Title

CAGI, the Critical Assessment of Genome Interpretation, establishes progress and prospects for computational genetic variant interpretation methods

Journal

Genome Biology, 25(1)

ISSN

1474-760X

Authors

Jain, Shantanu

Bakolitsa, Constantina

Brenner, Steven E

et al.

Publication Date

2024

DOI

10.1186/s13059-023-03113-6

Peer reviewed

The Critical Assessment of Genome Interpretation Consortium
Genome Biology (2024) 25:53
https://doi.org/10.1186/s13059-023-03113-6

Genome Biology

RESEARCH Open Access

CAGI, the Critical Assessment of Genome Interpretation, establishes progress and prospects for computational genetic variant interpretation methods

The Critical Assessment of Genome Interpretation Consortium[1*]

*Correspondence:
brenner@berkeley.edu;
predrag@northeastern.edu;
jmoult@umd.edu[1] https://genomeinterpretation.org

Abstract

Background: The Critical Assessment of Genome Interpretation (CAGI) aims to advance the state-of-the-art for computational prediction of genetic variant impact, particularly where relevant to disease. The five complete editions of the CAGI community experiment comprised 50 challenges, in which participants made blind predictions of phenotypes from genetic data, and these were evaluated by independent assessors.

Results: Performance was particularly strong for clinical pathogenic variants, including some difficult-to-diagnose cases, and extends to interpretation of cancer-related variants. Missense variant interpretation methods were able to estimate biochemical effects with increasing accuracy. Assessment of methods for regulatory variants and complex trait disease risk was less definitive and indicates performance potentially suitable for auxiliary use in the clinic.

Conclusions: Results show that while current methods are imperfect, they have major utility for research and clinical applications. Emerging methods and increasingly large, robust datasets for training and assessment promise further progress ahead.

Background

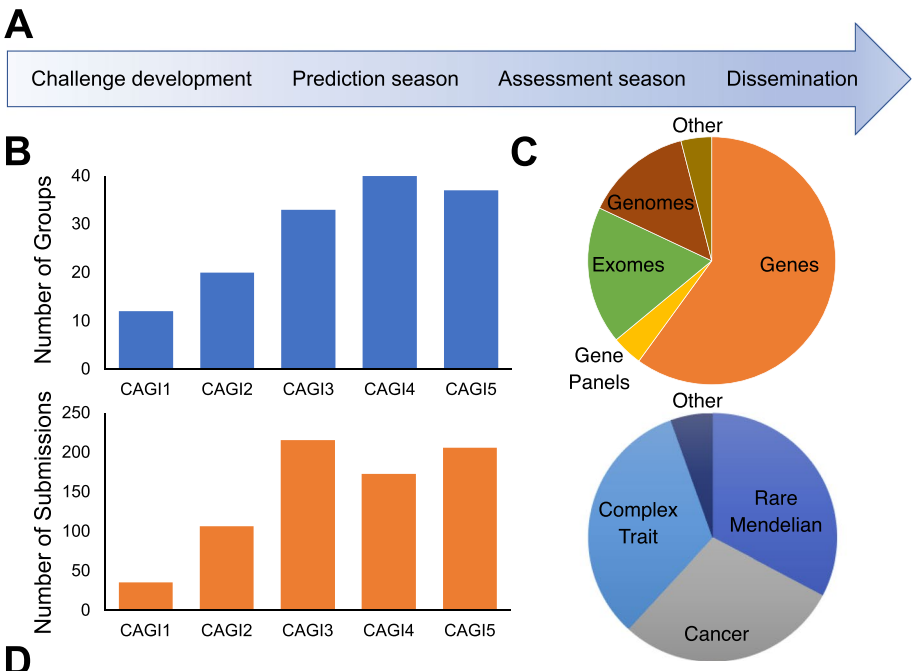
Rapidly accumulating data on individual human genomes hold the promise of revolutionizing our understanding and treatment of human disease [1, 2]. Effectively leveraging these data requires reliable methods for interpreting the impact of genetic variation. The DNA of unrelated individuals differs at millions of positions [3], most of which make negligible contribution to disease risk and phenotypes. Therefore, interpretation approaches must be able to identify the small number of variants with phenotypic

significance, including those causing rare disease such as cystic fibrosis [4], those contributing to increased risk of cancer [5] or acting as cancer drivers [6], those contributing to complex traits such as type II diabetes [7], and those affecting the response of individuals to drugs such as warfarin [8]. Identifying the relationship between variants and phenotype can also lead to new biological insights and new therapeutic strategies. Until recently, interpretation of the role of specific variants has either been acquired by empirical observations in the clinic, thus slowly accumulating robust knowledge [9], or by meticulous and often indirect in vitro experiments whose interpretation may be challenging. Computational methods offer a third and potentially powerful approach, and over one hundred have been developed [10], but their power, reliability, and clinical utility have not been established. Some computational approaches aim to directly relate sequence variation to disease or other organismal phenotypes; for example, using evolutionary conservation [11]. Other methods suggest impact on disease only secondarily, as they aim to relate genetic variants to functional properties such as effects on protein stability [12], intermolecular interactions [13], splicing [14], expression [15], or chromatin organization [16].

The Critical Assessment of Genome Interpretation (CAGI) is an organization that conducts community experiments to assess the state-of-the-art in computational interpretation of genetic variants. CAGI experiments are modeled on the protocols developed in the Critical Assessment of Structure Prediction (CASP) program, [17] adapted to the genomics domain. The process is designed to assess the accuracy of computational methods, highlight methodological innovation and reveal bottlenecks, guide future research, contribute to the development of new guidelines for clinical practice, and provide a forum for the dissemination of research results. Participants are periodically provided with sets of genetic data and asked to relate these to unpublished phenotypes. Independent assessors evaluate the anonymized predictions, promoting a high level of rigor and objectivity. Assessment outcomes together with invited papers from participants have been published in special issues of the journal *Human Mutation* [18, 19]. Since CAGI has stewardship of genetic data from human research participants, an essential part of the organizational structure is its Ethics Forum composed of ethicists and researchers, together with patient advocates. Further details are available at https://genomeinterpretation.org/.

Over a period of a decade, CAGI has conducted five rounds of challenges, 50 in all, attracting 738 submissions worldwide (Fig. 1, Additional file 1: Table S1, and Additional file 1). Challenge datasets have come from studies of variant impact on protein stability [20, 21] and functional phenotypes such as enzyme activity [22, 23], cell growth [24], and whole-organism fitness [25], with examples relevant to rare monogenic disease [26], cancer [27], and complex traits [28, 29]. Variants in these datasets have included those affecting protein function, gene expression, and splicing and have comprised single base changes, short insertions or deletions (indels), as well as structural variation. Genomic scale has ranged from single nucleotides to complete genomes, with inclusion of some complementary multiomic and clinical information (Fig. 1).

In this work, we analyze the first decade of CAGI challenges in a consistent clinically relevant framework, and we identify emergent themes and unifying principles



D

CAGI1	CAGI2	CAGI3	CAGI4	CAGI5
CBS	Asthma twins	BRCA1 & BRCA2	Bipolar disorder	Annotate all missense
CHEK2	Breast cancer PGX	Crohn's disease	Crohn's disease	CALM1
PGP	CBS	FCH	eQTL causal SNPs	CHEK2
	Crohn's disease	HA	Hopkins panel	Clotting disease
	Mouse exomes	MR-1	NAGLU	ENIGMA BRCA1-2
	MR-1	MRE11	NPM-ALK	Frataxin
	p53	NBS1	PGP	GAA
	PGP	p16	Pyruvate kinase	ID panel
	RAD50	PGP	SickKids	MaPSy
	riskSNPs	riskSNPs	SUMO ligase	PCM1
	SCN5A	TP53 splicing	Warfarin exomes	PTEN
				Regulation saturation
				SickKids
				TPMT
				Vex-Seq

Fig. 1 CAGI timeline, participation, and range of challenges. **A** Stages in a round of CAGI, typically extending over 2 years. Each round includes a set of challenges with similar timelines. **B** Number of participating unique groups (in blue) and submissions (in orange) across CAGI rounds. **C** Scale of the genetic data (top) and phenotypic characterization (bottom) of CAGI challenges. Some challenges belong to more than one category and are included more than once. **D** CAGI challenges, listed by round. Coloring is by scale of genetic data and phenotypic characterization according to **C**. See Supplemental Table 1 for more details

across the range of genome variation interpretation. Results are presented from three perspectives to provide (i) the clinical community with an assessment of the usefulness and limitations of computational methods, (ii) the biomedical research community with information on the current state-of-the-art for predicting variant impact on a range of biochemical and cellular phenotypes, and (iii) the developers of computational methods with data on method performance with the aim of spurring further

innovation. This latter perspective is particularly important because of the recent successes of artificial intelligence approaches in related fields [30, 31]. For each theme, specific examples of performance are provided, based on particular ranking criteria. As always in CAGI, these should not be interpreted as identifying winners and losers−−other criteria might result in different selections. Further, these selections were made by authors of this paper, some of whom have been CAGI participants, rather than by independent assessors. However, the examples shown are consistent with the assessors' earlier rankings.

Results

Biochemical effect predictions for missense variants are strongly correlated with the experimental data, but individual predicted effect size accuracy is limited

The pathogenicity of missense variants implicated in monogenic disease and cancer is often supported by in vitro experiments that measure effects on protein activity, cell growth, or various biochemical properties [32]. Thirteen CAGI challenges have assessed the ability of computational methods to estimate these functional effects using datasets from both high- and low-throughput experimental assays, and ten of these have been reanalyzed here.

Figure 2 shows selected results for two challenges, each with a different type of biochemical effect. In the NAGLU challenge [22], participants were asked to estimate

(See figure on next page.)

Fig. 2 Predicting the effect of missense variants on protein properties: Results for two example CAGI challenges. Each required estimation of continuous phenotype values, enzyme activity in a cellular extract for NAGLU and intracellular protein abundance for PTEN, for a set of missense variants. Selection of methods is based on the average ranking over four metrics for each participating method: Pearson's correlation, Kendall's tau, ROC AUC, and truncated ROC AUC; see "Methods" for definitions. **A** Relationship between observed and predicted values for the selected method in each challenge. "Benign" variants are yellow and "pathogenic" are purple (see text). The diagonal line represents exact agreement between predicted and observed values. Dashed lines show the thresholds for pathogenicity for observed (horizontal) and predicted biochemical values (vertical). For NAGLU, below the pathogenicity threshold, there are 12 true positives (lower left quadrant) and three false positives (upper left quadrant), suggesting a clinically useful performance. Bars below each plot show the boundaries for accuracy meeting the threshold for Supporting (green), Moderate (blue), and Strong (red) clinical evidence, with 95% confidence intervals. **B** Two measures of overall agreement between computational and experimental results, for the two selected performing methods and positive and negative controls, with 95% confidence intervals. An older method, PolyPhen-2, provides a negative control against which to measure progress over the course of the CAGI experiments. Estimated best possible performance is based on experimental uncertainty and provides an empirical upper limit positive control. The color code for the selected methods is shown in panel **C**. **C** ROC curves for the selected methods with positive and negative controls, using estimated pathogenicity thresholds. **D** Truncated ROC curves showing performance in the high true positive region, most relevant for identifying clinically diagnostic variants. The true positive rate and false positive rate thresholds for the Supporting, Moderate, and Strong evidential support are shown for one selected method. **E** Estimated probability of pathogenicity (left *y*-axis) and positive local likelihood ratio (right *y*-axis) as a function of one selected method's score. Predictions with probabilities over the red, blue, and green thresholds provide Strong, Moderate, and Supporting clinical evidence, respectively. Solid lines show smoothed trends. Prior probabilities of pathogenicity are the estimated probability that any missense variant in these genes will be pathogenic. For NAGLU, the probabilities of pathogenicity reach that needed for a clinical diagnosis of "likely pathogenic." For predicted enzyme activity less than 0.11, the probability provides Strong evidence, below 0.17 Moderate evidence, and below 0.42, Supporting evidence. The percent of variants encountered in the clinic expected to meet each threshold are also shown. Performance for PTEN shows that the results are consistent with providing Moderate and Supporting evidence levels for some variants

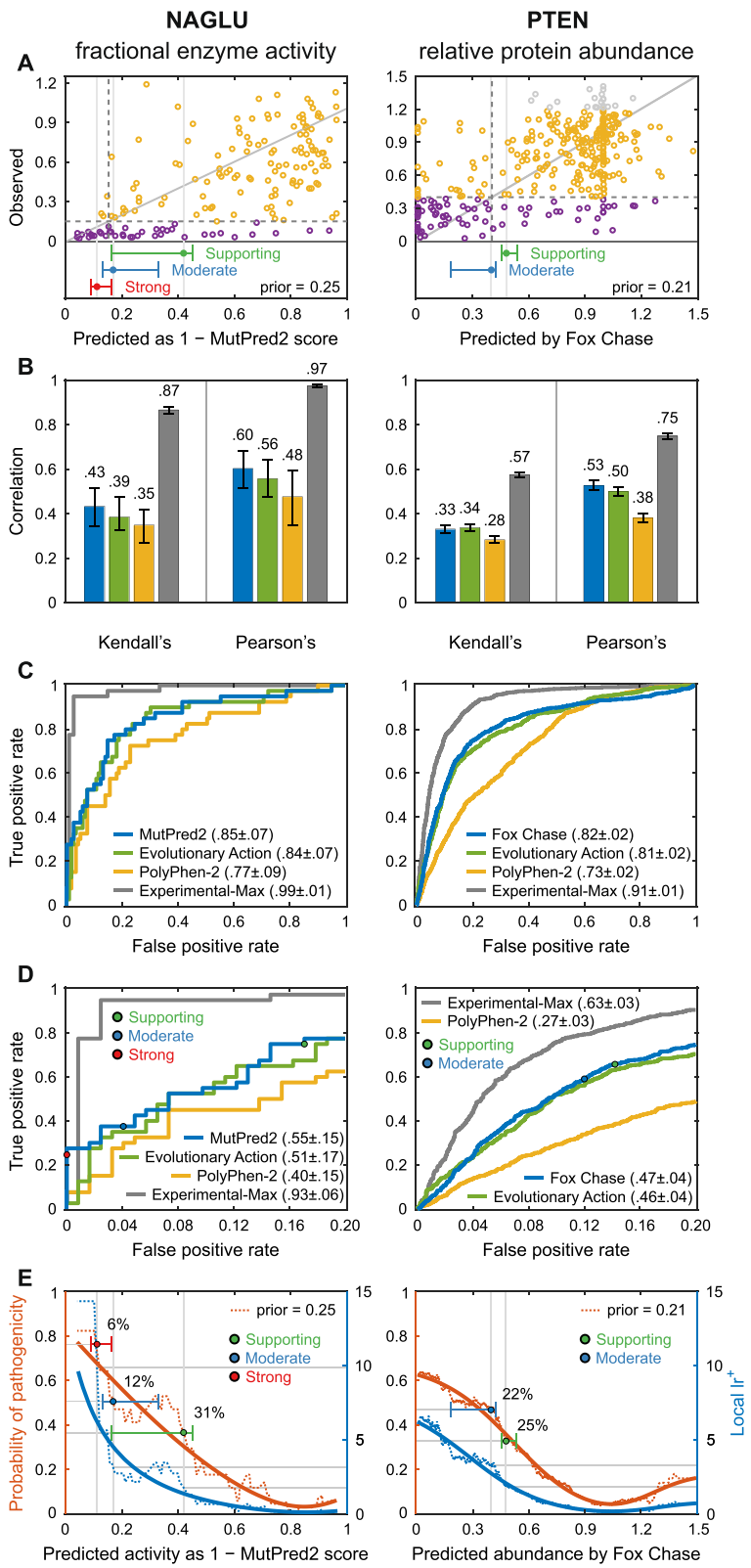


Fig. 2 (See legend on previous page.)

the relative enzyme activity of 163 rare missense variants in N-acetyl-glucosaminidase found in the ExAC database [33]. In the PTEN challenge [20], participants were asked to estimate the effects of a set of 3716 variants in the phosphatase and tensin homolog on the protein's stability as measured by relative intracellular protein abundance in a high-throughput assay [34]. For both challenges, the relationship between estimated and observed phenotype values shows high scatter (Fig. 2A). There is modest improvement with respect to a well-established older method, PolyPhen-2 [35], which we consider a baseline. This is a trend consistently seen in other missense challenges (Additional file 2: Table S2). How much of this improvement is due to the availability of larger and more reliable training sets rather than methodological improvements is unknown. Consistent with the scatter plots, there is moderate agreement between predicted and experimental values as measured by Pearson's correlation and Kendall's tau (Fig. 2B).

Over all ten analyzed missense functional challenges (Additional file 3: Table S3, Additional file 1: Figures S1-S6), Pearson's correlation for the selected methods ranges between 0.24 and 0.84 (average correlation $\overline{r} = 0.55$) and Kendall's tau ranges between 0.17 and 0.63 ($\overline{\tau} = 0.40$), both showing strong statistical significance over the random model ($\overline{r} = 0$, $\overline{\tau} = 0$). The PolyPhen-2 baseline achieves $\overline{r} = 0.36$ and $\overline{\tau} = 0.23$. Direct agreement between observed and predicted values is measured by R^2, which is 1 for a perfect method and 0 for a control method that assigns the mean of the experimental data for every variant. For NAGLU, the highest R^2 achieved is 0.16, but for PTEN it is only -0.09. Over the ten biochemical challenges, the highest R^2 value ranges between -0.94 and 0.40, with an average of -0.19. The relatively poorer performance shown by this criterion compared with Pearson's and Kendall's correlation metrics suggests that the methods are often not well calibrated to the experimental value, reflecting the fact that they are rarely designed for predictions of continuous values and scales of this kind. Overall, performance is far above random but modest in terms of absolute accuracy.

Diversity of methods

A diverse set of methods was used to address the biochemical effect challenges, varying in the type of training data, input features, and statistical framework. Most were trained on pathogenic versus presumed benign variants [10, 36]. At first glance, a binary classification approach appears ill-suited to challenges which require prediction across a full range of phenotype values. In practice, function and pathogenicity are related [37], and so these methods performed as well as the few trained specifically to identify alteration of function [38].

Many methods are based on measures that reflect the evolutionary fitness of substitutions and population dynamics, rather than pathogenicity or functional properties. The relationship between fitness, pathogenicity, and function is complex, perhaps limiting performance. To partly address this, some methods also exploit functional roles of specific sequence positions, particularly by utilizing UniProtKB annotations and predicted structural and functional properties [38–41].

Current methods typically address the effect of single variants in isolation from possible epistatic factors although many apparently monogenic diseases are influenced by modifier variants. For example, severity of cystic fibrosis is affected by several genes beyond CFTR [4], and studies of loss of function variants in general populations revealed cases where a strong disease phenotype is expected but not observed, implying the presence of compensating variants [42].

Despite the broad range of algorithms, training data, features, and the learning setting, there is a strong correlation between results of the leading methods (Pearson's correlation ranges from 0.6 to 0.9), almost always stronger than the correlation between specific methods and experiment (Additional file 1: Figure S8). The level of inter-method correlation is largely unrelated to the level of correlation with experiment, which varies widely from about 0.24 (CALM1) to 0.6 (NAGLU). Why correlation between methods is stronger than with experiment is unclear, though it may be affected by the relatedness of functional disruption, evolutionary conservation, and pathogenicity as well as common training data and experimental bias. The assessor for the NAGLU challenge identified 10 variants where experiment disagrees strongly with predicted values for all methods [22]. When these are removed, the correlation between the leading methods' results and experiment increases from 0.6 to 0.73 (Additional file 1: Figure S8), although it is still lower than the correlation between the two leading methods (0.82), of which, surprisingly, one is supervised [40] and the other is not [43]. It could be that these 10 variants are cases where the computational methods systematically fail, or it could be that most are some form of experimental artifact. In situations like this, follow-up experiments are needed.

Structure-informed approaches

Some methods use only biophysical input features, and in some cases are trained on the effect of amino acid substitutions on protein stability, rather than pathogenicity or functional impact. Benchmarking suggests that a large fraction of rare disease-causing and cancer driver missense mutations act through destabilization [44, 45], so there is apparently considerable potential for these approaches. These methods have been effective on challenges directly related to stability, being selected as first and second for the PTEN and TMPT protein abundance challenges and first for the Frataxin change of free energy of folding challenge. They have been among top performers in a few other challenges, sometimes in combination with sequence feature methods, for example, cancer drivers in CDKN2A and rescue mutations in TP53 [21]. Generally, however, these methods, along with the structure-based machine learning methods, have not been as successful as expected compared to the methods that are primarily sequence-based. Three factors may improve their performance in future. First, better combination with the sequence methods will likely mitigate the problem of false negatives; that is, pathogenic variants that are not stability related. Second, until recently, use of structure has been restricted by low experimental structural coverage of the human proteome (only about 20% of residues). Because of recent dramatic improvements in the accuracy of computed protein structures [46], variants in essentially all residues in ordered structural regions are now amenable to this approach. Third, better estimation of the effect of missense mutations on stability [47] should improve accuracy. An advantage of biophysical and related

methods is that they can sometime provide greater insight into underlying molecular mechanisms (Additional file 1: Figure S13).

Domain-level information had the potential to deliver improved performance in other instances, such as CBS, where the heme-binding domain present in humans was absent from the yeast ortholog, and RAD50, for which assessment showed that restricting predictions of deleteriousness to the specific domain involved in DNA repair would have substantially improved the accuracy of several methods.

Computational methods can substantially enhance clinical interpretation of missense variants

The most direct test of the clinical usefulness of computational methods is to assess their ability to correctly assign pathogenic or benign status for clinically relevant variants. CAGI challenges have addressed this for rare disease variant annotations and for germline variants related to cancer risk.

Results for predicting the biochemical effects of missense mutations inform clinical applications

For some biochemical challenges, it is possible to relate the results to clinical utility of the methods. For NAGLU, some rare variants in the gene cause recessive Sanfilippo B disease. Disease strongly correlates with variants conferring less than 15% enzyme activity [22], allowing variants in the study to be classified as pathogenic or benign on that basis (purple and yellow circles in Fig. 2A). Figure 2A shows that 12 out of the 15 variants with less than 15% predicted activity using the selected method also have less than 15% experimental activity, suggesting high positive predictive value and clinical usefulness for assigning pathogenicity. On the other hand, 28 of the 40 variants with measured activity below 15% are predicted to have higher activity so there are also false negatives. For PTEN, information on the relationship to disease is less well established, but data fall into low and high abundance distributions [20], and the assessor suggested a pathogenicity threshold at the distribution intersection.

Performance in correctly classifying variants as pathogenic is often represented by ROC curves (Fig. 2C), showing the tradeoff between true positive (y-axis) and false positive (x-axis) rates as a function of the threshold used to discretize the phenotype value returned by a prediction method, and summarized by the area under that curve (AUC). The selected methods return AUCs greater than 0.8 for both challenges. Over all reanalyzed biochemical effect challenges, the top AUC ranges from 0.68 to 1.0, with an average of $\overline{\text{AUC}} = 0.83$, and with high statistical significance over a random model (AUC = 0.5). The PolyPhen-2 baseline has $\overline{\text{AUC}} = 0.74$, see Additional file 3: Table S3. However, all models fall well short of the empirical limit ($\overline{\text{AUC}} = 0.98$) estimated from variability in experimental outputs. Because the experimental uncertainties are based on technical replicates, the experimental AUCs are likely overestimated, so it is difficult to judge how much further improvement might be possible. The full ROC curve areas provide a useful metric to measure the ability of the methods to separate pathogenic from other variants. In a clinical setting though, the left portion of the curve is often the most relevant; that is, the fraction of pathogenic variants identified without incurring too high a level of false positives, where the level of tolerated false positives is application dependent.

Figure 2D uses truncated ROC curves to show the performance in this region, with the selected methods' AUCs reaching 0.55 for NAGLU and 0.47 for PTEN. The smaller value for the PTEN truncated ROC curve AUC reflects the higher fraction of false positives at the left of the PTEN scatter plot, particularly those variants predicted to have near-zero protein abundance but with high observed values.

For use in the clinic, the quantity of most interest is the probability that a variant is diagnostic of disease (i.e., can be considered pathogenic), given the available evidence. In addition to the information provided by a computational method, initial evidence is also provided by knowledge of how likely any variant in a particular gene is to be diagnostic of the disease of interest [48]. For example, for NAGLU, about 25% of the rare missense variants in ExAC were found to have less than 15% enzyme activity, [49] suggesting that there is an approximately 25% prior probability that any rare missense variant found in the gene will be pathogenic (a prior odds of pathogenicity of 1:3). To obtain the desired posterior probability of pathogenicity, which is also the local positive predictive value at the score s returned by a method, one can use a standard Bayesian odds formulation [50]

$$\text{posterior odds of pathogenicity} = \text{lr}^+ \times \text{prior odds of pathogenicity},$$

where the local positive likelihood ratio, lr^+, is the slope of the ROC curve at the score value s; see "Methods" for a formal discussion.

