MVP: predicting pathogenicity of missense variants by deep neural networks

Hongjian Qi [1, 2, *], Chen Chen [1, 3, *], Haicang Zhang[1], John J. Long[2], Wendy K. Chung[4], Yongtao Guan[5], Yufeng Shen[1, 6, 7, #]

1. Department of Systems Biology, Columbia University, New York, NY, USA
2. Department of Applied Mathematics and Applied Physics, Columbia University, New York, NY, USA
3. Department of Biological Sciences, Columbia University, New York, NY, USA
4. Departments of Pediatrics and Medicine, Columbia University, New York, NY, USA
5. Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA
6. Department of Biomedical Informatics, Columbia University, New York, NY, USA
7. JP Sulzberger Columbia Genome Center, Columbia University, New York, NY, USA

* Equal contribution

Correspondence should be addressed to Y.S. (ys2411@cumc.columbia.edu)

Abstract

Precise prediction of pathogenicity of missense variants is critical to improve power in genetic studies and yield in clinical genetic testing. Here we describe a new prediction method, MVP, which uses deep neural networks to leverage large training data sets and many correlated predictors. Using cancer mutation hotspots and *de novo* germline mutations from developmental disorders for benchmarking, MVP achieved better performance in prioritizing pathogenic missense variants than previous methods.

Main Text

Missense variants are the most abundant type of coding genetic variants and a major class of genetic risk in a broad range of common and rare diseases. Previous studies have estimated that there is substantial contribution from de novo missense mutations to structural birth defects [1-5], and neurodevelopmental disorders [6-8]. However, only a small fraction of missense de novo mutations are pathogenic [6]. As a result, the statistical power of detecting individual risk genes based on missense variants or mutations is limited [9]. In clinical genetic testing, many of missense variants in well-established risk genes are classified as variants of uncertain significance ("VUS"), unless they are highly recurrent in the patient population. Previously published *in silico* prediction methods have facilitated the interpretation of missense variants, such as Polyphen [10], SIFT [11], CADD[12], metaSVM[13], M-CAP[14], and REVEL[15]. However, based on recent de novo mutation data [2,6], they all have limited performance (Table S1).

Here we hypothesize that missense variants pathogenicity prediction can be improved in a few directions. First, conventional machine learning approaches have limited capacity to leverage large amount of training data comparing to recently developed deep learning methods[16]. Second, databases of pathogenic variants curated from literature are known to have substantial rate of false positives[17,18], which are likely caused by common issues across databases and therefore introduce inflation of benchmark performance. Developing new benchmark data and methods can further improve real performance. Finally, previous methods do

not consider gene dosage sensitivity, which can modulate the pathogenicity of deleterious missense variants, as hypomorphic variants are pathogenic only in dosage sensitive genes [8]. With recently published metrics of mutation intolerance [19,20], it is now feasible to consider gene dosage sensitivity in predicting pathogenicity. Based on these ideas, we describe a new method, MVP, to improve Missense Variant Pathogenicity prediction.

MVP uses many correlated predictors, broadly grouped in four categories (Supplementary Table S2): (a) variant level; (b) gene mutation intolerance; (c) protein structure and modification; (d) published scores from selected previous methods. To capture the disruptive effect of key variants on protein structure and posttranslational modification, we included features such as protein complex formation score (CORUM, BioPlex etc.), protein interaction interface score (prePPI etc), phosphorylation scores (see Methods and Table S2 for details). Since the variants in constrained genes (based on pLI >= 0.5) and non-constrained genes (pLI < 0.5) have different mode of action on pathogenicity and different information gain among features, we trained our models separately for the two gene sets. We included 38 features for constrained gene model and 21 features for non-constrained gene model where we removed most published prediction methods features (Supplementary Table S2).

MVP uses a deep neural network method, more specifically, the residual neural network (Resnet)[21]. There are 2 layers of residual blocks consisted of convolutional filters and 2 fully connected layers (Supplementary Fig S1). The convolutional filters can exploit spatial locality by enforcing a local connectivity pattern between "neurons" of adjacent layers and identify nonlinear interaction at higher levels of the network. To take advantage of this character, we ordered the predictors based on their correlation as the local connected observations, highly correlated predictors are clustered together (Supplementary Fig S2). Notably new protein predictors are weakly correlated with previous scores, suggesting that they may include additional information and can help improve the overall prediction accuracy. A random order of features as input for MVP cannot learn the pattern effectively.

