Genome interpretation using in silico predictors of variant impact

Panagiotis Katonis1 · Kevin Wilhelm2 · Amanda Williams1 · Olivier Lichtarge1,3,4,5

Received: 2 August 2021 / Accepted: 17 April 2022 / Published online: 30 April 2022 © The Author(s) 2022

Abstract
Estimating the effects of variants found in disease driver genes opens the door to personalized therapeutic opportunities. Clinical associations and laboratory experiments can only characterize a tiny fraction of all the available variants, leaving the majority as variants of unknown significance (VUS). In silico methods bridge this gap by providing instant estimates on a large scale, most often based on the numerous genetic differences between species. Despite concerns that these methods may lack reliability in individual subjects, their numerous practical applications over cohorts suggest they are already helpful and have a role to play in genome interpretation when used at the proper scale and context. In this review, we aim to gain insights into the training and validation of these variant effect predicting methods and illustrate representative types of experimental and clinical applications. Objective performance assessments using various datasets that are not yet published indicate the strengths and limitations of each method. These show that cautious use of in silico variant impact predictors is essential for addressing genome interpretation challenges.

Introduction
The need of estimating variant impact
The drastic reduction in the cost of genome sequencing over the last decade (DNA Sequencing Costs: Data 2020) led to a proliferation of large-scale next-generation sequencing (NGS) datasets (Pereira et al. 2020). The NCBI dbGaP database (Mailman et al. 2007) hosts a vast collection of genotype–phenotype data in a common repository, but many other data sets of broad interest reside elsewhere and not all have simple access (Gutierrez-Sacristan et al. 2021). Currently, dbGaP contains more than 500 NGS case–control studies, including studies of such diseases as Alzheimer’s (Beecham et al. 2017), Parkinson’s (Rosenthal et al. 2016), autism spectrum disorder (ASD) (Fischbach and Lord 2010), and others (Taliun et al. 2021). Additional case–control studies on the same or similar traits are available through other independent initiatives (Marek et al. 2018; Petersen et al. 2010). Other sources, called biobanks, contain sequencing data and clinical diagnostics from individuals representing a population, without a focus on any particular disease. Examples include the UK Biobank, which currently holds genetic data from about 500,000 individuals (Backman et al. 2021; Bycroft et al. 2018), as well as the All of Us Research Program (All of Us Research Program et al. 2019) among others (Wichmann et al. 2011). Some datasets focus on ethnic differences, such as The 1000 Genomes Project (The 1000 Genomes Project Consortium 2015) and various population-specific cohorts (Genome of the Netherlands Consortium 2014; Jeon et al. 2020; Kim et al. 2018; Nagaasaki et al. 2015). Each of these sources contains rich data on human genetic variations over which we can address questions of disease etiology for mechanistic understanding, risk for prevention and early screening, personalized therapy for precision medicine, epistatic effects for complex epidemiology, pharmacogenomics for patient stratification, and ethnic
diversity for health equity. Accurate metrics for the impact of individual variants is critical to guide answers to these questions.

There are many types of genetic variants, broadly grouped by the region in which the variant occurs and the number of nucleotides affected. Variants can occur in protein coding regions or non-coding regions of the genome. Although the protein coding region represents only 1.2% of the human genome (Encode Project Consortium 2012), past variant interpretation efforts focused on these variants due to their effects in protein synthesis. However, a large proportion of the non-coding genome is functional and harbors variants that drive diseases by influencing regulatory regions controlling gene expression and untranslated regions affecting mRNA translation (French and Edwards 2020). Variants can also encompass single or multiple nucleotides. Substitutions that add or remove nucleotides are called insertions or deletions, respectively, and they are rare compared to single nucleotide variants (SNVs) that account for about 90% of the variants (1000 Genomes Project Consortium et al. 2010). Protein coding insertion and deletion (indel) variants may be pathogenic when they shift the reading frame of mRNA or map to functionally important sites (Lin et al. 2017; Mullaney et al. 2010). Protein coding SNVs may truncate the protein (stop gain and start loss), cause no change to the protein sequence (synonymous), or alter one amino acid (non-synonymous/missense). Stop gain and start loss variants exert profound effects on protein function, resulting in strong selection against them (Bartha et al. 2015). Synonymous variants are assumed to be benign, although they can cause various pre-translational changes (Zeng and Bromberg 2019) and affect codon usage bias (Plotkin and Kudla 2011). The impact of non-synonymous variants is challenging to predict, since they may affect a number of protein characteristics, such as folding (Wang and Moult 2001), protein interactions (Teng et al. 2009), dynamics (Uversky et al. 2008), post-translational modifications (Yang et al. 2019), solubility (Monplaisir et al. 1986), and others (Steffl et al. 2013), with approximately 30% of variants having a strong impact (Chasman and Adams 2001). Purifying selection accounts for all the above effects and reduces diversity within species (Cvijovic et al. 2018), allowing only beneficial and nearly neutral variants to spread and become fixed (Fu and Akey 2013; Patwa and Wahl 2008). Assuming the variant effects differ little between homologous proteins, the genetic differences between the species offer valuable information for estimating the overall effect of a variant (Ng and Henikoff 2001). The 3D structures of proteins can also provide additional complementary insights (Orengo et al. 1999; Ramensky et al. 2002). Consequently, homology and structure information have been the two main types of input for estimating the effects of coding variants on protein function.

### Available methods for predicting variant effects

Many computational methods estimate variant effects (Hu et al. 2019), but their aims may differ. Some methods focus on specific aspects of protein function, such as folding stability of the mutated protein (Capriotti et al. 2005, 2008; Cheng et al. 2006; Dehouck et al. 2011; Fariselli et al. 2015; Guerois et al. 2002; Parthiban et al. 2006; Pires et al. 2014; Quan et al. 2016; Worth et al. 2011; Zhou and Zhou 2002) or a combination of folding stability and binding affinity (Berliner et al. 2014). Typically, these folding prediction methods estimate the free energy change of folding ($\Delta\Delta G$) due to mutation from 3D structures in addition to scores derived from different force-fields or evolutionary information. Encouragingly, about three-fourths of variants that cause Mendelian disorders affect protein stability (Wang and Moult 2001; Yue et al. 2014), suggesting that folding stability prediction methods can prioritize candidate disease drivers (Bocchini et al. 2016; Pey et al. 2007; Siekierska et al. 2012) for chaperone treatment (Chaudhuri and Paul 2006). However, these methods are partially limited by the availability of protein structure data. Although the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) currently contains more than 50,000 human protein structures, many are redundant and 30% of the human proteins have no PDB structure that corresponds to their sequence or any homologous sequence with as low as 30% sequence identity (Somody et al. 2017). Alternatively, stability prediction methods may take advantage of the recent development of high-quality protein structure predictors, such as AlphaFold (Jumper et al. 2021) and RosettaFold (Baek et al. 2021), which broaden the available number of protein structures. However, evolution-based stability prediction methods have competitive performance, despite the fact they do not use any structure information (Fariselli et al. 2015; Montanucci et al. 2019).

