Assessing computational variant effect predictors via a prospective human cohort

Da Kuang,[1,2,3,4,6] Roujia Li,[1,2,3,4,6] Yingzhou Wu,[1,2,3,4] Jochen Weile,[1,2,3,4] Robert A. Hegele,[5] Frederick P. Roth[1,2,3,4,*]

[1] Donnelly Centre, University of Toronto, Toronto, ON M5S 3E1, Canada

[2] Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada

[3] Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON M5G 1X5, Canada

[4] Department of Computer Science, University of Toronto, Toronto, ON M5T 3A1, Canada

[5] Departments of Medicine and Biochemistry, Robarts Research Institute, Schulich School of Medicine and Dentistry, Western University, London, ON, Canada

[6] These authors contributed equally to this work

* Correspondence: fritz.roth@utoronto.ca

Abstract

Computational predictors can help interpret pathogenicity of human genetic variants, especially for the majority of variants where no experimental data are available. However, because we lack a high-quality unbiased test set, identifying the best-performing predictors remains a challenge. To address this issue, we evaluated missense variant effect predictors using genotypes and traits from a prospective cohort. We considered 139 gene-trait combinations with rare-variant burden association based on at least one of four systematic studies using phenotypes and whole-exome sequences from ~200K UK Biobank participants. Using an evaluation set of 35,525 rare missense variants and the relevant associated traits, we assessed the correlation of participants' traits with scores derived from 20 computational variant effect predictors. We found that two predictors—VARITY and REVEL—outperformed all others according to multiple performance measures. We expect that this study will help in selecting variant effect predictors, for both research and clinical purposes, while providing an unbiased benchmarking strategy that can be applied to additional cohorts and predictors.

Introduction

Rapidly increasing availability and use of sequencing in research and clinical genetic diagnostics has yielded millions of rare human genetic variants. Of particular interest are missense variants, which alter the coding sequence of human proteins, potentially altering protein functions[1] and thus contributing to human diseases[2].

Where genetic disease is suspected, clinical variant interpretation commonly uses the American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) framework, in which variants are classified as benign, likely benign, likely pathogenic, pathogenic, or of uncertain significance[3]. Of the 4.6 million missense variants in gnomAD[4], only 2% have been clinically interpreted. The majority of these have been classified as a variant of uncertain significance (VUS)[5], which, according to the ACMG/AMP guidelines[3], should not be included in clinical decision making. Moreover, most VUS variants are missense variants, highlighting an unmet need for reliable evidence to support unambiguous clinical interpretation and thereby guide the diagnosis and treatment of disease[6,7].

Well-established functional assays can provide strong evidence for classification of a variant as either pathogenic or benign[3,8], but such results are typically unavailable for rare missense variants. Although “variant effect mapping” technologies have been established to proactively determine the functional effects of many variants in parallel[9–11], a variant effect map is available for only ~1% of human disease-associated proteins[12].

A less resource-intensive means of proactively assembling evidence for all possible variants is computational variant effect prediction. Widely used computational variant effect predictors developed over the last two decades include PROVEAN[13], PolyPhen-2[14] and SIFT[15]. More recently, a new wave of variant effect predictors, e.g. DeepSequence[16] and PrimateAI[17], has benefited from advances in deep learning. Some 'meta-predictors', e.g. REVEL[18] and VARITY[19], have benefited in large part by combining the results of many evidence sources, including the results of other prediction algorithms.

Current ACMG/AMP guidelines consider computational prediction methods to (at best) provide weak evidence for clinical interpretation[3]. However, as predictors improve, objective evaluation of evidentiary value may justify increased reliance on computational prediction and thus ultimately enable improved clinical interpretation.

Although there have been several benchmarking studies[20–22], it has been difficult to address certain challenges inherent to the assessment of computational predictors. Chief among these is the establishment of "ground truth" test sets that are independent of the data used in training the predictors being assessed. Where test data have previously been used in training, performance estimates for a computational model may be artificially inflated[23].

To avoid this circularity issue, Livesey and Marsh[20] used variants measured by variant effect mapping experiments, but could only assess variant predictions for a few dozen proteins for which variant effect maps were available. Furthermore, they had limited ability to evaluate predictors (e.g., DeepSequence) that had been optimized using

variant effect map data. Although Livesey and Marsh found that variant effect maps were typically more accurate predictors of pathogenicity than any computational method, using variant effect maps as a source of ground truth does have the caveat that functional assays carried out in cultured cells may not perfectly capture variant impacts on human traits.

Population-based cohorts with genotyped and phenotyped participants have the potential to provide independent data that have not previously been used in training predictors. For example, the UK Biobank[24] has assembled in-depth genetic and trait information for a prospective cohort of 500,000 participants. Whole-exome sequences for >200,000 participants have been released widely to researchers, and information on >7,000 traits is available for a large fraction of participants. Because no variant effect predictors have yet been trained on these data, using the UK Biobank dataset as a source of “ground truth” human trait data sidesteps the risk of performance inflation due to circularity.

Here, as shown in Figure 1, we examined 139 gene-trait combinations (involving 35,525 rare missense variants in 99 genes, and 56 traits) for which it has been reported that the burden of rare missense variation in the gene depends on the trait[25–28]. For each gene-trait combination, we assessed the correspondence between variant scores and human phenotypes for 20 computational predictors. Results were then compiled to identify computational methods that were top performers for the greatest number of gene-trait combinations.

Material and methods

Gene-trait combinations

We assembled gene-trait combinations for which a significant burden-of-variation association had been reported by at least one of four systematic 'burden scan' studies[25–28] of the UK Biobank cohort. From the initial set of 162 gene trait combinations, we excluded combinations for which the trait had been ascertained in fewer than 10 participants or for which the gene IDs are currently unrecognized or not linked to any proteins in Ensembl database version 104[29]. We also excluded the *TTN* gene as non-representative given its extreme size and enormous number of reported variants.