We obtained large curated pathogenic variants datasets as positives and random rare missense variants from populations as negatives for training (Supplementary Table S3). To evaluate the generalization and predictive power of MVP model, we first applied 6-fold cross-validation on the training set (Supplementary Fig S3). We achieved mean area under the curve (AUC) of 0.99 in constrained genes and 0.97 in non-constrained genes.

To compare with other methods [15,22-25], we then evaluated the performance in curated testing dataset, VariBench dataset [13,26] (Table S3). In both datasets, MVP outperformed all other scores with an AUC of 0.96 and 0.92 in in constrained and non-constrained genes, respectively (Supplementary Fig S4). A few recently published methods (REVEL, M-CAP, and meta-SVM) were among the second-best predictors and achieved AUC of 0.91. To address potential issue of performance inflation due to systematic errors caused by similar factors across training and testing data sets, we obtained two additional types of data for further evaluation.

First, we complied cancer somatic mutation data, including missense mutations located in hotspots inferred based on statistical evidence from a recent study [27] as positives, and randomly selected variants from COSMIC database[28] as negatives. In this dataset, the performance of all methods decreased in the cancer hotspot dataset but MVP still achieved the best performance, AUC of 0.88 and 0.85 in constrained and non-constrained genes, respectively (Fig. 1).

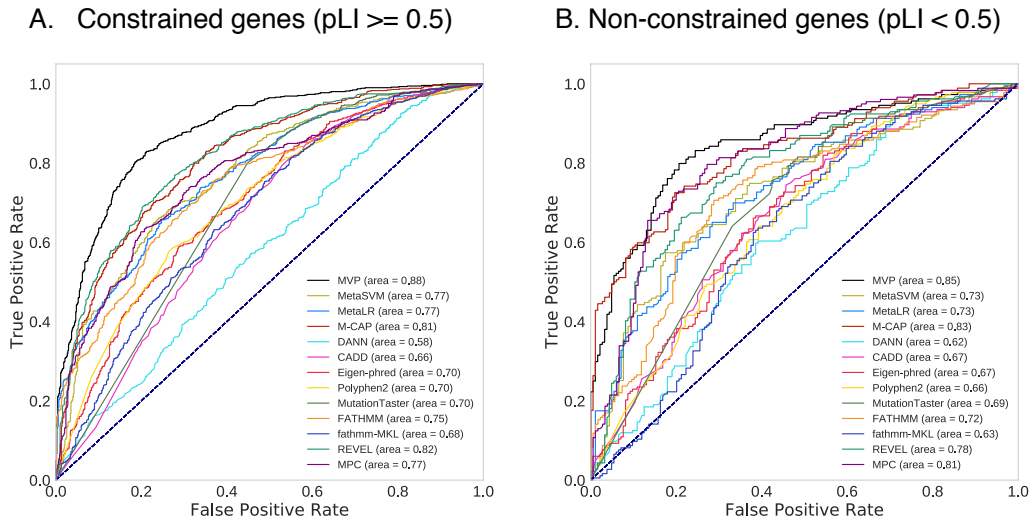

Figure 1. ROC curves using cancer somatic mutation data sets. (A) Constrained genes: evaluation on 699 cancer mutations located in hotspots from 150 cancer driver genes, and the same number of somatic mutations randomly selected from 13152 somatic mutations not located in hotspots in COSMIC database. (B) Non-constrained genes: evaluation on 177 cancer mutations located in hotspots from 55 cancer driver genes and same number of mutations randomly selected from 24343 somatic mutations not located in hotspots. The performance of each method is evaluated by the ROC curve and AUC score indicated in parenthesis. Higher AUC score indicates better performance.

Second, we compiled germline *de novo* missense variants (DNMs) from 2645 cases in a congenital heart disease (CHD) study [2] and 3953 cases in autism spectrum disorder (ASD) studies [2,6,7], and DNMs from 1911 controls (unaffected siblings) in Simons Simplex Collection (SSC) [2,6,7]. Most of the de novo mutations are not recurrent, therefore, we cannot use “ground truth” in the data to directly evaluate the performance of prediction methods. Instead, we calculated enrichment of missense DNMs in the cases compared to the controls, and then estimated precision rate and recall based on enrichment rate and observed number of DNMs in cases (Supplementary Notes). We compared the performance of MVP to other prediction methods by estimated precision and recall (Fig 2). Using threshold of 0.05, we observed an enrichment of 2.1 in CHD versus controls and an enrichment of 1.81 in ASD versus controls among constrained gene (Fig 2A, 2D, Supplementary Table S4), indicating more than 40% of the variants in the predicted damaging variants contribute to the disease. We also observed an enrichment of 1.35 in CHD and 1.25 in ASD versus controls among non-constrained genes (Fig 2B, 2E, Supplementary Table S5), indicating more than 25% of the variants in the predicted