Figure 2E shows lr^+ and the posterior probability of pathogenicity for NAGLU and PTEN. For NAGLU, at low predicted enzyme activities lr^+ rises sharply to about 15. The corresponding posterior probability is 0.8. For PTEN, lr^+ reaches a value of about 6. Using a pathogenicity prior of 0.21 (Additional file 1), the corresponding posterior probability of pathogenicity is 0.6. ACMG/AMP sequence variant interpretation guidelines recommend a probability of pathogenicity of ≥ 0.99 to label a variant "pathogenic" and ≥ 0.90 to label one "likely pathogenic", the thresholds for clinical action [32, 51, 52]. So for these and other biochemical challenges, the computational evidence alone is not sufficiently strong to classify the variants other than as variant of uncertain significance.

However, the clinical guidelines integrate multiple lines of evidence to contribute to meeting an overall probability of pathogenicity threshold, so that it is not necessary (or indeed possible) for computational methods alone to provide a pathogenicity assignment. The guidelines provide rules that classify each type of evidence as Very Strong, Strong, Moderate, and Supporting [32]. For example, a null mutation in a gene where other such mutations are known to cause disease is considered Very Strong evidence, while at the other extreme, a computational assignment of pathogenicity for a missense mutation is currently considered only Supporting. Although these guidelines were originally defined in terms of evidence types, Tavtigian et al. [52] have shown that the rules can be approximated using a Bayesian framework, with each threshold corresponding to reaching a specific positive likelihood ratio; e.g., $\text{lr}^+ = 2.08$ for Supporting evidence when the prior probability is 0.1 (Methods). The resulting thresholds for each level of evidence are shown below the scatter plots in Fig. 2A and in the posterior probability of pathogenicity plots in Fig. 2E. For NAGLU, for the selected method, predicted enzyme activities lower than 0.11 correspond to Strong evidence, below 0.17 to Moderate, and below 0.42 to Supporting. These thresholds correspond to approximately 31% of rare variants in this gene providing Supporting evidence, 12% Moderate, and 6% Strong. The

top-performing methods for the ten biochemical missense challenges all reach Supporting, and sometimes Moderate and Strong evidential support (Additional file 1: Figures S1-S6, Additional file 2: Table S2 and Additional file 3: Table S3). These results are encouraging in that they suggest this framework can supply a means of quantitatively evaluating the clinical relevance of computational predictions and that under appropriate circumstances, computational evidence can be given more weight than at present. The next section explores these properties further.

Identifying rare disease variants

The ClinVar [9] and HGMD [53] databases provide an extensive source of rare disease-associated variants against which to test computational methods. A limitation is that most methods have used some or all of these data in training, making it difficult to perform unbiased assessments. The prospective "Annotate All Missense" challenge assessed the accuracy of those predictions on all missense variants that were annotated as pathogenic or benign in ClinVar and HGMD after May 2018 when predictions were recorded, through December 2020, so avoiding training contamination. All predictions directly submitted for the challenge as well as all precomputed predictions deposited in the dbNSFP database [54] before May 2018 (dbNSFP v3.5) were evaluated, predictions from a total of 26 groups.

All selected methods, including PolyPhen-2, achieved high AUCs, ranging from 0.85 to 0.92 for separating "pathogenic" from "benign" variants, and only slightly lower values (maximum AUC 0.88) when "likely pathogenic" and "likely benign" are included (Fig. 3A). The two metapredictors, REVEL [55] and Meta-LR [56], tools that incorporate predictions from multiple other methods, perform slightly better than primary methods, although VEST3 and VEST4 [39] outperformed Meta-LR. There is a substantial improvement over the performance of PolyPhen-2, especially in the left part of the ROC curve (Fig. 3B), though as with the biochemical effect challenges, some of that may be

(See figure on next page.)

Fig. 3 Performance of computational methods in correctly identifying pathogenic variants in the two principal rare disease variant databases, HGMD and ClinVar. The left panels show data for variants labeled as "pathogenetic" in ClinVar and "DM" in HGMD together with "benign" in ClinVar. The right panels add variants labeled as "likely pathogenic" and "likely benign" in ClinVar as well as "DM?" in HGMD. Meta and single method examples were selected on the basis of the average ranking of each method for the ROC and truncated ROC AUCs. See Additional file 1 for more details and selection criteria. **A** ROC curves for the selected metapredictors and single methods, together with a baseline provided by PolyPhen-2. Particularly for pathogenic variants alone, impressively high ROC areas are obtained, above 0.9, and there is a substantial improvement over the older method's performance. **B** Blowup of the left-hand portion of the ROC curves, most relevant to high confident identification of pathogenic variants. Clinical thresholds for Supporting, Moderate, and Strong clinical evidence are shown. **C** Local positive likelihood ratio as a function of the confidence score returned by REVEL. Very high values (> 100) are obtained for the most confident pathogenic assignments. **D** Local posterior probability of pathogenicity; that is, probability that a variant is pathogenic as a function of the REVEL score for the two prior probability scenarios. For a prior probability of 0.1, typical of a single candidate gene situation (solid line) and database pathogenic and benign variants (left panel) the highest-scoring variants reach posterior probability above 0.9, strong enough evidence for a clinical assignment of "likely pathogenic." In both panels, variants with a score greater than 0.45 provide Supporting clinical evidence (green threshold), and scores greater than 0.8 provide Strong evidence (red threshold). The estimated % of variants encountered in a clinical setting expected to meet each threshold are also shown. For example, about 14% of variants provide Supporting evidence. Dotted lines show results obtained with a prior probability of 0.01

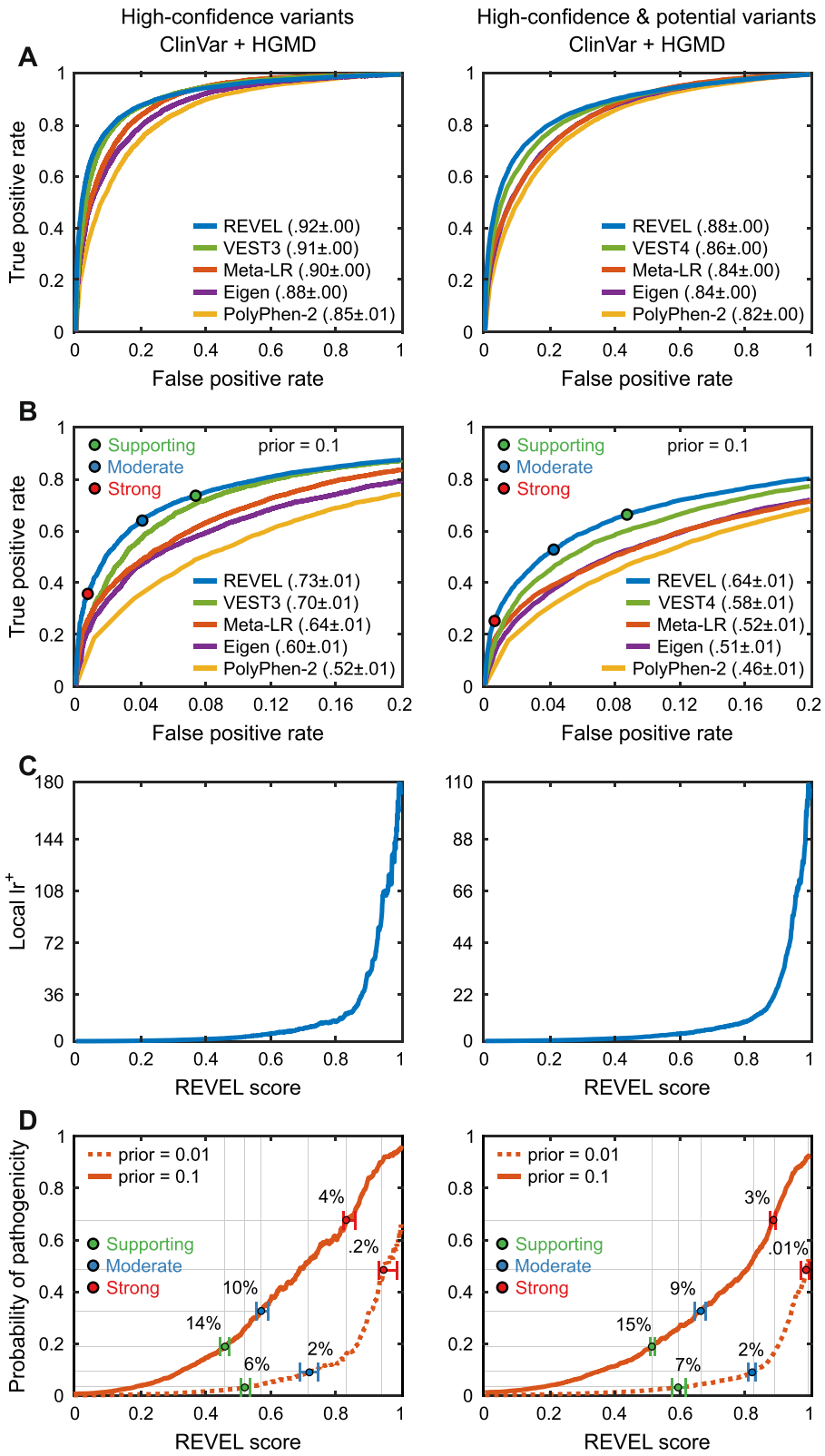


Fig. 3 (See legend on previous page.)

due to the availability of larger and more reliable training sets. Additional file 4: Table S4 shows slightly higher performance on ClinVar pathogenic variants than HGMD; however, these resources use different criteria for assigning pathogenicity.

The lower panels in Fig. 3 show positive likelihood ratios and posterior probability of pathogenicity for a selected metapredictor, REVEL [55]. With a prior probability of pathogenicity of 0.1 (approximating the prior when examining possible pathogenic variants in one or a few genes in a diagnostic setting; see "Methods") 14, 10, and 4% of variants reach the Supporting, Moderate, and Strong evidence thresholds, respectively. With the much smaller prior of 0.01, representative of screening for possible secondary variants, about 6% of variants will provide Supporting evidence and 2% reach Moderate. These estimates are not exact since there may be significant differences between the distribution and properties of variants in these databases and those encountered in the clinic. For example, some genes have only benign variant assignments in the databases, and these might be excluded from consideration in the clinic. Performing the analysis on only genes with both pathogenic and benign assignments slightly reduced performance—highest AUC on the confident set of variants is 0.89 instead of 0.92. The selected methods also change slightly; see Additional file 1: Figure S9. In spite of this and other possible caveats, the overall performance of the computational methods is encouraging and, as with biochemical effect challenges, suggests that the computational methods can provide greater benefit in the clinic than recognized by the current standards.

Identifying germline cancer risk variants

About a quarter of CAGI experiments have involved genes implicated in cancer (Fig. 1) and have included variants in BRCA1, BRAC2, PTEN, TPMT, NSMCE2 (coding for SUMO-ligase), CHEK2, the MRN complex (RAD50, MRE11, and NBS1), FXN, NPM-ALK, CDKN2A, and TP53. An additional challenge addressed breast cancer pharmacogenomics. From a cancer perspective, the most informative of these is a challenge provided and assessed by members of the ENIGMA consortium [27], using a total of 321 germline BRCA1/BRCA2 missense and in-frame indel variants. Performance on this challenge was impressively high, with four groups providing submissions that gave AUCs greater than 0.9 and two with AUCs exceeding 0.97. In the other BRCA1/BRCA2 variant challenge, the highest AUC is 0.88 on a total of 10 missense variants. The strong results may reflect the fact these are highly studied genes. More and larger scale challenges with a variety of genes are required in order to draw firm conclusions. Further details of cancer challenges are provided in Additional file 1: Figure S10 and Additional file 5: Table S5.

Assessing methods that estimate the effect of variants on expression and splicing is difficult, but results show these can contribute to variant interpretation

Variants that regulate the abundance and isoforms of mRNA either through altered splicing or through altered rates of transcription play a significant role in disease, particularly complex traits. CAGI has included four challenges using data from high-throughput assays of artificial gene constructs, two for splicing and two for expression. For all four, evaluation of the results is limited by a combination of small

effect sizes (changes larger than twofold in splicing or expression are rare in these challenges) and experimental uncertainty, but some interesting properties can be identified.

Splicing

The CAGI splicing challenges used data from high-throughput minigene reporter assays [57].

The MaPSy challenge asked participants to identify which of a set of 797 exonic single-nucleotide HGMD disease variants affect splicing and by how much. Two experimental assays were available, one in vitro on a cell extract and the other by transfection into a cell line. Only variants that produced a statistically robust change of at least 1.5-fold were considered splicing changes. Figure 4 summarizes the results. The top-performing groups achieved moderately high AUCs of 0.84 and 0.79 and the highest lr^+ is about 6. Notably, very few variants qualify as significant splicing changes, and there are inconsistences between the two assays, with a number of variants appearing to have a fold change substantially greater than 1.5 in one assay but not the other. Additionally, the experimental noise significantly overlaps with many splicing differences. For these reasons, it is unclear what maximum AUC could be achieved by a perfect method.

The Vex-Seq challenge required participants to predict the extent of splicing change introduced by 1098 variants in the vicinity of known alternatively spliced exons [57]. Additional file 1: Figure S11 shows that performance was rather weak for identification of variants that increase splicing (top AUC = 0.71), but that may be because many points classified as positive are experimental noise. There are more variants that show a

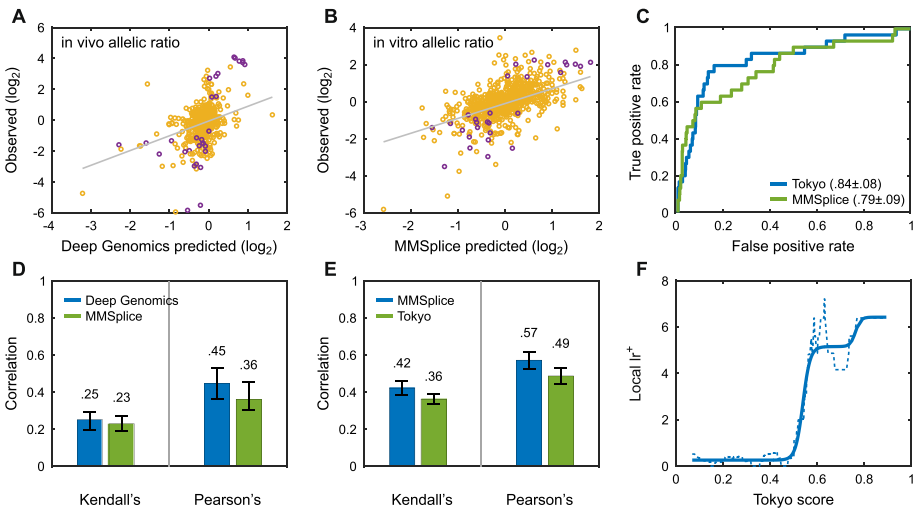


Fig. 4 Performance of computational methods in identifying variants that affect splicing in the MaPSy challenge. Methods were selected based on the average ranking over three metrics: Pearson's correlation, Kendall's tau, and ROC AUC. Scatter plots, Kendell's tau, and Pearson's correlation results are shown for in vivo (**A**, **D**) and in vitro assays (**B**, **E**) separately. The small number of purple points in the scatter plots represent splicing fold changes greater than 1.5-fold. The ROC curve (**C**) shows performance in variant classification for the two selected methods. The maximum local positive likelihood ratio (lr^+, **F**) may be large enough for use as auxiliary information, see "Discussion" (solid line is smoothed fit to the data)

statistically robust decrease in splicing, and prediction performance is correspondingly stronger (top AUC = 0.78). The assessor noted additional nuances in performance [57].

The selected high-performing method for both challenges (MMSplice [58]) decomposes the sequence surrounding alternatively spliced exons into distinct regions and evaluates each region using separate neural networks [58]. More detailed splicing results are provided in Additional file 6: Table S6.

Transcription

Several CAGI challenges have assessed the ability of computational methods to identify single base changes that affect the expression level of specific genes. The CAGI4 eQTL challenge assessed whether methods could find causative variants in a set of eQTL loci, using a massively parallel reporter assay [59]. Because of linkage disequilibrium and sparse sampling, variants associated with an expression difference in an eQTL screen are usually not those directly causing the observed expression change. Rather, a nearby variant will be. The challenge had two parts. Participants were asked to predict whether insertion of the section of genomic DNA around each variant position into the experimental construct produced any expression. Additional file 1: Figure S12A shows that the top-performing methods were effective at this—the largest AUC is 0.81. The second part of the challenge required participants to predict which variants affect expression levels. Here the results are much less impressive (Additional file 1: Figure S12B). The scatter plot shows a weak relationship between observed and predicted expression change and Pearson's or Kendall's correlation are also small. The best AUC is only 0.66 and the maximum lr^+ is about 5. Most of the experimental expression changes are small (less than twofold) and may be largely experimental noise, partly accounting for the apparent poor performance. But as the scatter plot shows, a subset of the variants with largest effects could not be identified by the top-performing method. A combination of experimental and computational factors contributed to poor performance, and more challenges of this sort are needed.

The CAGI5 regulation saturation mutagenesis challenge examined the impact on expression of every possible base pair variant within five disease-associated human enhancers and nine disease-associated promoters [60]. As shown in Fig. 5, performance is stronger in promoters than enhancers and stronger for decreases in expression compared with increases. Fewer variants show experimental increases and these tend to be less well distinguished from noise. Performance for small expression changes is hard to evaluate because of overlap with experimental noise. Nevertheless, the highest AUC for promoter impact prediction is 0.81 while the highest AUC for enhancer impact prediction is 0.79, relatively respectable values. In addition, the scatter plots show that large decreases in expression are well predicted, suggesting the methods are quite informative for the most significant effects.

The CAGI splicing and expression challenges are not as directly mappable to disease and clinical relevance as in some other challenge areas. Variants have been a mixture of common and rare and the use of artificial constructs in high-throughput experiments limits relevance of challenge performance in the whole-genome context. Nevertheless, the results do have potential applications. In complex trait disease genome-wide association studies (GWAS), the variants found to be associated with a phenotype are usually

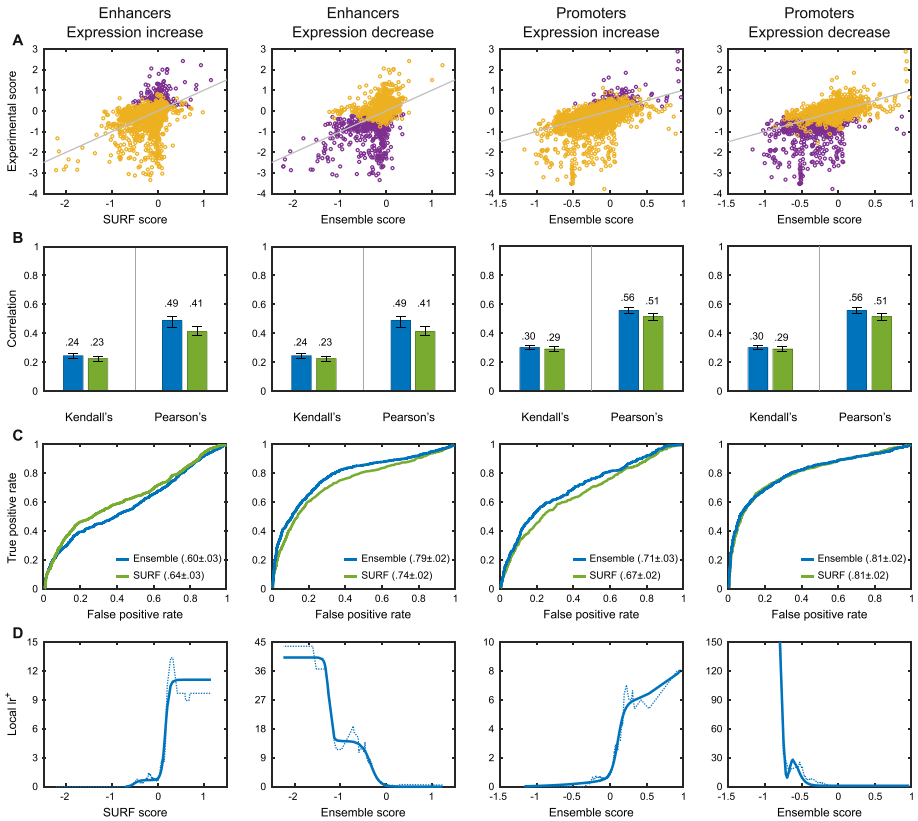


Fig. 5 Performance on the regulation saturation expression challenge. The two left columns show performance in predicting increased (left) and decreased (right) expression in a set of enhancers (purple points represent variants that significantly change expression). The right pair of columns show equivalent results for promoters. The scatter plots (**A**) show strong performance in identifying decreases in expression (purple points), but weaker results for expression increases. Performance on promoters is stronger than on enhancers. Overlap of changed and non-changed experimental expression points suggests that experimental uncertainty reduces the apparent performance of the computational methods. Panel **B** shows correlation coefficients for selected methods. Panel **C** shows ROC curves for predicting under and overexpression. Panel **D** shows local lr^+, where the solid lines are smoothed fits to the data

not those causing the effect. Identifying the functional variants is not straightforward, and current regulatory prediction methods can provide hypotheses as to possible effects on expression or splicing.

CAGI participants identified diagnostic variants that were not found by clinical laboratories

A major goal of CAGI is to test the performance of computational methods under as close to clinical conditions as possible. In the area of rare disease diagnosis, four challenges have addressed this by requiring participants to identify diagnostic variants in sets of clinical data. The Johns Hopkins and intellectual disability (ID) challenges employed diagnostic panels, covering a limited set of candidate genes in particular disease areas. As compared with genome-wide data, diagnostic panels inherently restrict the search to only variants belonging to a known set of relevant genes.

For a number of genetically undiagnosed cases in the Johns Hopkins panel, CAGI participants found high-confidence deleterious variants in genes associated with a different disease from that reported, suggesting physicians may have misdiagnosed the symptoms

[61]. However, because of the clinical operating procedures of the diagnostic laboratory, it has not been possible to further investigate these cases. In the ID panel, some plausible calls were made on novel variants that had not been reported to the patient partly because the majority of standard computational methods returned assignments of "benign" [62].

The two SickKids challenges (SickKids4 and SickKids5) were based on whole-genome sequence data for children with rare diseases from the Hospital for Sick Children in Toronto. These are all cases that were undiagnosed by the state-of-the-art SickKids pipeline [26], and so were particularly challenging compared with those normally encountered in the clinic. In the SickKids4 challenge, variants proposed by challenge participants were deemed diagnostic by the referring physicians for two of the cases in part due to matching detailed phenotypes. This was the first instance of the CAGI community directly contributing in the clinic. In SickKids5, two of the highest confidence nominated diagnostic variants provided correct genome-patient matches. While not meeting ACMG/AMP criteria for pathogenicity [32], these were considered interesting candidates for further investigation, again potentially resolving previously intractable cases.

These clinical challenges required participants to develop full analysis pipelines, including quality assessment for variant calls, proper inclusion of known pathogenetic variants from databases such as HGMD [53] and ClinVar [9], and an evaluation scheme for weighing the evidence. The SickKids challenges also required compilation of a set of candidate genes. Varying success in addressing these factors will have influenced the results, so it is not possible to effectively compare the core computational methods. Overall, current approaches have limitations in this setting––they tend to ignore or fail to reliably evaluate synonymous and noncoding variants; if the relevant gene is not known its variants will usually not be examined; and data for epigenetic causes are not available. Nevertheless, the CAGI results for these challenges again make it clear that current state-of-the-art computational approaches can make valuable contributions in real clinical settings.

Complex trait interpretation is often complicated by confounders in the data

Many common human diseases, such as Alzheimer's disease, asthma, and type II diabetes, are complex traits and as with monogenic disorders, genetic information should in principle be useful for both diagnosis and prognosis. Individual response to drugs (pharmacogenomics) also often has a complex trait component. Complex traits have relatively small contributions from each of many variants, collectively affecting a broad range of molecular mechanisms, including gene expression, splicing, and multiple aspects of protein function. Environmental factors also play a substantial role so that phenotype prediction based on genetic information alone has inherently limited accuracy. Also, most CAGI complex trait challenges have been based on exome data, whereas at many GWAS risk loci lie outside coding regions [63]. To some extent, the status of relevant common variants not present in the exome data can be imputed on the basis of linkage disequilibrium, but this places an unclear limit on achievable accuracy. Limited or no availability of training data also restricted method performance and phenotypes tend to be less precise than for other types of disease. Altogether, these factors make this a difficult CAGI area. Nevertheless, these challenges have been informative and have drawn new investigators into

the rapidly developing area of Polygenic Risk Score (PRS) estimation [64]. One challenge, CAGI4 Crohn's, has yielded apparently robust conclusions on the performance of methods in this area.

