As the accuracy of the machine learning model is directly proportional to the quality of the training data, we introduced filtering rules based on linkage disequilibrium and functional relevance to make sure that our training dataset is of high quality. However, there will nearly always be a trade-off between quality and quantity, and it is possible that the screening may have exclude some valid interactions, leading to fewer training data. To overcome the small sample size problem, we applied five times repeated tenfold cross-validation to assess the prediction confidence.

The annotated information for each two-variant combination of the training and control datasets was then used as training input for the MDVarP. For each variant combination, MDVarP outputs a class label, *i.e.* “Associated-pair” or “Nonrelevant-pair”, and two prediction scores: a classification score (C-score) and a support score (S-score) to indicate how strong the prediction is for two variant loci to be an “Associated-pair”. The C-score is defined as the median of the “Associated-pair” class

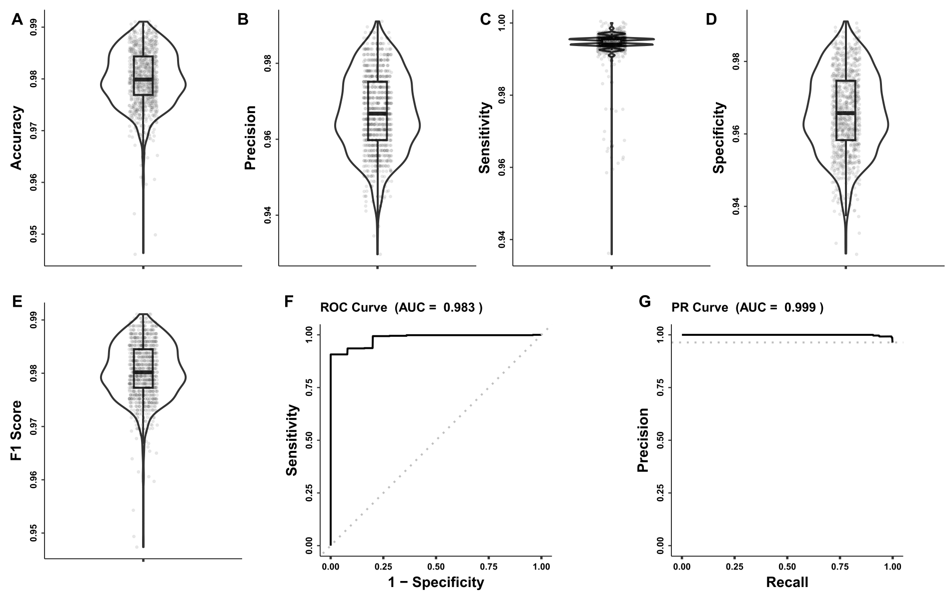

Fig. 2 Performance of MDVarP. MDVarP performance measured by Accuracy (**A**), Precision (**B**), Sensitivity (**C**), Specificity (**D**) and F1 score (**E**), Receive operator characteristic (ROC) curve (**F**) and Precision-Recall (PR) curve (**G**)

probabilities over all 1000 independent models and the S-score is defined as the percentage of models that agree on the "Associated-pair" class (see Methods). Higher C- and S-scores indicate that the predictor is more confident about the classification of two variant loci as "associated".

Adjusting the probability decision thresholds can improve the performance of a multi-label classifier [40], and this could also be applied to our two-label classification. The best prediction performance was obtained for a C-score of 0.49 and an S-score of 0.5, for which the highest accuracy and precision are achieved. Higher C-scores and S-scores provide a stronger indication of association for a two-variant combination. As the predictor is based on a majority vote, two variants will be predicted to be an "Associated-pair" when the S-score is greater than 50 and the C-score is greater than 0.49.