Toward broader genome interpretation of variant effect, other prediction methods rely on protein sequence homology primarily, supported when available by functional annotations and structural information. Table 1 lists well-established representatives of these methods, grouped according to the type of input data they use and whether the input includes scores from available predicting methods (Ensemble). Homology-based methods hypothesize that the frequent substitutions across clades are benign, under the implied assumption that homologous proteins share identical functions. Inversely, substitutions that do not occur across phylogenetic branches indicate negative selection and possible pathogenicity (Ng and Henikoff 2001). To overcome limitations that stem from this hypothesis, homology-based prediction methods employ...
Table 1  Representative predictors of variant fitness effects

| Type          | Method name          | Input features                                      | Training dataset | Validating datasets                                                                 | Paper                              | Citations |
|---------------|----------------------|-----------------------------------------------------|------------------|--------------------------------------------------------------------------------------|------------------------------------|-----------|
| Homology      | Align-GVGD           | Multiple sequence alignment                         | –                | Exp: p53  
Clin: IARC p53  
Tavtigian et al. (2006)  | 655                                                                 |           |
| Homology      | DeMaSk               | Multiple sequence alignment                         | Amino acid substitutions from 18 deep mutational scanning studies | Leave one out  
Munro and Singh (2020) | 6                                                                 |           |
| Homology      | EFIN                 | Sequence conservation features                      | UniProt, HumDiv  | SwissProt  
Zeng et al. (2014) | 25                                                                 |           |
| Homology      | Evolutionary Action  | Multiple sequence alignment                         | –                | Exp: lacI, lysozyme, HIV protease, RecA, p53  
Clin: IARC p53, UniProt, CFTR, GAA  
Katsonis and Lichtarge (2014) | 105                                                                 |           |
| Homology      | FATHMM               | Multiple sequence alignment                         | –                | Exp: VariBench  
Clin: HGMD, UniProt, SwissVar  
Shihab et al. (2013) | 932                                                                 |           |
| Homology      | likelihood ratio test (LRT) | Multiple sequence alignment | – | SNPs of three individuals exomes  
OMIM database  
Chun and Fay (2009) | 920                                                                 |           |
| Homology      | MAPP                 | Multiple sequence alignment Phylogenetic Tree       | –                | Exp: lacI, lysozyme, HIV protease, HIV RT  
Clin: pyruvate kinase, G6PD, HBB, IARC p53  
Stone and Sidow (2005) | 371                                                                 |           |
| Homology      | MutationAssessor     | Multiple sequence alignment                         | –                | UniProt, IARC p53, COSMIC  
Reva et al. (2011) | 1687                                                                |           |
| Homology      | PANTHER              | Multiple sequence alignment                         | –                | Clin: HGMD, dbSNP, Pop: SNPs sampled from healthy individuals  
Thomas et al. (2003) | 2924                                                                |           |
| Homology      | PhD-SNP              | Multiple sequence alignment Local sequence environment | Swiss-Prot, HumVar | Newer HumVar variants  
Capriotti et al. (2006) | 732                                                                 |           |
| Homology      | PrimateAI            | Sequence alignments and sequence-based predictions of structure features | Path: SNVs absent from ExAC  
Ben: common SNVs and primate changes  
Sundaram et al. (2018) | Withheld variants from training | 183                                                                 |           |
| Homology      | PROVEAN              | Multiple sequence alignment                         | –                | Exp: LacI, TP53, ABCA1  
Clin: UniProtKB/Swiss-Prot | Choi et al. (2012) | 2445       |
| Homology      | SIFT/SIFT4G          | Multiple sequence alignment                         | –                | Exp: lacI, lysozyme, HIV protease  
Clin: HumDiv, HumVar  
Ng and Henikoff (2003) | 5242                                                                |           |
|               |                      |                                                     |                  |                                                                       | Ng and Henikoff (2001) | 2543       |
|               |                      |                                                     |                  |                                                                       | Vaser et al. (2016) | 691        |
| Type               | Method name             | Input features                                                                 | Training dataset          | Validating datasets                        | Paper                                | Citations |
|-------------------|-------------------------|-------------------------------------------------------------------------------|----------------------------|--------------------------------------------|--------------------------------------|-----------|
| Multiple features | DeepSAV                 | 16 conservation, 28 structure and 1 dynamics score                            | ClinVar and UniProt       | ClinVar and UniProt                       | Pei et al. (2020)                     | 5         |
| Multiple features | ENTPRISE                | Alignment and structure related features                                      | multiple sources, including: HGMD, UniProt, SwissProt | Clin: COSMIC, TCGA, COBR Pop: VarSNP, 1000 Genomes | Zhou et al. (2016)                    | 22        |
| Multiple features | Envision                | homology, structure, amino acid properties, and more                          | large-scale experimental mutagenesis data | Withheld variants from training             | Gray et al. (2018)                    | 107       |
| Multiple features | FATHMM-MKL              | 10 feature groups (conservation and function annotations)                     | Path: HGMD Ben: 1000 Genomes | ClinVar, newer HGMD variants               | Shihab et al. (2015)                  | 451       |
| Multiple features | FATHMM-XF               | 31 feature groups (conservation, ENCODE, Epigenomics, and more)               | Path: HGMD Ben: 1000 Genomes | ClinVar                                    | Rogers et al. (2018)                 | 174       |
| Multiple features | MutationTaster/Mutation- | Conservation, sequence, and protein features                                 | Path: HGMD, OMM, and more Ben: 1000 Genomes, HapMap, and more | Withheld variants from training             | Schwarz et al. (2010)                 | 2675      |
| Taster2           |                         |                                                                               |                            |                                            | Schwarz et al. (2014)                 | 2778      |
| Multiple features | MutPred/MutPred2         | Structure, function annotation, and evolutionary features                    | Path: HGMD, SwissProt, Cancer drivers Ben: SwissProt | Withheld variants from training/ClinVar, SwissVar, experimental data | Li et al. (2009)                      | 773       |
| Multiple features | PolyPhen/PolyPhen-2     | Sequence and structure based features                                        | HGVbase / Path: UniProt (HumDiv, HumVar) Ben: Differences between orthologs | HumVar                                    | Ramensky et al. (2002)                | 2538      |
| Multiple features | PON-P2                  | Conservation, ontology functional, and structural                            | VariBench                  | Withheld variants from training            | Niroula et al. (2015)                | 161       |
| Multiple features | SNAP/SNAP2              | Conservation, structure prediction, and more                                 | Path: Protein Mutant Database Ben: Swiss-Prot | Exp: lacI, lysozyme, HIV protease, Melanocortin-4 | Bromberg and Rost (2007)              | 807       |
| Multiple features | SNPs&GO                 | 50 features of mutation, sequence profile, PANTHER output, and GO classification | Swiss-Prot                  | Disease classification from Goh et al. 2007 | Hecht et al. (2015)                   | 3.39      |
| Multiple features | SNPs3D                  | Structure related features/alignment related features                         | Path: HGMD Ben: Differences of homologous proteins | Clin: HGMD Pop: dbSNP                      | Yue et al. (2005)                     | 452       |
| Multiple features | VEST                    | 86 features of SNVBox                                                         | Path: HGMD Ben: ESP        | Pop: SwissProt, 1000 Genomes               | Yue and Moult (2006)                  | 283       |
| Multiple features | VIPUR                   | 20 protein sequence and structure-based features                             | UniProt                    | withholding variants from training         | Carter et al. (2013)                  | 342       |
| Type               | Method name  | Input features                                                                 | Training dataset                                                                 | Validating datasets                                                                        | Paper                                      | Citations |
|-------------------|--------------|--------------------------------------------------------------------------------|---------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------|-------------------------------------------|-----------|
| Functional genomic| fitCons      | DNase-seq data, RNA-seq data, ChIP-seq data                                    | –                                                                               | Transcription factor binding sites expression quantitative trait loci (eQTL) enhancers based on characteristic chromatin marks | Gulko et al. (2015)                        | 215       |
| Ensemble +        | CADD         | 63 features, including: predictors, conservation scores, and function annotation | Human-chimp changes simulated de novo variants                                  | Exp: ALDOB, ECR11, HBB Clin: HGMD, MML2, Clin-Var, IARC p53 Pop: ESP, GVS, 11 individuals | Kircher et al. (2014)                        | 4681      |
| Ensemble +        | CAPICE       | 63 features, including: predictors, conservation scores, and function annotation | Path: ClinVar, VKGL Ben: ClinVar, VKGL, ExAC                                      | ClinVar, VKGL and ExAC                                                                    | Li et al. (2020)                           | 9         |
| Ensemble +        | ClinPred     | 16 predictors and allele frequencies of gnomAD                                  | ClinVar                                                                        | Exp: BRCA1 Clin: mutagenetix database, DoCM,                                              | Alirezaie et al. (2018)                     | 79        |
| Ensemble          | Condel       | 5 predictors (Logre, MAPP, MutationAssessor, Polyphen2, and SIFT)              | HumVar, HumDiv, COSMIC, p53                                                    | –                                                                                         | Gonzalez-Perez and Lopez-Bigas (2011)     | 788       |
| Ensemble +        | DANN         | 949 features, including: predictors, conservation scores, and function annotation | Human-chimp changes simulated de novo variants                                  | Clin: ClinVar Pop: ESP                                                                    | Quang et al. (2015)                        | 689       |
| Ensemble +        | DEOGEN / DEOGEN2 | PROVEAN scores, alignment, Network, pathway, gene essentiality and more features | UniProt                                                                        | Clin: UniProt, p53, F8, BRCA1                                                            | Raimondi et al. (2016)                     | 26        |
| Ensemble +        | Eigen        | Predictors, conservation scores and allele frequencies of the 1000 Genomes     | ClinVar                                                                        | De novo variants in several studies                                                       | Ionita-Laza et al. (2016)                  | 408       |
| Ensemble          | InMeRF       | 28 predictors and 9 conservation scores                                        | Path: HGMD Ben: variants with MAF > 0.1%                                        | VariBench, PredictSNP, SwissVar                                                           | Takeda et al. (2020)                       | 3         |
| Ensemble +        | M-CAP        | 9 predictors, 7 conservation scores, and 298 alignment-based features          | Path: HGMD (AF < 1%) Ben: ExAC (AF < 1%)                                         | Withheld variants from training                                                           | Jagadeesh et al. (2016)                    | 531       |
| Ensemble          | MetaLR       | 15 predictors and 3 conservation scores                                        | UniProt                                                                        | VariBench, CHARGE database, and publications                                              | Dong et al. (2015)                         | 769       |
| Ensemble          | MetaSVM      | 15 predictors and 3 conservation scores                                        | UniProt                                                                        | Newer SwissVar variants                                                                    | Capriotti et al. (2013)                    | 176       |
| Ensemble +        | Meta-SNP     | 4 predictors: PANTHER, PhD-SNP, SIFT, and SNAP                                  | SwissVar                                                                        | Newer SwissVar variants                                                                    | Capriotti et al. (2013)                    | 176       |
| Ensemble +        | MISTIC       | 7 predictors, 8 conservation scores, MAF, and genetic and protein function     | Path: ClinVar Ben: gnomAD                                                      | Clin: new ClinVar Pop: SweGen, UK10K, and more                                            | Chennen et al. (2020)                      | 12        |
| Type     | Method name | Input features                                                                 | Training dataset                                      | Validating datasets                                                                 | Paper                                 | Citations |
|----------|-------------|--------------------------------------------------------------------------------|-------------------------------------------------------|-------------------------------------------------------------------------------------|---------------------------------------|-----------|
| Ensemble | MPC         | PolyPhen2 and other deleteriousness metrics                                     | Path: ClinVar, Ben: common ExAC variants               | 5620 neurodevelopmental disorder cases and 2078 controls                             | Samocha et al. (2017)                | 146       |
| Ensemble | MutScore    | 5 Predictors: SIFT, SIFT4G, LRT, PROVEAN, GERP + + RS and 9 conservation scores | ClinVar                                               |ClinVar                                                                              | Quinodoz et al. (2022)              | 0         |
| Ensemble | MVP         | 15 predictors, 6 conservation scores, structure, interactions, gene intolerance, | Path: HGMD, UniProt, ClinVar, Ben: UniProt, and more   | Path: VariBench, Cancer hotspots, Ben: DiscovEHR                                      | Qi et al. (2021)                     | 25        |
| Ensemble | PON-P       | 5 predictors: PhD-SNP, SIFT, PolyPhen-2, SNAP, I-Mutant                        | Path: PhenCode, Idbases, and more, Ben: dbSNP with AF > 0.1 |Protein Mutant Database                                                              | Olatubosun et al. (2012)            | 108       |
| Ensemble | PredictSNP/PredictSNP2 | 8 predictors / 6 predictors                                        | UniProt and training datasets of: SNPs&GO, MutPred, and PON-P / ClinVar, GWAS catalog, COSMIC, VarSNP |Protein Mutant Database, and experimental studies/Mendelian disease and cancer driver variants | Bendl et al. (2014)                  | 485       |
| Ensemble | REVEL       | 10 predictors and 8 conservation scores                                       | Path: HGMD, Ben: ESP, ARIC, 1000 Genomes              | ClinVar and SwissVar                                                                 | Ioannidis et al. (2016)             | 853       |
| Ensemble | Rhapsody    | 4 sequence, 1 structure, and 4 dynamic scores                                | HumVar, ExoVar, PredictSNP, VariBench, and SwissVar  | HumVar, ExoVar, PredictSNP, VariBench, and SwissVar                                 | Ponzoni et al. (2020)               | 32        |