Human Variants

This study was conducted with whole-exome sequencing data from 200,619 participants in the UK Biobank cohort, for which variants were retrieved from the OQFE version[30] of the whole-exome VCF files (field ID: 23156). The canonical isoform of each gene product we examined was defined according to the Ensembl database (GRCh38)[29], with coding exonic regions defined according to the CCDS database[31]. Coding variants corresponding to these coding regions were extracted from raw VCF files. Adapting the filtering criteria used by the UK Biobank[32], we examined only coding single-nucleotide variants (SNVs) having a Phred quality score > 20, individual missingness < 10%, minimum read coverage depth of 7, and carried by at least one participant passed the allele balance threshold (i.e. the proportion of reads covering a variant's location that support the variant) > 0.15. Because evidence to support clinical interpretation is typically more abundant for common variants (e.g., from genome-wide association

studies), the most critical context for computational predictors to perform well is for rare variation. Therefore, we further restricted our analysis to rare variants, as defined here by having a global MAF < 0.1% in gnomAD[33]. If a variant was not found in the gnomAD database, we assumed it to be rare (MAF < 0.1%).

Variant effect predictions

We considered 20 computational variant effect predictors (Table S1). Scores for predictors were obtained either from a pre-existing repository[34], or by running the predictor code directly. If a predictor did not provide predictions for at least 10 rare missense variants of the gene for a given gene-trait combination, it was not included in the comparison for that combination.

Predictors comparison

We first split the 139 gene-trait combinations into two categories, depending on whether the associated traits are binary or quantitative (i.e. where measurements are continuous).

Gene-trait combinations where the trait was binary were subjected to precision vs. recall analysis, where precision at a given score threshold is defined as fraction of variants that were correctly assigned as having come from an individual with the trait, and recall is the fraction of variants from individuals with the trait that were identified as such. Because precision depends on the prior probability, i.e., the prevalence of the trait, we used balanced precision (the precision expected if the test set were to have been balanced) and calculated the area under the balanced precision-recall curve (AUBPRC)[19].

For each variant effect predictor and gene-trait combination, we rescaled predictor scores to range from 0 to 1 (with 0 corresponding to neutral variants and 1 corresponding to damaging variants). To reduce the impact of outliers, we set the lowest and highest 5% of scores to 0 and 1, respectively.

We subsequently computed a person-centric variant score: the sum of all variant scores for a given participant. For each predictor, this allowed us to rank participants by the participant-centric variant score, plot a balanced precision-recall curve, and calculate AUBPRC for each gene-trait combination. To better understand uncertainty in the calculated AUBPRC values, we used 1000-iteration bootstrapping (random sampling of variants with replacement) to compute a distribution of AUBPRCs for each predictor with each gene-trait combination. We then empirically determined the mean AUBPRC and the 95% confidence interval of the resampled distribution. We compared variant effect predictors in terms of mean AUBPRCs, and calculated an empirical p-value for each pair of computational predictors, i.e., the fraction of pairs of resampled AUBPRC distributions, one from each predictor, in which one predictor achieved a higher AUBPRC than the other predictor. From the p-values, false discovery rates (FDRs) were subsequently calculated to account for multiple hypotheses testing[35]. For each gene-trait combination, predictors were ranked by performance and any predictor that was not significantly different from the numerically best-performing predictor (using an FDR threshold of 10%) was considered tied for best. We then ranked predictors by the number of gene-trait combinations for which the predictor was tied for best.

In contrast, for gene-trait combinations with quantitative traits, we employed the Pearson correlation coefficient (PCC) to compare predictor performance. Where

multiple participants carried the same variant, we averaged their quantitative trait values. Each variant effect predictor was examined individually. For each predictor and a given gene-trait combination, we examined variants for which the trait was measured and a variant effect score was available. To better understand uncertainty in the calculated PCC values, we used 1000-iteration bootstrapping (i.e. random sampling of variants with replacement) to compute a distribution of PCCs for each predictor with each gene-trait combination. We then empirically determined the mean PCC and the 95% confidence interval of the resampled distribution. To eliminate negative values, we used PCC^2 instead of PCC, and calculated an empirical p-value (i.e., the fraction of pairs of resampled PCC distributions, one from each predictor, in which one predictor achieved a higher PCC^2 than the other predictor) for each pair of computational predictors. To account for multiple hypothesis testing, we next derived FDRs from these empirical p-values[35]. For each gene-trait combination, predictors were ranked by performance and any predictor that was not significantly different from the numerically best-performing predictor (FDR < 10%) was considered tied for best. We then ranked predictors by the number of gene-trait combinations in which the predictor was tied for best.

To assess the statistical significance of performance differences between methods considering all gene-trait combinations, we carried out a two-tailed Wilcoxon signed-rank test comparing the arrays of performance measures (AUBPRC or PCC^2) for each pair of predictors. FDR values for each comparison were derived as above.

Results

Extracting rare missense variants from the UK Biobank cohort for trait-associated genes

To select gene-trait combinations for which rare (MAF < 0.1%) variation is associated with traits, we compiled a set of 139 gene-trait combinations that were collectively identified by four systematic burden testing studies performed using data from the UK Biobank cohort[25–28]. Table S2 lists the genes and the associated traits in the format of the UK Biobank field ID (FID). Of the 99 trait-associated genes, 73 (74%) were associated with only one trait. The remaining 26 genes were linked to multiple traits. For example, *LDLR*, a gene that encodes the low-density lipoprotein (LDL) receptor and is associated with autosomal dominant familial hypercholesterolemia (FH [MIM: 143890])[36], was previously found to be associated with five traits: 1) blood LDL cholesterol level (mmol/L; FID: 30780); 2) self-reported high cholesterol (FID: 20002-1473); 3) taking atorvastatin (a cholesterol-lowering drug; FID: 20003-1141146234); 4) taking any cholesterol-lowering medication (FID: 6153-1); and 5) atherosclerotic heart disease of native coronary artery (FID: 41270-I25.1).

Having obtained whole-exome sequencing data released for over 200K participants in the UK Biobank cohort (transfer of human data was approved and overseen by The UK Biobank Ethics Advisory Committee [Project ID: 51135]), we extracted SNVs for the 99 associated genes. We focused all subsequent analyses on the subset of rare variants (MAF < 0.1%). Because classifications of rare clinical variants are more likely to be VUS and would therefore benefit from improved computational predictor evidence[5], we were primarily interested in predictor performance for these variants. To determine whether

the variant alters the amino acid sequence of the encoding protein, we mapped each variant to the canonical transcript of its corresponding gene in the Ensembl and CCDS database[31,37], yielding an evaluation set of 35,525 rare human missense variants from 170,368 UK Biobank participants.

