Human Genetics (2019) 138:673–679
https://doi.org/10.1007/s00439-019-02021-9

ORIGINAL INVESTIGATION

Gene pathogenicity prediction of Mendelian diseases via the random forest algorithm

Sijie He[1,2,3,4] · Weiwei Chen[1,2] · Hankui Liu[2,3,4] · Shengting Li[2] · Dongzhu Lei[5] · Xiao Dang[2,3,4] · Yulan Chen[2,3,4] · Xiuqing Zhang[2,4] · Jianguo Zhang[2,3,4]

Received: 29 October 2018 / Accepted: 19 April 2019 / Published online: 8 May 2019

Abstract
The study of Mendelian diseases and the identification of their causative genes are of great significance in the field of genetics. The evaluation of the pathogenicity of genes and the total number of Mendelian disease genes are both important questions worth studying. However, very few studies have addressed these issues to date, so we attempt to answer them in this study. We calculated the gene pathogenicity prediction (GPP) score by a machine learning approach (random forest algorithm) to evaluate the pathogenicity of genes. When we applied the GPP score to the testing gene set, we obtained an accuracy of 80%, recall of 93% and area under the curve of 0.87. Our results estimated that a total of 10,384 protein-coding genes were Mendelian disease genes. Furthermore, we found the GPP score was positively correlated with the severity of disease. Our results indicate that GPP score may provide a robust and reliable guideline to predict the pathogenicity of protein-coding genes. To our knowledge, this is the first trial to estimate the total number of Mendelian disease genes.

Introduction

Mendelian diseases (MD), also called as single-gene disorders, refer to the phenotypes caused by a mutation (or mutations) in a single gene (Bamshad et al. 2011). In recent years, the identification of pathogenic genes for MD develops rapidly. Wenger et al. indicated that on average 266 OMIM phenotypes with known molecular bases and 241 gene–disease associations for MD have been reported annually (Wenger et al. 2017). Currently, 5316 single-gene disorders have been reported and 3666 genes are responsible for them (from OMIM, updated March 8th, 2019, http://omim.org/statistics/geneMap). Two of the key points in MD research are how to efficiently evaluate the pathogenicity of each gene and how many genes among the ~20,000 protein-coding genes may cause MD.

Variant-level prediction methods are widely used in the identification of MD genes. There are dozens of tools for performing variant-level prediction, such as SIFT (Ng and Henikoff 2003), Polyphen (Adzhubei et al. 2010), GERP++ (Davydov et al. 2010), MutationTaster (Schwarz et al. 2014), CADD (Kircher et al. 2014) and REVEL (Ioannidis et al. 2016). In contrast to the numerous variant prediction tools, only a few tools focus on gene-level prediction, such as RVIS (Petrovski et al. 2013), DNE (Samocha et al. 2014), EvoTol (Rackham et al. 2015) and GDI (Itan et al. 2015), which may play an irreplaceable supplementary role in MD gene identification. These existing gene-level prediction tools mainly rely on a single characteristic of genes, i.e., the intolerance to functional variants, to make predictions. Although the intolerance values produced by these tools may reflect the pathogenicity of genes, they did not give a cutoff for deeming them pathogenic.

Sijie He and Weiwei Chen contributed equally to the work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00439-019-02021-9) contains supplementary material, which is available to authorized users.

✉ Xiuqing Zhang
zhangxq@genomics.cn

✉ Jianguo Zhang
zhangjg@genomics.cn

1 BGI Education Center, University of Chinese Academy of Sciences, Shenzhen 518083, China

2 BGI-Shenzhen, Shenzhen 518083, China

3 Shenzhen Key Laboratory of Neurogenomics, BGI-Shenzhen, Shenzhen 518083, China

4 China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China

5 Center of Prenatal Diagnosis, ChenZhou No. 1 People's Hospital, Hunan 423000, China

The pathogenicity of genes is decided by more than one kind of factor. If we combine multiple characteristics instead of just one to do the prediction, we may get a better result. Machine learning is an efficient method for classification and prediction. It may combine many features from the studied objects to provide a comprehensive judgment for new objects. The evaluation of variant-level prediction tools indicates that the ones using machine learning tend to show better performance. The random forest algorithm is a well-developed machine learning model that can handle multiple features and tolerate samples with missing features, which is very suitable for predicting the pathogenicity of genes.