Methods citations (Adzhubei et al. 2010; Alirezaie et al. 2018; Baugh et al. 2016; Bendel et al. 2014, 2016; Bromberg and Rost 2007; Calabrese et al. 2009; Capriotti et al. 2013, 2006; Carter et al. 2013; Chernen et al. 2020; Choi et al. 2012; Chun and Fay 2009; Dong et al. 2015; Gonzalez-Perez and Lopez-Bigas 2011; Gray et al. 2018; Gulko et al. 2015; Hecht et al. 2015; Ioannidis et al. 2016; Ionita-Laza et al. 2016; Jugadeesh et al. 2016; Katonis and Lichtarge 2014; Kircher et al. 2014; Li et al. 2009, 2020; Munro and Singh 2020; Ng and Henikoff 2001, 2003; Niroula et al. 2015; Olatubosun et al. 2012; Pei et al. 2020; Pejaver et al. 2020; Ponzoni et al. 2020; Qi et al. 2021; Quang et al. 2015; Quinodoz et al. 2022; Raimondi et al. 2016, 2017; Ramensky et al. 2002; Reva et al. 2011; Rogers et al. 2018; Samocha et al. 2017; Schwarz et al. 2010; Shihab et al. 2013, 2014, 2015; Stone and Sidow 2005; Sundaram et al. 2018; Takeda et al. 2020; Tavtigian et al. 2006; Thomas et al. 2003; Vaser et al. 2016; Yue et al. 2005, 2006; Yue and Moult 2006; Zeng et al. 2014; Zhou et al. 2016)