Assessing the performance of computational variant effect predictors

For each rare human missense variant, we obtained variant effect scores from 20 computational variant effect predictors (see Table S1 for a complete list of predictors compared in this study). Some predictors (e.g. PROVEAN) assign low scores to predicted-damaging variants, while others (e.g. PolyPhen-2) assign high scores to such variants. To better compare rankings from different predictors, we negated scores of the former type so that the highest scores always corresponded to the most predicted-damaging variants.

For 12 (60%) of the 20 predictors we examined, scores were available for every missense variant in our evaluation set. For over 90% of the genes of interest, all predictors provided scores for at least 10 variants. Table S3 shows the prediction coverage for the 99 genes included in this study. To reduce the effect of extreme values on this performance evaluation, we applied a floor and ceiling at 5th and 95th percentile predictor scores, respectively, and applied a transformation to give all variant effect predictors the same 0-1 range of score values.

Next, we deployed two approaches to assess the performance of variant effect predictors, depending on whether the trait was binary or quantitative.

For gene-trait combinations with binary traits (including categorical traits that could be simplified and considered binary), we applied an AUBPRC approach. Because binary trait measurements were made on the participant level, i.e. one measurement per participant in the UK Biobank cohort, we wished to obtain a participant level summary of predicted variant effects. On average, about 1% of participants had more than one variant in a given gene (Figure S1). Therefore, we summed the total predictor score for all missense variants observed in each gene of interest in each participant. We note that this approach models variant effects as additive, e.g., two mildly damaging variants (score = 0.5) combined will show a more damaging effect (total score = 1.0). To illustrate this approach with the computational predictor VARITY, Figure 2 shows that UK Biobank participants taking cholesterol-lowering medication (FID: 6153-1) are three times more likely to have a damaging (≥1) total missense variant impact than those that are not.

For every gene-trait combination and each predictor, we analyzed the tradeoff between precision (i.e. fraction of participants above a given total variant impact score threshold that had the trait value associated with heightened rare-variant burden) and recall (i.e. fraction of all participants with the rare-variant-burden-associated trait value that were detected at a given total variant impact score threshold). More specifically, because precision depends on the prior probability of the trait value, we evaluated balanced precision, which represents precision where the prior probability of the trait value is 50%[19]. This enabled us to evaluate, for each combination of gene and binary trait, the AUBPRC for each computational predictor.

Because AUBPRC analysis is only appropriate for binary traits, we examined quantitative traits using a PCC-based approach, by which we assessed the

correspondence between variant impact score and trait value. Where multiple participants carried the same variant, we averaged the quantitative trait values. Variant predictors were considered better performing if they had a higher PCC value.

Thus, for each gene-trait combination, we obtained either a PCC or AUBPRC measure of performance. To estimate uncertainty in each performance measure, we carried out bootstrap resampling in which variants of a given gene were resampled with replacement and PCC or AUBPRC values were re-calculated for each sample. For each gene-trait combination, this yielded a distribution of either PCC or AUBPRC values for each computational predictor. From each of these distributions, we extracted a mean and a 95% confidence interval (CI) that reflects our uncertainty in the performance measure.

To illustrate this approach, Figure 3 shows AUBPRC and PCC values and CIs for each of 20 computational predictors for each of the five gene-trait combinations involving *LDLR*. Here, the numerically top-performing variant effect predictors were VARITY and MPC. To assess whether numerical differences were statistically significant, we computed empirical p-values between every pair of computational predictors for each gene-trait combination, and used the distribution of p-values for each gene-trait combination to derive corresponding FDR values. We considered a predictor Y to significantly outperform a predictor X if the comparison exhibited an FDR < 10%.

As an illustration, FDRs for all predictor pairs are shown for the gene-trait combination of *LDLR* with the binary trait of whether the participant was taking cholesterol-lowering medication (Figure 4). Although VARITY exhibited the highest AUBPRC, a larger group

of predictors (VARITY, REVEL, MutationAssessor, Eigen, MPC, SIFT, PolyPhen-2, CADD, MetaLR and PrimateAI) were statistically indistinguishable and therefore all considered as best-performing predictors for this gene-trait combination.

To summarize similar comparisons over all 139 gene-trait combinations, we counted the number of combinations in which a computational predictor was considered best-performing (Figure 5). These results showed that VARITY performed best across the largest number of gene-trait combinations: 135 (97%) of 139, while the next best predictor, REVEL, was a top predictor for 129 (93%) of 139 gene-trait combinations. The performance among the remaining 18 predictors ranged from 66 to 88%, with mean 71%. The best variant predictor according to this ranking was VARITY[19], followed by REVEL[18] and Eigen[38]. Because VARITY or REVEL were statistically indistinguishable (FDR = 0.32; Wilcoxon signed-rank test), we considered both VARITY and REVEL to be the best performing in this evaluation. Similar analysis (Figure S2) showed that both VARITY and REVEL each significantly (FDR < 0.1) out-performed all other methods.

Discussion

We used a large UK Biobank cohort to assess the performance of 20 computational predictors of missense variant effects. Combining four systematic burden-test studies, collectively based on whole-exome sequences of the 50K and 200K participants in the UK Biobank cohort, yielded a large set of gene-trait combinations (139) to enable a robust comparison of predictors.

Because none of the computational predictors we examined had been trained using UK Biobank data, our evaluation approach has the marked advantage of independence and

avoidance of the performance inflation that can arise when predictors are assessed using training data. Although every predictor was a top predictor for at least one gene-trait combination, counting the number of gene-trait combinations in which a predictor was best enabled an overall ranking.

It is interesting that the top three predictors overall – VARITY[19], REVEL[18], and Eigen[38] – are all meta-predictors, i.e., they combine prediction scores from other variant effect predictors. For meta-predictors, it can be especially difficult to establish ground truth sets of variants that had not been used for training any of the input predictors. That said, VARITY only exploited scores from other predictors that were unsupervised, i.e., made no direct use of variant pathogenicity annotations. The fourth-ranked predictor, MPC[39], exploited observations of depleted missense variation within particular sub-genic regions. That none of the top-three-ranked predictors exploited this kind of information suggests the possibility that combining these approaches could yield still-better results.