damaging variants contribute to the disease. If we assume the estimated number of all missense de novo mutations contributing to CHD and autism as “condition true positives”, then a MVP score of 0.15 reaches 80% recall rate for both CHD and autism with estimated precision at 40% and 30% (Figure 2C and 2F), respectively. The second-best methods reached 25% (M-CAP) and 20% (MPC and REVEL) given the same recall rate for CHD and autism, respectively. Overall, the results indicate that MVP can be used to prioritize the pathogenic variants and genes in birth defects and neurodevelopmental disorders.

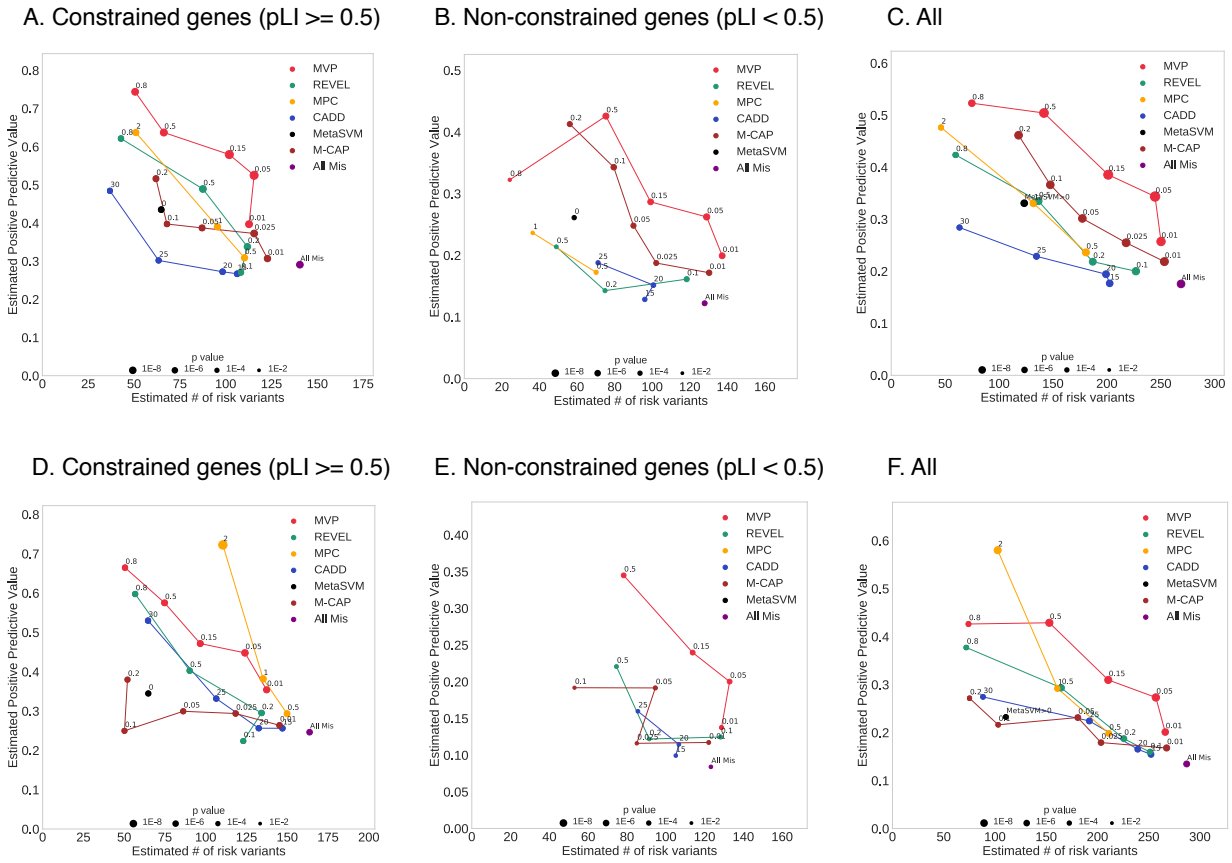

Figure 2. Comparison of MVP and published methods using de novo mutilations from CHD and ASD studies by precision-recall-like curves. Numbers on each data point indicate method thresholds. The size the of points are proportional to –log (p-value). P-value is calculated by binomial test, only points with p value less than 0.05 are plotted. (A) Performance in CHD de novo data in constrained genes. (B) Performance in CHD de novo data in non-constrained genes. (C) Performance in CHD de novo data in all genes. (D) Performance in ASD de novo data in constrained genes. (E) Performance in ASD de novo data in non-constrained genes. (F) Performance in ASD de novo data in all genes.