*Method inclusion criteria:
(i) be applicable on missense mutations
(ii) provide a single value for the impact of mutations
(iii) the impact should represent the overall effect on the protein function

*Path* pathogenic, *Ben* benign, *Exp* experimental associations, *Clin* clinical associations, *Pop* population data
Machine-learning approaches are typically used for such purposes. However, some of these methods have used pre-existing prediction method scores together with additional features in their training, so we labeled them as "Ensemble+". The method DANN (Quang et al. 2015) used the same feature set and training data as the method CADD (Kircher et al. 2014), but a different learning approach, and we subsequently labeled it "Ensemble+". Figure 1 reflects the popularity for a large collection of available methods, indicated by the citations of the original articles as a function of the publication year.

The aim of this review

In view of this diversity and importance of variant impact prediction methods, several reviews discuss the most common tools for predicting pathogenic variations, focusing on their underlying principles (Cline and Karchin 2011; Hassan et al. 2019; Katonis et al. 2014; Tang and Thomas 2016), their clinical value (Ghosh et al. 2017; Yazar and Ozbek 2021; Zhou et al. 2018b), their mutual agreement (Castellana and Mazza 2013), or the integration of large-scale data (Cardoso et al. 2015; Chakravorty and Hegde 2018). Here, this review assesses the agreement of variant effect prediction methods with experimental and clinical data, summarizes their performance in objective blind challenges, and presents some of their successful applications to interpret variant impact and link genes to phenotypes. We find that, despite common concerns regarding performance (Castellana and Mazza 2013; Flanagan et al. 2010; Mahmood et al. 2019), numerous practical applications show that variant effect prediction methods prove reliable when used cautiously in the proper context. Thus, objective assessment exercises (Andreoletti et al. 2019; Hoskins et al. 2017) are critical to define appropriate use cases for each method, to set expectations for accuracy, and to evaluate performance improvements due to methodological refinements. Together

| Type          | Method name    | Input features                                      | Validating datasets                              | Paper                          |
|---------------|----------------|-----------------------------------------------------|--------------------------------------------------|--------------------------------|
| Homology      | ConSurf        | Multiple sequence alignment                         | SH2 and PTB signaling domains/Bcl-XL/Bak peptide complex | Armon et al. (2001)            |
| Homology      | Evolutionary Trace | Multiple sequence alignment                          | ligand binding sites (SH2, SH3 domains, DNA binding/PDB structures with ligands bound | Lichtarge et al. (1996)        |
| Homology      | GERP + +       | Multiple sequence alignment                         | PolII binding regions (ENCODE)                   | Davydov et al. (2010)          |
| Homology      | PhastCons      | Multiple sequence alignment                         | –                                                 | Siepel et al. (2005)           |
| Homology      | PSIC           | Multiple sequence alignment                         | –                                                 | Sunyaev et al. (1999)          |
| Homology      | Rate4Site      | Multiple sequence alignment                         | Src SH2 domain                                    | Pupko et al. (2002)            |
| Homology      | SiPhy          | Multiple sequence alignment                         | ENCODE regions                                    | Garber et al. (2009)           |
| Ensemble      | PhyloP         | 4 conservation scores: LRT, SCORE, SPH, GERP         | ENCODE                                            | Pollard et al. (2010)          |

Methods citations (Armon et al. 2001; Davydov et al. 2010; Garber et al. 2009; Glaser et al. 2003; Lichtarge et al. 1996; Mihalek et al. 2004; Pollard et al. 2010; Pupko et al. 2002; Siepel et al. 2005; Sunyaev et al. 1999)
with the interpretation of other variant types (Liu et al. 2019; Spurdle et al. 2008; Zeng and Bromberg 2019) and the integration with multi-omics data (Gomez-Cabrero et al. 2014; Huang et al. 2017a; Subramanian et al. 2020), variant effect prediction methods can help in understanding the genotype–phenotype relationship.

Main text

Method training and validation

Most methods estimate variant impact by weighing homology and/or structural information in unique ways. Homology-based predictors use multiple sequence alignments (MSA) to achieve their goal, which may either be narrow sets of orthologous sequences or larger sets that include distant homologs and paralogs. The choice of MSA affects the performance of the prediction methods and using the MSA provided by each method does not guarantee optimal performance (Hicks et al. 2011). MSA are also useful in predictors that go beyond basic homology and infer additional properties. Distinct properties are captured by conservation scores (Sunyaev et al. 1999), phylogenetic correlations of residues (Lichtarge et al. 1996), amino acid substitution frequency (Henikoff and Henikoff 1992), biochemical properties at the position of interest (Grantham 1974) and whole substitution profiles. These properties may define the context in which the mutation is observed, together with protein structure variables when a protein structure is available. Solvent accessibility (Ahmad et al. 2004; Lee and Richards 1971), secondary structure (Kabsch and Sander 1983), B-factors (Sun et al. 2019), intrinsic disorder (Dunker et al. 2002), and ligand binding sites (Hendlich et al. 1997; Yang et al. 2013) provide insights into changes occurring in the protein that can only be minimally inferred by homology-based methods. Given the plethora of features to choose and combine, differences between variant impact predictors are not surprising. In addition, many methods make use of various training data to refine predictions and select the most important features. The type of data chosen to train a method will likely influence the predicted impact to be more relevant to experimental, clinical, or evolutionary effects.

Experimental data to train and validate prediction methods

To validate, benchmark, and compare methods, predictors may turn to the experimental outcomes of large mutagenesis studies. For this purpose, four readily accessible datasets gained special popularity and were frequently used: (i) the repression activity of 4041 lac repressor mutations in E. coli (Markiewicz et al. 1994), (ii) the transactivation activity of 2314 human p53 mutations (Kato et al. 2003),
(iii) the break-up of the host cell walls due to 2015 lysozyme mutations in bacteriophage T4 (Rennell et al. 1991), and (iv) the cleavage of Gag and Gag-Pol due to 336 HIV-1 protease mutations (Loeb et al. 1989). More broadly, some dedicated databases curate experimental variation data for benchmarking (Kawabata et al. 1999; Sasidharan Nair and Vihinen 2013). In addition, many more large mutagenesis studies are now available and could be used for additional benchmarking (Gray et al. 2017; Livesey and Marsh 2020; Sruthi et al. 2020). Overall, variant impact prediction methods using these large datasets perform well, if not superbly, according to developer benchmark analyses. Pearson’s correlations were nearly perfect when the experimental data were binned (Katsonis and Lichtarge 2014), showing excellent agreement with the overall trend despite point-by-point fluctuations. Accuracy was about 70% for SNAP (Bromberg and Rost 2007), 68–80% for SIFT (Ng and Henikoff 2001), and slightly better for MAPP (Stone and Sidow 2005). Areas under the ROC curve (AUC) have been reported to fluctuate. Accuracy was about 70% for SNAP (Bromberg and Rost 2007), 68–80% for SIFT (Ng and Henikoff 2001), and slightly better for MAPP (Stone and Sidow 2005). Areas under the ROC curve (AUC) have been reported to be 81–87% for PROVEAN (Choi et al. 2012) and from 86 to 89% for Evolutionary Action (Katsonis and Lichtarge 2014).