One limitation of our study was that we did not evaluate all published computational variant effect predictors. This was in part due to the vast number of such predictors but many predictors were excluded due to non-functional websites. We were particularly interested in evaluating EVmutation[40] given its excellent reported performance. However, EVmutation provided a score for only 7,021 (20%) of the 35,525 missense variants we examined, perhaps because the EVmutation method requires a deep multiple sequence alignment. However, we were able to examine variant effect predictors that are 1) widely used to predict variant effects (e.g. PolyPhen2, PROVEAN) and 2) novel and claimed better-performing than conventional predictors (e.g.

PrimateAI, REVEL). We suggest that this analysis should be repeated periodically to benchmark and test evaluate predictors as they emerge.

Another limitation of this study is that we did not consider correlations between traits. For example, multiple gene-trait combinations involved *LDLR*, and these traits were correlated with one another, so our analysis was influenced by some genes and traits more than others. That said, body mass index (BMI; FID: 21001) was the most recurrent trait (appearing in 22 combinations) and *LDLR* was the most recurrent gene (appearing in 5 gene-trait combinations). Thus, no one gene or trait dominated the collection of 139 gene-trait combinations examined here.

We also did not adjust trait values to account for dependencies on other participant variables. For example, LDL cholesterol levels are known to correlate with both age and sex[41]. In future analyses, LDL cholesterol measures adjusted for age and sex, or more precise variables such as apolipoprotein B, could have a variation that is more attributable to genetic variation and therefore show greater correlation with predictor scores. However, because the same adjusted LDL cholesterol values would be used to evaluate all predictors, such an adjustment would be expected to have limited impact on the relative rankings amongst predictors.

One possible criticism of our study is that the UK Biobank dataset we used contains variants that may have been used in training predictors. For this reason, we considered excluding variants that had been previously reported as pathogenic or benign, e.g. in ClinVar[42], or as disease-associated or disease-causing by the Human Gene Mutation Database (HGMD)[43], as many predictors will have trained on variants from these

resources. However, vanishingly few of the variants used in predictor training sets will have been deposited in HGMD or ClinVar *on the basis of analysis of the UK Biobank data*, especially given that we have excluded common variants from our analysis. More obviously, neither the selection of UK Biobank participants, nor their traits or genotypes, was determined by variant annotations in ClinVar, HGMD or elsewhere.

Our analysis should in future be expanded to evaluate additional burden-test associated gene-trait combinations beyond the 139 examined here, as they emerge. Moreover, release of whole-exome sequences for an additional 250,000 UK Biobank participants is expected in 2022, and it will be important to revisit these comparisons with the expanded dataset.

In conclusion, this study provides an independent assessment of several computational variant effect predictors based on their correspondence with human traits in a large prospective cohort. Given the critical need for improving performance of computational predictors for both clinical and research applications, our benchmarking method is likely to be applicable to future human cohorts and methods for inferring the pathogenic impact of human genetic variants.

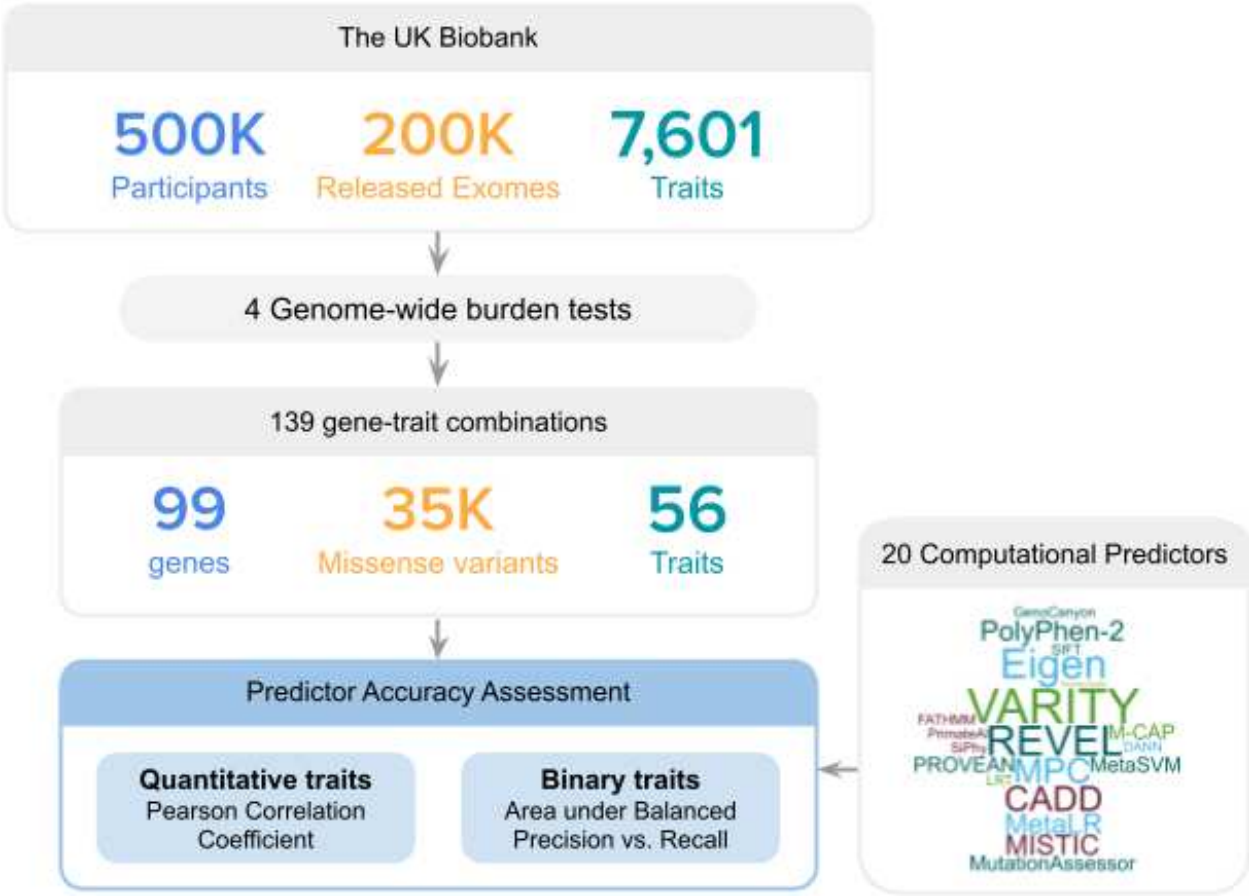

Figure 1. A schematic overview of the study. 139 gene-trait combinations were selected from the UK Biobank and used to assess the accuracy of 20 computational variant effect predictors (see Methods for detail).