The performance of MDVarP was evaluated on a set of different classification parameters (Fig. 2A-E), *i.e.* accuracy, precision, sensitivity, specificity, and F1 score (a balanced measure between precision and recall), by comparing the predicted label with the actual label. MDVarP successfully distinguished actual modifier ~ disease-causing variant combinations from randomly selected two-variant combinations with an average accuracy of 0.980, an average precision of 0.967, an average sensitivity of 0.994, an average specificity of 0.966, and an average F1 score of 0.980. We further investigated the area under the ROC curve (AUC) value. The AUC parameter quantifies the classification accuracy; the value of 1.0 represents the perfect prediction. MDVarP achieved a high AUC value of up to 0.983 (Fig. 2F). Figure 2G shows the precision and recall when MDVarP is tested on the modifier ~ disease-causing variant combinations together with a set of randomly selected two-variant combinations of the same size, MDVarP gets a much higher AUC of 0.999, indicating an efficient classification. Our results show that MDVarP performs well, achieving a true positive rate of 0.996 and a false positive rate of 0.037.

Validation on independent data confirms the MDVarP's predictive success

To evaluate the performance of MDVarP, we validated the MDVarP on a set of 108 interacting pairs of variants with published evidence of genetic interactions. These independent two variant combinations contain variant pairs relating to 75 modifier variants and 55 disease-causing variants corresponding to 43 modifier genes and 40 disease-causing genes. MDVarP worked well on the validation set, the true positive rate was 89% and most of the new variant pairs (83%) were correctly labeled as "Associated-pair" with a high confidence (S-score > 80).

Candidates of modifier ~ disease-causing variant combinations identified by MDVarP

We used the MDVarP to calculate the C- and S-scores for each of the two-variant combinations in the random sets and we found 25 novel candidates of modifier ~ disease-causing variant combinations under the threshold of the S-score greater than 50 and the C-score greater than 0.49 (Table S3). Among the 25 novel candidates, the two genes of 15 combinations have been reported to be associated with the same disease(s) based on the annotations inHGMD [6] or DisGeNet [36]. The putative novel combination of variants in gene C8A and FAT1 with the highest S- and C-scores are both associated with the autism spectrum disorder and developmental disorder. The variant FAT1:p.V1597E (chr4:187,549,329) was reported to be an autism-causing mutation (DM?) in HGMD [6]. It has been reported that FAT1 plays a contributing role in the development of autism [41], and that the carbonyl level of C8A protein product is significantly higher in the plasma of autistic children than in non-autism controls [42].

To obtain additional information supporting the predictions, we further investigated the characteristics which could indicate the associations in the context of functional networks [7, 8, 37] and/or genetic linkage association [12, 38]. By searching the MSigDB database [43] we checked whether there was any functional gene set in which the two associated genes co-existed, and we found this to be the case for eleven gene pairs (Table S3). Evidence that supported the relationship between the disorder and the potential function associated with the two genes was also extracted from several publications (Table S3). We next computed the normalized linkage disequilibrium value D' for all the 2-locus haplotypes of the 25 candidate variant pairs based on the 1KGP Data, and found that seven allele combinations co-occur in high linkage disequilibrium with a D' value equal to 1 (Table S3). Taken together, the investigation on co-existing functional gene set as well as the linkage disequilibrium analysis added additional support for non-random associations for sixteen of the 25 candidate pairs.

Although more functional studies are needed for verification, these preliminary results suggest that our approach for candidate modifier ~ disease-causing variant combinations is promising.

The synergy of different biological features

We trained MDVarP with a random forest algorithm on combined molecular features derived from 57 annotations at the variant, gene, and gene-pair level to identify which variant combinations are potentially interacting. The importance of the original 57 features was analyzed through a feature selection procedure. The result suggests that a