**Clinical data to train and validate prediction methods**

For further training and validation, prediction methods can also call upon databases of clinical annotations. A well-known source of clinical associations is the ClinVar repository which reports associations between genetic variation and clinical phenotypes (Landrum et al. 2018). Variants in this database were found in patient samples and annotated according to the guidelines of the American College of Medical Genetics (ACMG) and of the Association for Molecular Pathology (AMP) (Richards et al. 2015), with the goal to provide clinicians with the most robust consensus assessment. The ClinGen group convened Variant Curation Expert Panels to validate genetic annotations for specific genes or sets of genes and update variant annotations in ClinVar (Rehm et al. 2015; Rivera-Munoz et al. 2018). An alternative dataset that is frequently used in training variant impact predicting methods is the Human Gene Mutation Database (HGMD) (Stenson et al. 2020). HGMD is a manually curated, proprietary collection of published germline mutations in nuclear genes associated with human inherited disease. Also, HumSavar is an open access collection of missense variant reports, curated from literature according to ACMG/AMP guidelines (Richards et al. 2015), and it is available through UniProtKB/Swiss-Prot (Wu et al. 2006). Moreover, Cancer-specific sources are available for somatic and germline variants, such as COSMIC (Tate et al. 2019), International Agency for Research on Cancer (IARC) (Petitjean et al. 2007), Database of Curated Mutations (DoCM) (Ainscough et al. 2016), CanProVar (Li et al. 2010), and BRCA Exchange (Cline et al. 2018). Other sources include LOVD (Fokkema et al. 2005), HuVarBase (Ganesan et al. 2019), and disease-specific collections, such as Pompe disease (Kroos et al. 2008), Wilson disease (Kumar et al. 2020), and others (Gout et al. 2007; Peltomaki and Vasen 2004). The mouse-specific database Mutagenetix (Wang et al. 2018, 2015) also contains genotype–phenotype correlations, using various phenotype assays. Using these databases, many studies showed that the disease driver variants are strongly enriched for pathogenic predictions. Performance, however, varies with the choice of reference dataset and the confidence level of the clinical association, and typically ranges between AUCs of 70% and 90% (Choi et al. 2012; Dong et al. 2015; Ioannidis et al. 2016; Niroula et al. 2015; Pejaver et al. 2020; Qi et al. 2021). Nearly perfect AUCs are occasionally reported (Alirezaie et al. 2018; Ghosh et al. 2017), but such outstanding performances may be overly optimistic and prone to indirect circularity (Grimm et al. 2015).

**Population fitness effect data to train and validate prediction methods**

Population-wide sequencing efforts are also used to train or validate predictions. The pioneering 1000 Genomes Project ran from 2008 to 2015 and sequenced more than 2500 individuals across 26 populations from five continental groups (The 1000 Genomes Project Consortium 2015). Since 2015, the International Genome Sample Resource (IGSR) maintains The 1000 Genomes Project (Clarke et al. 2017) and added, thus far, samples from sources such as The Gambian Genome Variation Project (Malaria Genomic Epidemiology Network 2019), The Simons Genome Diversity Project (Mallick et al. 2016), and The Human Genome Diversity Project (Bergstrom et al. 2020). LARGER-population-wide sequencing datasets, such as the Exome Aggregation Consortium (ExAC) (Karczewski et al. 2017; Lek et al. 2016) and the more recent Genome Aggregate Database (gnomAD) (Karczewski et al. 2020), seek to aggregate variants from various projects. These projects mostly include data from disease-specific studies. ExAC and gnomAD provide allele frequency information instead of the individuals’ genomes and may contain more disease driver variants than expected in healthy individuals. Other databases describing population-wide variation include the UK Biobank (Sudlow et al. 2015), dbSNP (Markiewicz et al. 1975), the Exome Variant Server (NHLBI Exome Sequencing Project 2011), HapMap (The International HapMap Consortium 2003), and various population-specific data sets (All of Us Research Program et al. 2019; Ameur et al. 2017; GenomeAsia 100K Consortium 2019; Jain et al. 2021; John et al. 2018; Jung et al. 2020). These sources of human polymorphisms often provide the benign set of variants for
training many predictors (such as FATHMM –MKL and –XF, M-CAP, MISTIC, MPC, MutationTaster, PON-P, PrimateAI, REVEL, and VEST). The underlying hypothesis is that human polymorphisms have been under negative selection pressure that may eliminate or prevent the spread of many pathogenic variants. Therefore, most of the observed variants, especially those with high allele frequency should have nearly neutral or positive effects on protein function (Kimura 1979). Prediction scores that were not trained on human polymorphisms supported this hypothesis, showing that human polymorphisms were enriched in low pathogenicity scores, with the enrichment becoming stronger for variants with higher minor allele frequency (Katsonis and Lichtarge 2014). Another study used VEST to show that sources based on complete genomes, such as The 1000 Genomes Project, contain less pathogenic-predicted variants than sources of compiled clinical data, such as SwissProt (Carter et al. 2013). However, pathogenic variants exist within clinical genomes and studies suggest that predictors of protein function effects can prioritize them to identify candidate disease variants (Chennnen et al. 2020; Ioannidis et al. 2016; Jagadeesh et al. 2016).