In this study, we applied a machine learning approach (random forest algorithm) that combined 201 gene-level characteristics to produce the gene pathogenicity prediction (GPP) score. The GPP score showed better performance than residual variation intolerance score (RVIS) and gene damage index (GDI) in distinguishing MD genes. Our results estimated that a total of 10,384 protein-coding genes are MD genes. The characteristics of GPP score were also analyzed. Gene dominance prediction (GDP) score and gene recessiveness prediction (GRP) score were calculated by the same method to evaluate the inheritance model of MD genes. The three kinds of scores were integrated into a list called the gene catalog of Mendelian diseases (GCMD). Our results may be applied to MD research, especially for the identification of pathogenic genes. This is the first trial to provide a clear cutoff for judging MD genes and to estimate the total number of them.

Methods

Data collection and gene standardization

Gene sets data were extracted from ClinVar (Landrum et al. 2018), OMIM (http://omim.org/), MalaCards (Rappaport et al. 2017), the "super hero project" (Chen et al. 2016) and our internal database. Variants data were extracted from 1000 Genomes Project (KG) (Sudmant et al. 2015) and Genome Aggregation Database (containing whole-genome data (GAD) and whole-exome data (ExAC)) (Lek et al. 2016). Gene-level characteristics were extracted from Refgene, STRING (Szklarczyk et al. 2017), HomoloGene, GTEx, ANNOVAR (Wang et al. 2010), DOMINO (Quinodoz et al. 2017) and our internal data. The annotation results of variants and genes were involved as well. Data from different sources may have inconsistent gene names, which may lead to some confusion. In our study, all the gene names from different data sets were standardized according to the HUGO Gene Nomenclature Committee (HGNC) (see Supplementary methods). The data used in our study and their download links are listed in Supplementary Table 1.

Gene sets selection and gene-level characteristics filtration

To explore a comprehensive method to predict the pathogenicity of genes, we applied a machine learning approach. The gene sets and gene-level characteristics used in machine learning were produced as follows.

Gene sets selection: The loss-of-function (LOF) variant tolerant genes were obtained from the KG, GAD and ExAC databases, among which 630 high-quality genes were used as the training set of negative genes, and the remaining 850 genes were used as the testing set of negative genes. To ensure a balance between the number of positive and negative genes used in our model, we extracted and selected 630 and 850 MD genes as the training and testing sets of positive genes, respectively.

Gene-level characteristic filtration: We extracted many gene-level characteristics from ANNOVAR, STRING, Refgene, HomoloGene, GTEx and DOMINO, as well as several characteristics calculated by ourselves, including the gene intolerance scores, variant damaging scores, conservation scores, expression data and etc. In total, we enrolled 405 characteristics. We excluded genes with missing values for more than 50% of all characteristics. If any gene missed value for a characteristic, we used the median value of other genes on this characteristic to fill it. Then, we performed some filtration and finally kept 201 characteristics (Supplementary Table 2).

The details are provided in the Supplementary methods.

Gene pathogenicity prediction score produced by random forest algorithm

After the gene sets and gene-level characteristics were obtained, we built a model of random forest algorithm and adjusted the parameters (ntree and mtry) with the training gene sets and gene-level characteristics. We applied the model to the testing gene set to evaluate its effect and analyzed the 10 most important characteristics of the model. Then, the GPP score of all protein-coding genes were calculated by the model, and the number of predicted pathogenic and non-pathogenic genes (Np and Nn) of MD was obtained according to their scores (the default cutoff of the score was 0.5). Considering the positive predicted value (PPV) and negative predicted value (NPV), the real number of MD genes (Nm) was calculated by the formula: $Nm = Np \times PPV + Nn \times (1 - NPV)$. Furthermore, we collected several disease gene sets and analyzed the prediction accuracy and score distribution of them. The detailed information of gene sets and the gene selection method are listed in Supplementary Table 3.