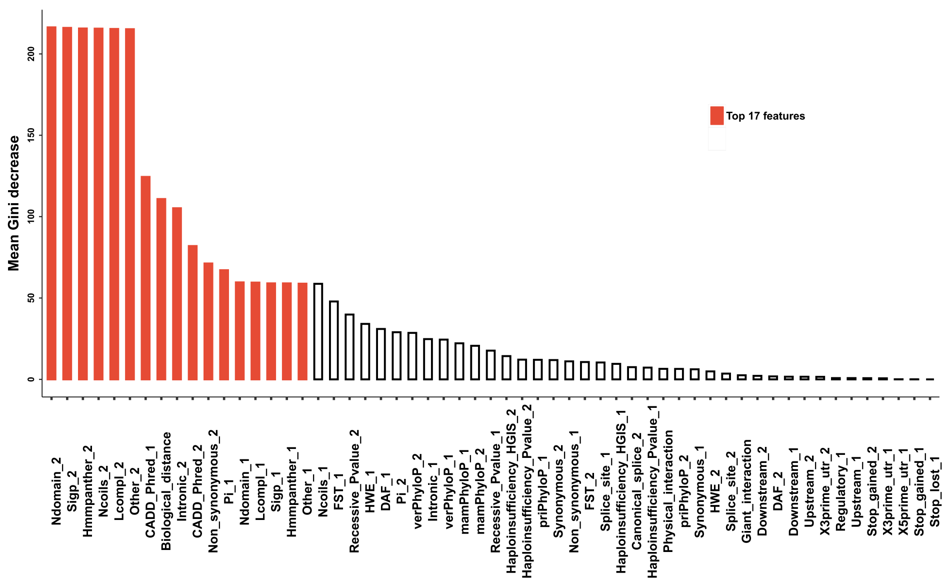

Fig. 3 Mean Gini decrease over all the 1000 predictors of the MDVarP using the training dataset and nonrelevant control dataset. The 17 biological features sufficient for making high-quality predictions are labeled in red

subset of 17 biological features (labeled in red in Fig. 3) is sufficient for making high-quality predictions, giving an error rate of 0.0089, yet it worth noting that the lowest error rate of 0.0077 was obtained when using all the original 57 features (Table S4).

For each of these 57 features, we calculated a mean Gini decrease score, which quantifies the importance of each feature in classification, with higher values indicating higher importance. As shown in Fig. 3, protein domain annotations of the disease-causing variants, CADD Phred score, nucleotide diversity (pi) of the modifier variants, the probability of the gene being recessively disease-causing and the biological relatedness between genes, are the most important features for separating associated pairs from non-relevant pairs. We took a closer, more detailed look at the top 25 features out of the 57 features that are ranked in the increasing order of feature weights generated by Gini decrease score, and found that all of the features contribute to the differentiation between the associated pairs in the training dataset and non-relevant pairs in the control data sets (Fig. S1).

By adding features, we see an improvement in classification, and it is the addition of the complete biological information that provides the best performance (Fig. 4). Therefore, it is the synergy of all features that results in the most optimal classification. This advantage of integrating information from a large number of data points, which is difficult to achieve for a human actor, is what underlies the efficiency of a tool like the MDVarP, compared with solely relying on combinations based on single-variant information.

Evidence suited to evaluate the cooperation between modifier and disease-causing variants is typically less available compared to pathogenicity evaluation of single variants. Experimental research has only produced a limited amount of empirical evidence for interacting variants. For these reasons, there is a particularly strong motivation to develop computational methods for inferring the co-operations between modifier variant and disease-causing variant.