Previous studies have estimated that de novo coding mutations, including loss of function variants and damaging missense varints, have much smaller contribution to isolated CHD than syndromic CHD [2,3]. Here, we used MVP to revise the estimated fraction of isolated CHD cases attributable to de novo mutations. With damaging missense variants defined by metaSVM algorithm, which was used in Jin et al 2017 [2], the estimated DNMs contributes to 4.3% to isolated CHD cases. Using MVP score larger than 0.15 to define damaging missense variants, we estimated DMSs contributes to 7.4%(95% CI = [6.1%, 8.7%]) of isolated CHD cases, nearly doubling the previous estimate [2].

In summary, we developed a new method, MVP, to predict pathogenicity of missense variants. MVP is based on deep residual network, a supervised deep learning approach, and was trained using large number of reported pathogenic variants curated by clinical databases, separately on genes constrained of function variants and the ones that not constrained. Using cancer mutation hotspots and de novo mutations from congenital heart disease and autism, we showed that MVP achieved overall better performance than published methods, and would improve power of detecting novel risk genes in genetic studies. Currently, MVP is limited by size and potentially high false positive rate in the training data. Systematic efforts such as ClinVar[29] will eventually produce better training data to improve prediction performance.

Methods and materials

Training data sets

We compiled 22,390 missense mutations from Human Gene Mutation Database Pro version 2013 (HGMD)[30] database under the disease mutation (DM) category, 12,875 deleterious variants from UniProt[13,31] and 4,424 pathogenic variants from ClinVar database[29] as true positive (TP), altogether there are 32,074 unique positive training variants. The negative training sets

included 5,190 neutral variants from Uniprot[13,31], and randomly selected 42,415 rare variants rom DiscovEHR database[32], and 39,593 observed human-derived changes[12] as true negative, altogether there are 86,620 unique negative training variants(Supplementary Table S3).

Testing data sets

We have three categories of testing data sets (Supplementary Table S3) in which variants overlapping with the training data sets are removed. The three categories are Benchmark data sets from VariBench [13,26], cancer somatic missense mutations located in hotspots from recent study[27] and random selected somatic missense mutations from COSMIC database[28], and de novo missense mutation data sets from recent published exome-sequencing studies[2,6,7].

To compare the performance of different methods in different mode of action, we tested the performance based on gene property and evaluate the metrics in constrained genes (ExAC pLI ≥ 0.5) and non-constrained gene (ExAC pLI < 0.5)[19] separately.

To focus on rare variants with large effect, we focused on ultra-rare variants and we used a cutoff of MAF $<10^{-4}$ based on gnomAD database to filter variants in both training and testing data sets. We applied additional filter of MAF $< 10^{-6}$ for variants in constrained genes in both cases and controls for comparison based on a recent study [33].

Features used in MVP model

For each variant, we generated four groups of features in training and prediction: 1) variant level features, 2) gene level features 3) biochemical annotation features, and 4) published predictors (Supplementary Table S2).

Since variants in constrained genes and non-constrained genes have different mode of action on pathogenicity and different information gain among features, we trained our models on constrained and non-constrained variants separately with different sets of features (38 features used in constrained model, 21 features used in non-constrained model, Supplementary Table S2).

The predicted secondary structure and predicted accessible surface areas were obtained from dbPTM[34], interfacial residual prediction are the combination of three algorithms C-PPISP, PINUP, PredUS from PrePPI database[35], protein interactions status are from the protein interactions in the BioPlex 2.0 Network[36], the protein complexes information are from CORUM database[37], SUMO scores within 7 amino acids of peptide are obtained from GPS-SUMO software[38], phosphorylation sites predictions within 7 amino acids of peptide are from GPS3.0 software[39], ubiquitination within 14 amino acids of peptide are obtained from UbiProber prediction tools[40], regional constraint are from ExAC database, other features are obtained through dbNSFPv3.3a[41]. For consistence and to avoid one missense variant corresponds to different mutations, we used canonical transcripts in the analysis.[42] Missing values of features are filled with -1 except for the missing values of allele frequency in 1000 Genome or ExAC or gnomAD database, which is filled with 0. REVEL scores[15] and MPC scores [42] are included for method comparison.