Prediction of dominant and recessive models of Mendelian disease genes

To distinguish dominant genes and recessive genes, we applied the same algorithm to calculate the GDP and GRP scores, respectively. Briefly, we collected genes from the autosome that showed dominant, recessive and both inheritance patterns. In our model, 1243 positive genes and 1584 negative genes were used for calculating GDP score, while 1985 positive genes and 842 negative genes were used for calculating GRP score. After the filtration, 183 gene-level characteristics were kept for the calculation of both scores. We did not divide the genes into a training and a testing set, and we applied 4× cross-testing to do the calculation (see Supplementary methods).

The two scores (GDP and GRP) and GPP score are integrated in Supplementary Table 4.

Results

The performance of GPP score produced by the random forest algorithm

We applied a machine learning approach (random forest algorithm) to calculate gene pathogenicity prediction (GPP) score. After the adjustment of parameters, 201 gene-level characteristics were used, and we found ntree = 500 and mtry = 46 to be suitable parameters. When we applied the GPP score produced by this model to the testing gene set, we obtained an accuracy of 80%, recall of 93%, FPR of 34% and FNR of 7% (the default cutoff 0.5 was used). The receiver operating characteristic (ROC) curve showed an area under the curve (AUC) of 0.87. We compared the performance of the GPP score with that of RVIS and GDI on the testing gene set and found that the GPP score performed significantly better than the other two tools (Fig. 1a). The ten most important characteristics determined by the mean decrease in the Gini coefficient of the model mainly belonged to two categories (gene intolerance scores and variant damage prediction results) (Fig. 1b). Then, we evaluated the performance of the