Crohn's disease (CAGI4)

Participants were provided with exome data for 111 individuals and asked to identify the 64 who had been diagnosed with Crohn's disease. A variety of computational approaches were used, including clustering by genotypes, analysis of variants in pathways related to the disease, and evaluation of SNPs in known disease-associated loci. The highest-scoring method (AUC 0.72; Fig. 6A) used the latter approach together with conventional machine learning, and trained on data from an earlier GWAS [65]. Fig. 6B shows case and control score distributions for that method. A perfect method would have no distribution overlap. These results are far from that, but there is clear signal at the extremes, and as Fig. 6C shows, that translates into a positive likelihood ratio with an approximately 20-fold range (0.3 to 6), only a little lower than that obtained for the biochemical effect and clinical missense

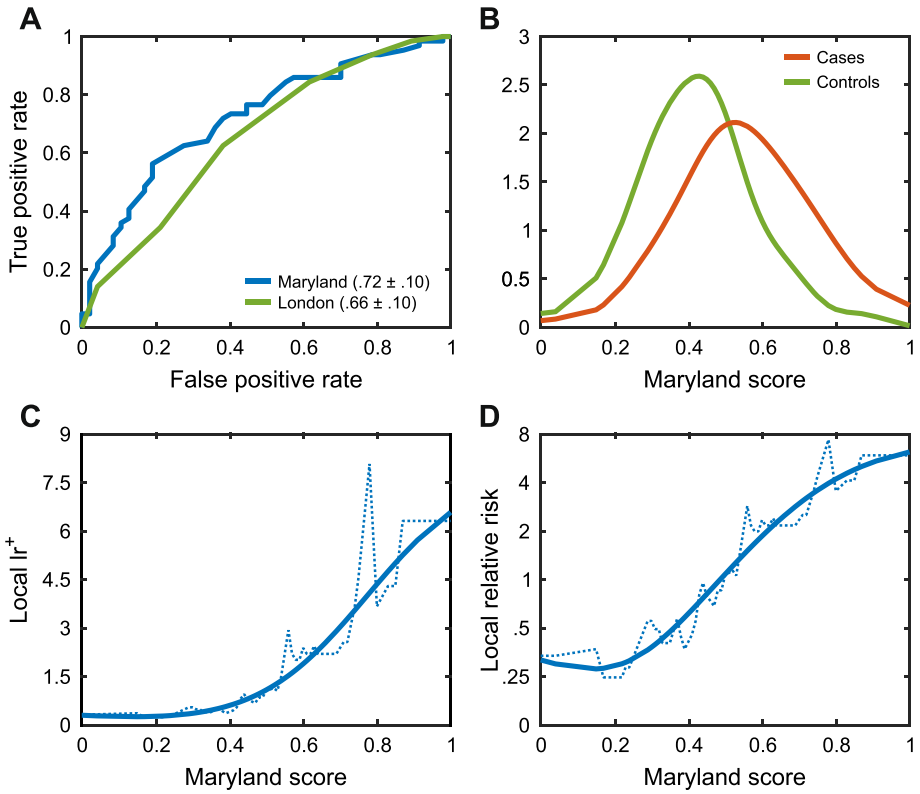


Fig. 6 Identifying which of a set of individuals are most at risk for Crohn's disease, given exome data. Examples were selected on the basis of ranking by ROC AUC. **A** ROC curves for two selected methods. Statistically significant but relatively low ROC areas are obtained. **B** Distributions of disease prediction scores for individuals with the disease (red) and without (green) for the method with the highest AUC (kernel density representation of the data). **C** Local positive likelihood ratio (lr^+) as a function of prediction score for the method with the highest AUC. **D** Relative risk of disease (log_2 scale), compared to that in the general population as a function of prediction score. Individuals with the lowest risk scores have approximately 1/3 the average population risk, while those with the highest scores have risk exceeding fourfold the average, a 12-fold total range. Depending on the disease, identifying individuals with higher than threefold the average risk may be sufficient for clinical action

challenges. With a prior probability of disease of 1.3% [66], relative risk (see "Methods") also has a range of about 20-fold (Fig. 6D), with the highest-risk individuals having sixfold higher risk compared to that estimated for the population average. For some complex trait diseases, for example coronary heart disease [67], this is discriminatory enough to support clinical action, and for many diseases would provide a valuable additional factor to more standard risk measures such as age and sex. Newer PRS methods, which aim to incorporate many weak contributions from SNPs, were not evaluated in this CAGI challenge.

Other complex trait CAGI challenges (Additional file 1) revealed batch effects (CAGI2 Crohn's and bipolar disorder) and population structure effects (CAGI3 Crohn's) in the data, leaked clinical data (Warfarin and VET challenges), or discrepancies between self-reported traits and those predicted from genetic data (PGP challenges), thereby complicating assessment. Performance in matching genomes and disease phenotypes, an additional component in some challenges, was poor. Additional complex trait challenge results are provided in Additional file 7: Table S7.

The CAGI Ethics Forum has guided responsible data governance

Data used in CAGI challenges are diverse in terms of sensitivity (e.g., with respect to participant reidentification risk, potential for stigmatization, potential impact of pre-publication data disclosure), collected under a broad variety of participant consent understandings and protection frameworks, and analyzed by predictors with varying degrees of familiarity with local and international biomedical regulations. This heterogeneity calls for a nuanced approach to data access and the tailored vetting of CAGI experiments. The CAGI Ethics Forum was launched in 2015 to proactively address these concerns. Incorporating input from bioethicists, researchers, clinicians, and patient advocates, it has developed policies for responsible data governance (e.g., assisting in revision of the general CAGI data use agreement, to safeguard human data and also protect all CAGI participants, including data providers for unpublished data), cautioned against overinterpretation of findings (e.g., highlighting the contribution of social and environmental risk factors to disease, and the potential negative consequences, such as stigma, of associating particular disease variants with a specific population), and provided input on a variety of guidelines and procedures, including CAGI's participant vetting process (e.g., how to identify a bona fide researcher) and a system of tiered access conditions for datasets, depending on their sensitivity. Future directions include investigating the scalability of current user validation and data access models, exploring implications for family members of unexpected challenge findings, discussing policies to ensure proper credit attribution for constituent primary methods used by metapredictors, and identifying additional means of ensuring accountability options with respect to responsible data sharing.

Discussion

Over CAGI's first decade, five rounds of CAGI challenges have provided a picture of the current state-of-the-art in interpreting the impact of genetic variants on a range of phenotypes and provided a basis for the development of improved methods as well as for more calibrated use in clinical settings.

A key finding is that for most missense challenges it is possible to relate phenotype values to a pathogenicity threshold, and so deduce potential performance in a clinical

setting, particularly for rare Mendelian diseases. The results suggest that such computational methods are generally more reliable than recognized in the current clinical guidelines [32]. Several challenges have directly assessed the usefulness of variant impact prediction under clinical conditions, highlighting the fact that successful application in the clinic requires integration of the computational methods into a comprehensive pipeline.

Computational challenges and methods for identifying the effects of splicing and regulatory variants have been less well represented and issues with data availability have limited insights. Nevertheless, the results suggest potential for providing evidence for informing pathogenicity and offering mechanistic insights.

CAGI relies on the same factors as other critical assessment community experiments: a willingness of the relevant research groups to participate, clearly defined metrics for success, the availability of large enough and accurate enough sets of experimental data to provide a gold standard, and independent objective assessment of the predictions. Participation in CAGI has been strong in most areas and a vibrant and interactive community has developed. New researchers have been attracted to the field and new collaborations have resulted in the development of creative algorithms with broad applicability [68–70].

The biggest obstacle to clear assessment has been and continues to be data diversity and quality, a key difference between CAGI and related community endeavors. Other initiatives, such as CASP [71], deal primarily with one type of data (protein structure) and the data are usually of high quality and directly relevant to the goals of the computational methods. By contrast, CAGI deals with many different settings, including studies of biochemical effects with a broad range of phenotypes, the pathogenicity of variants both germline and somatic, clinical phenotypes, and statistical relationships. Also, while genome variant calling is reliable, it does have limits [72]. For example, in the SickKids challenges, some variants suggested as diagnostic by CAGI participants had been found to be incorrect calls and so eliminated in the clinical pipeline, using sequence validation data CAGI participants did not have access to. Adapting available data to form suitable challenges is difficult and compromises are sometimes needed to devise a challenge where the results can be objectively assessed. For example, in one of the SickKids and in the Johns Hopkins clinical challenges, assessment hinged on requiring participants to match genomes to phenotypes. But that makes it much harder to identify diagnostic variants than in the real-life situation. Conversely, challenge providers have sometimes benefited from the detailed scrutiny of their data by CAGI participants prior to its publication. In some cases, interpretation of clinical challenge results is also complicated by there being no conclusive diagnoses. Numerical assessments can also have limitations, as clinicians may often use other considerations while evaluating patients. CAGI participants have similarly sometimes used unanticipated information to improve performance on challenges. For example, in a PGP challenge requiring matching of full genome sequences to extensive phenotype profiles, a participant made use of information in the PGP project blog [73]. Though these sometimes subvert the intended challenge, in some ways this reflects what happens in a clinical setting—all relevant information is used, however obscure.

Experimental biotechnology platforms have become widely available [74] and the genomic data collection in the clinic has greatly increased. While the next generation of computational tools should benefit from these developments, they also pose new challenges. Deeper characterization of experimental approaches is needed to address data uncertainty and biases. Potential circularity between computation-assisted variant annotation and method assessment also needs to be considered. CAGI has committed to employing a diversity of evaluation metrics rather than any single primary performance indicator, and new assessment metrics may need to be introduced [75–77]. Future rounds of CAGI will address these issues by using assessment methods that mitigate or eliminate problems with data, by developing and promoting practices and standards for application of methods, by working with experimental groups to provide sufficiently large and high-quality datasets, and by effectively following up on disagreements at the intersection of high-throughput functional experiments, genetic association studies, and outputs of computational prediction methods. CAGI also plans to increase evaluation of the combinatorial effect of variants, either in single-molecule biochemical assays or in clinical applications with whole-genome studies for both rare and complex phenotypes.

Conclusions

Results from the first decade of CAGI have highlighted current abilities and limitations of computational methods in genome interpretation and indicate future research directions. The current performance levels for missense variation in Mendelian disorders, combined with rapidly accumulating data and a recent breakthrough in protein structure prediction [71], suggest that upcoming methods should consistently achieve Strong and potentially Very Strong clinical evidence levels. Progress is also expected in method performance for other types of genome variation and complex disorders, assisted by improvements in experimental and statistical methodologies, as well as new clinical standards [78]. CAGI's independent assessment rigorously ascertains the performance characteristics of computational methods for variant interpretation; this assessment approach therefore offers a model framework for evaluating clinical validity of diagnostics and screening. Genomic science has been tremendously advanced by policies ensuring rapid release of data, and to help promote the development and assessment of analytical methods these must be crafted to support portions of the datasets being incorporated into evaluations like CAGI. These developments will enable computational approaches to further narrow the gap between basic and clinical research, advancing our understanding over the entire breadth of genome variation.

Methods

We describe different evaluation scenarios considered in the Critical Assessment of Genome Interpretation (CAGI), motivate the selection of performance measures, and discuss ways to interpret the results.

Terminology and notation

Let $(x, y) \in \mathcal{X} \times \mathcal{Y}$ be a realization of an input–output random vector (X, Y). The input space $\mathcal{X}$ may describe variants, gene panels, exomes, or genomes in different CAGI scenarios. Similarly, the output space $\mathcal{Y}$ can describe discrete or continuous targets; e.g., it can

be a binary set {pathogenic, benign} when the task is to predict variant pathogenicity, or a continuous set $[0, \infty)$ representing a percent of enzymatic activity of the wildtype protein (a mutated molecule can have increased activity compared to the wildtype) or a cell growth rate relative to that with the wildtype gene.

Let $s : \mathcal{X} \to \mathbb{R}$ be a predictor that assigns a real-valued score to each input and $f : \mathcal{X} \to \mathcal{Y}$ be a predictor that maps inputs into elements of the desired output space; i.e., f is real-valued when predicting a continuous output and discrete when predicting a discrete output. When predicting binary outcomes (e.g., $\mathcal{Y} = \{0, 1\}$), $s(x)$ is often a soft prediction score between 0 and 1, whereas $f(x)$ can be obtained by discretizing $s(x)$ based on some decision threshold τ. Scores $s(x)$ can also be discretized based on a set of thresholds $\{\tau_j\}_{j=1}^m$, as discussed later. In a binary classification scenario, $f(x) = 1$ (pathogenic prediction) when $s(x) \geq \tau$ and $f(x) = 0$ (benign prediction) otherwise. We shall sometimes denote a discretized binary model as $f_\tau(x)$ to emphasize that the predictor was obtained by thresholding a soft scoring function $s(x)$ at τ. The target variable Y can similarly be obtained by discretizing the continuous space $\mathcal{Y}$ using a set of thresholds $\{\tau'_k\}$, that are different from $\{\tau_j\}$ used for the scoring function. In one such case, discretizing the continuous space $\mathcal{Y}$ of functional impact of an amino acid variant into {damaging, nondamaging} transforms a regression problem into classification, which may provide additional insights during assessment. With a minor abuse of notation, we will refer to both continuous and the derived discrete space as $\mathcal{Y}$. The exact nature of the target variable Y and the output space will be clear from the context.

Finally, let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ be a test set containing n input–output pairs on which the predictors are evaluated. Ideally, this data set is representative of the underlying data distribution and non-overlapping with the training data for each evaluated predictor. Similarly, we assume the quality of the measurement of the ground truth values $\{y_i\}_{i=1}^n$ is high enough to ensure reliable evaluation. While we took multiple steps to ensure reliable experiments and blind assessments, it is difficult to guarantee complete enforcement of either of these criteria. For example, an in vitro assay may be an imperfect model of an in vivo impact or there might be uncertainty in collecting experimental read-outs. Additionally, the notion of a representative test set may be ambiguous and cause difficulties when evaluating a model that was developed with application objectives different from those used to assess its performance in CAGI.

Evaluation for continuous targets

Evaluating the prediction of continuous outputs is performed using three primary measures (R^2, Pearson's correlation coefficient, and Kendall's tau) and two secondary measures (root mean square error and Spearman's correlation coefficient). R^2 is defined as the difference between the variance of the experimental values and the mean-squared error of the predictor, normalized by the variance of the experimental values. It is also referred to as the fraction of variance of the target that is explained by the model. $R^2 \in (-\infty, 1]$ is estimated on the test set $\mathcal{D}$ as

$$R^2 = 1 - \frac{\sum_{i=1}^n \big(f(x_i) - y_i\big)^2}{\sum_{i=1}^n \big(y_i - \bar{y}\big)^2}, \tag{1}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ is the mean of the target values in $\mathcal{D}$ (observe that each value y_i may itself be an average over technical or biological replicates, if available in the experimental

data). The R^2 values above 0 indicate that the predictor is better than the trivial predictor—one that always outputs the mean of the target variable—and values close to 1 are desired. The values below 0 correspond to models with inferior performance to the trivial predictor. Maximizing the R^2 metric requires calibration of output scores; that is, a high correlation between predictions and target values as well as the proper scaling of the prediction outputs. For example, a predictor outputting a linear transformation of the target such as $f(X) = 10 \cdot Y$ or a monotonic nonlinear transformation of the target such as $f(X) = \log Y$ may have a high correlation, but a low R^2. R^2, therefore, can be seen as the strictest metric used in CAGI. However, this metric can adversely impact methods outputting discretized prediction values. Such methods are preferred by some tool developers as they simplify interpretation by clinicians, experimental scientists, or other users.

In some cases, it may be useful to also report the root mean square error (RMSE), estimated here as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - y_i\right)^2}. \tag{2}$$

RMSE can offer a useful interpretation of the performance and is provided as a secondary measure in CAGI evaluations.

The correlation coefficient between the prediction $f(X)$ and target Y is defined as a normalized covariance between the prediction output and the target variable. Pearson's correlation coefficient $-1 \leq r \leq 1$ is estimated on $\mathcal{D}$ as

$$r = \frac{\sum_{i=1}^{n}(f_i - \bar{f})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}\left(f_i - \bar{f}\right)^2 \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}, \tag{3}$$

where $f_i = f(x_i)$ and $\bar{f} = \frac{1}{n}\sum_{i=1}^{n} f_i$ is the mean of the predictions. Pearson's correlation coefficient does not depend on the scale of the prediction, but it is affected by the extent of a linear relationship between predictions and the target. That is, a predictor outputting a linear transformation of the target such as $f(X) = 10 \cdot Y$ will have a perfect correlation. However, a monotonic nonlinear transformation of the target such as $f(X) = \log Y$ may have a relatively low r. Although not our main metric, we also explored Spearman's rank correlation as a secondary metric. Spearman's correlation is defined as Pearson's correlation on the rankings.

We also computed Kendall's tau, which is the probability of a concordant pair of prediction-target points linearly scaled to the $[-1, 1]$ interval instead of $[0, 1]$. Assuming that all prediction and target values are distinct, a pair of points $\left(f(x_i), y_i\right)$ and $(f(x_j), y_j)$ is concordant if either ($f(x_i) > f(x_j)$ and $y_i > y_j$) or ($f(x_i) < f(x_j)$ and $y_i < y_j$). Otherwise, a pair of points is discordant. Kendall's tau was estimated on $\mathcal{D}$ as

$$\tau = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\text{sign}(f(x_i) - f(x_j)) \cdot \text{sign}(y_i - y_j). \tag{4}$$

It ranges between -1 and 1, with 1, indicating that all pairs are concordant, 0 indicating half of the concordant pairs (e.g., a random ordering) and -1 indicating that all pairs are

discordant. A predictor outputting a linear transformation of the target $f(X) = 10 \cdot Y$ and a monotonic nonlinear transformation of the target $f(X) = \log Y$ will both have a perfect tau of 1. Compared to Pearson's correlation, Kendall's tau can be seen as less sensitive to the scale but more sensitive to the ordering of predictions. Equation 4 is defined under the assumption that both the predictions and the outputs are unique. However, this assumption is not satisfied by all biological datasets and predictors. To address this issue, we use Kendall's tau-b, a widely accepted correction for ties,

$$\tau_b = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\text{sign}(f(x_i) - f(x_j)) \cdot \text{sign}(y_i - y_j)}{\sqrt{\left(\beta(n) - \sum_{i=1}^{T}\beta(u_i)\right)\left(\beta(n) - \sum_{i=1}^{S}\beta(v_i)\right)}}, \tag{5}$$

where $\beta(n) = n(n-1)/2$, u_i (v_i) is the size of the i th group of ties in the predictions (outputs) and T (S) is the number of such groups in the predictions (outputs) [79].

Evaluation for binary targets

Evaluating binary outputs is performed using standard protocols in binary classification [80]. We compute the Receiver Operating Characteristic (ROC) curve, which is a 2D plot of the true positive rate $\gamma = P(f_\tau(X) = 1|Y = 1)$ as a function of the false positive rate $\eta = P(f_\tau(X) = 1|Y = 0)$, where τ is varied over the entire range of prediction scores. The area under the ROC curve can be mathematically expressed as $\text{AUC} = \int_0^1 \gamma d\eta$ and is the probability that a randomly selected positive example x_+ will be assigned a higher score than a randomly selected negative example x_- by the model [81]. That is, assuming no ties in prediction scores $\text{AUC} = P(s(X_+) > s(X_-))$. In the presence of ties, AUC is given by $P(s(X_+) > s(X_-)) + \frac{1}{2}P(s(X_+) = s(X_-))$ [82]. The AUC is estimated on the test set $\mathcal{D}$ using the standard numerical computation that allows for ties [83]. Although AUC does not serve as a metric that directly translates into clinical decisions, it is useful in that it shows the degree of separation of the examples from the two groups of data points (positive vs. negative). Another useful property of the AUC is its insensitivity to class imbalance.

Though AUC is a useful measure for capturing the overall performance of a classifier's score function, it has limitations when applied to a decision-making setting such as the one encountered in the clinic. Typically, clinically relevant score thresholds that determine the variants satisfying Supporting, Moderate or Strong evidence [32] lie in a region of low false positive rate (FPR). A measure well-suited to capture clinical significance of a predictor ought to be sensitive to the variations in the classifier's performance in the low FPR region (when predicting pathogenicity). However, the contribution of the low FPR region to AUC is relatively small. This is because it not only represents a small fraction of the entire curve, but also because the TPR values in that region are relatively small. Thus, AUC is not sensitive enough to the variation in a predictor's performance in the low FPR region. To mitigate this problem, we also provide area under the ROC curve truncated to the [0, 0.2] FPR interval. What constitutes low FPR is not well defined; however, it appears that the [0, 0.2] FPR interval combined with the [0, 1] TPR interval is a reasonable choice in CAGI applications; see Figs. 2 and 3. We normalize the truncated AUC to span the entire [0, 1] range by dividing the observed value by 0.2, the maximum possible area below the ROC truncated at FPR = 0.2.

CAGI evaluation of binary classifiers also involves calculation of the Matthews correlation coefficient [84]. The Matthews correlation coefficient (MCC) was computed as Pearson's correlation coefficient between binary predictions and target values on the test set $\mathcal{D}$. Efficient MCC estimation was carried out from the confusion matrix [84].

Evaluation for clinical significance

Current guidelines from the American College for Medical Genetics and Genomics (ACMG) and Association for Molecular Pathology (AMP) established a qualitative framework for combining evidence in support of or against variant pathogenicity for clinical use [32]. These guidelines point to five different levels of pathogenicity and (effectively) nine distinct types of evidence in support of or against variant pathogenicity. The five pathogenicity levels involve classifications into pathogenic, likely pathogenic, variant of uncertain significance (VUS), likely benign, and benign variants, whereas the nine levels of evidential support are grouped into Very Strong, Strong, Moderate, and Supporting for either pathogenicity or benignity, as well as indeterminate evidence that supports neither pathogenicity nor benignity.

Richards et al. [32] have manually categorized different types of evidence and also listed twenty rules for combining evidence for a variant to be classified into one of the five pathogenicity-benignity groups. For example, variants that accumulate one Very Strong and one Strong line of evidence of pathogenicity lead to the classification of the variant as pathogenic; variants that accumulate one Strong and two Supporting lines of evidence lead to the classification of the variant as likely pathogenic, etc. [32] The guidelines allow for the use of computational evidence such that a computational prediction of pathogenicity can be considered as the weakest (Supporting) line of evidence. Thus, combined with other evidence, these methods can presently contribute to a pathogenicity assertion for a variant, but in a restricted and arbitrary way [50]. Since Supporting evidence is qualitatively the smallest unit of contributory evidence in the ACMG/AMP guidelines, we refer to any computational model that reaches the prediction quality equivalent of Supporting evidence and higher as a model that provides contributory evidence in the clinic.

Numerically, a variant that is classified as pathogenic should have at least a 99% probability of being pathogenic given all available evidence, whereas a variant that is likely pathogenic should have at least a 90% probability of being pathogenic given the evidence [32, 52]. Variants that cross the 90% probability threshold for pathogenicity are considered clinically actionable [32]. Analogously, variants with sufficient support for benignity will typically be ruled out from being diagnostic in a clinical laboratory. Note that, though the guidelines provide a probabilistic interpretation of the pathogenicity assertions, they do not provide any general quantitative interpretation of the evidence. Consequently, any framework designed to express the evidence levels quantitatively, must tie such quantitative evidential support to the pathogenicity probabilities, mediated by the ACMG/AMP rules for combining evidence.

The possibility of incorporating computational methods into clinical decision making in a properly calibrated manner presents interesting opportunities and unique challenges. In particular, since the evidence levels are only described qualitatively, it is not obvious how to determine what values of a predictor's output score qualify as providing a given level of evidence. Thus, to apply a computational line of evidence in

the clinic in a principled manner, and consistent with the guidelines, there is a need for a framework that assigns a quantitative interpretation to each evidence level.

Tavtigian et al. [52] proposed such a framework to provide numerical support for each type of evidential strength for its use in ACMG/AMP guidelines for or against variant pathogenicity. This approach is based on the relationship between prior and posterior odds of pathogenicity as well as on independence of all lines of evidential support for a given variant. We briefly review this approach.