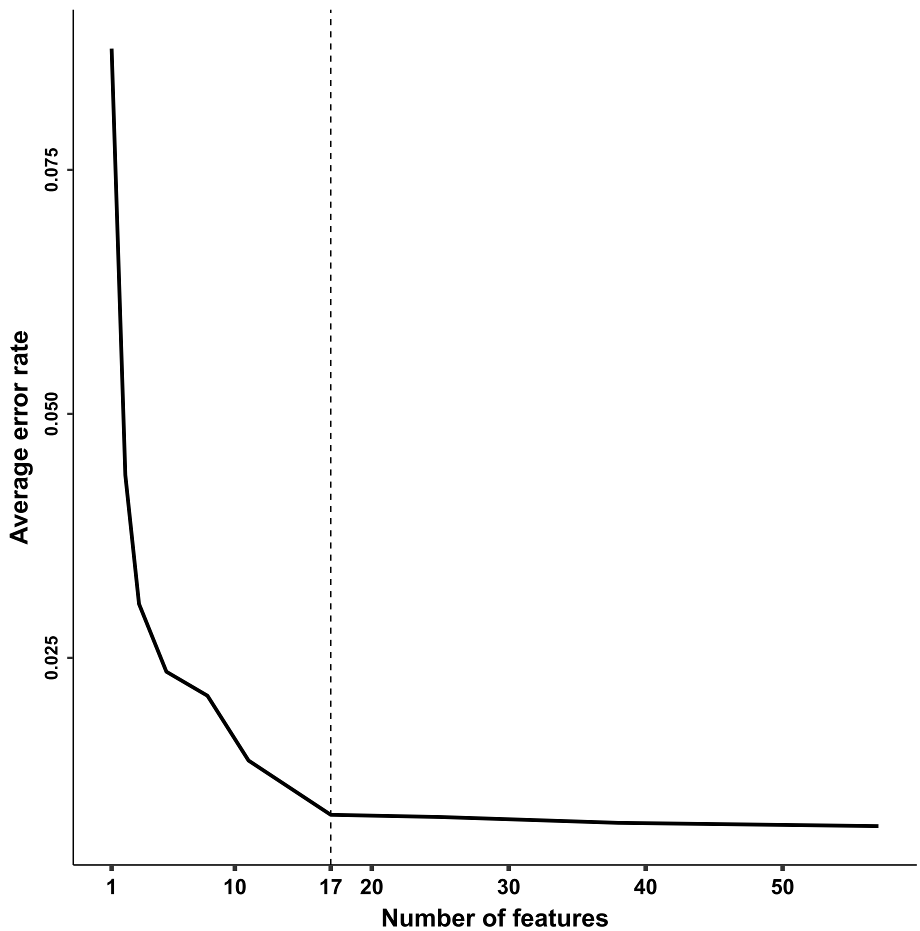

Fig. 4 Effect of the number of features on the error rate obtained by random forest cross-validation for feature selection. The x-axis represents the number of features included for prediction during each elimination procedure, while the y axis represents the average error rate during the five times repeated tenfold cross-validation for that number of features

MDVarP offers a novel way to explore the modifier-disease variant interactions relating to a patient's phenotype, nevertheless, room for improvement still exists. The training dataset of modifier ~ disease-causing variant combinations depends on the manually curated data, which is slow to come. Increasing model complexity or using more informative features will contribute to the future to improve MDVarP performance and reduce overfitting due to limited training data. More thorough feature engineering such as mutation burden analysis may help identifying modifiers as some modifier effects are latent depending on genetic background and environmental context [12, 44]. It is also hard to give good predictions for variant pairs that have "anomalous features" previously not reported. However, improvement of MDVarP is likely to go hand in hand with experimental work as it is being reported in the literature. Thus, "anomalous features" that are found to be common to a set of modifier-gene combinations can be included as prediction parameters, and will cease to be anomalous, whereas as anomalous features that are specific to only one (or some very few) modifier-gene combinations will possibly always be out of reach for a tool like MDVarP. Thus, the limitations posed inadequate knowledge of the defining features of modifier-gene combination are likely to be further relaxed as more data on biological features of variants become available. In addition, using deep learning framework based on multiple sectional views with different learning strategies like transfer learning would also help improving the performance in the future tasks.

MDVarP is a tool that is trained on known modifier genes associated with mendelian disorders for predicting modifier ~ disease-causing variant combinations. Predicting modifier effect for complex, non-mendelian disease is a possible future development for a tool like MDVarP. However, modifier genetics for common complex diseases is full of significant challenge when compared with rare mendelian disorders, as modifier effects on complex, non-mendelian diseases are likely to be far more subtle. A possible future expansion of MDVarP to deal with complex, non-mendelian diseases will need to consider that the variant filtering criteria will differ from the "strict" criteria that are necessary to identify disease-causing variants in rare mendelian diseases. Similarly, while it is widely assumed that genes involved in the same disease are likely to belong to the same molecular pathway or biological process, and this may not necessarily apply to complex diseases. Thus, generalizing MDVarP into a tool for prediction of modifier effects of complex, non-mendelian disease is likely to be a process that will require different resources, from training dataset to features, that help training machine learning model to make classification.