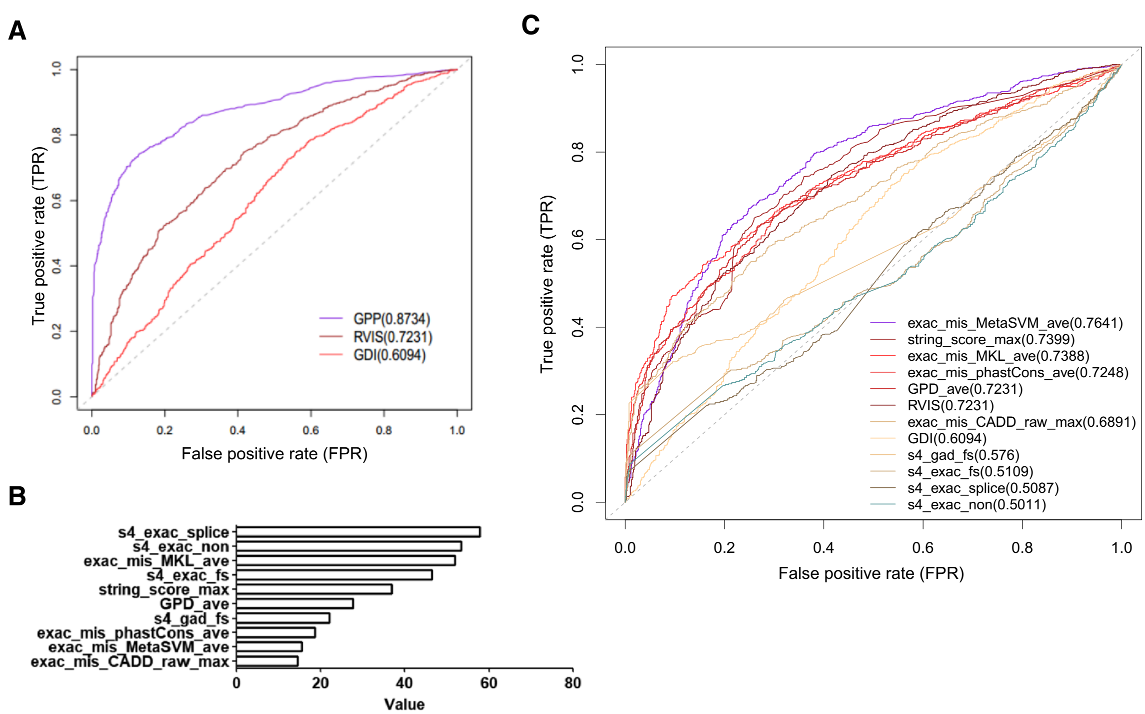

Fig. 1 The ROC curve of GPP score and the ten most important characteristics of the model. **a** The ROC curve on the testing gene set. The AUCs of the GPP score and two similar tools, RVIS and GDI, are 0.87, 0.72 and 0.61, respectively. **b** The ten most important characteristics identified by the mean decrease in the Gini coefficient, ranked by their values. GPD_ave belonged to gene potential damage score, s4_exac_splice, s4_exac_non, s4_exac_fs and s4_gad_fs belonged to gene intolerance score, exac_mis_MKL_ave, exac_mis_phastCons_ave, exac_mis_MetaSVM_ave and exac_mis_CADD_raw_max belonged to variant damage prediction result, string_score_max belonged to gene interaction score. **c** The ROC curve on the testing gene set. The AUCs of the ten most important characteristics and 2 other tools (GDI and RVIS) are listed on the right, ranked by their values

ten characteristics on the testing gene set, respectively. The ROC curves showed that the AUC ranged from 0.5 to 0.76 (Fig. 1c), which was much lower than that of the GPP score.

We collected reported MD genes and susceptible genes from OMIM, LOF variants intolerant genes from ExAC and known pathogenic genes from several resources to evaluate the prediction accuracy of the GPP score. The genes without GPP score were excluded. The cutoff to distinguish pathogenic and non-pathogenic genes of MD was set as 0.5. MD genes showed higher accuracy (92.8%) than susceptible genes (74.5%), LOF variants intolerant genes showed even higher accuracy (96.7%) and pathogenic genes showed a bit lower accuracy (87.4%) (Table 1). It was comprehensible that susceptible genes showed lower prediction accuracy because a part of them were not MD genes. The accuracy of pathogenic gene set was not that high because a part of pathogenic genes of non-Mendelian diseases were involved. LOF intolerance genes showed the highest prediction accuracy because gene intolerance scores were most important characteristics used in our model. The results indicated that LOF variants intolerant genes in ExAC may contain massive potential MD genes waiting to be confirmed.

The estimation for the number of Mendelian disease genes and the GO analysis

Among 18,226 high-quality protein-coding genes involved in our model, 13,401 genes were predicted to be pathogenic (the default cutoff 0.5 was used) (Supplementary Table 4). Considering the PPV and NPV of our model, we estimated that approximately 10,384 genes were MD genes.

We extracted 2984 reported MD genes (reported_M) from OMIM, 13,401 predicted MD genes (predicted_M) and 4825 predicted non-MD genes (predicted_NM) from our prediction result, then we performed gene ontology (GO) analysis of them. The number of genes with at least one GO entry of these gene sets was 2974, 13,154 and 4377, respectively. The average numbers of GO entries per gene of the three gene sets were 22.9, 17.2 and 8.5, respectively.

Table 1 Prediction accuracy of GPP score for several kinds of pathogenic gene sets

Gene set	Gene number	Percent of genes in training set (%)	Prediction accuracy (%)
LOF intolerance genes	3183	4.5	96.7
Mendelian disease genes in OMIM	2946	20.4	92.8
Pathogenic genes	4893	12.7	87.4
Susceptible genes	98	0	74.5

In general, the reported_M and the predicted_M sets showed little difference, while the predicted_NM set showed larger difference compared to them (Supplementary Fig. 1). We found that the predicted_M and predicted_NM sets showed little difference compared to the reported_M set in some GO pathways, such as membrane part (GO:0044425), extracellular region (GO:0005576) and immune system process (GO:0002376). In particular, the predicted_NM set showed higher percentage of enrichment compared to the other two sets in two pathways, which were molecular transducer activity (GO:0060089) and cell killing (GO:0001906).