Let E be a random variable indicating evidence that can be used in support of or against variant pathogenicity. The positive likelihood ratio (LR^+) given concrete evidence e is defined as

$$LR^+ = \frac{\text{posterior odds of pathogenicity}}{\text{prior odds of pathogenicity}} \tag{6}$$

or equivalently

$$LR^+(e) = \frac{P(Y=1|E=e)}{1-P(Y=1|E=e)} \cdot \frac{1-P(Y=1)}{P(Y=1)}, \tag{7}$$

where the first term on the right corresponds to the posterior odds of pathogenicity given the evidence and the second term on the right corresponds to the reciprocal of the prior odds of pathogenicity. The prior odds of pathogenicity depend solely on the class prior $P(Y=1)$; that is, the fraction of pathogenic variants in the selected reference set. The expression for LR^+ also allows for an easy interpretation as the increase in odds of pathogenicity given evidence e compared to the situation when no evidence whatsoever is available. The likelihood ratio of 2, for example, states that a variant with evidence e is expected to have twice as large odds of being pathogenic than a variant picked uniformly at random from a reference set. As CAGI only considers computational evidence, we will later replace the posterior probability $P(Y=1|E=e)$ by $P\big(Y=1|f(X)=1\big)$ for discretized predictors or by $P(Y=1|s(X)=s)$ for the predictors that output a soft numerical score s. The probability $P\big(Y=1|f(X)=1\big)$ is the positive predictive value (or precision) of a binary classifier, whereas the probability $P(Y=1|s(X)=s)$ can be seen as the local positive predictive value, defined here in a manner analogous to the local false discovery rate [85].

It can be shown [86] that the positive likelihood ratio can also be stated as

$$LR^+(e) = \frac{P(E=e|Y=1)}{P(E=e|Y=0)} \tag{8}$$

thus clarifying that LR^+ can be seen as the ratio of the true positive rate and false positive rate when $P(E=e|Y=1)$ is replaced by $P\big(f(X)=1|Y=1\big)$ and $P(E=e|Y=0)$ by $P\big(f(X)=1|Y=0\big)$.

Tavtigian et al. [52] give an expression relating the posterior $P(Y=1|E=e)$ to LR^+ and the prior $P(Y=1)$ as

$$P(Y=1|E=e) = \frac{LR^+(e)P(Y=1)}{\big(LR^+(e)-1\big)P(Y=1)+1} \tag{9}$$

which itself is obtained from Eq. 7. They also present a framework that allows for assigning probabilistic interpretations to different types of evidential strength (Supporting, Moderate, Strong, and Very Strong) and combining them in a manner consistent with the rules listed in Richards et al. [32] and the probabilistic interpretation of likely pathogenic and pathogenic classes. Their formulation is given in terms of the positive likelihood ratio LR^+ in an exponential form. We restate this model using a notion of the total (or combined) positive likelihood ratio $\mathrm{LR}^+_{\mathrm{T}}$, based on all available evidence, E_{T}, of a variant that is expressed as a product of LR^+ factors from different strengths of evidence as

$$\mathrm{LR}^+_{\mathrm{T}} = \mathrm{c}^{\frac{n_{\mathrm{su}}}{8}+\frac{n_{\mathrm{mo}}}{4}+\frac{n_{\mathrm{st}}}{2}+\frac{n_{\mathrm{vs}}}{1}}, \tag{10}$$

where n_{su}, n_{mo}, n_{st}, and n_{vs} are the counts of Supporting (su), Moderate (mo), Strong (st), and Very Strong (vs) lines of evidence present in E_{T}, and c is the LR^+ value assigned to a single line of Very Strong evidence. It is easy to show that $\sqrt[8]{c}$, $\sqrt[4]{c}$, and $\sqrt[2]{c}$ correspond to the LR^+ for a single line of Supporting, Moderate, and Strong line of evidence, respectively. In other words, the model from Eq. 10 enforces that if a Very Strong piece of evidence increases $\mathrm{LR}^+_{\mathrm{T}}$ by a multiplicative factor of c, then a Supporting, Moderate, or a Strong piece of evidence increases $\mathrm{LR}^+_{\mathrm{T}}$ by a factor of $\sqrt[8]{c}$, $\sqrt[4]{c}$, and $\sqrt[2]{c}$, respectively. For a reasonable consistency with Richards et al. [32] this model also explicitly encodes that one line of Very Strong evidence is equal to the two lines of Strong evidence, four lines of Moderate evidence, and eight lines of Supporting evidence.

The appropriate value of c, however, depends on the class prior. It is the smallest number for which the $\mathrm{LR}^+_{\mathrm{T}}$ values computed for the qualitative criteria from the likely pathogenic class in Richards et al. [32] reach $P(Y = 1|E_{\mathrm{T}} = e)$ values of at least 0.9 and, similarly, for those in the pathogenic class, reach a $P(Y = 1|E_{\mathrm{T}} = e)$ value of at least 0.99. The dependence on the class prior is due to the conversion between $\mathrm{LR}^+_{\mathrm{T}}$ and $P(Y = 1|E_{\mathrm{T}} = e)$ governed by Eq. 9. If the class prior is small, a larger value of $\mathrm{LR}^+_{\mathrm{T}}$ will be required to achieve the same posterior level, thereby requiring a larger value of c (Additional file 1: Figure S14).

Tavtigian et al. [52] also proposed that two rules from Richards et al. [32] be revised; that is, one of the rules was proposed to be "demoted" from pathogenic to likely pathogenic, whereas another rule was proposed to be "promoted" from the likely pathogenic to pathogenic. For a class prior of 0.1 that was selected based on the experience from the clinic, the value $c = 350$ was found to be suitable. This, in turn, suggests that the Supporting, Moderate, and Strong lines of evidence should require the likelihood ratio values of $\sqrt[8]{c} = 2.08$, $\sqrt[4]{c} = 4.32$, and $\sqrt[2]{c} = 18.7$, respectively. However, note again that for different priors, these values will be different; see next section and Additional file 1: Figure S14. Moreover, while the level of posterior for the combined evidence (Eq. 13) is required to be at least 0.9 to satisfy the likely pathogenic rule and 0.99 for pathogenic, this does not mean that the posterior level for a single line of evidence is the same for all values of c. This is a consequence of the fact that the framework provides intuitive interpretation only at the level of the combined posterior.

When drawing evidence from a pathogenicity predictor, it is necessary to further clarify what evidence is in the first place. At least two options are available: (i) the evidence is the score $s(x)$; that is, a raw prediction of pathogenicity, or (ii) the evidence

is a discretized prediction $f_\tau(x)$, obtained by thresholding $s(x)$. These approaches, referred to here as local and global, respectively, lead to different interpretations because all evaluation metrics hold only on average, either over all variants with a score $s(x)$ or all variants satisfying $f_\tau(x) = 1$; i.e., having a score above τ. When both $s(x)$ and $f_\tau(x)$ are available, this leads to difficulties in interpreting the results of the global approach because all scores $s(x)$ that map into $f_\tau(x) = 1$ will be treated identically. Unfortunately, this implies that scores s greater than but still close to τ most likely do not meet the levels of evidential strength for the interval. At the same time, scores close to the high end of the range almost certainly make the levels of evidential strength above the designated level. This means that a clinician seeing a variant with score slightly above τ would have to interpret this prediction as contributory to pathogenicity, yet this interpretation would almost certainly be incorrect. Based on the recommendations from the ClinGen's Sequence Variant Interpretation group, [50] we focus on the local view as well as local performance criteria to define levels of evidential strength and assess whether methods achieve these levels. In the end, however, we also provide global estimates to understand the performance of each tool more comprehensively.

We define the local positive likelihood ratio as

$$\mathrm{lr}^+(s) = \frac{p(s|Y=1)}{p(s|Y=0)}, \tag{11}$$

where $p\big(s|Y=y\big)$, for $y \in \{0,1\}$ are class-conditional densities.

We obtain an estimate of the local positive likelihood ratio $\widehat{\mathrm{lr}}^+$ from the test data as described in the section titled "Computing clinically relevant measures." Now, the threshold to determine the variants with Supporting level of evidence is given as the minimum score above which all variants achieve local positive likelihood ratio value greater than or equal to $\sqrt[8]{c}$; i.e.,

$$\tau_{\mathrm{su}} = \min\left\{\tau : \forall s \geq \tau, \widehat{\mathrm{lr}}^+(s) \geq \sqrt[8]{c}\right\} \tag{12}$$

though we note that Pejaver et al. [50] incorporated an additional factor based on the confidence interval for $\widehat{\mathrm{lr}}^+(s)$ to result in more stringent recommendations for score thresholding. Similarly, the thresholds for variants with Moderate, Strong, and Very Strong evidence are given by $\tau_{\mathrm{mo}} = \min\left\{\tau : \forall s \geq \tau, \widehat{\mathrm{lr}}^+(s) \geq \sqrt[4]{c}\right\}$, $\tau_{\mathrm{st}} = \min\left\{\tau : \forall s \geq \tau, \widehat{\mathrm{lr}}^+(s) \geq \sqrt[2]{c}\right\}$ and $\tau_{\mathrm{vs}} = \min\left\{\tau : \forall \mathrm{s} \geq \tau, \widehat{\mathrm{lr}}^+(s) \geq \mathrm{c}\right\}$.

Once the threshold set $\{\tau_{\mathrm{su}}, \tau_{\mathrm{mo}}, \tau_{\mathrm{st}}, \tau_{\mathrm{vs}}\}$ is determined, we can compute either the global LR^+ (e.g., $s \geq \tau_{\mathrm{su}}$) or the LR^+ corresponding to an interval of scores (e.g., $\tau_{\mathrm{su}} \leq s < \tau_{\mathrm{mo}}$) by computing the true positive rate and false positive rate for a given set of scores. A global positive predictive value can be similarly estimated once the class prior is known.

In all CAGI evaluations, a predictor is considered to provide contributory evidence in a clinical setting if it reaches any one of the evidence levels according to the ACMG/AMP guidelines, and according to the model by Tavtigian et al. [52] and recommendations by Pejaver et al. [50]. Among predictors that reach the desired levels of evidential support, the ones that reach higher levels are generally considered favorably. However,

we have not considered any criterion to rank the predictors that reach the same levels of evidential support.

Selection of class priors for variant pathogenicity

Different clinical scenarios require the use of different class priors of variant pathogenicity. We generally distinguish between two clinical situations.

In the first setting, a clinician is presented with a proband with specific phenotypic expression and the objective is to find variants responsible for the clinical phenotype. In certain monogenic disorders with Mendelian inheritance patterns, the fraction of rare variants found to be pathogenic can be as high as 25%, as in the case of the NAGLU challenge. Similarly, Tavtigian et al. [52] report an experience-based prior of 10% based on their work with BRCA genes, which we adopted in this work.

The second setting reflects situations such as screening for potential secondary variants. Here we have used an estimate by Pejaver et al. [40] that up to 1.5% of missense variants in an apparently healthy individual could be disease-causing.

Overall, prior probability of pathogenicity was set to 1 and 10% to demonstrate the distinction in the level of evidential support necessary. These resulted in $c = 8511$ and $c = 351$, respectively (note that $c = 351$ was selected instead of $c = 350$ to avoid rounding errors in finding a c that best models ACMG/AMP rules). In each functional missense challenge, the level of prior probability observed for each gene based on experimental data was further considered. For large class priors such as 50% or above, the Tavtigian et al. [52] framework holds only when an additional rule from Richards et al. [32] is removed; that is, we ignored that two Supporting lines of evidence for benignity assert a likely benign variant.

Performance measures for clinical application

Diagnostic odds ratio

The diagnostic odds ratio (DOR) is commonly used in biomedical sciences to measure the increase in odds of pathogenicity in the presence of evidence e compared to the odds of pathogenicity in the absence of e; [86] that is,

$$\mathrm{DOR}(e) = \frac{P(Y=1|E=e)}{1-P(Y=1|E=e)} \cdot \frac{1-P(Y=1|E\neq e)}{P(Y=1|E\neq e)}. \tag{13}$$

The difference between Eq. 7 and Eq. 13 is that the prior odds, those governed by the prior $P(Y=1)$ and used in Eq. 7, are replaced by the odds governed by the probability $P(Y=1|E\neq e)$; that is, odds of pathogenicity when the evidence e was not the one that was observed. The quantity $P(Y=0|E\neq e) = 1-P(Y=1|E\neq e)$ is referred to as the negative predictive value when the observed evidence is $f(X)=0$. $\mathrm{DOR}\in[0,\infty)$ can also be expressed as

$$\mathrm{DOR}(e) = \frac{\mathrm{LR}^+(e)}{\mathrm{LR}^-(e)}, \tag{14}$$

where $\mathrm{LR}^+(e)$ is defined in Eq. 8 and

$$\mathrm{LR}^-(e) = \frac{P(E\neq e|Y=1)}{P(E\neq e|Y=0)}. \tag{15}$$

In contrast to typical studies of variant risk assessment [87] and polygenic risk scores [64], DOR was calculated without adjustments for usual confounders such as race and ethnicity that are generally not available in CAGI challenges and, technically, produce conditional odds ratios [86]. However, the DOR values estimated in our experiments have an identical interpretation as the results of logistic regression run with a single independent variable (co-variate) at a time. Glas et al. [86] give a broader coverage of diagnostic odds ratios that further connect some of the quantities discussed here (e.g., AUC vs. DOR).

We only consider DOR with the computational evidence of the "global" type; that is, when $s(x) \geq \tau$. Consequently, DOR at τ can be expressed as

$$\mathrm{DOR}(\tau) = \frac{\mathrm{LR}^{+}(\tau)}{\mathrm{LR}^{-}(\tau)} = \frac{P(s(X) \geq \tau | Y = 1)}{P(s(X) \geq \tau | Y = 0)} \frac{P(s(X) < \tau | Y = 0)}{P(s(X) < \tau | Y = 1)}. \tag{16}$$

Unlike positive likelihood ratio, DOR does not have a "local" version. This is because one cannot define a local negative likelihood ratio.

Percent of variants predicted as pathogenic

In addition to finding whether a method reaches Supporting, Moderate, or Strong levels of evidence, it is important to also quantify the proportion of variants in the reference set for which a given evidence level is reached. To this end, for a given score threshold τ, we define the percent of variants in the reference set that the method assigns a score as high as or higher than τ, and refer to it as "probability of pathogenic (positive) predictions," or PPP. Mathematically, it can be expressed as the following probability

$$\mathrm{PPP}(\tau) = P(s(X) \geq \tau). \tag{17}$$

The probability (equivalently, percent) of variants reaching a given level of evidence can now be quantified as $\mathrm{PPP}(\tau)$, where τ is the score threshold at which a variant is declared to meet the desired evidential support.

Posterior probability of pathogenicity

Given a method, the posterior probability of pathogenicity or the absolute risk for a variant is defined as the probability that the variant is pathogenic based on the score it is assigned by the method. It is expressed as

$$\rho(s) = P(Y = 1 | s(X) = s). \tag{18}$$

We also refer to this quantity as a local positive predictive value or local precision.

Relative risk

Given a method, the relative risk (RR) of pathogenicity of a variant is defined as the posterior probability of pathogenicity (based on the score assigned by the method) relative to the prior probability of pathogenicity. It is expressed as the following ratio

$$\mathrm{RR}(s) = \frac{P(Y=1|s(X)=s)}{P(Y=1)}. \tag{19}$$

The prior probability of pathogenicity can also be interpreted as the average of the posterior probability over all variants in the reference set; that is

$$\begin{aligned}\mathbb{E}[P(Y=1|s(x)=s)] &= \int_{\mathcal{X}} P(Y=1|s(x)=s)p(x)dx \\ &= \int_{\mathbb{R}} P(Y=1|s)p(s)ds \\ &= \int_{\mathbb{R}} p(s|Y=1)P(Y=1)ds \\ &= P(Y=1),\end{aligned} \tag{20}$$

where the last step follows since $p(s|Y=1)$ is a density function and its integral over $\mathbb{R}$ is 1. Observe that our definition of relative risk is an extension of the "global" version used in clinical applications where the denominator would be $P(Y=1|s(X)\neq s)$, which effectively equals $P(Y=1)$ for all predictors outputting continuous scores.

Computing clinically relevant measures

We show here how the measures for evaluation of binary targets and clinically relevant measures are computed from the test data $\mathcal{D}$. It is necessary to be cautious when making decisions on a reference (target) population based on the measures computed on the test set $\mathcal{D}$. Some of the measures computed on $\mathcal{D}$ accurately represent the corresponding values on the target population. However, other measures are biased because the test data set for many challenges is not representative of the target population. In particular, the proportions of positives (e.g., pathogenic variants) in the test set $\alpha_{\mathcal{D}} = P_{\mathcal{D}}(Y=1)$ may be vastly different from that in the target population $\alpha = P(Y=1)$. Consequently, the class-prior dependent measures, when estimated directly from the test set, are incorrectly calibrated to the test set class priors.

Fortunately, the class-prior dependent measures can be corrected using an estimate of the target population's class priors if known or if estimated using a principled approach [88, 89]. The correction is derived under the assumption that the reference population and the test set are distributionally identical, except for the differences in class priors. To elaborate, the target distribution of inputs $p(x)$ can be expressed in terms of the class-conditional distributions, $p\big(x|Y=y\big)$ for $y \in \{0,1\}$, and the class priors as follows

$$p(x) = \alpha \cdot p(x|Y=1) + (1-\alpha)\cdot p(x|Y=0). \tag{21}$$

We assume that the test set distribution of inputs might have different class priors, but the same class-conditional distributions as the target population. Precisely,

$$\begin{aligned}p_{\mathcal{D}}(x) &= \alpha_{\mathcal{D}} p_{\mathcal{D}}(x|Y=1) + (1-\alpha_{\mathcal{D}})p_{\mathcal{D}}(x|Y=0) \\ &= \alpha_{\mathcal{D}} p(x|Y=1) + (1-\alpha_{\mathcal{D}})p(x|Y=0).\end{aligned} \tag{22}$$

It is easy to see that any of the clinical and non-clinical measures that only depend on the class-conditional distributions, but not class priors, when computed on the

test set is an unbiased estimate of the measure on the target population. However, if a measure also depends on the class priors, it needs to be corrected to reflect the reference population's class prior. All the class-prior independent measures used in this paper can be expressed in terms of class-conditional derived quantities such as the true positive rate (TPR), the false positive rate (FPR), and the local positive likelihood ratio $\mathrm{lr}^+(s)$. The class-prior dependent measures additionally have the class-prior in their expressions.

TPR, FPR, and $lr^+(s)$

Formally, TPR is defined as the proportion of positive inputs that are correctly predicted to be positive. Mathematically,

$$\mathrm{TPR}(\tau) = P(s(x) \geq \tau | Y = 1), \tag{23}$$

where $s(x)$ is a continuous score function of a classifier and τ is a threshold such that an input scoring above τ is predicted to be positive. Similarly, FPR is defined as the proportion of negative inputs that are incorrectly predicted to be positive. Mathematically,

$$\mathrm{FPR}(\tau) = P(s(x) \geq \tau | Y = 0). \tag{24}$$

TPR and FPR can be computed from the test data as the proportion of positive and negative test inputs scoring $\geq \tau$, respectively. That is,

$$\begin{aligned} \widehat{\mathrm{TPR}}(\tau) &= \frac{\sum_{x \in \mathcal{D}_+} I[s(x) \geq \tau]}{|\mathcal{D}_+|} \\ \widehat{\mathrm{FPR}}(\tau) &= \frac{\sum_{x \in \mathcal{D}_-} I[s(x) \geq \tau]}{|\mathcal{D}_-|}, \end{aligned} \tag{25}$$

where $\mathcal{D}_+$ and $\mathcal{D}_-$ are the subsets of points in the test set $\mathcal{D}$ labeled as positive and negative, respectively.

Some of the clinically relevant measures used in our study are "local" in nature in the sense that they are derived from a local neighborhood around a score value instead of the entire range of scores above (or below) the threshold. Such measures can be expressed in terms of the local positive likelihood ratio $\mathrm{lr}^+(s)$. To compute $\mathrm{lr}^+(s)$, we exploit its relationship to the posterior probability at score s; that is,

$$\begin{aligned} P(Y = 1 | s(X) = s) &= \frac{p(s(X) = s | Y = 1) P(Y = 1)}{p(s(X) = s)} \\ &= \frac{p(s(X) = s | Y = 1) P(Y = 1)}{p(s(X) = s | Y = 1) P(Y = 1) + p(s(X) = s | Y = 0) P(Y = 0)} \\ &= \frac{\mathrm{lr}^+(s) P(Y = 1)}{\mathrm{lr}^+(s) P(Y = 1) + P(Y = 0)} \\ &= \frac{\mathrm{lr}^+(s) P(Y = 1)}{\big(\mathrm{lr}^+(s) - 1\big) P(Y = 1) + 1}. \end{aligned} \tag{26}$$

Similarly, the test data posterior probability can be expressed as

$$P_{\mathcal{D}}(Y=1|s(X)=s)=\frac{\mathrm{lr}^{+}(s)P_{\mathcal{D}}(Y=1)}{\big(\mathrm{lr}^{+}(s)-1\big)P_{\mathcal{D}}(Y=1)+1}. \tag{27}$$

Note that since $\mathrm{lr}^{+}(s)$ only depends on the class-conditional distribution, it does not change when defined on the target population. Unlike the target population's posterior, the test data posterior can be estimated from the test data as described below. Once the test posterior is estimated, the equation above can be inverted to estimate lr^{+} as

$$\widehat{\mathrm{lr}}^{+}(s)=\frac{\widehat{P}_{\mathcal{D}}(Y=1|s(X)=s)}{1-\widehat{P}_{\mathcal{D}}(Y=1|s(X)=s)}\cdot\frac{1-P_{\mathcal{D}}(Y=1)}{P_{\mathcal{D}}(Y=1)}, \tag{28}$$

where the $\widehat{P}_{\mathcal{D}}(Y=1|s(X)=s)$ is an estimate of the test data posterior and $P_{\mathcal{D}}(Y=1)$ is the proportion of positives in the test data, which may differ from the true prior for a randomly picked variant in the gene of interest or another reference sample. Note that though the formula above expresses $\widehat{\mathrm{lr}}^{+}(s)$ in terms of the prior odds, suggesting a dependence on the class prior, $\widehat{\mathrm{lr}}^{+}(s)$ is class-prior independent, as discussed earlier. In theory, $P_{\mathcal{D}}(Y=1|s(X)=s)$ is the proportion of pathogenic variants among all variants in $\mathcal{D}$ having a score s. Therefore, estimating the local posterior efficiently would require observing the same score many times in the set of variants with known labels. This is unlikely since we only have scores for a finite set of variants and thus the posterior cannot be estimated without making further assumptions. However, assuming that the posterior is a smooth function of the score—similar scores correspond to similar local posterior values—we estimate the posterior as the proportion of pathogenic variants in a small window around the score; that is, $[s-\epsilon, s+\epsilon]$, where ϵ was selected to be 5% of the range of the predictor's outputs, with the range considered to be an interval between the 5th and 95th percentile of predicted values on the dataset, selected as such to minimize the influence of outliers. In addition, for stable estimates, we required that at least 10% of the variants, up to a maximum of 50 variants, from the data set are within a window; therefore, the final window size was dependent on score s and data set $\mathcal{D}$.

Measures that do not require correction

Among the measures considered in this paper, TPR, FPR, ROC curve, AUC, LR^{+}, LR^{-}, DOR, and lr^{+} do not require correction. Class-prior independence of TPR, FPR, and lr^{+} is obvious from their definitions as discussed earlier. ROC curve is obtained by plotting TPR against FPR and consequently, it is also class-prior independent. By extension AUC, being the area under the ROC curve, is also class-prior independent. The global positive likelihood ratio LR^{+}, formulated with the evidence of the type $s(x)\geq\tau$, is given by $\mathrm{TPR}(\tau)/\mathrm{FPR}(\tau)$. Similarly, the global LR^{-} is given by $(1-\mathrm{TPR}(\tau))/(1-\mathrm{FPR}(\tau))$. Since DOR is the ratio of LR^{+} and LR^{-}, it is by extension class-prior independent.

Measures that require correction

Among the measures considered in this paper, probability of pathogenic predictions (PPP), positive predictive value (PPV), posterior probability (ρ), and relative risk (RR), being class-prior dependent, require corrections to be properly applied to the target population. To show that the measures are indeed class-prior dependent, we re-formulate them by separating the class-prior from the class-conditional dependent terms.

$$\begin{aligned}\mathrm{PPP}(\tau) &= P(s(X) \geq \tau) \\ &= P(s(X) \geq \tau | Y = 1)P(Y = 1) + P(s(X) \geq \tau | Y = 0)P(Y = 0) \\ &= \alpha \mathrm{TPR}(\tau) + (1 - \alpha)\mathrm{FPR}(\tau)\end{aligned} \tag{29}$$

$$\begin{aligned}\mathrm{PPV}(\tau) &= P(Y = 1 | S(X) \geq \tau) \\ &= \frac{P(s(X) \geq \tau | Y = 1)P(Y = 1)}{P(s(X) \geq \tau)} \\ &= \frac{\alpha \mathrm{TPR}(\tau)}{\alpha \mathrm{TPR}(\tau) + (1 - \alpha)\mathrm{FPR}(\tau)}\end{aligned} \tag{30}$$

$$\begin{aligned}\rho(s) &= P(Y = 1 | s(X) = s) \\ &= \frac{\alpha \mathrm{lr}^+(s)}{\alpha\left(\mathrm{lr}^+(s) - 1\right) + 1},\end{aligned} \tag{31}$$

where the derivation is the same as that for Eq. 26.

$$\begin{aligned}\mathrm{RR}(s) &= \frac{P(Y = 1 | s(X) = s)}{\alpha} \\ &= \frac{\mathrm{lr}^+(s)}{\alpha\left(\mathrm{lr}^+(s) - 1\right) + 1}.\end{aligned} \tag{32}$$

We use the expressions above to correctly calculate class-prior dependent metrics on the target domain by first computing the class-conditional dependent terms (TPR, FPR, or lr^+) using the test data $\mathcal{D}$ and then using an estimate of the class prior of the target distribution in the corresponding expression.