**Performance assessment of variant impact prediction methods – CAGI challenges**

The performance of variant impact prediction methods is hard to assess unambiguously. Independent studies (Chan et al. 2007; Dong et al. 2015; Ghosh et al. 2017; Gunning et al. 2020; Leong et al. 2015; Li et al. 2018; Livesey and Marsh 2020; Michels et al. 2019; Miosge et al. 2015; Suybeng et al. 2020; Tian et al. 2019; Yadegari and Majidzadeh 2019) have compared a limited number of methods each, using specific sets of variants and evaluation tests, but ignoring potential training circularities, and making cutoff assumptions that may not fit each method equally well. In contrast, efforts to systematically and objectively assess variant impact prediction methods come from the Critical Assessment of Genome Interpretation (CAGI) community. CAGI, so far, has organized five assessment experiments that include many different challenges, with a new, sixth assessment currently in progress. The challenges may evaluate prediction performance with respect to the impact of variants on a specific protein function (Clark et al. 2019; Kasak et al. 2019a; Xu et al. 2017), clinical sequalae (Carraro et al. 2017; Cline et al. 2019; Voskanian et al. 2019), or global fitness, such as in yeast competition assays (Zhang et al. 2019, 2017a). Critically, all CAGI challenges use new and unpublished data, developer groups make predictions blind to pathogenic associations, and independent judges use multiple criteria to score success blind to the developer identity. The aims of these challenges are to recognize advantageous strategies used by the developers and bottlenecks that prevent the field from advancing. This goal is achieved through direct comparison of each method’s performance and the features they use. In our view, the performance on the CAGI challenges did not point to obvious links between the type of predictor and the type of challenge, because it was subject to several confounding factors (including input data availability, participation, predictor adjustments and approximations, assessor choices, assay or clinical data interpretation). Some methods clearly performed better than others according to multiple assessment metrics, but often different metrics indicated different top methods for the same challenge, highlighting the need for combining multiple metrics. Consistently top-ranked predictions come from the Evolutionary Action (Katsonis and Lichtarge 2017, 2019), MutPred (Pejaver et al. 2017), SNAP (Kasak et al. 2019a), and ensemble methods (Yin et al. 2017). It is worth noting that many less well-known participating methods showed better performance than PolyPhen2 and SIFT, which are very popular and widely used variant impact prediction methods (Katsonis and Lichtarge 2017). Very simple predictive models, such as baseline sequence conservation predictors, may perform on par or better than sophisticated methods (Zhang et al. 2019, 2017a). Also, for a given approach (submissions from the Yang & Zhou lab on the cell proliferation rates upon CDKN2A variants), gradually adding features to the prediction model led to gradual performance improvements (Carraro et al. 2017), indicating that future development should focus on the aggregation, refinement, and validation of new features. In the PCM1 challenge that evaluated 38 human missense mutations implicated in schizophrenia with a zebrafish model assay (Monzon et al. 2019), all submitted predictions had poor performance and yielded a nearly random distribution of balanced accuracy (Katsonis and Lichtarge 2019), leaving questions about the source of disagreement (Miller et al. 2019). In the CALM1 gene challenge, the performance of all submitted predictions improved when the yeast complementation assay data points were limited to those with gradually smaller experimental standard deviation (Katsonis and Lichtarge 2019), indicating a potential underestimation of the performance in CAGI challenges due to experimental errors.

In addition to these variant-level challenges, exome-level CAGI challenges have suggested that variant effect prediction methods can help in finding the disease risk of individuals (Cai et al. 2017; Chandonia et al. 2017; Daneshjou et al. 2017). However, addressing such challenges requires weighing the contribution of genes to each trait and combining the effects of multiple variants, which are not straightforward, so the successes were limited. Three CAGI challenges were related to predicting the risk of Crohn’s disease, with two of them resulting in unreliable performance evaluations due to sample stratification issues and the third one showing AUCs of up to 0.7 (Giollo et al. 2017). The performance
was slightly worse for predicting the risk for venous thromboembolism, with AUCs up to 0.65 and accuracies up to 0.63 (McInnes et al. 2019). In a more complex challenge, two methods performed significantly better than chance for matching the clinical descriptions of undiagnosed patients (Kasak et al. 2019b). However, it proved harder to distinguish between individuals of different intellectual disability phenotypes (Carraro et al. 2019). Overall, predicting the risk of individuals for complex diseases remains challenging and methods predicting variant effect may complement current efforts to solve the etiology of more cases.

Applications of variant impact prediction methods

The methods predicting variant impact are already used in numerous practical applications, as indicated by the thousands of citations of PolyPhen2 (Adzhubei et al. 2010), SIFT (Kumar et al. 2009), and CADD (Kircher et al. 2014), amongst other methods. Typically, they are used to assess the impact of new variants of unknown significance, narrow down driver candidates, and to support evidence for pathogenic effects. The ACMG and AMP guidelines encourage using multiple lines of computational evidence to support pathogenic or benign classification (Richards et al. 2015). In addition, variant impact prediction methods have been used to guide mutagenesis studies and associate genes to phenotype. Next, we discuss representative practical applications of predicting methods that illustrate their value in genome interpretation. Although these applications regard specific predicting methods each, the CAGI experiments suggest that other methods could be equally or more successful in addressing the same scientific questions.

Guiding mutagenesis studies

Targeted mutagenesis experimental studies can take advantage of the variant impact predicting methods to reduce experimental cost and effort while maximizing return without missing key results (Sruthi and Prakash 2020). However, such applications are rare and typically complement other prioritization strategies that account for the available protein structures and methods that predict protein residue importance, such as the methods in Table 2. For example, evolutionarily important residues were used to identify functional motifs in the DNA-dependent protein kinase catalytic subunit and the analysis of variant impact unveiled functional insights and implications (Lees-Miller et al. 2021). Variant effect prediction methods indicated human NAGK polymorphisms that reduced its binding to the dynein subunit DYNLRB1, an interaction that promotes cellular growth and other functions (Dash et al. 2021). Mutation impact and important residues also guided mutagenesis studies aiming to uncover the interaction between the RecA and LexA protein in E. coli, which controls antibiotic resistance (Adikesavan et al. 2011; Marciano et al. 2014). In G-protein-coupled receptors, variant impact scores correlated with the phenotypic change (Gallion et al. 2017; Schonegge et al. 2017), and they were used in selecting targeted mutations that recode the allosteric pathway specificity (Peterson et al. 2015; Rodriguez et al. 2010; Schonegge et al. 2017). More recently, variant impact score analysis guided the development of a mutant esterase that gained stereospecificity properties while maintaining a 53-substrate repertoire (CeaRama et al. 2021). These examples show that variant effect prediction methods can effectively prioritize and reduce the workload of mutagenic experimental studies.

Supporting clinical associations

Methods predicting variant impact can help with the association of variants to traits, which currently relies on observational statistics in family studies (Borecki and Province 2008) and genome-wide association studies (GWAS) (Bush and Moore 2012; MacArthur et al. 2017). Because of linkage disequilibrium (Pritchard and Przeworski 2001), GWAS variants do not necessarily indicate causal effects (Cooper and Shendure 2011), highlighting the need for a transition to functional associations (Gallagher and Chen-Plotkin 2018). Currently, predicting methods are used in both monogenic and polygenic traits, supporting pathogenic effects for variants observed in cases and showing that candidate disease drivers are enriched in pathogenic scores, collectively. Any predicting method or the consensus of multiple methods can serve this supporting role. Few examples are variants causing FARS2 deficiency (Almannai et al. 2018), ANO5 variants causing Gnathodiaphyseal Dysplasia (Otaiy et al. 2018), DHCR24 variants causing Desmosterolosis (Schaaf et al. 2011), variants in multiple genes causing ALS (Gibson et al. 2017; Kenna et al. 2016), Parkinson’s disease (Oluwole et al. 2020), and Alzheimer’s disease (Vardarajan et al. 2015). Similar analyses for somatic mutations in tumors revealed strong selection patterns, either on numerous candidate genes (Bailey et al. 2018) or on specific genes such as in PTPN12 (Nair et al. 2018), BAP1 (Sharma et al. 2019), and TP53 (Li et al. 2016). Therefore, computational approaches can prioritize somatic variants for their role in cancer, using variant effect predictors and gene features (Kaminker et al. 2007). Predicting methods may also prioritize germline variants of trait-associated genes for further examination, such as in cardiovascular diseases (Rababa’h et al. 2013; Suryavanshi et al. 2018; Wang et al. 2021b). With the spread of SARS-CoV-2, computational prediction methods have presented a functional site overview for all SARS-CoV-2 proteins (Wang et al. 2021a) and suggested that their mutational hotspots can alter protein stability and binding affinity (Teng et al. 2021; Wu et al. 2021; Zou et al. 2020). These
studies show that predicting methods have practical value in a variety of clinical associations, in both Mendelian and complex diseases, including cancer.