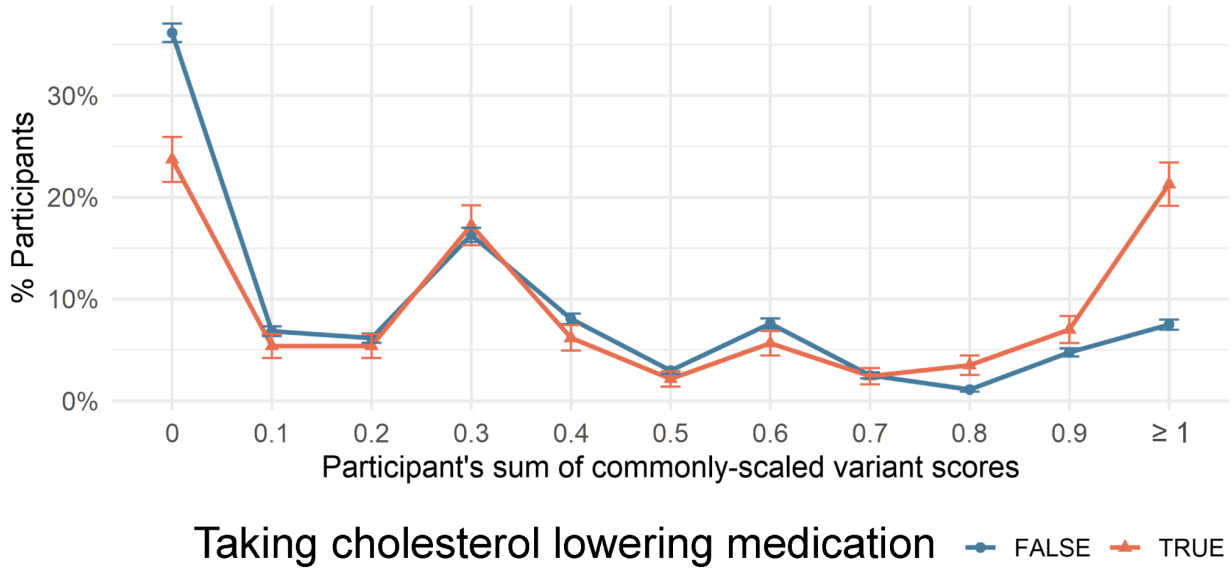

Figure 2. Percentages of UK Biobank participants with different sums of variant predictor scores for participants either taking or not taking cholesterol-lowering medication. This example used commonly-scaled VARITY scores in which higher scores indicate more damaging variants. Error bars represent one standard error of proportions.