Deep neural network model

MVP is a deep neural network based method that integrated features from variant level, gene level, protein structure and modification level as well as published predictors to predict pathogenicity of a certain variant. It is trained on a synthetic training sets including disease mutation in HGMD, Uniprot and ClinVar database and putatively neutral variants from Uniprot, DiscovEHR and observed human-derived alleles (Supplementary Table S3).

Convolution neural network can exploit spatial locality by enforcing a local connectivity pattern between neurons of adjacent layers and identify nonlinear interaction at higher levels of the network, it has been an important breakthrough in pattern recognition in recent years. To preserve the structured features in training data, we ordered the features according to their correlations (Fig 1). In training, we randomly partitioned the synthetic training data sets into two parts, 80% of the total training sets for training and 20% for validation.

We built a ResNet structure model[21] in this study. First a convolutional layers of 3 x1 with 32 convolutional filters, then 2 computational residual units consisting of 2 convolutional layers of 3x1convolutions and 32 convolutional filters, we summed the out layers and input layer, using ReLU as the activation function. After the bottleneck units, 2 fully connected layers of 320 x 512 and 512 x1 are used (Supplementary Fig S1).

We trained the ResNet model with batch size of 64 of training data, used adam optimizer to perform gradient descent and calculated logarithmic loss between the predicted value and true value as metrics. After each epoch, we applied the latest model weights on validation data to compute validation loss.

To avoid over fitting, we used early stopping regularization during training. We computed the loss in training data and validation data after each epoch and stopped the process when validation loss is comparable to training loss and do not decrease after 5 more epochs or the loss starts to increase, and then we saved model weights with last lowest validation loss. We applied the same model weights on testing data to obtain MVP scores for further analysis.

Previously published methods for comparison

We compared MVP score to 12 previously published prediction scores, namely, M-CAP[14], DANN[43], Eigen[44], Polyphen2[45], MutationTaster[46], FATHMM, Fathmm-MKL[47], REVEL[15], CADD[12], metaSVM[13], metaLR[13], and MPC[42], in the following two ways.

ROC curves

We plotted Receiver operating characteristic (ROC) curves and calculated Area Under the Curve (AUC) values to compare the performance among prediction scores in training data with 6-fold cross validation (Supplementary Fig S2), curated benchmark testing datasets (Supplementary Fig S3) and cancer hotspot mutation dataset (Fig 2). For each prediction method, we varied the threshold for calling pathogenic mutations in a certain range and computed the corresponding sensitivity and specificity based on true positive, false positive, false negative and true negative predictions. ROC curve was therefore generated by plotting sensitivity against 1 – specificity at each threshold.

Precision-recall-like curves

We used the excess of predicted pathogenic missense *de novo* variants in Autism Spectrum Disorder (ASD) and congenital heart disease (CHD) cases compared to controls as proxies of sensitivity and specificity since the de novo data do not have ground truth. For various thresholds of different scores, we can calculate the estimated number of risk variants and estimated positive prediction value based on enrichment of predicted damaging variants in cases compared to controls. We adjusted the number of missense *de novo* mutation in controls by the synonymous rate ratio in cases verses controls as the data sets were sequenced and processed separately)(Table S6), which partly reduced the signal but ensures that our results were not inflated by the technical difference in data processing. Using number of case individuals (M) and number of predicted damaging variants (X) in cases, number of control individuals (N) and number of predicted damaging variants (Y) in

controls and adjusted rate (α), we can use the formula below to obtain the enrichment, predicted positive value as well as estimated number of risk variants.

$$\alpha = \frac{\text{number of synonymous variants in cases}}{\text{number of synonymous variants in control}}$$

$$enrichment = \frac{\frac{X}{M}}{\frac{Y}{N} \times \alpha}$$

$$PPV = \frac{enrichment - 1}{enrichment} = \frac{\frac{X}{M} - \frac{Y}{N} \times \alpha}{\frac{X}{M}}$$

$$estimated\ number\ of\ risk\ variants = X \times PPV = X - \frac{M}{N} \times Y \times \alpha$$

Acknowledgements

We thank Na Zhu, Ben Lai, Itsik Pe'er, Emily Gao, and Jiayao Wang for helpful discussions. We thank Pediatric Cardiac Genomics Consortium (PCGC) for CHD data access, and the patients and their families for their generous contribution to the PCGC study.