GPP score is positively correlated with the severity of disease

Two peaks were observed in the GPP score distribution (Fig. 2a). The proportion of overlap between the genes with different GPP scores and reported MD genes was analyzed. We found that genes with higher scores showed a higher proportion of overlap with reported MD genes, and the proportion ranged from 4.4% (GPP scores under 0.5) to 30.8% (GPP scores above 0.9) (Fig. 2b). We also examined the distribution of predicted and reported MD genes on each chromosome, but we did not find obvious hot or cold spots (Supplementary Fig. 2).

The distribution of the GPP score for several kinds of pathogenic gene set was observed. We found that genes with known inheritance models showed significantly higher scores than susceptibility genes (Fig. 3a). We found the scores of LOF genes were higher than those of gain-of-function (GOF) genes (Wilcoxon rank sum test, $p \leq 0.001$) (Fig. 3b). To determine whether the GPP score could reflect the age of onset or severity of the disease, we examined the score distribution of several pathogenic gene sets accordingly. Among several kinds of neurological disease, intellectual disability (ID) and autism spectrum disorder (ASD) tended to show an early age of onset, the onset of schizophrenia (SCZ) mainly occurred in late adolescence, Alzheimer's disease (AD) often occurred in the elderly and epilepsy (EP) occurred in a wide range of age. The scores of these gene sets showed little difference (Fig. 3c). The pathogenic genes of diseases with severe phenotypes (neuromuscular disease, metabolic disease, congenital heart disease and hereditary tumors) showed higher scores, while the genes of diseases with mild phenotypes (skin diseases such as psoriasis and ichthyosis) showed lower scores. These results indicated GPP score was positively correlated with the severity of disease. Genes of complex diseases (hypertension, obesity and diabetes) and male infertility showed scores between those of severe and mild diseases (Fig. 3d).