Conclusion

Here, we presented the MDVarP, a predictor of modifier ~ disease-causing variant combinations. MDVarP is precise and sensitive both in cross validation settings (89% correct predictions) and also when tested on independent data. MDVarP tackles a crucial but under-explored area, and is reliable and lightweight for identifying potential interactions between modifier and disease-causing variants, thus providing a framework for interpreting genetic interactions related to phenotypic variations in human diseases. The predictive ability of MDVarP would be further improved with the advent of new data, the inclusion of additional biological information and the adoption of complex machine learning strategy.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13040-024-00392-y.

Supplementary Material 1.

Authors' contributions

H.S. designed the study; Y.C. developed the program; L.M. performed data analysis and prepared the figures; H.S. and Y.C. wrote the manuscript. All authors reviewed the manuscript.

Funding

This work was supported by funding from National Natural Science Foundation of China (32070661 and 32100531).

Availability of data and materials

The R implementation of MDVarP is present at https://github.com/BioAI2024/MDVarP/releases/tag/v1.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Competing interests
The authors declare no competing interests.

Received: 20 August 2024 Accepted: 28 September 2024
Published online: 08 October 2024

References
1. Zschocke J, Byers PH, Wilkie AOM. Mendelian inheritance revisited: dominance and recessiveness in medical genetics. Nat Rev Genet. 2023;24(7):442–63.
2. Ashley EA. Towards precision medicine. Nat Rev Genet. 2016;17(9):507–22.
3. Kutalik Z, Whittaker J, Waterworth D, Beckmann JS, Bergmann S. Novel method to estimate the phenotypic variation explained by genome-wide association studies reveals large fraction of the missing heritability. Genet Epidemiol. 2011;35(5):341–9.
4. The Genomes Project C. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
5. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91.
6. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, Hussain M, Phillips AD, Cooper DN. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet. 2017;136(6):665–77.
7. Sackton TB, Hartl DL. Genotypic Context and Epistasis in Individuals and Populations. Cell. 2016;166(2):279–87.
8. Mackay TFC. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. Nat Rev Genet. 2014;15(1):22–33.
9. Turner H, Jackson L. Evidence for penetrance in patients without a family history of disease: a systematic review. Eur J Hum Genet. 2020;28(5):539–50.
10. Niemi MEK, Martin HC, Rice DL, Gallone G, Gordon S, Kelemen M, McAloney K, McRae J, Radford EJ, Yu S, et al. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. Nature. 2018;562(7726):268–71.
11. Kingdom R, Wright CF. Incomplete Penetrance and Variable Expressivity: From Clinical Studies to Population Cohorts. Front Genet. 2022;13:920390.
12. Riordan JD, Nadeau JH. From Peas to Disease: Modifier Genes, Network Resilience, and the Genetics of Health. Am J Hum Genet. 2017;101(2):177–91.
13. Schwartz MB, Williams MS, Murray MF. Adding protective genetic variants to clinical reporting of genomic screening results: Restoring balance. JAMA. 2017;317(15):1527–8.
14. Schmidt A, Röner S, Mai K, Klinkhammer H, Kircher M, Ludwig KU. Predicting the pathogenicity of missense variants using features derived from AlphaFold2. Bioinformatics. 2023;39(5):btad280.
15. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, Pritzel A, Wong LH, Zielinski M, Sargeant T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381(6664):eadg7492.
16. Versbraegen N, Gravel B, Nachtegael C, Renaux A, Verkinderen E, Nowé A, Lenaerts T, Papadimitriou S. Faster and more accurate pathogenic combination predictions with VarrCoPP2.0. BMC Bioinformatics. 2023;24(1):023–05291.
17. Papadimitriou S, Gazzo A, Versbraegen N, Nachtegael C, Aerts J, Moreau Y, Van Dooren S, Nowé A, Smits G, Lenaerts T. Predicting disease-causing variant combinations. Proc Natl Acad Sci U S A. 2019;116(24):11878–87.
18. Sun H, Guo Y, Lan X, Jia J, Cai X, Zhang G, Xie J, Liang Q, Li Y, Yu G. PhenoModifier: a genetic modifier database for elucidating the genetic basis of human phenotypic variation. Nucleic Acids Res. 2020;48(D1):D977–82.
19. Sun H, Lan X, Ma L, Zhou J. Revealing modifier variations characterizations for elucidating the genetic basis of human phenotypic variations. Hum Genet. 2022;141(6):1223–33.
20. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. Evolution. 1984;38(6):1358–70.
21. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet. 2005;76(5):887–93.
22. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
23. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886–94.
24. 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.
25. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20(1):110–21.
26. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021;49(D1):D412–9.
27. Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and predicting haploinsufficiency in the human genome. PLoS Genet. 2010;6(10):1001154.
28. Steinberg J, Honti F, Meader S, Webber C. Haploinsufficiency predictions without study bias. Nucleic Acids Res. 2015;43(15):22.
29. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012;335(6070):823–8.