Funding

This work is supported by NIH grants R01GM120609 (Q.H., H.Z, W.K.C., and Y.S.) and R01HG008157 (Q.H. and Y.S.).

References

1. Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science (New York, N.Y.)* **350**, 1262-6 (2015).
2. Jin, S.C. *et al.* Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nature genetics* **49**, ng. 3970 (2017).
3. Sifrim, A. *et al.* Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat Genet* **48**, 1060-5 (2016).
4. Yu, L. *et al.* Increased burden of de novo predicted deleterious variants in complex congenital diaphragmatic hernia. *Hum Mol Genet* (2015).
5. Qi, H. *et al.* Genetic analysis of de novo variants reveals sex differences in complex and isolated congenital diaphragmatic hernia and indicates *MYRF* as a candidate gene. *bioRxiv* (2017).
6. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-221 (2014).
7. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-215 (2014).
8. McRae, J.F. *et al.* Prevalence, phenotype and architecture of developmental disorders caused by de novo mutation. *bioRxiv* (2016).
9. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* **111**, E455-64 (2014).
10. Adzhubei, I., Jordan, D.M. & Sunyaev, S.R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7 20 (2013).
11. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-81 (2009).
12. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**, 310-315 (2014).
13. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human molecular genetics* **24**, 2125-2137 (2014).
14. Jagadeesh, K.A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature genetics* **48**, 1581-1586 (2016).
15. Ioannidis, N.M. *et al.* REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics* **99**, 877-885 (2016).
16. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning*, (MIT Press, 2016).
17. Dorschner, M.O. *et al.* Actionable, pathogenic incidental findings in 1,000 participants' exomes. *The American Journal of Human Genetics* **93**, 631-640 (2013).
18. Wang, J. & Shen, Y. When a "disease-causing mutation" is not a pathogenic variant. *Clinical chemistry* **60**, 711-713 (2014).

19. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).
20. Cassa, C.A. *et al.* Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat Genet* **49**, 806-810 (2017).
21. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 770-778 (2016).
22. Kharchenko, P.V., Silberstein, L. & Scadden, D.T. Bayesian approach to single-cell differential expression analysis. *Nat Methods* **11**, 740-2 (2014).
23. Jagadeesh, K.A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* **48**, 1581-1586 (2016).
24. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics* **24**, 2125-2137 (2015).
25. Samocha, K.E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* (2017).
26. Nair, P.S. & Vihinen, M. VariBench: a benchmark database for variations. *Human mutation* **34**, 42-49 (2013).
27. Chang, M.T. *et al.* Accelerating discovery of functional mutant alleles in cancer. *Cancer discovery* (2017).
28. Forbes, S.A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic acids research* **45**, D777-D783 (2016).
29. Landrum, M.J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic acids research* **44**, D862-D868 (2015).
30. Stenson, P.D. *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics*, 1-13 (2017).
31. Consortium, U. Ongoing and future developments at the Universal Protein Resource. *Nucleic acids research* **39**, D214-D219 (2011).
32. Dewey, F.E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, aaf6814 (2016).
33. Kosmicki, J.A. *et al.* Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat Genet* **49**, 504-510 (2017).
34. Lee, T.-Y. *et al.* dbPTM: an information repository of protein post-translational modification. *Nucleic acids research* **34**, D622-D627 (2006).
35. Zhang, Q.C., Petrey, D., Garzón, J.I., Deng, L. & Honig, B. PrePPI: a structure-informed database of protein–protein interactions. *Nucleic acids research* **41**, D828-D833 (2012).
36. Huttlin, E.L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* (2017).
37. Ruepp, A. *et al.* CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic acids research* **38**, D497-D501 (2009).
38. Zhao, Q. *et al.* GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs. *Nucleic acids research* **42**, W325-W330 (2014).

39. Xue, Y. *et al.* GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. *Protein Engineering, Design & Selection* **24**, 255-260 (2010).
40. Chen, X. *et al.* Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites. *Bioinformatics* **29**, 1614-1622 (2013).
41. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3. 0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human mutation* **37**, 235-241 (2016).
42. Samocha, K.E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *bioRxiv*, 148353 (2017).
43. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761-763 (2014).
44. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J.D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature genetics* **48**, 214-220 (2016).
45. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-249 (2010).
46. Schwarz, J.M., Cooper, D.N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nature methods* **11**, 361-362 (2014).
47. Shihab, H.A. *et al.* Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Human genomics* **8**, 11 (2014).