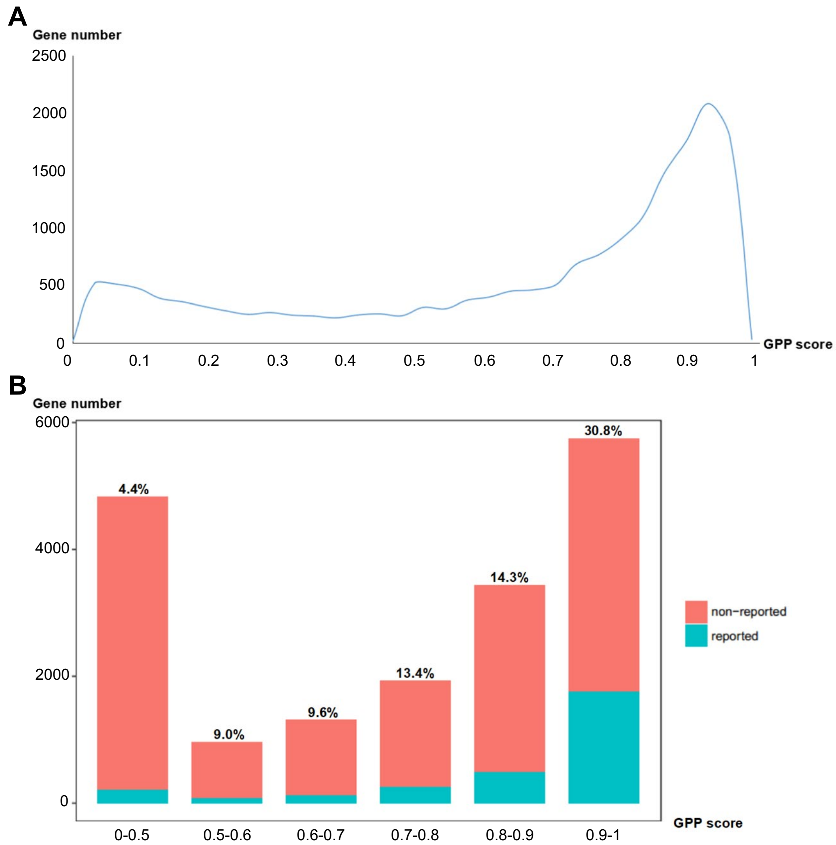

Fig. 2 Distribution of the GPP score and proportion of reported Mendelian disease genes for different scores. **a** The distribution of the GPP score of all protein-coding genes shows two peaks, and the majority of genes are predicted to be MD genes. The *X*-axis represents the GPP score, while the *Y*-axis represents the number of genes. **b** Genes are divided into six sets by the GPP score. Each bar consists of two parts. The blue part shows reported MD genes, and the red part shows non-reported ones. The proportion of reported genes is listed at the top of each bar

The dominant and recessive patterns for pathogenic genes

We applied the same machine learning approach to calculate GDP score and GRP score to predict dominant and recessive genes. After adjustment of the parameters, we obtained accuracy of 75%, recall of 64% and AUC of 0.81 for the GDP score and accuracy of 81%, recall of 97% and AUC of 0.81 for the GRP score. We also checked the score distribution of all MD genes and found that more genes tended to follow the recessive model (Supplementary Fig. 3). Our prediction results showed that there were 4942 autosomal dominant genes and 10,041 autosomal recessive genes. Considering the PPV and NPV, we estimated that the numbers of real dominant and recessive genes were 5756 and 8502, respectively.

Discussion

Our results estimate that 10,384 genes are MD genes, which is much more than the currently reported number. As far as we know, this is the first time to estimate the total number of MD genes. We further tried different cutoffs of GPP score to calculate the FPR and FNR of our model by the testing gene sets. When we chose 0.65 as cutoff, we obtained the lowest summation of false positive rate and false negative rate, and in this situation the number of MD genes was 10,173. We found that the number of MD genes was relatively stable (from 9155 to 10,479) under different cutoffs (Supplementary Table 5), which may indicate that the real number of MD genes is around 10,000. The estimated number may indicate that there are many pathogenic genes (or lethal genes) of MD waiting to be discovered. Lethal genes are those with important function and the dysfunction of them will lead to death before birth, which makes them rarely identified in case-dependent disease research. We extracted some lethal genes from the Mouse Genome Informatics (MGI) database. The quartile value of them was 0.888, which may suggest a cutoff of lethal genes.

Analysis of the score distributions of different gene sets showed that the GPP score was positively correlated with disease severity. Infertility is a special kind of disease, which may have a strong influence on the next generation but little influence on the patients themselves. We also analyzed the diseases with different ages of onset and found no significant difference between the scores of these gene sets. These results indicate that the GPP score may reflect disease severity, where severity means the degree of threat to the health or survival of the individuals themselves but not the next generation. The gene sets we used may contain some non-Mendelian disease genes that influence the results. So, we