30. Itan Y, Mazel M, Mazel B, Abhyankar A, Nitschke P, Quintana-Murci L, Boisson-Dupuis S, Boisson B, Abel L, Zhang S-Y, et al. HGCS: an online tool for prioritizing disease-causing gene variants by biological distance. BMC Genomics. 2014;15(1):256.
31. Oughtred R, Rust J, Chang C, Breitkreutz BJ, Stark C, Willems A, Boucher L, Leung G, Kolas N, Zhang F, et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. 2021;30(1):187–200.
32. Wong AK, Krishnan A, Troyanskaya OG. GIANT 2.0: genome-scale integrated analysis of gene networks in tissues. Nucleic Acids Res. 2018;46(W1):W65–70.
33. Breiman L. Random Forests. Mach Learn. 2001;45(1):5–32.
34. Wiener A. Classification and Regression by randomForest. R News. 2002;2(3):18–22.
35. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(1):016–0974.
36. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong LI. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2020;48(D1):D845–55.
37. John M, Grimm D, Korte A. Predicting Gene Regulatory Interactions Using Natural Genetic Variation. Methods Mol Biol. 2023;2698:3354–3350_3318.
38. McGee TL, Devoto M, Ott J, Berson EL, Dryja TP. Evidence that the penetrance of mutations at the RP11 locus causing dominant retinitis pigmentosa is influenced by a gene linked to the homologous RP11 allele. Am J Hum Genet. 1997;61(5):1059–66.
39. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43.
40. Fan R-E, Lin C-J. A Study on Threshold Selection for Multi-label Class. 2007. Available at: https://www.csie.ntu.edu.tw/~cjlin/papers/threshold.pdf.
41. Frei JA, Brandenburg C, Nestor JE, Hodzic DM, Plachez C, McNeill H, Dykxhoorn DM, Nestor MW, Blatt GJ, Lin YC. Postnatal expression profiles of atypical cadherin FAT1 suggest its role in autism. Biol Open. 2021;10(6):8.
42. Feng C, Chen Y, Pan J, Yang A, Niu L, Min J, Meng X, Liao L, Zhang K, Shen L. Redox proteomic identification of carbonylated proteins in autism plasma: insight into oxidative stress and its related biomarkers in autism. Clin Proteomics. 2017;14(1):2.
43. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
44. Paaby AB, Rockman MV. Cryptic genetic variation: evolution’s hidden substrate. Nat Rev Genet. 2014;15(4):247–58.

Publisher’s Note