Statistical significance and confidence interval estimation

All p-values and confidence intervals in CAGI evaluations were estimated using bootstrapping with 1000 iterations [90].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-03113-6.

Additional file 1. Description of the analysis framework, implementation details, analyzed and non-analyzed CAGI challenges and the corresponding data. It also contains all supplementary figures and a description of the supplementary tables [92–166].

Additional file 2: Table S2. Analysis of all biochemical effect challenges.

Additional file 3: Table S3. Meta-analysis of all biochemical effect challenges.

Additional file 4: Table S4. Analysis of the Annotate All Missense challenge.

Additional file 5: Table S5. Analysis of cancer challenges.

Additional file 6: Table S6. Analysis of splicing and transcription challenges.

Additional file 7: Table S7. Analysis of the complex trait challenge: Crohn's disease (CAGI4).

Additional file 8. Review history.

Acknowledgements

The authors are grateful for the contributions of Talal Amin, Patricia Babbitt, Eran Bachar, Stefania Boni, Kirstine Calloe, Ombretta Carlet, Shann-Ching Chen, Chien-Yuan Chen, Jun Cheng, Luigi Chiricosta, Alex Colavin, Qian Cong, Emma D'Andrea, Carla Davis, Xin Feng, Carlo Ferrari, Yao Fu, Alessandra Gasparini, David Goldgar, Solomon Grant, Steve Grossman, Todd Holyoak, Xiaolin Li, Quewang Liu, Beth Martin, Zev Medoff, Nasim Monfared, Susanna Negrin, Michael Parsons,

Nathan Pearson, Alexandra Piryatinska, Catherine Plotts, Jennifer Poitras, Clive Pulinger, Francesco Reggiani, Melvin M. Scheinman, George Shackelford, Vasily Sitnik, Fiorenza Soli, Qingling Tang, Nancy Mutsaers Thomsen, Jing Wang, Chenling Xiong, Lijing Xu, Shuhan Yang, Lijun Zhan, and Huiying Zhao.

Authors' Contributions

Analysis and writing

Shantanu Jain[1,+], Constantina Bakolitsa[2,+], Steven E. Brenner[2,$,*], Predrag Radivojac[1,3,*], John Moult[4,*]
[+]Contributed equally.
[$]This author was unable to fully contribute to the paper writing due to an injury and its sequelae.
*****Corresponding authors (brenner@berkeley.edu; predrag@northeastern.edu; jmoult@umd.edu).

CAGI organizers

John Moult[4], Susanna Repo[2], Roger A. Hoskins[2], Gaia Andreoletti[2], Daniel Barsky[2], Steven E. Brenner.[2]

Informatics infrastructure and support

Ajithavalli Chellapan[5], Hoyin Chu[1,6–7], Navya Dabbiru[5], Naveen K. Kollipara[5], Melissa Ly[2], Andrew J. Neumann[2], Lipika R. Pal[4], Eric Odell[2], Gaurav Pandey[2,8], Robin C. Peters-Petrulewicz[2], Rajgopal Srinivasan[5], Stephen F. Yee[2], Sri Jyothsna Yeleswarapu[5], Maya Zuhl[4,9]

Predictors

Ogun Adebali[10–11], Ayoti Patra[12–13], Michael A. Beer[12], Raghavendra Hosur[14–15], Jian Peng[14], Brady M. Bernard[16–17], Michael Berry[10], Shengcheng Dong[18], Alan P. Boyle[18], Aashish Adhikari[2,19], Jingqi Chen[2,20], Zhiqiang Hu[2], Robert Wang[2,21], Yaqiong Wang[2,20], Maximilian Miller[22], Yanran Wang[22–23], Yana Bromberg[22], Paola Turina[24], Emidio Capriotti[24], James J. Han[25], Kivilcim Ozturk[25], Hannah Carter[25], Giulia Babbi[24], Samuele Bovo[24], Pietro Di Lena[24], Pier Luigi Martelli[24], Castrense Savojardo[24], Rita Casadio[24], Melissa S. Cline[26], Greet De Baets[27], Sandra Bonache[28–29], Orland Díez[28,30], Sara Gutiérrez-Enríquez[28], Alejandro Fernández[28,30], Gemma Montalban[28,31], Lars Ootes[28], Selen Özkan[28], Natàlia Padilla[28], Casandra Riera[28], Xavier De la Cruz[28], Mark Diekhans[26], Peter J. Huwe[32–33], Qiong Wei[32,34], Qifang Xu[32], Roland L. Dunbrack[32], Valer Gotea[35], Laura Elnitski[35], Gennady Margolin[35], Piero Fariselli[36–37], Ivan V. Kulakovskiy[38–39], Vsevolod J. Makeev[38], Dmitry D. Penzar[38,40], Ilya E. Vorontsov[38–39], Alexander V. Favorov[12,38], Julia R. Forman[41–42], Marcia Hasenahuer[43–44], Maria S. Fornasari[43], Gustavo Parisi[43], Ziga Avsec[45], Muhammed H. Çelik[45–46], Thi Yen Duong Nguyen[45], Julien Gagneur[45], Fang-Yuan Shi[47], Matthew D. Edwards[14,48], Yuchun Guo[14,49], Kevin Tian[14,50], Haoyang Zeng[14,51], David K. Gifford[14], Jonathan Göke[52], Jan Zaucha[53–54], Julian Gough[55], Graham R.S. Ritchie[44,56], Adam Frankish[44,57], Jonathan M. Mudge[44,57], Jennifer Harrow[57–58], Erin L. Young[59], Yao Yu[60], Chad D. Huff[60], Katsuhiko Murakami[61–62], Yoko Nagai[61,63], Tadashi Imanishi[61,64], Christopher J. Mungall[65], Julius O.B. Jacobsen[57,66], Dongsup Kim[67], Chan-Seok Jeong[67–68], David T. Jones[69], Mulin Jun Li[70–71], Violeta Beleva Guthrie[12,54], Rohit Bhattacharya[12,72], Yun-Ching Chen[12,73], Christopher Douville[12], Jean Fan[12], Dewey Kim[7,12], David Masica[12], Noushin Niknafs[12], Sohini Sengupta[12,74], Collin Tokheim[6,12,75], Tychele N. Turner[12,76], Hui Ting Grace Yeo[12,52], Rachel Karchin[12], Sunyoung Shin[77], Rene Welch[78], Sunduz Keles[78], Yue Li[14,79], Manolis Kellis[7,14], Carles Corbi-Verge[80–81], Alexey V. Strokach[80], Philip M. Kim[80], Teri E. Klein[50], Rahul Mohan[82–83], Nicholas A. Sinnott-Armstrong[50], Michael Wainberg[82,84], Anshul Kundaje[50], Nina Gonzaludo[85–86], Angel C. Y. Mak[85,87], Aparna Chhibber[88–89], Hugo Y.K. Lam[88,90], Dvir Dahary[91], Simon Fishilevich[92], Doron Lancet[92], Insuk Lee[93], Benjamin Bachman[94], Panagiotis Katsonis[94], Rhonald C. Lua[94], Stephen J. Wilson[94–95], Olivier Lichtarge[94], Rajendra R. Bhat[96], Laksshman Sundaram[50], Vivek Viswanath[96], Riccardo Bellazzi[97], Giovanna Nicora[97–98], Ettore Rizzo[98], Ivan Limongelli[98], Aziz M. Mezlini[80], Ray Chang[99], Serra Kim[99], Carmen Lai[99], Robert O'Connor[99–100], Scott Topper[99], Jeroen van den Akker[99], Alicia Y. Zhou[99], Anjali D. Zimmer[99], Gilad Mishne[99], Timothy R. Bergquist[101–102], Marcus R. Breese[85,103], Rafael F. Guerrero[3,104], Yuxiang Jiang[3], Nikki Kiga[101], Biao Li[103,105], Matthew Mort[103,106], Kymberleigh A. Pagel[3], Vikas Pejaver[8,101], Moses H. Stamboulian[3], Janita Thusberg[103], Sean D. Mooney[101], Predrag Radivojac[1,3], Nuttinee Teerakulkittipong[4,107], Chen Cao[4,108], Kunal Kundu[4,109], Yizhou Yin[4], Chen-Hsin Yu[4], Maya Zuhl[4,9], Lipika R. Pal[4], John Moult[4], Michael Kleyman[4,110], Chiao-Feng Lin[21,111], Mary Stackpole[4,112], Stephen M. Mount[4], Gökcen Eraslan [7,113], Nikola S. Mueller[113], Tatsuhiko Naito[114], Aliz R. Rao[31], Johnathan R. Azaria[115–116], Aharon Brodie[115], Yanay Ofran[115], Aditi Garg[117], Debnath Pal[117], Alex Hawkins-Hooker[41,69], Henry Kenlay[41,118], John Reid[41,119], Eliseos J. Mucaki[120], Peter K. Rogan[120], Jana M. Schwarz[121], David B. Searls[122], Gyu Rie Lee[101,123], Chaok Seok[123], Andreas Krämer[124], Sohela Shah[124–125], ChengLai V. Huang[2,126], Jack F. Kirsch[2,†], Maxim Shatsky[65], Yue Cao[127], Haoran Chen[127–128], Mostafa Karimi[126–127], Oluwaseyi Moronfoye[127], Yuanfei Sun[127], Yang Shen[127], Ron Shigeta[129–130], Colby T. Ford[131], Conor Nodzak[131], Aneeta Uppal[131–132], Xinghua Shi[131,133], Thomas Joseph[5], Sujatha Kotte[5], Sadhna Rana[5], Aditya Rao[5], V. G. Saipradeep[5], Naveen Sivadasan[5], Uma Sunderam[5], Rajgopal Srinivasan[5], Mario Stanke[134], Andrew Su[135], Ivan Adzhubey[136–137], Daniel M. Jordan[8,138], Shamil Sunyaev[138], Frederic Rousseau[27], Joost Schymkowitz[27], Joost Van Durme[27], Sean V. Tavtigian[59], Marco Carraro[36], Manuel Giollo[36,126], Silvio C. E. Tosatto[36], Orit Adato[115], Liran Carmel[139], Noa E. Cohen[139–140], Tzila Fenesh[115], Tamar Holtzer[115], Tamar Juven-Gershon[115], Ron Unger[115], Abhishek Niroula[141], Ayodeji Olatubosun[142,†], Jouni Väliaho[142], Yang Yang[143], Mauno Vihinen[141–142], Mary E. Wahl[34,138], Billy Chang[144], Ka Chun Chong[144], Inchi Hu[145–146], Rui Sun[144,147], William Ka Kei Wu[144], Xiaoxuan Xia[144], Benny C. Zee[144], Maggie H. Wang[144], Meng Wang[47], Chunlei Wu[135], Yutong Lu[147], Ken Chen[147], Yuedong Yang[148–150], Christopher M. Yates[151–152], Anat Kreimer[2,22], Zhongxia Yan[2,14], Nir Yosef[2], Huying Zhao[150], Zhipeng Wei[153], Zhaomin Yao[154], Fengfeng Zhou[153], Lukas Folkman[149,155], Yaoqi Zhou.[149,156]

Challenge data providers

Roxana Daneshjou[50], Russ B. Altman[50], Fumitaka Inoue[85,157], Nadav Ahituv[85], Adam P. Arkin[2], Federica Lovisa[36,158], Paolo Bonvini[36,158], Sarah Bowdin[159], Stefano Gianni[160], Elide Mantuano[161], Velia Minicozzi[162], Leonore Novak[160], Alessandra Pasquo[163], Annalisa Pastore[164], Maria Petrosino[165–166], Rita Puglisi[42], Angelo Toto[160], Liana Veneziano[161], Roberta Chiaraluce[165], Mad P. Ball[138,167], Jason R. Bobe[138,168], George M. Church[138], Valerio Consalvi[160], Matthew Mort[103,106], David N. Cooper[106], Bethany A. Buckley[125], Molly B. Sheridan[169], Garry R. Cutting[169], Maria Chiara Scaini[170], Kamil J. Cygan[109,171],

Alger M. Fredericks[171], David T. Glidden[171], Christopher Neil[171–172], Christy L. Rhine[171–172], William G. Fairbrother[171], Aileen Y. Alontaga[173], Aron W. Fenton[173], Kenneth A. Matreyek[101,174], Lea M. Starita[101], Douglas M. Fowler[101], Britt-Sabina Löscher[175], Andre Franke[176], Scott I. Adamson[177], Brenton R. Graveley[177], Joe W. Gray[178], Mary J. Malloy[85], John P. Kane[85], Maria Kousi[179], Nicholas Katsanis[180–181], Max Schubach[121], Martin Kircher[121], Nina Gonzaludo[85–86], Angel C.Y. Mak[85,87], Paul L. F. Tang[85,182], Pui-Yan Kwok[85], Richard H. Lathrop[46,†], Wyatt T. Clark[183], Guoying K. Yu[183–184], Jonathan H. LeBowitz[183], Francesco Benedicenti[185], Elisa Bettella[36], Stefania Bigoni[186], Federica Cesca[36], Isabella Mammi[187], Cristina Marino-Buslje[188], Donatella Milani[189], Angela Peron[190–192], Roberta Polli[36], Stefano Sartori[36,158], Franco Stanzial[185], Irene Toldo[36], Licia Turolla[193], Maria C. Aspromonte[36], Mariagrazia Bellini[36], Emanuela Leonardi[36], Xiaoming Liu[194–195], Christian Marshall[80], W. Richard McCombie[196], Lisa Elefanti[170], Chiara Menin[170], M. Stephen Meyn[78,197], Alessandra Murgia[36], Kari C.Y. Nadeau[50], Lipika R. Pal[4], John Moult[4], Susan L. Neuhausen[198], Robert L. Nussbaum[125], Mehdi Pirooznia[73,199], James B. Potash[169], Dago F. Dimster-Denk[2], Jasper D. Rine[2], Jeremy R. Sanford[26], Michael Snyder[50], Sean V. Tavtigian[59], Atina G. Cote[80,200], Song Sun[80,200], Marta W. Verby[80,200], Jochen Weile[80,200], Frederick P. Roth[80,200], Ryan Tewhey[201], Pardis C. Sabeti[7], Joan Campagna[85], Marwan M. Refaat[85,202], Julianne Wojciak[85], Soren Grubb[203], Nicole Schmitt[203], Jay Shendure[101], Amanda B. Spurdle[204], Dimitri J. Stavropoulos[80], Nephi A. Walton[205–206], Peter P. Zandi[169], Elad Ziv.[85]

Ethics forum

Wylie Burke[101], Flavia Chen[85,138], Lawrence R. Carr[122], Selena Martinez[122], Jodi Paik[122], Julie Harris-Wai[85], Mark Yarborough[207], Stephanie M. Fullerton[208], Barbara A. Koenig.[85]

Assessors

Roxana Daneshjou[50], Gregory McInnes[50,209], Russ B. Altman[50], Dustin Shigaki[12], Michael A. Beer[12], Aashish Adhikari[2,19], John-Marc Chandonia[2,65], Mabel Furutsuki[2], Zhiqiang Hu[2], Laura Kasak[2,210], Changhua Yu[2,211], Vikas Pejaver[8,101], Yana Bromberg[22], Castrense Savojardo[24], Rui Chen[50,88], Melissa S. Cline[26], Qifang Xu[32], Roland L. Dunbrack[32], Gaurav Pandey[2,8], Iddo Friedberg[212], Gad A. Getz[7,137,213], Qian Cong[214], Lisa N. Kinch[214], Jing Zhang[214], Nick V. Grishin[214], Alin Voskanian[215], Maricel G. Kann[215], Wyatt T. Clark[183], Elizabeth Tran[82], Nilah M. Ioannidis[2], Maria C. Aspromonte[36], Mariagrazia Bellini[36], Emanuela Leonardi[36], Jesse M. Hunter[78,216], Rupa Udani[78,217], M. Stephen Meyn[78,197], Binghuang Cai[101], Sean D. Mooney[101], Alexander A. Morgan[50,218], Lipika R. Pal[4], John Moult[4], Stephen M. Mount[4], Alessandra Murgia[36], Robert L. Nussbaum[125], Jeremy R. Sanford[26], Artem Sokolov[26,137], Joshua M. Stuart[26], Shamil Sunyaev[138], Sean V. Tavtigian[59], Marco Carraro[36], Manuel Giollo[36,126], Giovanni Minervini[36], Alexander M. Monzon[36], Silvio C. E. Tosatto[36], Anat Kreimer[2,22], Nir Yosef.[2]

Advisory board and scientific council

Russ B. Altman[50], Serafim Batzoglou[19,219], Yana Bromberg[22], Atul J. Butte[85], George M. Church[138], Garry R. Cutting[169], Laura Elnitski[35], Marc S. Greenblatt[220], Reece K. Hart[221], Ryan Hernandez[79,85], Tim J. P. Hubbard[42], Scott Kahn[19,222], Rachel Karchin[12], Anne O'Donnell-Luria[7], M. Stephen Meyn[78,197], Sean D. Mooney[101], Alexander A. Morgan[50,218], Pauline C. Ng[52], Robert L. Nussbaum[125], John Shon[19,223], Michael Snyder[50], Shamil Sunyaev[138], Sean V. Tavtigian[59], Scott Topper[99], Joris Veltman[224], Justin M. Zook.[225]

CAGI chairs

John Moult[4], Steven E. Brenner.[2]

Affiliations

1. Northeastern University, Boston, Massachusetts, USA; 2. University of California, Berkeley, California, USA; 3. Indiana University, Bloomington, Indiana, USA; 4. University of Maryland, College Park, Maryland, USA; 5. Tata Consultancy Services, Hyderabad, India; 6. Dana-Farber Cancer Institute, Boston, Massachusetts, USA; 7. Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 8. Icahn School of Medicine at Mount Sinai, New York City, New York, USA; 9. ICF International, Cambridge, Massachusetts, USA; 10. University of Tennessee, Knoxville, Tennessee, USA; 11. Sabanci University, Tuzla, Turkey; 12. Johns Hopkins University, Baltimore, Maryland, USA; 13. Intel Corporation, Santa Clara, California, USA; 14. Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; 15. Encoded Genomics, South San Francisco, California, USA; 16. Institute for Systems Biology, Seattle, Washington, USA; 17. Earle A. Chiles Research Institute, Providence Health & Services, Portland, Oregon, USA; 18. University of Michigan, Ann Arbor, Michigan, USA; 19. Illumina, San Diego, California, USA; 20. Fudan University, Shanghai, China; 21. University of Pennsylvania, Philadelphia, Pennsylvania, USA; 22. Rutgers University, New Brunswick, New Jersey, USA; 23. Genentech, South San Francisco, California, USA; 24. University of Bologna, Bologna, Italy; 25. University of California at San Diego, San Diego, California, USA; 26. University of California at Santa Cruz, Santa Cruz, California, USA; 27. Katholieke Universiteit Leuven, Leuven, Belgium; 28. Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain; 29. Germans Trias i Pujol Hospital, Badalona, Spain; 30. University Hospital of Vall d'Hebron, Barcelona, Spain; 31. Université Laval, Québec, Québec, Canada; 32. Fox Chase Cancer Center, Philadelphia, Pennsylvania, USA; 33. Belmont University, Nashville, Tennessee, USA; 34. Microsoft, Redmond, Washington, USA; 35. National Human Genome Research Institute, Bethesda, Maryland, USA; 36. University of Padova, Padova, Italy; 37. University of Turin, Turin, Italy; 38. Vavilov Institute of General Genetics, Moscow, Russia; 39. Institute of Protein Research, Pushchino, Russia; 40. Lomonosov Moscow State University, Moscow, Russia; 41. University of Cambridge, Cambridge, UK; 42. King's College London, London, UK; 43. Universidad Nacional de Quilmes, Bernal, Argentina; 44. European Molecular Biology Laboratory—European Bioinformatics Institute, Hinxton, UK; 45. Technical University of Munich, Munich, Germany; 46. University of California at Irvine, Irvine, California, USA; 47. Peking University, Beijing, China; 48. Verily Life Sciences, South San Francisco, California, USA; 49. CAMP4 Therapeutics, Cambridge, Massachusetts, USA; 50. Stanford University, Stanford, California, USA; 51. Insitro, South San Francisco, California, USA; 52. Genome Institute of Singapore, Singapore; 53. University of Bristol, Bristol, UK; 54. AstraZeneca, Cambridge, UK; 55. MRC Laboratory of Molecular Biology, Cambridge, UK; 56. Amazon Web Services, Seattle, Washington, USA; 57. Wellcome Sanger Institute, Hinxton, UK. 58. ELIXIR, Hinxton, UK; 59. University of Utah, Salt Lake City, Utah, USA; 60. University of Texas MD Anderson Cancer Center, Houston, Texas, USA; 61. National Institute of Advanced Industrial Science and Technology, Tokyo, Japan; 62. Fujitsu Ltd., Tokyo, Japan; 63. Varinos Inc., Tokyo, Japan; 64. Tokai University