Human Genetics (2019) 138:673–679

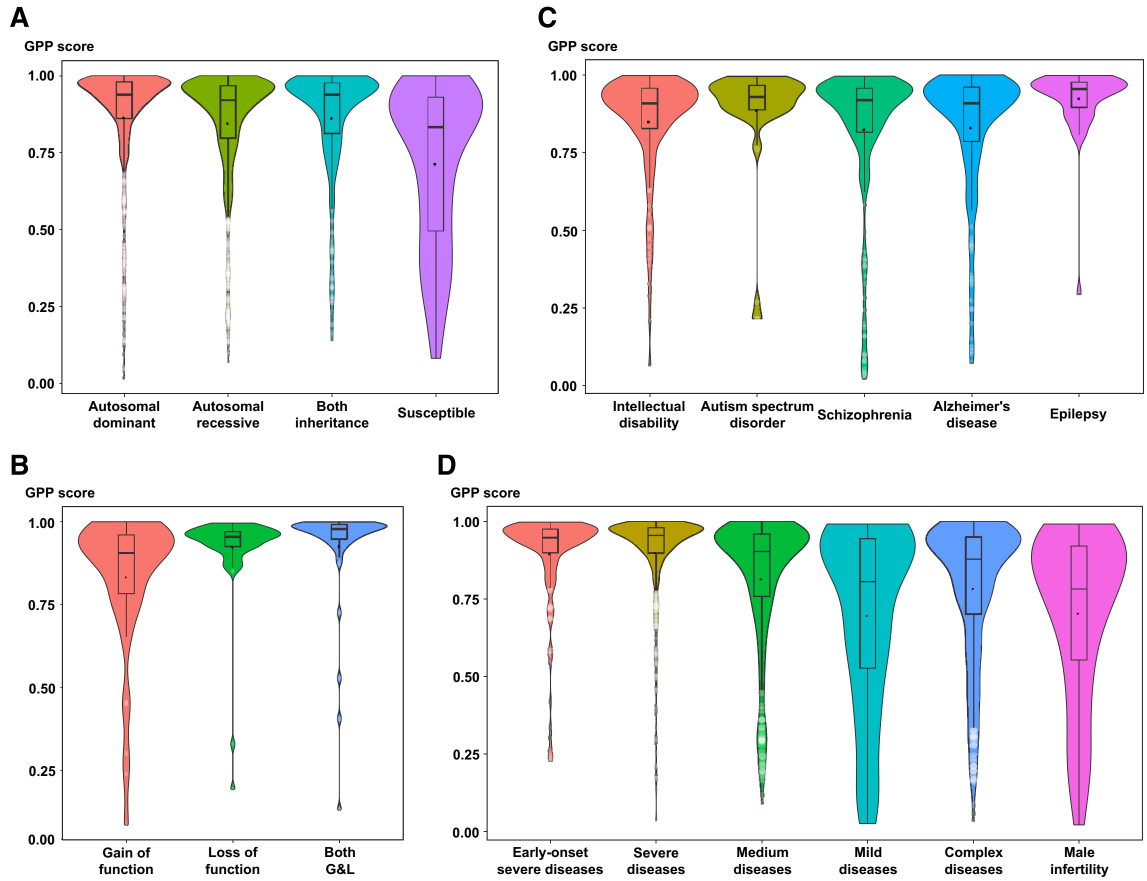

Fig. 3 Distribution of the GPP score for different pathogenic gene sets. **a** The distribution of genes with different inheritance models, including dominant, recessive, both dominant and recessive genes in autosome and susceptible genes. **b** The distribution of gain-of-function, loss-of-function and both G and L genes. **c** The distribution of pathogenic genes of neurological diseases with different ages of onset: intellectual disability (ID), autism spectrum disorder (ASD), schizophrenia (SCZ), Alzheimer's disease (AD) and epilepsy (EP), from left to right. **d** The distribution of genes responsible for diseases with different severities. Early-onset severe diseases contain genes of some early-onset severe Mendelian diseases, severe diseases contain genes of diseases with severe phenotype (neuromuscular disease, metabolic disease, congenital heart disease and hereditary tumors), medium diseases contain genes of some diseases with medium phenotype (ophthalmic diseases and hearing disorders), mild diseases contains genes of some diseases with mild phenotype (skin diseases such as psoriasis and ichthyosis), complex diseases contain genes of some complex diseases (hypertension, obesity and diabetes) and male infertility contain genes of male infertility

excluded genes not in the reported MD gene pool and do the analysis again, and similar results are obtained.

In this study, we ultimately calculated three kinds of gene-level score to evaluate the pathogenicity of all protein-coding genes and to assign the dominant or recessive model to each MD gene. The GPP score may help to identify MD genes, while the GDP and GRP scores may help to identify which genes follow dominant or recessive models. To explore a broader application of our scores, we applied the GPP score to some genes related to schizophrenia identified by GWAS (Li et al. 2017), and a high proportion (87.3%) of the genes were predicted to be pathogenic. Some diseases are highly related to copy number variations (CNVs), but there may be several genes involved when a responsible CNV is identified. We find the GPP score can significantly distinguish core genes from background genes of CNV, which may help us identify the core genes so that we can obtain accurate targets for later research.