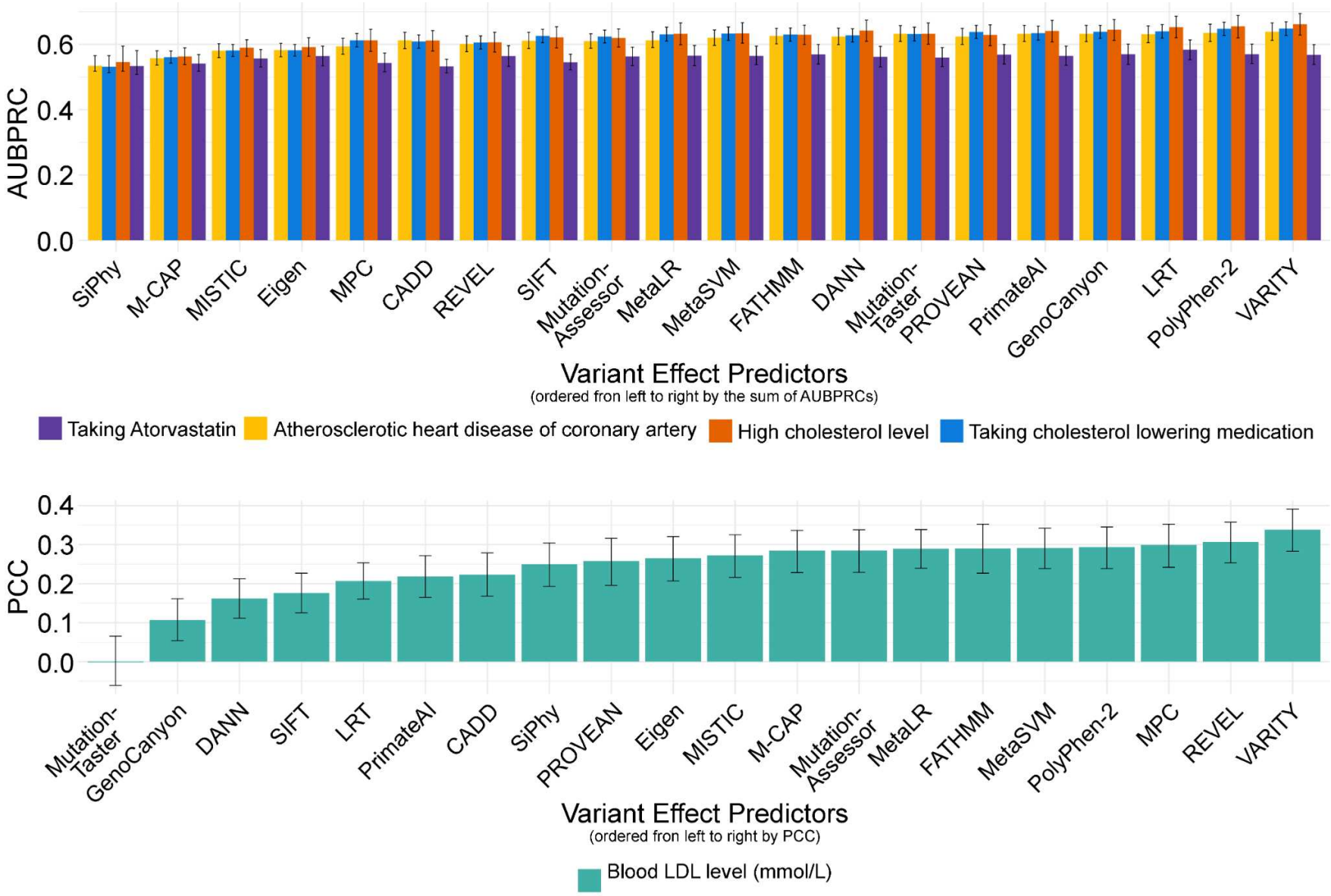

Figure 3. The performance of twenty computational methods in predicting the effect of rare *LDLR* missense variants. Performance comparisons used mean AUBPRC for binary traits and mean PCC values for quantitative traits, respectively. Error bars indicate the 95% confidence intervals of performance measures.

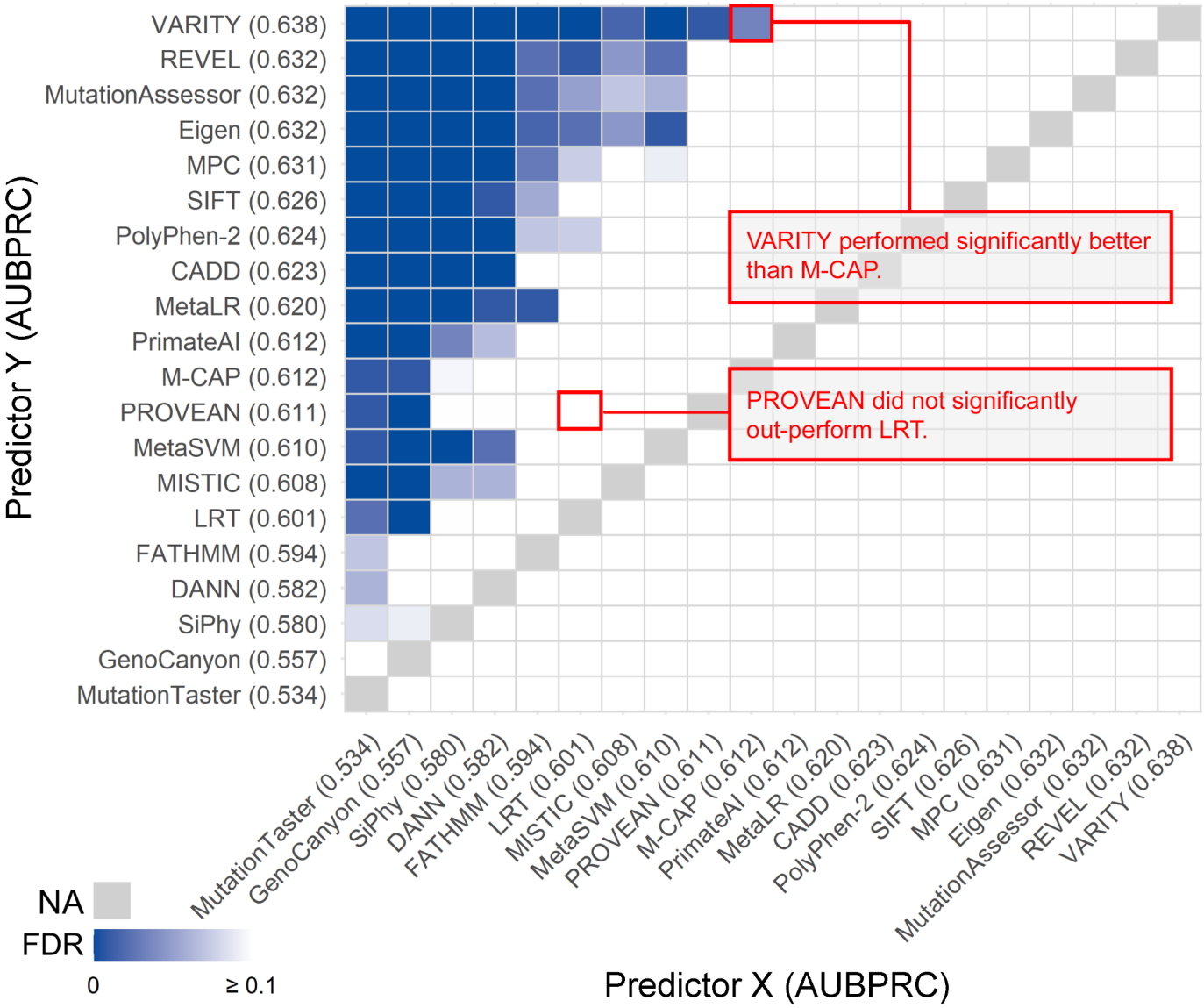

Figure 4. Example summary of comparisons between all pairs of computational predictors, evaluating how well predictor scores for rare *LDLR* missense variants correspond to whether the participant is taking cholesterol-lowering medication. Variant effect predictors are ranked top-to-bottom and right-to-left based on decreasing performance (AUBPRC). Comparisons in which one predictor Y outperforms another predictor X (with FDR < 0.1) are indicated in blue.