School of Medicine, Tokyo, Japan; 65. Lawrence Berkeley National Laboratory, Berkeley, California, USA; 66. Queen Mary University, London, UK; 67. Korea Advanced Institute of Science and Technology, Daejeon, South Korea; 68. Korea Institute of Science and Technology Information, Daejeon, South Korea; 69. University College London, London, UK; 70. University of Hong Kong, Hong Kong; 71. Tianjin Medical University, Tianjin, China; 72. Williams College, Williamstown, Massachusetts, USA; 73. Johnson & Johnson, New Brunswick, New Jersey, USA; 74. PierianDx, Creve Coeur, Missouri, USA; 75. Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA; 76. Washington University School of Medicine, St. Louis, Missouri, USA; 77. Pohang University of Science and Technology, Pohang, South Korea; 78. University of Wisconsin-Madison, Madison, Wisconsin, USA; 79. McGill University, Montreal, Canada; 80. University of Toronto, Toronto, Ontario, Canada; 81. Cyclica Inc., Toronto, Ontario, Canada; 82. Stanford University School of Medicine, Stanford, California, USA; 83. Dashworks, Boston, Massachusetts, USA; 84. Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health, Toronto, Ontario, Canada; 85. University of California at San Francisco, San Francisco, California, USA; 86. Pacific Biosciences, Menlo Park, California, USA; 87. CytomX Therapeutics, South San Francisco, California, USA; 88. Roche, Basel, Switzerland; 89. Bristol Myers Squibb, Redwood City, California, USA; 90. HypaHub, Sunnyvale, California, USA; 91. LifeMap Sciences Inc., Alameda, California, USA; 92. Weizmann Institute, Rehovot, Israel; 93. Yonsei University, Seoul, South Korea; 94. Baylor College of Medicine, Houston, Texas, USA; 95. Calm, San Francisco, California, USA; 96. University of Florida, Gainesville, Florida, USA; 97. University of Pavia, Pavia, Italy; 98. enGenome, Pavia, Italy; 99. Color Genomics, Burlingame, California, USA; 100. Syapse, San Francisco, California, USA; 101. University of Washington, Seattle, Washington, USA; 102. Sage Bionetworks, Seattle, Washington, USA; 103. Buck Institute for Research on Aging, Novato, California, USA; 104. North Carolina State University, Raleigh, North Carolina, USA; 105. Gilead, Foster City, California, USA; 106. Institute of Medical Genetics, Cardiff University, Cardiff, UK; 107. Burapha University, Chonburi, Thailand; 108. Google LLC, Mountain View, California, USA; 109. Regeneron, Tarrytown, New York, USA; 110. Moderna, Cambridge, Massachusetts, USA; 111. DNANexus, Mountain View, California, USA; 112. EarlyDiagnostics, Los Angeles, California, USA; 113. Helmholtz Zentrum Muenchen, Neuherberg, Germany; 114. University of Tokyo, Tokyo, Japan; 115. Bar-Ilan University, Ramat-Gan, Israel; 116. Imperva, San Mateo, California, USA; 117. Indian Institute of Science, Bengaluru, India; 118. University of Oxford, Oxford, UK; 119. Blue Prism, Warrington, UK; 120. University of Western Ontario, London, Ontario, Canada; 121. Charité—Universitätsmedizin Berlin, Berlin, Germany; 122. No affiliation; 123. Seoul National University, Seoul, South Korea; 124. Qiagen, Germantown, Maryland, USA; 125. Invitae, San Francisco, California, USA; 126. Amazon, Seattle, Washington, USA; 127. Texas A&M University, College Station, Texas, USA; 128. Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; 129. IndieBio, San Francisco, California, USA; 130. iAccelerate, North Wollongong, Australia; 131. University of North Carolina at Charlotte, Charlotte, North Carolina, USA; 132. Vindara, Orlando, Florida, USA; 133. Temple University, Philadelphia, Pennsylvania, USA; 134. University of Greifswald, Greifswald, Germany; 135. Scripps Research Institute, San Diego, California, USA; 136. Brigham and Women's Hospital, Boston, Massachusetts, USA; 137. Harvard Medical School, Boston, Massachusetts, USA; 138. Harvard University, Cambridge, Massachusetts, USA; 139. Hebrew University of Jerusalem, Jerusalem, Israel; 140. School of Software Engineering and Computer Science, Azrieli College of Engineering, Jerusalem, Israel; 141. Lund University, Lund, Sweden; 142. University of Tampere, Tampere, Finland; 143. Soochow University, Suzhou, China; 144. Chinese University of Hong Kong, Hong Kong; 145. Hong Kong University of Science and Technology, Hong Kong; 146. George Mason University, Virginia, USA; 147. Sun Yat-sen University, Guangzhou, China; 148. Indiana University-Purdue University Indianapolis, Indianapolis, Indiana, USA; 149. Griffith University, Queensland, Australia; 150. Sun Yat-sen Memorial Hospital, Guangzhou, China; 151. Imperial College, London, UK; 152. Vertex Pharmaceuticals, Boston, Massachusetts, USA; 153. Jilin University, Changchun, China; 154. Northeastern University, China; 155. CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria; 156. Shenzhen Bay Laboratory, Shenzhen, China; 157. Kyoto University, Kyoto, Japan; 158. Paediatric Research Institute Città della Speranza, Padova, Italy; 159. Addenbrookes Hospital, University of Cambridge, Cambridge, UK; 160. Sapienza University of Rome, Rome, Italy; 161. Institute of Translational Pharmacology CNR, Rome, Italy; 162. University of Rome Tor Vergata, Rome, Italy; 163. ENEA—Frascati Research Centre, Rome, Italy; 164. Crick Institute, London, UK; 165. University of Rome, Rome, Italy; 166. University of Fribourg, Fribourg, Switzerland; 167. Open Humans Foundation, Sanford, North Carolina, USA; 168. PersonalGenomes.org, Sanford, North Carolina, USA; 169. Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; 170. Veneto Institute of Oncology, San Giovanni Rotondo, Italy; 171. Brown University, Providence, Rhode Island, USA; 172. Remix Therapeutics, Cambridge, Massachusetts, USA; 173. University of Kansas Medical Center, Kansas City, Kansas, USA; 174. Case Western Reserve University, Cleveland, Ohio, USA; 175. Kiel University, Kiel, Germany; 176. Christian-Albrechts-University of Kiel, Kiel, Germany; 177. UConn Health, Farmington, Connecticut, USA; 178. Oregon Health and Science University, Portland, Oregon, USA; 179. Third Rock Ventures, Boston, Massachusetts, USA; 180. Northwestern University, Evanston, Illinois, USA; 181. Rescindo Therapeutics Inc., Durham, North Carolina, USA; 182. AccuraGen Inc., Menlo Park, California, USA; 183. BioMarin, Novato, California, USA; 184. Global Blood Therapeutics, South San Francisco, California, USA; 185. Regional Hospital of Bolzano, Bolzano, Italy; 186. Ferrara University Hospital, Ferrara, Italy; 187. Mirano Hospital, Venice, Italy; 188. Fundacion Instituto Leloir, Buenos Aires, Argentina; 189. Milan Polyclinic, Milan, Italy; 190. University of Milan, Milan, Italy; 191. San Paolo Hospital, ASST Santi Paolo e Carlo, Milan, Italy; 192. University of Utah School of Medicine, Salt Lake City, Utah, USA; 193. AULSS 2 Marca Trevigiana, Treviso, Italy; 194. School of Public Health, University of Texas, Dallas, Texas, USA; 195. University of South Florida, Tampa, Florida, USA; 196. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA; 197. Hospital for Sick Children, Toronto, Ontario, Canada; 198. Beckman Research Institute of City of Hope, Duarte, California, USA; 199. National Institutes of Health, Bethesda, Maryland, USA; 200. Mount Sinai Health System, New York City, New York, USA; 201. Jackson Laboratory, Bar Harbor, Maine, USA; 202. American University of Beirut Medical Center, Beirut, Lebanon; 203. University of Copenhagen, Copenhagen, Denmark; 204. QIMR Berghofer Medical Research Institute, Brisbane, Australia; 205. Geisinger Genomic Medicine Institute, Danville, Pennsylvania, USA; 206. Intermountain Healthcare Precision Genomics, St. George, Utah, USA; 207. University of California at Davis, Davis, California, USA; 208. University of Washington School of Medicine, Seattle, Washington, USA; 209. Empirico Inc., San Diego, California, USA; 210. University of Tartu, Tartu, Estonia; 211. California Institute of Technology, Pasadena, California, USA; 212. Iowa State University, Ames, Iowa, USA; 213. Massachusetts General Hospital, Boston, Massachusetts, USA; 214. University of Texas Southwestern

Medical Center, Dallas, Texas, USA; 215. University of Maryland, Baltimore County, Baltimore, Maryland, USA; 216. Nationwide Children's Hospital, Columbus, Ohio, USA; 217. Sema4, Stamford, Connecticut, USA; 218. Khosla Ventures, Menlo Park, California, USA; 219. Seer.bio, Redwood City, California, USA; 220. University of Vermont, Burlington, Vermont, USA; 221. MyOme Inc, Palo Alto, California, USA; 222. LunaPBC, San Diego, California, USA; 223. SerImmune, Goleta, California, USA; 224. Newcastle University, Newcastle upon Tyne, UK; 225. National Institute of Standards and Technology, Portland, Oregon, USA.
† Deceased.

Review history
The review history is available as Additional file 8.

Peer review information
Kevin Pang was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Funding
The content of this work is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. The CAGI experiments and conferences were supported by National Institutes of Health (NIH) awards U41 HG007346, U24 HG007346, and R13 HG006650 to SEB, as well as a Research Agreement with Tata Consultancy Services (TCS) to SEB, and a supplement to NIH U19HD077627 to RLN. NIH U41HG007346 supported ANA, GA, CB, SEB, J-MC, RAH, ZH, LK, BAK, MKL, ZM, JM, AJN, RCP, YW, SFY; NIH U24 HG003467 supported CB, SEB, J-MC, ZH, SMF, RCP, PR, SJ; the TCS research agreement supported CAGI activities of ANA, CB, DB, J-MC, ZH, LK, and SFY; NIH U19 HD077627 supported RAH and SEB; NIH R13 HG006650 provided CAGI travel support to SA, ANA, GA, JRA, CB, DB, MAB, BB, SEB, BAB, YC, EC, MC, HC, J-MC, JC, MSC, RD, DD, RLD, MDE, BF, AWF, IF, SMF, MF, NG, MSG, NVG, JJH, RH, ZH, CLVH, TJPH, SK, RK, MK, PK, DK, BAK, AK, RHL, MKL, DM, GM, MSM, SDM, AAM, JM, SMM, AJN, AO, KAP, LRP, VRP, PR, SR, GS, AS, JMS, SS, RST, CT, AV, MEW, RW, SJW, ZY, SFY, JZ, MZ; NIH U41HG007346 provided travel support to RK, ZM, and SDM; UC Berkeley funds additionally provided CAGI travel support for YB and LS. CAGI projects and participants were further supported as follows: OA by an EMBO Installation Grant (No: 4163), TÜBİTAK (Grant IDs: 118C320, 121E365), TÜSEB (Grant ID: 4587); IA by NIH R01 GM078598, R35 GM127131, R01 HG010372; RBA by NIH GM102365; APA by the Office of Science, Office of Biological and Environmental Research of the US Department of Energy, under contract DE-AC02-05CH11231; ZA, MHÇ, JC, JG, and TYDN by a Competence Network for Technical, Scientific High Performance Computing in Bavaria KONWIHR, a Deutsche Forschungsgemeinschaft fellowship through the Graduate School of Quantitative Biosciences Munich and an NVIDIA hardware grant; MAB, AP, and DS by NIH U01HG009380; SB by the Asociación Española contra el Cáncer; PB and FL by the Fondazione CARIPARO, Padova, Italy; APB and SD by NIH U24 HG009293; WB by the NIH and the Greenwall Foundation; AJB by the Barbara and Gerson Bakar Foundation, and Priscilla Chan and Mark Zuckerberg; KC, SG, NS, and NMT by the Danish National Research Foundation Centre for Cardiac Arrhythmia; EC and PT by MIUR 201744NR8S; RC and VC by the 201744NR8S/Ministero Istruzione, Università e Ricerca, PRIN project; MSC by NIH U01 CA242954; MM and DNC by Qiagen through a License Agreement with Cardiff University; RD by a Paul and Daisy Soros Fellowship; XdlC by Spanish Ministerio de Ciencia e Innovación (PID2019-111217RB-I00) and Ministerio de Economía y Competitividad (SAF2016-80255-R and BIO2012-40133) and a European Regional Development Fund (Pirepred-EFA086/15); MD by NIH U41 HG007234; OD by FIS PI12/02585 and PI15/00355 from the Spanish Instituto de Salud Carlos III (ISCIII) funding, an initiative of the Spanish Ministry of Economy and Innovation partially supported by European Regional Development FEDER Funds; MDE, DKG, YG, KT, and HZ by NIH R01 HG008363, U01 HG007037, U41 HG007346, R13 HG006650; LE, VG, and MSG by a grant from the Intramural Research Program of the NHGRI to LE (1ZIAHG200323-14); WGF by NIH R01 GM127472; PF by MIUR 201744NR8S and 20182022D15D18000410001; MSF, MH, and GP by the Comision Nacional de Investigaciones Cientificas y Tecnicas (CONICET) [Grant ID: PIP 112201101–01002]; Universidad Nacional de Quilmes [Grant ID: 1402/15]; DMF by NIH R01 GM109110, R01 HG010461; AF, JH, and JMM by Wellcome Trust grant [098051]; and NIH U54 HG004555; AG and DP by intramural funding; NVG by NIH GM127390 and the Welch Foundation I-1505; SGE by PI13/01711, PI16/01218, PI19/01303, and PI22/01200 from the Spanish Instituto de Salud Carlos III (ISCIII) funding, an initiative of the Spanish Ministry of Economy and Innovation partially supported by European Regional Development FEDER Funds. SGE was also supported by the Miguel Servet Program [CP16/00034] and Government of Catalonia 2021SGR01112; TI by JSPS KAKENHI Grant Number 16HP8044; JOBJ, CJM by NIH R24OD011883; SK by NIH U01 HG007019, R01 HG003747; MK by NIH AG054012, AG058002, AG062377, NS110453, NS115064, AG062335, AG074003, NS127187, AG067151, MH 109978, MH 119509, HG 008155, DA 053631; IVK by RSF 20–74-10075; CL, JVdA by Color Genomics; KAM by NIH R35 GM142886; GM by NIH T32 LM012409; JM by NIH R01 GM120364, R01 GM104436; LO, SÖ, NP, CR by the Spanish Ministerio de Ciencia e Innovación (PID2019-111217RB-I00) and Ministerio de Economía y Competitividad (SAF2016-80255-R and BIO2012-40133); European Regional Development Fund (Pirepred-EFA086/15); VP by NIH K99 LM012992; SDM and PR by R01 LM009722 and R01 MH105524; PR by U01 HG012022 and Precision Health Initiative of Indiana University; AR by NIH T32 HG002536; SR by a Marie Curie International Outgoing Fellowship PIOF-GA-2009–237751; PKR by Natural Sciences and Engineering Research Council of Canada 371758–2009, Canadian Breast Cancer Foundation, Canada Foundation For Innovation, Canada Research Chairs, Compute Canada and Western University; FR, JVD, JS by VIB and KU Leuven; FPR by One Brave Idea Initiative, the NIH HG004233, HG010461, the Canada Excellence Research Chairs, and a Canadian Institutes of Health Research Foundation Grant; PCS by NIH UM1 HG009435; JRS by NIH R35 GM130361; CS by MUR PRIN2017 2017483NH8_002; CS by NRF-2020M3A9G7103933; YS by NIH R35 GM124952; LMS by NIH RM1 HG010461; RT by NIH 1UM1 HG009435; JMS by NCI R01CA180778; SVT by NCI R01 CA121245; MV by Finnish Academy, Swedish Research Council, Swedish Cancer Society; MEW by NSF GRF, NIH GM068763; MHW by the National Natural Science Foundation of China (NSFC) [31871340, 71974165]; FZ by The Senior and Junior Technological Innovation Team (20210509055RQ), the Jilin Provincial Key Laboratory of Big Data Intelligent Computing (20180622002JC), and the Fundamental Research Funds for the Central Universities, JLU; YZ by the Australia Research Council [DP210101875]; EZ by NCI K24 CA169004, California Initiative to Advance Precision Medicine.

Availability of data and materials
The code used for the analyses in this paper is available on GitHub (https://github.com/genomeinterpretation/CAGI50) and zenodo (doi.org/10.5281/zenodo.8436229) [91] under an MIT Open Source license.

Data availability:
The answer keys for all reanalyzed CAGI challenges are available on the CAGI website. Access requires registration including acceptance of a data use agreement. The CAGI website is an archival venue that has been in operation longer than resources such as Zenodo. Papers containing publicly available answer keys are also referenced in the table.

Challenges	CAGI Year	Data source type	DOI
NAGLU	CAGI 4	Publication	https://doi.org/10.1371/journal.pone.0200008
PTEN/TPMT	CAGI 5	Publication	https://doi.org/10.1038/s41588-018-0122-z
Annotate All Missense	CAGI 5	CAGI website*,**	
CALM1	CAGI 5	Publication	https://doi.org/10.15252/msb.20177908
GAA	CAGI 5	CAGI website	
CBS	CAGI 1 & 2	CAGI website	
SUMO-ligase	CAGI 4	CAGI website	
PCM1	CAGI 5	Publication	https://doi.org/10.1038/s41467-020-19637-5
L-PYK	CAGI 4	CAGI website	
p53 rescue	CAGI 2	Data file (training set) CAGI website (answer key)	https://doi.org/10.24432/C5T89H
Frataxin	CAGI 5	Publication	https://doi.org/10.1002/humu.23843
p16	CAGI 3	Publication	https://doi.org/10.1002/humu.22550
ENIGMA	CAGI 5	CAGI website	
BRCA	CAGI 3	CAGI website	
Vex-Seq	CAGI 5	CAGI website	
eQTL	CAGI 4	Publication	https://doi.org/10.1002/humu.23197
MaPSy	CAGI 5	Publication	https://doi.org/10.1186/s13059-019-1653-z
Regulation-Satutation	CAGI 5	CAGI website	
Crohn's	CAGI 4	Publication*** CAGI website (answer key)	https://doi.org/10.1097/MIB.0000000000001235 (pediatric IBD cohort). https://doi.org/10.1016/j.ebiom.2016.08.037 (healthy controls) https://portal.popgen.de
SickKids	CAGI 4 & 5	Publication***	https://doi.org/10.1002/humu.23874

* The CAGI website URL is https://genomeinterpretation.org

** A subset of variants used in the Annotate All Missense challenge were obtained from the a proprietary version of the HGMD, database. These variants (those in HGMD 2020.4 but not HGMD 2019) are excluded from the answer key on the CAGI website but can be obtained under the HGMD Professional license (https://www.hgmd.cf.ac.uk).

*** The full patient data for the Crohn's and SickKids challenges are not maintained on the CAGI website because of patient privacy issues. However, approved users may obtain these from the authors of the corresponding publications in the above table.

Declarations

Ethics approval and consent to participate
This work was not Human Subject research. As described in the main text, the CAGI organizational structure includes an Ethics Forum that provides detailed advice on relevant aspects of study design and responsible data governance.

Competing interests
Principal authors of this paper participated as predictors in many of the CAGI challenges reported. The unified numerical framework employed for reanalysis of the challenges yields results that are consistent with those obtained by the independent assessors of each challenge and in particular selected methods are the highest ranked in the independent assessments. Nevertheless, while every care was taken to mitigate any potential biases in this work, the authors' participation in CAGI may have affected the presentation of findings, including the selection of challenges, metrics, assessment criteria, and emphasis given on particular results.
VBG is a current employee and shareholder of AstraZeneca; RB is a shareholder of enGenome; AJB is a co-founder and consultant to Personalis and NuMedii as well as a consultant to Samsung, Mango Tree Corporation and in the recent past, 10 × Genomics, Helix and Pathway; Carles Corbi-Verge is a computational scientist at the drug discovery company;

Cyclica INC and is compensated with income and equity; KC is one of the Regeneron authors and owns options and/or stock of the company; DD is Chief Scientist at Geneyx Genomex Ltd; CD is a consultant to Exact Sciences and is compensated with income and equity; GAG receives research funds from IBM and Pharmacyclics and is an inventor on patent applications related to MSMuTect, MSMutSig, MSIDetect, POLYSOLVER, and SignatureAnalyzer-GPU, and is a founder, consultant, and holds privately held equity in Scorpion Therapeutics; NG is an employee and stockholder at Pacific Biosciences; RH is a paid consultant for Invitae and Scientific Advisory Board member for Variant Bio; AK is a consultant at Illumina Inc., Scientific Advisory Board member of OpenTargets; KK is one of the Regeneron authors and owns options and/or stock of the company; IL is an employer and stockholder of enGenome; MSM owns stock in PhenoTips; GN is an employee of enGenome; AOD-L is a member of the Scientific Advisory Board of Congenica; ER is a shareholder of enGenome; PKR is the founder of CytoGnomix; FPR is a shareholder in Ranomics and SeqWell, an advisor for SeqWell, BioSymetrics, and Constantiam BioSciences, and has received research sponsorships from Biogen, Alnylam, Deep Genomics, and Beam Therapeutics; PCS is the co-founder and shareholder of Sherlock Biosciences, a board member and shareholder of Danaher Corporation, and has filed patents related to this work; PLFT is an employer and stockholder in AccuraGen; RT has filed patents related to this work; MHW is a shareholder of Beth Bioinformatics Co., Ltd.; CMY is an employee and shareholder of Vertex Pharmaceuticals; JZ is an employee of AstraZeneca; SEB receives support at the University of California, Berkeley from a research agreement from TCS.

Received: 21 January 2023 Accepted: 17 November 2023
Published online: 22 February 2024

References

1. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, Kathiresan S, Kenny EE, Lindgren CM, MacArthur DG, North KN, Plon SE, Rehm HL, Risch N, Rotimi CN, Shendure J, Soranzo N, McCarthy MI. A brief history of human disease genetics. Nature. 2020;577(7789):179–89.
2. Gibbs RA. The human genome project changed everything. Nat Rev Genet. 2020;21(10):575–6.
3. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061–73.
4. Cutting GR. Cystic fibrosis genetics: from molecular understanding to clinical application. Nat Rev Genet. 2015;16(1):45–56.
5. Nielsen FC, van Overeem HT, Sorensen CS. Hereditary breast and ovarian cancer: new genes in confined pathways. Nat Rev Cancer. 2016;16(9):599–612.
6. ICGC Tcga Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020;578(7793):82–93.
7. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, Ma C, Fontanillas P, Moutsianas L, McCarthy DJ, Rivas MA, Perry JRB, Sim X, Blackwell TW, Robertson NR, Rayner NW, Cingolani P, Locke AE, Tajes JF, Highland HM, Dupuis J, Chines PS, Lindgren CM, Hartl C, Jackson AU, Chen H, Huyghe JR, van de Bunt M, Pearson RD, Kumar A, Muller-Nurasyid M, Grarup N, Stringham HM, Gamazon ER, Lee J, Chen Y, Scott RA, Below JE, Chen P, Huang J, Go MJ, Stitzel ML, Pasko D, Parker SCJ, Varga TV, Green T, Beer NL, Day-Williams AG, Ferreira T, Fingerlin T, Horikoshi M, Hu C, Huh I, Ikram MK, Kim BJ, Kim Y, Kim YJ, Kwon MS, Lee J, Lee S, Lin KH, Maxwell TJ, Nagai Y, Wang X, Welch RP, Yoon J, Zhang W, Barzilai N, Voight BF, Han BG, Jenkinson CP, Kuulasmaa T, Kuusisto J, Manning A, Ng MCY, Palmer ND, Balkau B, Stancakova A, Abboud HE, Boeing H, Giedraitis V, Prabhakaran D, Gottesman O, Scott J, Carey J, Kwan P, Grant G, Smith JD, Neale BM, Purcell S, Butterworth AS, Howson JMM, Lee HM, Lu Y, Kwak SH, Zhao W, Danesh J, Lam VKL, Park KS, Saleheen D, So WY, Tam CHT, Afzal U, Aguilar D, Arya R, Aung T, Chan E, Navarro C, Cheng CY, Palli D, Correa A, Curran JE, Rybin D, Farook VS, Fowler SP, Freedman BI, Griswold M, Hale DE, Hicks PJ, Khor CC, Kumar S, Lehne B, Thuillier D, Lim WY, Liu J, van der Schouw YT, Loh M, Musani SK, Puppala S, Scott WR, Yengo L, Tan ST, Taylor HA Jr, Thameem F, Wilson G Sr, Wong TY, Njolstad PR, Levy JC, Mangino M, Bonnycastle LL, Schwarzmayr T, Fadista J, Surdulescu GL, Herder C, Groves CJ, Wieland T, Bork-Jensen J, Brandslund I, Christensen C, Koistinen HA, Doney ASF, Kinnunen L, Esko T, Farmer AJ, Hakaste L, Hodgkiss D, Kravic J, Lyssenko V, Hollensted M, Jorgensen ME, Jorgensen T, Ladenvall C, Justesen JM, Karajamaki A, Kriebel J, Rathmann W, Lannfelt L, Lauritzen T, Narisu N, Linneberg A, Melander O, Milani L, Neville M, Orho-Melander M, Qi L, Qi Q, Roden M, Rolandsson O, Swift A, Rosengren AH, Stirrups K, Wood AR, Mihailov E, Blancher C, Carneiro MO, Maguire J, Poplin R, Shakir K, Fennell T, DePristo M, de Angelis MH, Deloukas P, Gjesing AP, Jun G, Nilsson P, Murphy J, Onofrio R, Thorand B, Hansen T, Meisinger C, Hu FB, Isomaa B, Karpe F, Liang L, Peters A, Huth C, O'Rahilly SP, Palmer CNA, Pedersen O, Rauramaa R, Tuomilehto J, Salomaa V, Watanabe RM, Syvanen AC, Bergman RN, Bharadwaj D, Bottinger EP, Cho YS, Chandak GR, Chan JCN, Chia KS, Daly MJ, Ebrahim SB, Langenberg C, Elliott P, Jablonski KA, Lehman DM, Jia W, Ma RCW, Pollin TI, Sandhu M, Tandon N, Froguel P, Barroso I, Teo YY, Zeggini E, Loos RJF, Small KS, Ried JS, DeFronzo RA, Grallert H, Glaser B, Metspalu A, Wareham NJ, Walker M, Banks E, Gieger C, Ingelsson E, Im HK, Illig T, Franks PW, Buck G, Trakalo J, Buck D, Prokopenko I, Magi R, Lind L, Farjoun Y, Owen KR, Gloyn AL, Strauch K, Tuomi T, Kooner JS, Lee JY, Park T, Donnelly P, Morris AD, Hattersley AT, Bowden DW, Collins FS, Atzmon G, Chambers JC, Spector TD, Laakso M, Strom TM, Bell GI, Blangero J, Duggirala R, Tai ES, McVean G, Hanis CL, Wilson JG, Seielstad M, Frayling TM, Meigs JB, Cox NJ, Sladek R, Lander ES, Gabriel S, Burtt NP, Mohlke KL, Meitinger T, Groop L, Abecasis G, Florez JC, Scott LJ, Morris AP, Kang HM, Boehnke M, Altshuler D, McCarthy MI. The genetic architecture of type 2 diabetes. Nature. 2016;536(7614):41–7.
8. Wadelius M, Pirmohamed M. Pharmacogenetics of warfarin: current status and future challenges. Pharmacogenomics J. 2007;7(2):99–111.

9. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, Jang W, Katz K, Ovetsky M, Riley G, Sethi A, Tully R, Villamarin-Salomon R, Rubinstein W, Maglott DR. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44(D1):D862–868.
10. Hu Z, Yu C, Furutsuki M, Andreoletti G, Ly M, Hoskins R, Adhikari AN, Brenner SE. VIPdb, a genetic variant impact predictor database. Hum Mutat. 2019;40(9):1202–14.
11. Katsonis P, Wilhelm K, Williams A, Lichtarge O. Genome interpretation using in silico predictors of variant impact. Hum Genet. 2022;141(10):1549–77.
12. Sanavia T, Birolo G, Montanucci L, Turina P, Capriotti E, Fariselli P. Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. Comput Struct Biotechnol J. 2020;18:1968–79.
13. Backwell L, Marsh JA. Diverse molecular mechanisms underlying pathogenic protein mutations: beyond the loss-of-function paradigm. Annu Rev Genomics Hum Genet. 2022;23:475–98.
14. Riolo G, Cantara S, Ricci C. What's wrong in a jump? Prediction and validation of splice site variants. Methods Protoc. 2021;4(3):62.
15. Avsec Z, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P, Kelley DR. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18(10):1196–203.
16. Ibrahim DM, Mundlos S. Three-dimensional chromatin in disease: what holds us together and what drives us apart? Curr Opin Cell Biol. 2020;64:1–9.
17. Moult J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. Proteins. 1995;23:ii–iv.
18. Hoskins RA, Repo S, Barsky D, Andreoletti G, Moult J, Brenner SE. Reports from CAGI: the critical assessment of genome interpretation. Hum Mutat. 2017;38(9):1039–41.
19. Andreoletti G, Pal LR, Moult J, Brenner SE. Reports from the fifth edition of CAGI: the Critical Assessment of Genome Interpretation. Hum Mutat. 2019;40(9):11907–1201.
20. Pejaver V, Babbi G, Casadio R, Folkman L, Katsonis P, Kundu K, Lichtarge O, Martelli PL, Miller M, Moult J, Pal LR, Savojardo C, Yin Y, Zhou Y, Radivojac P, Bromberg Y. Assessment of methods for predicting the effects of PTEN and TPMT protein variants. Hum Mutat. 2019;40(9):1495–506.
21. Savojardo C, Petrosino M, Babbi G, Bovo S, Corbi-Verge C, Casadio R, Fariselli P, Folkman L, Garg A, Karimi M, Katsonis P, Kim PM, Lichtarge O, Martelli PL, Pasquo A, Pal D, Shen Y, Strokach AV, Turina P, Zhou Y, Andreoletti G, Brenner SE, Chiaraluce R, Consalvi V, Capriotti E. Evaluating the predictions of the protein stability change upon single amino acid substitutions for the FXN CAGI5 challenge. Hum Mutat. 2019;40(9):1392–9.
22. Clark WT, Kasak L, Bakolitsa C, Hu Z, Andreoletti G, Babbi G, Bromberg Y, Casadio R, Dunbrack R, Folkman L, Ford CT, Jones D, Katsonis P, Kundu K, Lichtarge O, Martelli PL, Mooney SD, Nodzak C, Pal LR, Radivojac P, Savojardo C, Shi X, Zhou Y, Uppal A, Xu Q, Yin Y, Pejaver V, Wang M, Wei L, Moult J, Yu GK, Brenner SE, LeBowitz JH. Assessment of predicted enzymatic activity of alpha-N-acetylglucosaminidase variants of unknown significance for CAGI 2016. Hum Mutat. 2019;40(9):1519–29.
23. Zhang J, Kinch LN, Cong Q, Weile J, Sun S, Cote AG, Roth FP, Grishin NV. Assessing predictions of fitness effects of missense mutations in SUMO-conjugating enzyme UBE2I. Hum Mutat. 2017;38(9):1051–63.
24. Carraro M, Minervini G, Giollo M, Bromberg Y, Capriotti E, Casadio R, Dunbrack R, Elefanti L, Fariselli P, Ferrari C, Gough J, Katsonis P, Leonardi E, Lichtarge O, Menin C, Martelli PL, Niroula A, Pal LR, Repo S, Scaini MC, Vihinen M, Wei Q, Xu Q, Yang Y, Yin Y, Zaucha J, Zhao H, Zhou Y, Brenner SE, Moult J, Tosatto SCE. Performance of in silico tools for the evaluation of p16INK4a (CDKN2A) variants in CAGI. Hum Mutat. 2017;38(9):1042–50.
25. Zhang J, Kinch LN, Cong Q, Katsonis P, Lichtarge O, Savojardo C, Babbi G, Martelli PL, Capriotti E, Casadio R, Garg A, Pal D, Weile J, Sun S, Verby M, Roth FP, Grishin NV. Assessing predictions on fitness effects of missense variants in calmodulin. Hum Mutat. 2019;40(9):1463–73.
26. Kasak L, Hunter JM, Udani R, Bakolitsa C, Hu Z, Adhikari AN, Babbi G, Casadio R, Gough J, Guerrero RF, Jiang Y, Joseph T, Katsonis P, Kotte S, Kundu K, Lichtarge O, Martelli PL, Mooney SD, Moult J, Pal LR, Poitras J, Radivojac P, Rao A, Sivadasan N, Sunderam U, Saipradeep VG, Yin Y, Zaucha J, Brenner SE, Meyn MS. CAGI SickKids challenges: assessment of phenotype and variant predictions derived from clinical and genomic data of children with undiagnosed diseases. Hum Mutat. 2019;40(9):1373–91.
27. Cline MS, Babbi G, Bonache S, Cao Y, Casadio R, de la Cruz X, Diez O, Gutierrez-Enriquez S, Katsonis P, Lai C, Lichtarge O, Martelli PL, Mishne G, Moles-Fernandez A, Montalban G, Mooney SD, O'Conner R, Ootes L, Ozkan S, Padilla N, Pagel KA, Pejaver V, Radivojac P, Riera C, Savojardo C, Shen Y, Sun Y, Topper S, Parsons MT, Spurdle AB, Goldgar DE, ENIGMA Consortium. Assessment of blind predictions of the clinical significance of BRCA1 and BRCA2 variants. Hum Mutat. 2019;40(9):1546–56.
28. Carraro M, Monzon AM, Chiricosta L, Reggiani F, Aspromonte MC, Bellini M, Pagel K, Jiang Y, Radivojac P, Kundu K, Pal LR, Yin Y, Limongelli I, Andreoletti G, Moult J, Wilson SJ, Katsonis P, Lichtarge O, Chen J, Wang Y, Hu Z, Brenner SE, Ferrari C, Murgia A, Tosatto SCE, Leonardi E. Assessment of patient clinical descriptions and pathogenic variants from gene panel sequences in the CAGI-5 intellectual disability challenge. Hum Mutat. 2019;40(9):1330–45.
29. Daneshjou R, Wang Y, Bromberg Y, Bovo S, Martelli PL, Babbi G, Lena PD, Casadio R, Edwards M, Gifford D, Jones DT, Sundaram L, Bhat RR, Li X, Pal LR, Kundu K, Yin Y, Moult J, Jiang Y, Pejaver V, Pagel KA, Li B, Mooney SD, Radivojac P, Shah S, Carraro M, Gasparini A, Leonardi E, Giollo M, Ferrari C, Tosatto SCE, Bachar E, Azaria JR, Ofran Y, Unger R, Niroula A, Vihinen M, Chang B, Wang MH, Franke A, Petersen BS, Pirooznia M, Zandi P, McCombie R, Potash JB, Altman RB, Klein TE, Hoskins RA, Repo S, Brenner SE, Morgan AA. Working toward precision medicine: predicting phenotypes from exomes in the critical assessment of genome interpretation (CAGI) challenges. Hum Mutat. 2017;38(9):1182–92.
30. Callaway E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. Nature. 2020;588(7837):203–4.

31. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Zidek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D. Improved protein structure prediction using potentials from deep learning. Nature. 2020;577(7792):706–10.
32. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL, ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405–24.
33. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91.
34. Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, Kircher M, Khechaduri A, Dines JN, Hause RJ, Bhatia S, Evans WE, Relling MV, Yang W, Shendure J, Fowler DM. Multiplex assessment of protein variant abundance by massively parallel sequencing. Nat Genet. 2018;50(6):874–82.
35. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9.
36. Peterson TA, Doughty E, Kann MG. Towards precision medicine: advances in computational approaches for the analysis of human variants. J Mol Biol. 2013;425(21):4047–63.
37. Pejaver V, Mooney SD, Radivojac P. Missense variant pathogenicity predictors generalize well across a range of function-specific prediction challenges. Hum Mutat. 2017;38(9):1092–108.
38. Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res. 2007;35(11):3823–35.
39. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14(Suppl 3):S3.
40. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam HJ, Mort M, Cooper DN, Sebat J, Iakoucheva LM, Mooney SD, Radivojac P. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun. 2020;11(1):5918.
41. Capriotti E, Martelli PL, Fariselli P, Casadio R. Blind prediction of deleterious amino acid variations with SNPs&GO. Hum Mutat. 2017;38(9):1064–71.
42. Narasimhan VM, Hunt KA, Mason D, Baker CL, Karczewski KJ, Barnes MR, Barnett AH, Bates C, Bellary S, Bockett NA, Giorda K, Griffiths CJ, Hemingway H, Jia Z, Kelly MA, Khawaja HA, Lek M, McCarthy S, McEachan R, O'Donnell-Luria A, Paigen K, Parisinos CA, Sheridan E, Southgate L, Tee L, Thomas M, Xue Y, Schnall-Levin M, Petkov PM, Tyler-Smith C, Maher ER, Trembath RC, MacArthur DG, Wright J, Durbin R, van Heel DA. Health and population effects of rare gene knockouts in adult humans with related parents. Science. 2016;352(6284):474–7.
43. Katsonis P, Lichtarge O. A formal perturbation equation between genotype and phenotype determines the evolutionary action of protein-coding variations on fitness. Genome Res. 2014;24(12):2050–8.
44. Wang Z, Moult J. SNPs, protein structure, and disease. Hum Mutat. 2001;17(4):263–70.
45. Lugo-Martinez J, Pejaver V, Pagel KA, Jain S, Mort M, Cooper DN, Mooney SD, Radivojac P. The loss and gain of functional amino acid residues is a common mechanism causing human inherited disease. PLoS Comput Biol. 2016;12(8): e1005091.
46. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Zidek A, Bridgland A, Cowie A, Meyer C, Laydon A, Velankar S, Kleywegt GJ, Bateman A, Evans R, Pritzel A, Figurnov M, Ronneberger O, Bates R, Kohl SAA, Potapenko A, Ballard AJ, Romera-Paredes B, Nikolov S, Jain R, Clancy E, Reiman D, Petersen S, Senior AW, Kavukcuoglu K, Birney E, Kohli P, Jumper J, Hassabis D. Highly accurate protein structure prediction for the human proteome. Nature. 2021;596(7873):590–6.
47. Iqbal S, Li F, Akutsu T, Ascher DB, Webb GI, Song J. Assessing the performance of computational predictors for estimating protein stability changes upon missense mutations. Brief Bioinform. 2021;22(6):bbab184.
48. Rost B, Radivojac P, Bromberg Y. Protein function in precision medicine: deep understanding with machine learning. FEBS Lett. 2016;590(15):2327–41.
49. Clark WT, Yu GK, Aoyagi-Scharber M, LeBowitz JH. Utilizing ExAC to assess the hidden contribution of variants of unknown significance to sanfilippo type B incidence. PLoS ONE. 2018;13(7): e0200008.
50. Pejaver V, Byrne AB, Feng B-J, Pagel KA, Mooney SD, Karchin R, O'Donnell-Luria A, Harrison SM, Tavtigian SV, Greenblatt MS, Biesecker LG, Radivojac P, Brenner SE, ClinGen Sequence Variant Interpretation Working Group. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. Am J Hum Genet. 2022;109(12):2163–77.
51. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL, Plon SE, Ramos EM, Sherry ST, Watson MS, ClinGen. ClinGen–the Clinical Genome Resource. N Engl J Med. 2015;372(23):2235–42.
52. Tavtigian SV, Greenblatt MS, Harrison SM, Nussbaum RL, Prabhu SA, Boucher KM, Biesecker LG, ClinGen Sequence Variant Interpretation Working Group. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. Genet Med. 2018;20(9):1054–60.
53. Stenson PD, Mort M, Ball EV, Chapman M, Evans K, Azevedo L, Hayden M, Heywood S, Millar DS, Phillips AD, Cooper DN. The human gene mutation database (HGMD®): optimizing its use in a clinical diagnostic or research setting. Hum Genet. 2020;139(10):1197–207.

54. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 2020;12(1):103.
55. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, Cannon-Albright LA, Teerlink CC, Stanford JL, Isaacs WB, Xu J, Cooney KA, Lange EM, Schleutker J, Carpten JD, Powell IJ, Cussenot O, Cancel-Tassin G, Giles GG, MacInnis RJ, Maier C, Hsieh CL, Wiklund F, Catalona WJ, Foulkes WD, Mandal D, Eeles RA, Kote-Jarai Z, Bustamante CD, Schaid DJ, Hastie T, Ostrander EA, Bailey-Wilson JE, Radivojac P, Thibodeau SN, Whittemore AS, Sieh W. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet. 2016;99(4):877–85.
56. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24(8):2125–37.
57. Mount SM, Avsec Z, Carmel L, Casadio R, Celik MH, Chen K, Cheng J, Cohen NE, Fairbrother WG, Fenesh T, Gagneur J, Gotea V, Holzer T, Lin CF, Martelli PL, Naito T, Nguyen TYD, Savojardo C, Unger R, Wang R, Yang Y, Zhao H. Assessing predictions of the impact of variants on splicing in CAGI5. Hum Mutat. 2019;40(9):1215–24.
58. Cheng J, Nguyen TYD, Cygan KJ, Celik MH, Fairbrother WG, Avsec Z, Gagneur J. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. Genome Biol. 2019;20(1):48.
59. Kreimer A, Zeng H, Edwards MD, Guo Y, Tian K, Shin S, Welch R, Wainberg M, Mohan R, Sinnott-Armstrong NA, Li Y, Eraslan G, Amin TB, Tewhey R, Sabeti PC, Goke J, Mueller NS, Kellis M, Kundaje A, Beer MA, Keles S, Gifford DK, Yosef N. Predicting gene expression in massively parallel reporter assays: a comparative study. Hum Mutat. 2017;38(9):1240–50.
60. Shigaki D, Adato O, Adhikari AN, Dong S, Hawkins-Hooker A, Inoue F, Juven-Gershon T, Kenlay H, Martin B, Patra A, Penzar DD, Schubach M, Xiong C, Yan Z, Boyle AP, Kreimer A, Kulakovskiy IV, Reid J, Unger R, Yosef N, Shendure J, Ahituv N, Kircher M, Beer MA. Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay. Hum Mutat. 2019;40(9):1280–91.
61. Chandonia JM, Adhikari A, Carraro M, Chhibber A, Cutting GR, Fu Y, Gasparini A, Jones DT, Kramer A, Kundu K, Lam HYK, Leonardi E, Moult J, Pal LR, Searls DB, Shah S, Sunyaev S, Tosatto SCE, Yin Y, Buckley BA. Lessons from the CAGI-4 Hopkins clinical panel challenge. Hum Mutat. 2017;38(9):1155–68.
62. Testa U, Testa EP, Mavilio F, Petrini M, Sposi NM, Petti S, Samoggia P, Montesoro E, Giannella G, Bottero L, et al. Differential regulation of transferrin receptor gene expression in human hemopoietic cells: molecular and cellular aspects. J Recept Res. 1987;7(1–4):355–75.
63. Pal LR, Yu CH, Mount SM, Moult J. Insights from GWAS: emerging landscape of mechanisms underlying complex trait disease. BMC Genomics. 2015;16(Suppl 8):S4.
64. Wand H, Lambert SA, Tamburro C, Iacocca MA, O'Sullivan JW, Sillari C, Kullo IJ, Rowley R, Dron JS, Brockman D, Venner E, McCarthy MI, Antoniou AC, Easton DF, Hegele RA, Khera AV, Chatterjee N, Kooperberg C, Edwards K, Vlessis K, Kinnear K, Danesh JN, Parkinson H, Ramos EM, Roberts MC, Ormond KE, Khoury MJ, Janssens A, Goddard KAB, Kraft P, MacArthur JAL, Inouye M, Wojcik GL. Improving reporting standards for polygenic scores in risk prediction studies. Nature. 2021;591(7849):211–9.
65. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–78.
66. Dahlhamer JM, Zammitti EP, Ward BW, Wheaton AG, Croft JB. Prevalence of inflammatory bowel disease among adults aged ≥18 years - United States, 2015. MMWR Morb Mortal Wkly Rep. 2016;65(42):1166–9.
67. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, Kathiresan S. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50(9):1219–24.
68. Chen YC, Douville C, Wang C, Niknafs N, Yeo G, Beleva-Guthrie V, Carter H, Stenson PD, Cooper DN, Li B, Mooney S, Karchin R. A probabilistic model to predict clinical phenotypic traits from genome sequencing. PLoS Comput Biol. 2014;10(9): e1003825.
69. Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, Fritzilas N, Hakenberg J, Dutta A, Shon J, Xu J, Batzoglou S, Li X, Farh KK. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet. 2018;50(8):1161–70.
70. Wang Y, Miller M, Astrakhan Y, Petersen BS, Schreiber S, Franke A, Bromberg Y. Identifying crohn's disease signal from variome analysis. Genome Med. 2019;11(1):59.
71. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)-round XIV. Proteins. 2021;89(12):1607–17.
72. Wang RJ, Radivojac P, Hahn MW. Distinct error rates for reference and nonreference genotypes estimated by pedigree analysis. Genetics. 2021;217(1):1–10.
73. Cai B, Li B, Kiga N, Thusberg J, Bergquist T, Chen YC, Niknafs N, Carter H, Tokheim C, Beleva-Guthrie V, Douville C, Bhattacharya R, Yeo HTG, Fan J, Sengupta S, Kim D, Cline M, Turner T, Diekhans M, Zaucha J, Pal LR, Cao C, Yu CH, Yin Y, Carraro M, Giollo M, Ferrari C, Leonardi E, Tosatto SCE, Bobe J, Ball M, Hoskins RA, Repo S, Church G, Brenner SE, Moult J, Gough J, Stanke M, Karchin R, Mooney SD. Matching phenotypes to whole genomes: lessons learned from four iterations of the personal genome project community challenges. Hum Mutat. 2017;38(9):1266–76.
74. Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, Seelig G, Shendure J, Fowler DM. Variant interpretation: functional assays to the rescue. Am J Hum Genet. 2017;101(3):315–25.
75. Brenner SE, Chothia C, Hubbard TJ. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci U S A. 1998;95(11):6073–8.
76. Adams NM, Hand DJ. Comparing classifiers when the misallocation costs are uncertain. Pattern Recognit. 1999;32(7):1139–47.
77. Wu Y, Li R, Sun S, Weile J, Roth FP. Improved pathogenicity prediction for rare human missense variants. Am J Hum Genet. 2021;108(10):1891–906.
78. Bournazos AM, Riley LG, Bommireddipalli S, Ades L, Akesson LS, Al-Shinnag M, Alexander SI, Archibald AD, Balasubramaniam S, Berman Y, Beshay V, Boggs K, Bojadzieva J, Brown NJ, Bryen SJ, Buckley MF, Chong B, Davis MR,

Dawes R, Delatycki M, Donaldson L, Downie L, Edwards C, Edwards M, Engel A, Ewans LJ, Faiz F, Fennell A, Field M, Freckmann ML, Gallacher L, Gear R, Goel H, Goh S, Goodwin L, Hanna B, Harraway J, Higgins M, Ho G, Hopper BK, Horton AE, Hunter MF, Huq AJ, Josephi-Taylor S, Joshi H, Kirk E, Krzesinski E, Kumar KR, Lemckert F, Leventer RJ, Lindsey-Temple SE, Lunke S, Ma A, Macaskill S, Mallawaarachchi A, Marty M, Marum JE, McCarthy HJ, Menezes MP, McLean A, Milnes D, Mohammad S, Mowat D, Niaz A, Palmer EE, Patel C, Patel SG, Phelan D, Pinner JR, Rajagopalan S, Regan M, Rodgers J, Rodrigues M, Roxburgh RH, Sachdev R, Roscioli T, Samarasekera R, Sandaradura SA, Savva E, Schindler T, Shah M, Sinnerbrink IB, Smith JM, Smith RJ, Springer A, Stark Z, Strom SP, Sue CM, Tan K, Tan TY, Tantsis E, Tchan MC, Thompson BA, Trainer AH, van Spaendonck-Zwarts K, Walsh R, Warwick L, White S, White SM, Williams MG, Wilson MJ, Wong WK, Wright DC, Yap P, Yeung A, Young H, Jones KJ, Bennetts B, Cooper ST, Australasian Consortium for RNA Diagnostics. Standardized practices for RNA diagnostics using clinically accessible specimens reclassifies 75% of putative splicing variants. Genet Med. 2022;24(1):130–45.

79. Knight WR. A computer method for calculating kendall's tau with ungrouped data. J Am Stat Assoc. 1966;61(314):436–9.
80. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. New York, NY: Springer Verlag; 2001.
81. Hanley J, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.
82. Byrne S. A note on the use of empirical AUC for evaluating probabilistic forecasts. Electron J Stat. 2016;10(1):380–93.
83. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett. 2006;27:861–74.
84. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975;405(2):442–51.
85. Efron B. Size, power and false discovery rates. Ann Stat. 2007;35(4):1351–77.
86. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol. 2003;56(11):1129–35.
87. Breast Cancer Association Consortium, Dorling L, Carvalho S, Allen J, Gonzalez-Neira A, Luccarini C, Wahlstrom C, Pooley KA, Parsons MT, Fortuno C, Wang Q, Bolla MK, Dennis J, Keeman R, Alonso MR, Alvarez N, Herraez B, Fernandez V, Nunez-Torres R, Osorio A, Valcich J, Li M, Torngren T, Harrington PA, Baynes C, Conroy DM, Decker B, Fachal L, Mavaddat N, Ahearn T, Aittomaki K, Antonenkova NN, Arnold N, Arveux P, Ausems M, Auvinen P, Becher H, Beckmann MW, Behrens S, Bermisheva M, Bialkowska K, Blomqvist C, Bogdanova NV, Bogdanova-Markov N, Bojesen SE, Bonanni B, Borresen-Dale AL, Brauch H, Bremer M, Briceno I, Bruning T, Burwinkel B, Cameron DA, Camp NJ, Campbell A, Carracedo A, Castelao JE, Cessna MH, Chanock SJ, Christiansen H, Collee JM, Cordina-Duverger E, Cornelissen S, Czene K, Dork T, Ekici AB, Engel C, Eriksson M, Fasching PA, Figueroa J, Flyger H, Forsti A, Gabrielson M, Gago-Dominguez M, Georgoulias V, Gil F, Giles GG, Glendon G, Garcia EBG, Alnaes GIG, Guenel P, Hadjisavvas A, Haeberle L, Hahnen E, Hall P, Hamann U, Harkness EF, Hartikainen JM, Hartman M, He W, Heemskerk-Gerritsen BAM, Hillemanns P, Hogervorst FBL, Hollestelle A, Ho WK, Hooning MJ, Howell A, Humphreys K, Idris F, Jakubowska A, Jung A, Kapoor PM, Kerin MJ, Khusnutdinova E, Kim SW, Ko YD, Kosma VM, Kristensen VN, Kyriacou K, Lakeman IMM, Lee JW, Lee MH, Li J, Lindblom A, Lo WY, Loizidou MA, Lophatananon A, Lubinski J, MacInnis RJ, Madsen MJ, Mannermaa A, Manoochehri M, Manoukian S, Margolin S, Martinez ME, Maurer T, Mavroudis D, McLean C, Meindl A, Mensenkamp AR, Michailidou K, Miller N, Mohd Taib NA, Muir K, Mulligan AM, Nevanlinna H, Newman WG, Nordestgaard BG, Ng PS, Oosterwijk JC, Park SK, Park-Simon TW, Perez JIA, Peterlongo P, Porteous DJ, Prajzendanc K, Prokofyeva D, Radice P, Rashid MU, Rhenius V, Rookus MA, Rudiger T, Saloustros E, Sawyer EJ, Schmutzler RK, Schneeweiss A, Schurmann P, Shah M, Sohn C, Southey MC, Surowy H, Suvanto M, Thanasitthichai S, Tomlinson I, Torres D, Truong T, Tzardi M, Valova Y, van Asperen CJ, Van Dam RM, van den Ouweland AMW, van der Kolk LE, van Veen EM, Wendt C, Williams JA, Yang XR, Yoon SY, Zamora MP, Evans DG, de la Hoya M, Simard J, Antoniou AC, Borg A, Andrulis IL, Chang-Claude J, Garcia-Closas M, Chenevix-Trench G, Milne RL, Pharoah PDP, Schmidt MK, Spurdle AB, Vreeswijk MPG, Benitez J, Dunning AM, Kvist A, Teo SH, Devilee P, Easton DF. Breast cancer risk genes - association analysis in more than 113,000 women. N Engl J Med. 2021;384(5):428–39.
88. Jain S, White M, Radivojac P. *Estimating the class prior and posterior from noisy positives and unlabeled data*. Advances in Neural Information Processing Systems, 2016; pp. 2693–2701.
89. Jain S, White M, Trosset MW, Radivojac P. Nonparametric semi-supervised learning of class proportions. arXiv:1601.01944. 2016.
90. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat Sci. 1986;1(1):54–77.
91. Jain S. CAGI flagship software. 2022.https://doi.org/10.5281/zenodo.8436229.
92. Adhikari AN. Gene-specific features enhance interpretation of mutational impact on acid alpha-glucosidase enzyme activity. Hum Mutat. 2019;40(9):1507–18.
93. Kraus JP, Janosik M, Kozich V, Mandell R, Shih V, Sperandeo MP, Sebastio G, de Franchis R, Andria G, Kluijtmans LA, Blom H, Boers GH, Gordon RB, Kamoun P, Tsai MY, Kruger WD, Koch HG, Ohura T, Gaustadnes M. Cystathionine beta-synthase mutations in homocystinuria. Hum Mutat. 1999;13(5):362–75.
94. Dimster-Denk D, Tripp KW, Marini NJ, Marqusee S, Rine J. Mono and dual cofactor dependence of human cystathionine beta-synthase enzyme variants in vivo and in vitro. G3. 2013;3(10):1619–28.
95. Geiss-Friedlander R, Melchior F. Concepts in sumoylation: a decade on. Nat Rev Mol Cell Biol. 2007;8(12):947–56.
96. Sun S, Yang F, Tan G, Costanzo M, Oughtred R, Hirschman J, Theesfeld CL, Bansal P, Sahni N, Yi S, Yu A, Tyagi T, Tie C, Hill DE, Vidal M, Andrews BJ, Boone C, Dolinski K, Roth FP. An extended set of yeast-based functional assays accurately identifies human disease mutations. Genome Res. 2016;26(5):670–80.
97. Schulz TJ, Thierbach R, Voigt A, Drewes G, Mietzner B, Steinberg P, Pfeiffer AF, Ristow M. Induction of oxidative metabolism by mitochondrial frataxin inhibits cancer growth: Otto Warburg revisited. J Biol Chem. 2006;281(2):977–81.