**Informing diagnoses and clinical decision making**

Variant impact predictors may extend from the bench to the bedside. First tier clinical tests typically use chromosomal microarrays, with reported diagnostic yield of 15–20% in patients with developmental disabilities or congenital anomalies (Miller et al. 2010). Unexplained cases may proceed to whole exome sequencing (WES) or whole genome sequencing (WGS) to fill this gap. Variant impact predictors can aid in the interpretation of the sequenced variants, offering significant increase in diagnostic yield (Grunseich et al. 2021; Stavropoulos et al. 2016). Pharmacogenomics studies may take advantage of predicting methods to interpret the impact of amino acid variations on drug metabolism (Isvoran et al. 2017; Matimba et al. 2009). Additionally, the predicted pathogenicity of somatic mutations in cancer was used in a classification system that may inform patient management (Sukhai et al. 2016). In a study of how TP53 variants affect the health of head and neck cancer patients, Evolutionary Action was able to stratify the overall survival and time to metastasis (Neskey et al. 2015), indicated resistance to cisplatin therapy (Osman et al. 2015b), and provoked suggestions for personalized treatment (Osman et al. 2015a). Similarly, survival stratification was obtained in two independent studies for colorectal liver metastases patients (Chun et al. 2019) and myelodysplastic syndrome patients (Kanagal-Shamanna et al. 2021). Therefore, predicting methods contribute in clinical diagnosis and can open paths toward precision medicine.

**Associating genes to phenotype**

Variant impact scores may lead to associations of genes to traits. Typically, gene-trait associations rely on detecting selection patterns within a group of individuals who share the trait (cases) compared to unaffected individuals (controls). These selection patterns arise because trait driver genes harbor several pathogenic variants in cases, in addition to non-pathogenic variants that may appear in either the cases or the controls. Current gene discovery methods may quantify patterns such as whether the gene has more mutations compared to the expected number (Lawrence et al. 2013), the mutations cluster in the protein structure or sequence compared to homogeneous spread (Tamborero et al. 2013), and the characteristic nucleotide context of the mutations differs from the context of all other mutations (Dietlein et al. 2020). Since the driver variants have larger predicted pathogenicity values compared to random nucleotide substitutions, methods that predict protein function effects offer an additional pattern toward pointing to candidate trait-driver genes. This selection pattern is orthogonal and complementary to the aforementioned measures, making variant impact prediction methods valuable for gene discovery. Next, we note such applications to somatic, de novo, and inherited variants.

**Genes under selection in somatic mutations**

Many cancer studies use variant impact prediction methods either as supporting evidence for the pathogenicity of gene variants (Bailey et al. 2018; Cancer Genome Atlas Research Network 2011) or as the main evidence to establish a gene-cancer link through an automated discovery process (Davoli et al. 2013; Gonzalez-Perez and Lopez-Bigas 2012; Hsu et al. 2022; Parvandeh et al. 2022). The underlying hypotheses are that most somatic mutations are passengers (i.e. they do not contribute to oncogenesis) and that driver mutations (i.e. they contribute to the development of cancer) occur selectively in specific genes (Greenman et al. 2007; Stratton et al. 2009). Because the driver variants affect protein function, predicting methods should statistically score driver variants as more pathogenic than passenger variants (Carter et al. 2009; Chen et al. 2020; Cline et al. 2019; Mullany et al. 2015; Reva et al. 2011) and point to cancer driver genes. Moreover, protein effect prediction methods can inform regarding the role of each gene in cancer, with tumor suppressor genes having mostly loss-of-function variants with high impact scores and oncogenes having mostly gain-of-function variants with intermediate to high scores (Hsu et al. 2022; Shi and Moult 2011). Gene pathway information may complement variant impact prediction methods in finding cancer driver genes (Cancer Genome Atlas Research Network 2017), even for small patient sets, such as 29 patients with sporadic Parathyroid Cancer (Clarke et al. 2019). These applications suggest that variant impact prediction methods can help in finding candidate driver genes within whole cancer cohorts and within their cancer type divisions.

**Genes under selection in de novo mutations**

Variant impact prediction methods are commonly used in prioritizing the functional effects of de novo variants (Hu et al. 2016; Pejaver et al. 2020; Wang et al. 2019; Willsey et al. 2017). However, de novo variants are typically absent from the general population, with each individual harboring less than two coding de novo variants (Iossifov et al. 2012; Sevim Bayrak et al. 2020). This fact limits the gene-level analysis of de novo variants, even for large datasets, such as the Simons Simplex Collection (SSC) (Fischbach and Lord 2010), which contains sequencing data from more than 2500 families with at least one child diagnosed with autism spectrum disorder (ASD) (Iossifov et al. 2012; Lord et al. 2020, 2013; Gonzalez-Perez and Lopez-Bigas 2012; Hsu et al. 2022; Parvandeh et al. 2022). The underlying hypotheses are that most somatic mutations are passengers (i.e. they do not contribute to oncogenesis) and that driver mutations (i.e. they contribute to the development of cancer) occur selectively in specific genes (Greenman et al. 2007; Stratton et al. 2009). Because the driver variants affect protein function, predicting methods should statistically score driver variants as more pathogenic than passenger variants (Carter et al. 2009; Chen et al. 2020; Cline et al. 2019; Mullany et al. 2015; Reva et al. 2011) and point to cancer driver genes. Moreover, protein effect prediction methods can inform regarding the role of each gene in cancer, with tumor suppressor genes having mostly loss-of-function variants with high impact scores and oncogenes having mostly gain-of-function variants with intermediate to high scores (Hsu et al. 2022; Shi and Moult 2011). Gene pathway information may complement variant impact prediction methods in finding cancer driver genes (Cancer Genome Atlas Research Network 2017), even for small patient sets, such as 29 patients with sporadic Parathyroid Cancer (Clarke et al. 2019). These applications suggest that variant impact prediction methods can help in finding candidate driver genes within whole cancer cohorts and within their cancer type divisions.
Typically, such data are analyzed in the contexts of known gene-phenotype associations and the human interactome network (Chen et al. 2018). Variant impact prediction methods, such as MutPred2 (Pejaver et al. 2020) and VIPUR (Buja et al. 2018), have shown that de novo variants in ASD cases have a higher fraction of predicted pathogenic variants compared to healthy siblings. Going one step further, a study using the Evolutionary Action (EA) method and gene pathway information without prior knowledge of phenotype associations identified 398 genes (representing 23 pathways) as candidate drivers for ASD, based on the enrichment of de novo variants to pathogenic scores (Koire et al. 2021). The same study proposed polygenic risk scores based on the EA scores of either de novo or rare inherited variants on candidate genes and showed that these scores correlated with the Intelligence Quotient (IQ) of patients. These correlations were stronger when the contribution of each gene was weighted by Residual Variation Intolerance Scores (RVIS), a measure of genic intolerance to mutations (Petrovski et al. 2013). Similar analyses can be done for more phenotypes, such as congenital heart disease (Jin et al. 2017), where cases appear to have a higher fraction of predicted pathogenic variants compared to healthy controls (Qi et al. 2021). Such large family datasets provide de novo mutations that can use variant impact predictions together with other information to discover new genes toward decoding the genotype–phenotype relationship.