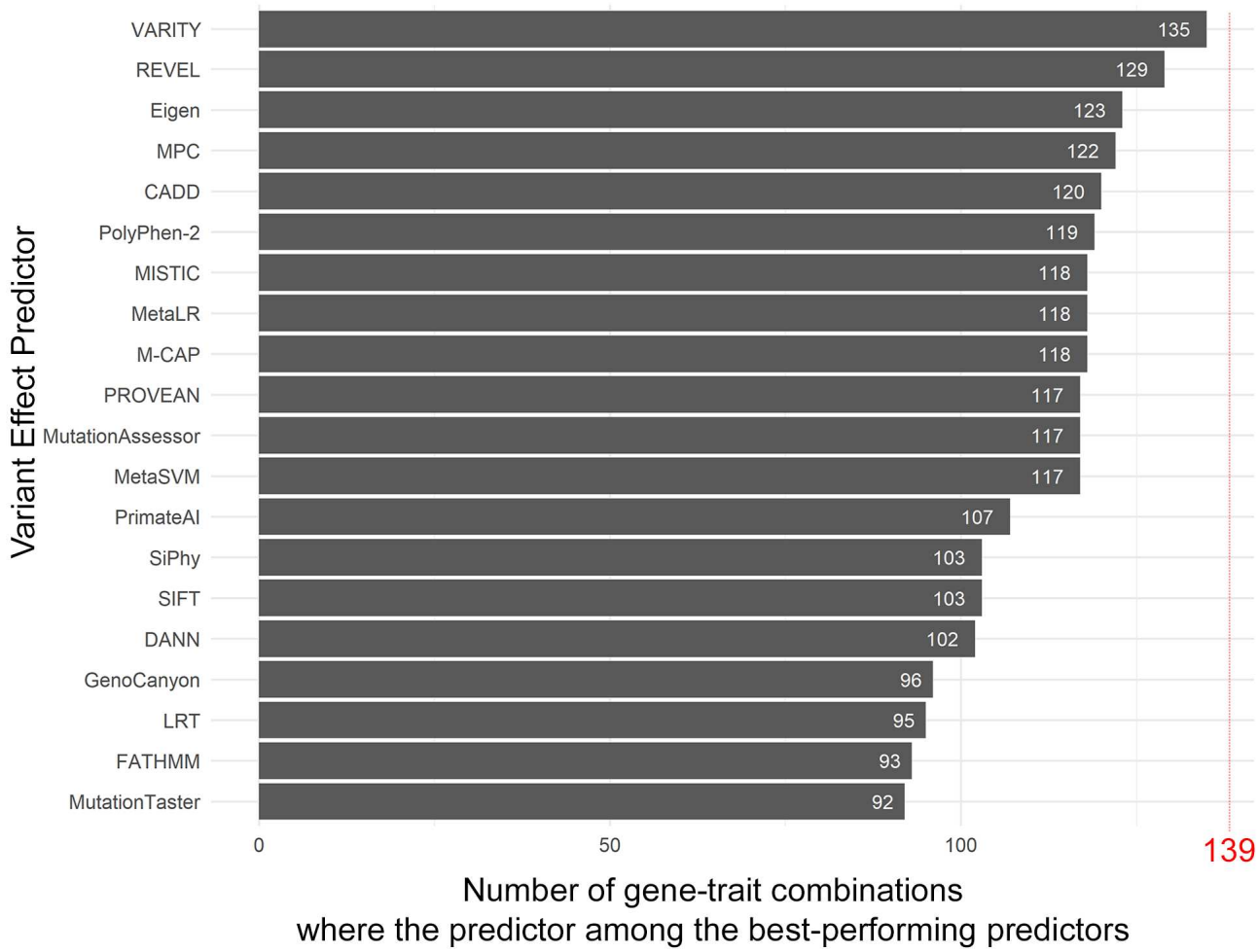

Figure 5. The number of gene-trait combinations for which each predictor is best-performing or indistinguishable from best-performing. The red line highlights the maximum value based on the 139 gene-trait combinations considered in this study.

Supplemental data

Supplemental data includes three tables and two figures.

Declaration of interests

F.P.R.is a scientific advisor and shareholder for Constantiam Biosciences and BioSymetrics, and a Ranomics shareholder. The authors declare no other competing interests.

Acknowledgements

We are grateful to J. Knapp and D. Sheykhkarimli for helpful comments. This work was supported by a Canadian Institutes of Health Research Foundation Grant (F.P.R.), by the National Human Genome Research Institute of the National Institutes of Health Center of Excellence in Genomic Science Initiative (RM1HG010461), the Canada Excellence Research Chairs Program (F.P.R.) and by the One Brave Idea Initiative (jointly funded by the American Heart Association, Verily Life Sciences LLC, and Astra-Zeneca, Inc.).

Data and code availability

The source code generated during this study is available at GitHub: https://github.com/kvnkuang/variant-effect-predictor-assessment. The UK Biobank dataset is available by application via https://www.ukbiobank.ac.uk/.

Reference

1. Vihinen, M. (2021). Functional effects of protein variants. Biochimie *180*, 104–120.

2. Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. *12*, 745–755.

3. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405–424.

4. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2021). Author Correction: The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *590*, E53.

5. Starita, L.M., Ahituv, N., Dunham, M.J., Kitzman, J.O., Roth, F.P., Seelig, G., Shendure, J., and Fowler, D.M. (2017). Variant Interpretation: Functional Assays to the Rescue. Am. J. Hum. Genet. *101*, 315–325.

6. Pavlíková, M., Sokolová, J., Janosíková, B., Melenovská, P., Krupková, L., Zvárová, J., and Kozich, V. (2012). Rare allelic variants determine folate status in an unsupplemented European population. J. Nutr. *142*, 1403–1409.

7. Momozawa, Y., and Mizukami, K. (2021). Unique roles of rare variants in the genetics of complex diseases in humans. J. Hum. Genet. *66*, 11–23.

8. Kanavy, D.M., McNulty, S.M., Jairath, M.K., Brnich, S.E., Bizon, C., Powell, B.C., and Berg, J.S. (2019). Comparative analysis of functional assay evidence use by ClinGen Variant Curation Expert Panels. Genome Med. *11*, 77.

9. Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nat. Methods *11*, 801–807.

10. Matreyek, K.A., Starita, L.M., Stephany, J.J., Martin, B., Chiasson, M.A., Gray, V.E., Kircher, M., Khechaduri, A., Dines, J.N., Hause, R.J., et al. (2018). Multiplex assessment of protein variant abundance by massively parallel sequencing. Nat. Genet. *50*, 874–882.