98. Guccini I, Serio D, Condo I, Rufini A, Tomassini B, Mangiola A, Maira G, Anile C, Fina D, Pallone F, Mongiardi MP, Levi A, Ventura N, Testi R, Malisan F. Frataxin participates to the hypoxia-induced response in tumors. Cell Death Dis. 2011;2: e123.
99. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009;25(21):2744–50.
100. Goldgar DE, Easton DF, Byrnes GB, Spurdle AB, Iversen ES, Greenblatt MS, IARC Unclassified Genetic Variants Working Group. Genetic evidence and integration of various data sources for classifying uncertain variants into a single model. Hum Mutat. 2008;29(11):1265–72.
101. Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FB, Hoogerbrugge N, Spurdle AB, Tavtigian SV, IARC Unclassified Genetic Variants Working Group. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum Mutat. 2008;29(11):1282–91.
102. Parsons MT, Tudini E, Li H, Hahnen E, Wappenschmidt B, Feliubadalo L, Aalfs CM, Agata S, Aittomaki K, Alducci E, Alonso-Cerezo MC, Arnold N, Auber B, Austin R, Azzollini J, Balmana J, Barbieri E, Bartram CR, Blanco A, Blumcke B, Bonache S, Bonanni B, Borg A, Bortesi B, Brunet J, Bruzzone C, Bucksch K, Cagnoli G, Caldes T, Caliebe A, Caligo MA, Calvello M, Capone GL, Caputo SM, Carnevali I, Carrasco E, Caux-Moncoutier V, Cavalli P, Cini G, Clarke EM, Concolino P, Cops EJ, Cortesi L, Couch FJ, Darder E, de la Hoya M, Dean M, Debatin I, Del Valle J, Delnatte C, Derive N, Diez O, Ditsch N, Domchek SM, Dutrannoy V, Eccles DM, Ehrencrona H, Enders U, Evans DG, Farra C, Faust U, Felbor U, Feroce I, Fine M, Foulkes WD, Galvao HCR, Gambino G, Gehrig A, Gensini F, Gerdes AM, Germani A, Giesecke J, Gismondi V, Gomez C, Gomez Garcia EB, Gonzalez S, Grau E, Grill S, Gross E, Guerrieri-Gonzaga A, Guillaud-Bataille M, Gutierrez-Enriquez S, Haaf T, Hackmann K, Hansen TVO, Harris M, Hauke J, Heinrich T, Hellebrand H, Herold KN, Honisch E, Horvath J, Houdayer C, Hubbel V, Iglesias S, Izquierdo A, James PA, Janssen LAM, Jeschke U, Kaulfuss S, Keupp K, Kiechle M, Kolbl A, Krieger S, Kruse TA, Kvist A, Lalloo F, Larsen M, Lattimore VL, Lautrup C, Ledig S, Leinert E, Lewis AL, Lim J, Loeffler M, Lopez-Fernandez A, Lucci-Cordisco E, Maass N, Manoukian S, Marabelli M, Matricardi L, Meindl A, Michelli RD, Moghadasi S, Moles-Fernandez A, Montagna M, Montalban G, Monteiro AN, Montes E, Mori L, Moserle L, Muller CR, Mundhenke C, Naldi N, Nathanson KL, Navarro M, Nevanlinna H, Nichols CB, Niederacher D, Nielsen HR, Ong KR, Pachter N, Palmero EI, Papi L, Pedersen IS, Peissel B, Perez-Segura P, Pfeifer K, Pineda M, Pohl-Rescigno E, Poplawski NK, Porfirio B, Quante AS, Ramser J, Reis RM, Revillion F, Rhiem K, Riboli B, Ritter J, Rivera D, Rofes P, Rump A, Salinas M, Sanchez de Abajo AM, Schmidt G, Schoenwiese U, Seggewiss J, Solanes A, Steinemann D, Stiller M, Stoppa-Lyonnet D, Sullivan KJ, Susman R, Sutter C, Tavtigian SV, Teo SH, Teule A, Thomassen M, Tibiletti MG, Tischkowitz M, Tognazzo S, Toland AE, Tornero E, Torngren T, Torres-Esquius S, Toss A, Trainer AH, Tucker KM, van Asperen CJ, van Mackelenbergh MT, Varesco L, Vargas-Parra G, Varon R, Vega A, Velasco A, Vesper AS, Viel A, Vreeswijk MPG, Wagner SA, Waha A, Walker LC, Walters RJ, Wang-Gohrke S, Weber BHF, Weichert W, Wieland K, Wiesmuller L, Witzel I, Wockel A, Woodward ER, Zachariae S, Zampiga V, Zeder-Goss C, Investigators KC, Lazaro C, De Nicolo A, Radice P, Engel C, Schmutzler RK, Goldgar DE, Spurdle AB. Large scale multifactorial likelihood quantitative analysis of BRCA1 and BRCA2 variants: an ENIGMA resource to support clinical variant classification. Hum Mutat. 2019;40(9):1557–78.
103. Lai C, Zimmer AD, O'Connor R, Kim S, Chan R, van den Akker J, Zhou AY, Topper S, Mishne G. LEAP: using machine learning to support variant classification in a clinical setting. Hum Mutat. 2020;41(6):1079–90.
104. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat. 2011;32(8):894–9.
105. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat. 2013;34(9):E2393–2402.
106. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v30: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. Hum Mutat. 2016;37(3):235–41.
107. Kircher M, Xiong C, Martin B, Schubach M, Inoue F, Bell RJA, Costello JF, Shendure J, Ahituv N. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. Nat Commun. 2019;10(1):3583.
108. Halme L, Paavola-Sakki P, Turunen U, Lappalainen M, Farkkila M, Kontula K. Family and twin studies in inflammatory bowel disease. World J Gastroenterol. 2006;12(23):3668–72.
109. Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, Anderson CA, Bis JC, Bumpstead S, Ellinghaus D, Festen EM, Georges M, Green T, Haritunians T, Jostins L, Latiano A, Mathew CG, Montgomery GW, Prescott NJ, Raychaudhuri S, Rotter JI, Schumm P, Sharma Y, Simms LA, Taylor KD, Whiteman D, Wijmenga C, Baldassano RN, Barclay M, Bayless TM, Brand S, Buning C, Cohen A, Colombel JF, Cottone M, Stronati L, Denson T, De Vos M, D'Inca R, Dubinsky M, Edwards C, Florin T, Franchimont D, Gearry R, Glas J, Van Gossum A, Guthery SL, Halfvarson J, Verspaget HW, Hugot JP, Karban A, Laukens D, Lawrance I, Lemann M, Levine A, Libioulle C, Louis E, Mowat C, Newman W, Panes J, Phillips A, Proctor DD, Regueiro M, Russell R, Rutgeerts P, Sanderson J, Sans M, Seibold F, Steinhart AH, Stokkers PC, Torkvist L, Kullak-Ublick G, Wilson D, Walters T, Targan SR, Brant SR, Rioux JD, D'Amato M, Weersma RK, Kugathasan S, Griffiths AM, Mansfield JC, Vermeire S, Duerr RH, Silverberg MS, Satsangi J, Schreiber S, Cho JH, Annese V, Hakonarson H, Daly MJ, Parkes M. Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. Nat Genet. 2010;42(12):1118–25.
110. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Schumm LP, Sharma Y, Anderson CA, Essers J, Mitrovic M, Ning K, Cleynen I, Theatre E, Spain SL, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491(7422):119–24.
111. Uhlig HH, Schwerd T, Koletzko S, Shah N, Kammermeier J, Elkadri A, Ouahed J, Wilson DC, Travis SP, Turner D, Klein C, Snapper SB, Muise AM, Group CIIS, Neopics. The diagnostic approach to monogenic very early onset inflammatory bowel disease. Gastroenterology. 2014;147(5):990–1007 e1003.
112. Ellinghaus D, Zhang H, Zeissig S, Lipinski S, Till A, Jiang T, Stade B, Bromberg Y, Ellinghaus E, Keller A, Rivas MA, Skieceviciene J, Doncheva NT, Liu X, Liu Q, Jiang F, Forster M, Mayr G, Albrecht M, Hasler R, Boehm BO, Goodall J, Berzuini CR, Lee J, Andersen V, Vogel U, Kupcinskas L, Kayser M, Krawczak M, Nikolaus S, Weersma RK, Ponsioen CY, Sans M, Wijmenga C, Strachan DP, McArdle WL, Vermeire S, Rutgeerts P, Sanderson JD, Mathew CG, Vatn MH, Wang

J, Nothen MM, Duerr RH, Buning C, Brand S, Glas J, Winkelmann J, Illig T, Latiano A, Annese V, Halfvarson J, D'Amato M, Daly MJ, Nothnagel M, Karlsen TH, Subramani S, Rosenstiel P, Schreiber S, Parkes M, Franke A. Association between variants of PRDM1 and NDP52 and crohn's disease, based on exome sequencing and functional studies. Gastroenterology. 2013;145(2):339–47.
113. Voskanian A, Katsonis P, Lichtarge O, Pejaver V, Radivojac P, Mooney SD, Capriotti E, Bromberg Y, Wang Y, Miller M, Martelli PL, Savojardo C, Babbi G, Casadio R, Cao Y, Sun Y, Shen Y, Garg A, Pal D, Yu Y, Huff CD, Tavtigian SV, Young E, Neuhausen SL, Ziv E, Pal LR, Andreoletti G, Brenner SE, Kann MG. Assessing the performance of in silico methods for predicting the pathogenicity of variants in the gene CHEK2, among hispanic females with breast cancer. Hum Mutat. 2019;40(9):1612–22.
114. Zakai NA, McClure LA. Racial differences in venous thromboembolism. J Thromb Haemost. 2011;9(10):1877–82.
115. Feero WG. Genetic thrombophilia. Prim Care. 2004;31(3):685–709.
116. McInnes G, Daneshjou R, Katsonis P, Lichtarge O, Srinivasan R, Rana S, Radivojac P, Mooney SD, Pagel KA, Stamboulian M, Jiang Y, Capriotti E, Wang Y, Bromberg Y, Bovo S, Savojardo C, Martelli PL, Casadio R, Pal LR, Moult J, Brenner SE, Altman R. Predicting venous thromboembolism risk from exomes in the critical assessment of genome interpretation (CAGI) challenges. Hum Mutat. 2019;40(9):1314–20.
117. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK biobank. Nat Genet. 2018;50(11):1593–9.
118. Wray NR, Yang J, Goddard ME, Visscher PM. The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet. 2010;6(2): e1000864.
119. Soria JM, Morange PE, Vila J, Souto JC, Moyano M, Tregouet DA, Mateo J, Saut N, Salas E, Elosua R. Multilocus genetic risk scores for venous thromboembolism risk assessment. J Am Heart Assoc. 2014;3(5): e001060.
120. Fairfield H, Gilbert GJ, Barter M, Corrigan RR, Curtain M, Ding Y, D'Ascenzo M, Gerhardt DJ, He C, Huang W, Richmond T, Rowe L, Probst FJ, Bergstrom DE, Murray SA, Bult C, Richardson J, Kile BT, Gut I, Hager J, Sigurdsson S, Mauceli E, Di Palma F, Lindblad-Toh K, Cunningham ML, Cox TC, Justice MJ, Spector MS, Lowe SW, Albert T, Donahue LR, Jeddeloh J, Shendure J, Reinholdt LG. Mutation discovery in mice by whole exome sequencing. Genome Biol. 2011;12(9):R86.
121. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat. 2009;30(8):1237–44.
122. Deutschbauer A, Price MN, Wetmore KM, Shao W, Baumohl JK, Xu Z, Nguyen M, Tamse R, Davis RW, Arkin AP. Evidence-based annotation of gene function in shewanella oneidensis MR-1 using genome-wide fitness profiling across 121 conditions. PLoS Genet. 2011;7(11): e1002385.
123. Lai R, Ingham RJ. The pathobiology of the oncogenic tyrosine kinase NPM-ALK: a brief update. Ther Adv Hematol. 2013;4(2):119–31.
124. Lu L, Ghose AK, Quail MR, Albom MS, Durkin JT, Holskin BP, Angeles TS, Meyer SL, Ruggeri BA, Cheng M. ALK mutants in the kinase domain exhibit altered kinase activity and differential sensitivity to small molecule ALK inhibitors. Biochemistry. 2009;48(16):3600–9.
125. Larsen CC, Karaviti LP, Seghers V, Weiss RE, Refetoff S, Dumitrescu AM. A new family with an activating mutation (G431S) in the TSH receptor gene: a phenotype discussion and review of the literature. Int J Pediatr Endocrinol. 2014;2014(1):23.
126. Robinson PN, Mundlos S. The human phenotype ontology. Clin Genet. 2010;77(6):525–34.
127. Pal LR, Kundu K, Yin Y, Moult J. CAGI4 SickKids clinical genomes challenge: a pipeline for identifying pathogenic variants. Hum Mutat. 2017;38(9):1169–81.
128. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The ensembl variant effect predictor. Genome Biol. 2016;17(1):122.
129. Budnitz DS, Lovegrove MC, Shehab N, Richards CL. Emergency hospitalizations for adverse drug events in older americans. N Engl J Med. 2011;365(21):2002–12.
130. International Warfarin Pharmacogenetics Consortium, Klein TE, Altman RB, Eriksson N, Gage BF, Kimmel SE, Lee MT, Limdi NA, Page D, Roden DM, Wagner MJ, Caldwell MD, Johnson JA. Estimation of the warfarin dose with clinical and pharmacogenetic data. N Engl J Med. 2009;360(8):753–64.
131. Daneshjou R, Klein TE, Altman RB. Genotype-guided dosing of vitamin K antagonists. N Engl J Med. 2014;370(18):1762–3.
132. Sundaram L, Bhat RR, Viswanath V, Li X. DeepBipolar: identifying genomic mutations for bipolar disorder via deep learning. Hum Mutat. 2017;38(9):1217–24.
133. Wang MH, Chang B, Sun R, Hu I, Xia X, Wu WKK, Chong KC, Zee BC. Stratified polygenic risk prediction model with application to CAGI bipolar disorder sequencing data. Hum Mutat. 2017;38(9):1235–9.
134. Niroula A, Vihinen M. PON-P and PON-P2 predictor performance in CAGI challenges: lessons learned. Hum Mutat. 2017;38(9):1085–91.
135. Katsonis P, Lichtarge O. CAGI5: objective performance assessments of predictions based on the evolutionary action equation. Hum Mutat. 2019;40(9):1436–54.
136. Garg A, Pal D. Exploring the use of molecular dynamics in assessing protein variants for phenotypic alterations. Hum Mutat. 2019;40(9):1424–35.
137. Kasak L, Bakolitsa C, Hu Z, Yu C, Rine J, Dimster-Denk DF, Pandey G, De Baets G, Bromberg Y, Cao C, Capriotti E, Casadio R, Van Durme J, Giollo M, Karchin R, Katsonis P, Leonardi E, Lichtarge O, Martelli PL, Masica D, Mooney SD, Olatubosun A, Radivojac P, Rousseau F, Pal LR, Savojardo C, Schymkowitz J, Thusberg J, Tosatto SCE, Vihinen M, Valiaho J, Repo S, Moult J, Brenner SE, Friedberg I. Assessing computational predictions of the phenotypic effect of cystathionine-beta-synthase variants. Hum Mutat. 2019;40(9):1530–45.
138. Katsonis P, Lichtarge O. Objective assessment of the evolutionary action equation for the fitness effect of missense mutations across CAGI-blinded contests. Hum Mutat. 2017;38(9):1072–84.
139. Savojardo C, Babbi G, Bovo S, Capriotti E, Martelli PL, Casadio R. Are machine learning based methods suited to address complex biological problems? lessons from CAGI-5 challenges. Hum Mutat. 2019;40(9):1455–62.
140. Wang Y, Bromberg Y. Identifying mutation-driven changes in gene functionality that lead to venous thromboembolism. Hum Mutat. 2019;40(9):1321–9.

141. Giollo M, Jones DT, Carraro M, Leonardi E, Ferrari C, Tosatto SCE. Crohn disease risk prediction-best practices and pitfalls with exome data. Hum Mutat. 2017;38(9):1193–200.
142. Pal LR, Kundu K, Yin Y, Moult J. CAGI4 crohn's exome challenge: marker SNP versus exome variant models for assigning risk of crohn disease. Hum Mutat. 2017;38(9):1225–34.
143. Cao Y, Sun Y, Karimi M, Chen H, Moronfoye O, Shen Y. Predicting pathogenicity of missense variants with weakly supervised regression. Hum Mutat. 2019;40(9):1579–92.
144. Padilla N, Moles-Fernandez A, Riera C, Montalban G, Ozkan S, Ootes L, Bonache S, Diez O, Gutierrez-Enriquez S, de la Cruz X. BRCA1- and BRCA2-specific in silico tools for variant interpretation in the CAGI 5 ENIGMA challenge. Hum Mutat. 2019;40(9):1593–611.
145. Zeng H, Edwards MD, Guo Y, Gifford DK. Accurate eQTL prioritization with an ensemble-based framework. Hum Mutat. 2017;38(9):1259–65.
146. Beer MA. Predicting enhancer activity and variant impact using gkm-SVM. Hum Mutat. 2017;38(9):1251–8.
147. Strokach A, Corbi-Verge C, Kim PM. Predicting changes in protein stability caused by mutation using sequence- and structure-based methods in a CAGI5 blind challenge. Hum Mutat. 2019;40(9):1414–23.
148. Petrosino M, Pasquo A, Novak L, Toto A, Gianni S, Mantuano E, Veneziano L, Minicozzi V, Pastore A, Puglisi R, Caprioti E, Chiaraluce R, Consalvi V. Characterization of human frataxin missense variants in cancer tissues. Hum Mutat. 2019;40(9):1400–13.
149. Kundu K, Pal LR, Yin Y, Moult J. Determination of disease phenotypes and pathogenic variants from exome sequence data in the CAGI 4 gene panel challenge. Hum Mutat. 2017;38(9):1201–16.
150. Aspromonte MC, Bellini M, Gasparini A, Carraro M, Bettella E, Polli R, Cesca F, Bigoni S, Boni S, Carlet O, Negrin S, Mammi I, Milani D, Peron A, Sartori S, Toldo I, Soli F, Turolla L, Stanzial F, Benedicenti F, Marino-Buslje C, Tosatto SCE, Murgia A, Leonardi E. Characterization of intellectual disability and autism comorbidity through gene panel sequencing. Hum Mutat. 2019;40(9):1346–63.
151. Chen J. A fully-automated event-based variant prioritizing solution to the CAGI5 intellectual disability gene panel challenge. Hum Mutat. 2019;40(9):1364–72.
152. Rhine CL, Neil C, Glidden DT, Cygan KJ, Fredericks AM, Wang J, Walton NA, Fairbrother WG. Future directions for high-throughput splicing assays in precision medicine. Hum Mutat. 2019;40(9):1225–34.
153. Cheng J, Celik MH, Nguyen TYD, Avsec Z, Gagneur J. CAGI 5 splicing challenge: improved exon skipping and intron retention predictions with MMSplice. Hum Mutat. 2019;40(9):1243–51.
154. Naito T. Predicting the impact of single nucleotide variants on splicing via sequence-based deep neural networks and genomic features. Hum Mutat. 2019;40(9):1261–9.
155. Yin Y, Kundu K, Pal LR, Moult J. Ensemble variant interpretation methods to predict enzyme activity and assign pathogenicity in the CAGI4 NAGLU (human N-acetyl-glucosaminidase) and UBE2I (human SUMO-ligase) challenges. Hum Mutat. 2017;38(9):1109–22.
156. Monzon AM, Carraro M, Chiricosta L, Reggiani F, Han J, Ozturk K, Wang Y, Miller M, Bromberg Y, Capriotti E, Savojardo C, Babbi G, Martelli PL, Casadio R, Katsonis P, Lichtarge O, Carter H, Kousi M, Katsanis N, Andreoletti G, Moult J, Brenner SE, Ferrari C, Leonardi E, Tosatto SCE. Performance of computational methods for the evaluation of pericentriolar material 1 missense variants in CAGI-5. Hum Mutat. 2019;40(9):1474–85.
157. Miller M, Wang Y, Bromberg Y. What went wrong with variant effect predictor performance for the PCM1 challenge. Hum Mutat. 2019;40(9):1486–94.
158. Tang Q, Fenton AW. Whole-protein alanine-scanning mutagenesis of allostery: a large percentage of a protein can contribute to mechanism. Hum Mutat. 2017;38(9):1132–43.
159. Tang Q, Alontaga AY, Holyoak T, Fenton AW. Exploring the limits of the usefulness of mutagenesis in studies of allosteric mechanisms. Hum Mutat. 2017;38(9):1144–54.
160. Xu Q, Tang Q, Katsonis P, Lichtarge O, Jones D, Bovo S, Babbi G, Martelli PL, Casadio R, Lee GR, Seok C, Fenton AW, Dunbrack RL Jr. Benchmarking predictions of allostery in liver pyruvate kinase in CAGI4. Hum Mutat. 2017;38(9):1123–31.
161. Dong S, Boyle AP. Predicting functional variants in enhancer and promoter elements using RegulomeDB. Hum Mutat. 2019;40(9):1292–8.
162. Kreimer A, Yan Z, Ahituv N, Yosef N. Meta-analysis of massively parallel reporter assays enables prediction of regulatory function across cell types. Hum Mutat. 2019;40(9):1299–313.
163. Pal LR, Kundu K, Yin Y, Moult J. Matching whole genomes to rare genetic disorders: identification of potential causative variants using phenotype-weighted knowledge in the CAGI SickKids5 clinical genomes challenge. Hum Mutat. 2020;41(2):347–62.
164. Gotea V, Margolin G, Elnitski L. CAGI experiments: modeling sequence variant impact on gene splicing using predictions from computational tools. Hum Mutat. 2019;40(9):1252–60.
165. Wang R, Wang Y, Hu Z. Using secondary structure to predict the effects of genetic variants on alternative splicing. Hum Mutat. 2019;40(9):1270–9.
166. Chen K, Lu Y, Zhao H, Yang Y. Predicting the change of exon splicing caused by genetic variant using support vector regression. Hum Mutat. 2019;40(9):1235–42.

Publisher's Note