There are some deficiencies in our study. When selecting non-pathogenic genes of MD (negative gene sets), we used the existing variants in public databases to select LOF variants tolerant genes. We did not find "reported non-pathogenic genes" because it's hard to deem a gene to be non-pathogenic. This might result in some errors, although our analysis verified that the selected genes were generally accurate. When we applied the GPP score to analyze gene sets of different diseases, we selected only a limited number of diseases and genes. If more diseases and genes had been involved, we might have obtained more convincing results. We also tried to perform the GOF and LOF prediction by the same method, but we did not collect enough GOF genes. We may finish this effort to complement our study when we obtain more GOF genes. Furthermore, we think there is heterogeneity for different kinds of disease, so distinguishing the pathogenic genes for specific diseases, for example, ophthalmic diseases, neurological disease and developmental diseases, may be a new research direction.

In conclusion, our study estimates the total number of MD genes. We introduce the gene pathogenicity prediction (GPP) score, which may provide robust and reliable guidance for the identification of pathogenic genes in MD research. We also provide two additional gene-level scores that may suggest the dominant or recessive inheritance model of genes. In addition, our results may promote the understanding of MD genes.

Acknowledgements We thank the providers and maintainers of the public databases we used in our study. We appreciate the support of Shenzhen Key Laboratory of Neurogenomics (CXB201108250094A). This study was supported by the National Natural Science Foundation of China (81771444) and the Shenzhen Municipal Government of China (NO.JCYJ20170412153248372).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Adzhubei IA, Schmidt S, Peshkin L et al (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249

Bamshad MJ, Ng SB, Bigham AW et al (2011) Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 12:745–755

Chen R, Shi L, Hakenberg J et al (2016) Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. Nat Biotechnol 34:531–538

Davydov EV, Goode DL, Sirota M et al (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6:e1001025

Ioannidis NM, Rothstein JH, Pejaver V et al (2016) REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet 99:877–885

Itan Y, Shang L, Boisson B et al (2015) The human gene damage index as a gene-level approach to prioritizing exome variants. Proc Natl Acad Sci USA 112:13615–13620

Kircher M, Witten DM, Jain P et al (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46:310–315

Landrum MJ, Lee JM, Benson M et al (2018) ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 46:D1062–D1067

Lek M, Karczewski KJ, Minikel EV et al (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285–291

Li Z, Chen J, Yu H et al (2017) Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. Nat Genet 49:1576–1583

Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814

Petrovski S, Wang Q, Heinzen EL et al (2013) Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet 9:e1003709

Quinodoz M, Royer-Bertrand B, Cisarova K et al (2017) DOMINO: using machine learning to predict genes associated with dominant disorders. Am J Hum Genet 101:623–629

Rackham OJ, Shihab HA, Johnson MR, Petretto E (2015) EvoTol: a protein-sequence based evolutionary intolerance framework for disease-gene prioritization. Nucleic Acids Res 43:e33

Rappaport N, Twik M, Plaschkes I et al (2017) MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucleic Acids Res 45:D877–D887

Samocha KE, Robinson EB, Sanders SJ et al (2014) A framework for the interpretation of de novo mutation in human disease. Nat Genet 46:944–950

Schwarz JM, Cooper DN, Schuelke M, Seelow D (2014) MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods 11:361–362

Sudmant PH, Rausch T, Gardner EJ et al (2015) An integrated map of structural variation in 2,504 human genomes. Nature 526:75–81

Szklarczyk D, Morris JH, Cook H et al (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 45:D362–D368

Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38:e164

Wenger AM, Guturu H, Bernstein JA, Bejerano G (2017) Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. Genet Med 19:209–214