11. Weile, J., Sun, S., Cote, A.G., Knapp, J., Verby, M., Mellor, J.C., Wu, Y., Pons, C., Wong, C., van Lieshout, N., et al. (2017). A framework for exhaustively mapping functional missense variants. Mol. Syst. Biol. *13*, 957.

12. Kuang, D., Weile, J., Kishore, N., Rubin, A.F., Fields, S., Fowler, D.M., and Roth,

F.P. (2021). MaveRegistry: a collaboration platform for multiplexed assays of variant effect. Bioinformatics.

13. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., and Chan, A.P. (2012). Predicting the functional effect of amino acid substitutions and indels. PLoS One *7*, e46688.

14. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

15. Vaser, R., Adusumalli, S., Leng, S.N., Sikic, M., and Ng, P.C. (2016). SIFT missense predictions for genomes. Nat. Protoc. *11*, 1–9.

16. Riesselman, A.J., Ingraham, J.B., and Marks, D.S. (2018). Deep generative models of genetic variation capture the effects of mutations. Nat. Methods *15*, 816–822.

17. Sundaram, L., Gao, H., Padigepati, S.R., McRae, J.F., Li, Y., Kosmicki, J.A., Fritzilas, N., Hakenberg, J., Dutta, A., Shon, J., et al. (2018). Predicting the clinical impact of human mutation with deep neural networks. Nat. Genet. *50*, 1161–1170.

18. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am. J. Hum. Genet. *99*, 877–885.

19. Wu, Y., Li, R., Sun, S., Weile, J., and Roth, F.P. (2021). Improved pathogenicity prediction for rare human missense variants. Am. J. Hum. Genet.

20. Livesey, B.J., and Marsh, J.A. (2020). Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. Mol. Syst. Biol. *16*, e9380.

21. Niroula, A., and Vihinen, M. (2019). How good are pathogenicity predictors in detecting benign variants? PLoS Comput. Biol. *15*, e1006481.

22. Thusberg, J., Olatubosun, A., and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. Hum. Mutat. *32*, 358–368.

23. Grimm, D.G., Azencott, C.-A., Aicheler, F., Gieraths, U., MacArthur, D.G., Samocha,

K.E., Cooper, D.N., Stenson, P.D., Daly, M.J., Smoller, J.W., et al. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum. Mutat. *36*, 513–523.

24. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. *12*, e1001779.

25. Cirulli, E.T., White, S., Read, R.W., Elhanan, G., Metcalf, W.J., Tanudjaja, F., Fath, D.M., Sandoval, E., Isaksson, M., Schlauch, K.A., et al. (2020). Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. Nat. Commun. *11*, 542.

26. Jurgens, S.J., Choi, S.H., Morrill, V.N., Chaffin, M., Pirruccello, J.P., Halford, J.L., Weng, L.-C., Nauffal, V., Roselli, C., Hall, A.W., et al. (2020). Rare Genetic Variation Underlying Human Diseases and Traits: Results from 200,000 Individuals in the UK Biobank.

27. Van Hout, C.V., Tachmazidou, I., Backman, J.D., Hoffman, J.D., Liu, D., Pandey, A.K., Gonzaga-Jauregui, C., Khalid, S., Ye, B., Banerjee, N., et al. (2020). Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature *586*, 749–756.

28. Curtis, D. (2020). Multiple Linear Regression Allows Weighted Burden Analysis of Rare Coding Variants in an Ethnically Heterogeneous Population. Hum. Hered. *85*, 1–10.

29. Howe, K.L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., Bhai, J., et al. (2021). Ensembl 2021. Nucleic Acids Res. *49*, D884–D891.

30. Krasheninina, O., Hwang, Y.-C., Bai, X., Zalcman, A., Maxwell, E., Reid, J.G., and Salerno, W.J. (2020). Open-source mapping and variant calling for large-scale NGS data from original base-quality scores.

31. Pujar, S., O'Leary, N.A., Farrell, C.M., Loveland, J.E., Mudge, J.M., Wallin, C., Girón, C.G., Diekhans, M., Barnes, I., Bennett, R., et al. (2018). Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. Nucleic Acids Res. *46*, D221–D228.

32. Szustakowski, J.D., Balasubramanian, S., Kvikstad, E., Khalid, S., Bronson, P.G., Sasson, A., Wong, E., Liu, D., Wade Davis, J., Haefliger, C., et al. (2021). Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. Nat. Genet. *53*, 942–948.

33. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

34. Liu, X., Li, C., Mou, C., Dong, Y., and Tu, Y. (2020). dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. *12*, 103.

35. Storey, J.D. (2002). A Direct Approach to False Discovery Rates. J. R. Stat. Soc. Series B Stat. Methodol. *64*, 479–498.

36. Hobbs, H.H., Brown, M.S., and Goldstein, J.L. (1992). Molecular genetics of the LDL receptor gene in familial hypercholesterolemia. Hum. Mutat. *1*, 445–466.

37. Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., et al. (2020). Ensembl 2020. Nucleic Acids Res. *48*, D682–D688.

38. Ionita-Laza, I., McCallum, K., Xu, B., and Buxbaum, J.D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat. Genet. *48*, 214–220.

39. Samocha, K.E., Kosmicki, J.A., Karczewski, K.J., O'Donnell-Luria, A.H., Pierce-Hoffman, E., MacArthur, D.G., Neale, B.M., and Daly, M.J. (2017). Regional missense constraint improves variant deleteriousness prediction.

40. Hopf, T.A., Ingraham, J.B., Poelwijk, F.J., Schärfe, C.P.I., Springer, M., Sander, C., and Marks, D.S. (2017). Mutation effects predicted from sequence co-variation. Nat. Biotechnol. *35*, 128–135.

41. Yi, S.-W., Yi, J.-J., and Ohrr, H. (2019). Total cholesterol and all-cause mortality by sex and age: a prospective cohort study among 12.8 million adults. Sci. Rep. *9*, 1596.

42. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. *46*, D1062–D1067.

43. Stenson, P.D., Mort, M., Ball, E.V., Chapman, M., Evans, K., Azevedo, L., Hayden, M., Heywood, S., Millar, D.S., Phillips, A.D., et al. (2020). The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. Hum. Genet. *139*, 1197–1207.