Genes under selection in inherited mutations

Case–control studies are routinely designed for the discovery of genes associated with a particular trait. For Mendelian traits, these associations are straightforward, and methods predicting protein function effects can help (Carter et al. 2013; Hu et al. 2013). For complex traits, the standard is gene-based GWAS for the trait of interest using all variants of unknown significance that are candidate disease drivers. This ability to translate genomic data into quantitative traits raises hope for improved diagnostic tests with polygenic risk scores that account for functional effects rather than relying only on observational statistics. A caveat is that the basis for most methods remains rooted in homology information. The scores will thus tend to assess long-term “evolutionary” effects. Generally, and depending on the prediction method and the test data, these effects will tend to align with clinical or experimental impact as shown by strong correlations through extensive validation studies and objective assessments. In other words, the fitness landscape may appear similar at different scales.

Discussion

This review of current computational estimates of protein function effects due to variants illustrates several practical applications. They routinely guide experimental studies of protein structure and function and clinical studies of variants of unknown significance that are candidate disease drivers. Most recently, they played a major role in identifying new genes associated with traits, for either somatic, de novo, or inherited variants. This ability to translate genomic data into polygenic risk scores that account for functional effects rather than relying only on observational statistics. A caveat is that the basis for most methods remains rooted in homology information. The scores will thus tend to assess long-term “evolutionary” effects. Generally, and depending on the prediction method and the test data, these effects will tend to align with clinical or experimental impact as shown by strong correlations through extensive validation studies and objective assessments. In other words, the fitness landscape may appear similar at different scales.

Criticism and value

In the past, variant impact prediction methods sustained pointed criticism (Flanagan et al. 2010; Mahmood et al. 2017; Tchernitchko et al. 2004) and this curtailed their use as prognostic tools. Most often, the criticism was fed on the one hand by a demand for nearly perfect accuracy in clinical diagnostics (Walters-Sen et al. 2015), and on the other hand by disagreements, first, between different methods (Chun and Fay 2009), and second, between prediction
methods and experimental data or clinical annotations (Mahmood et al. 2017; Miller et al. 2019). At some degree, these discrepancies are due to misalignments between the hypotheses adopted by the method developers and the data analysts: a key is useful only when it is properly applied to the right lock. In the light of epistatic interactions, inaccuracy is expected for single variant estimations, since each individual has a unique genetic, epigenetic, and environmental background that may modify the impact of this variant. These factors may result in incomplete penetrance, where two individuals with the same genetic variant can have either benign or disease phenotype linked to that variant (Cooper et al. 2013; Waalen and Beutler 2009; Zlotogora 2003). Predictors can explicitly capture residue dependencies between positions to improve accuracy (Hopf et al. 2017) and focused methods can detect covariation signals in multiple sequence alignments to identify residue pairs with epistatic effects (Jones et al. 2012; Morcós et al. 2011; Salinas and Ranganathan 2018; Shen and Li 2016). However, most predictors of protein function effects provide estimates in a broader view, as when individual background effects are averaged out over cohorts of individuals, suggesting they are more informative in high-penetrance genes and disorders. Literally, homology-based prediction methods ignore the context and answer whether a specific variant is pathogenic in an "evolutionary sense," which at best matches the human population at large rather than addressing the context-dependent effects of the variants (DiGiammarino et al. 2002). The choice of a multiple sequence alignment input defines the “average context” of the computation and its potential biases and errors will affect the accuracy of the predictions (Hicks et al. 2011). Each algorithmic approach weighs input features differently from the other methods, which may influence prediction accuracy dramatically. Since both the sequence alignment input and the algorithmic approach affect prediction accuracy, we should avoid generalizing the performance conclusions based on a single analysis. Moreover, the assessors should ensure their hypothesis does not conflict those underlying each prediction method. The CAGI challenges offer useful insights into the performance of different methods since method developers are able to modify their approach according to the needs of each challenge and independent assessors ensure objectivity. These assessments demonstrate progress in the field of variant impact prediction and the need to adjust predictors given specific tasks. Newer approaches achieve strong correlations with experimental assay data and perform consistently better than well-known methods (Katsonis and Lichtarge 2017b). Such correlations may improve when the impact of experimental noise is reduced, using only data points with small standard deviations (Katsonis and Lichtarge 2019) or combining multiple experimental assays (Gallion et al. 2017). This suggests that even systematic assessments may under-estimate the performance of predicting methods.

Predicting the impact of other variant types

Whole genome sequencing shows that non-synonymous variants are less than 0.3% of the total calls (Shen et al. 2013). There is therefore growing interest in prediction methods of other variant types. Stop-gain and frameshift insertion and deletion (fs-indel) variants result in protein sequence truncation and are traditionally viewed as pathogenic, but many of them appear frequently in human genomes even in a homozygous state (MacArthur and Tyler-Smith 2010). Non-frameshifting insertion and deletion (indel) variants are also of interest due to their link to diverse clinical effects and their substantial genetic load in most humans (Mullaney et al. 2010). Methods such as SIFT Indel (Hu and Ng 2012), DDIG-in (Folkman et al. 2015), VEST-Indel (Douville et al. 2016), and MutPred-LOF/Indel (Pagel et al. 2017, 2019) may use homology, structure, intrinsic disorder predictions, and gene importance features to prioritize nonsense and indel variants with reported balanced accuracy of 80–90% (Douville et al. 2016). PROVEAN (Choi et al. 2012) and MutationTaster2 (Schwarz et al. 2014) also provide predictions to non-frameshifting indel variants following the same framework they used for predicting the impact of missense variants. CADD (Kircher et al. 2014) is designed to predict the impact of all classes of genetic variation, including splice sites (Rentzsch et al. 2021) and non-coding variations. Methods that focus on predicting splicing effects use as input the genomic sequence of the pre-mRNA transcripts and include SpliceAI (Jaganathan et al. 2019), MutPred Splice (Mort et al. 2014), Human Splicing Finder (Desmet et al. 2009), SPICE (Leman et al. 2018), and Skippy (Woolfe et al. 2010). Methods that focus on predicting noncoding variant effects rely on functional genomics data, such as various sequence conservation and constraint scores (Dousse et al. 2016; Garber et al. 2009; Siepel et al. 2005), in silico predictions of transcription factor binding sites, enhancer regions, and long noncoding RNAs (lncRNAs) (Abugessaisa et al. 2021; Fu et al. 2014; Loots and Ovcharenko 2004; Pachkov et al. 2013), and experimental evidence provided by the Encyclopedia of DNA Elements (ENCODE) (Davis et al. 2018; Encode Project Consortium 2012), including transcription factor ChIP-seq, DNA methylation arrays, and small RNA-seq projects. Some non-coding functional impact predictors include, but are not limited to, LINSIGHT (Huang et al. 2017b), GenoCanyon (Lu et al. 2015), FATHMM-MKL (Shihab et al. 2015), FATHMM-XF (Rogers et al. 2018), PAFA (Zhou and Zhao 2018), DIVAN (Chen et al. 2016), and GWAVA (Ritchie et al. 2014). Synonymous variants,
Significance in personalized therapy

Genome interpretation relies on the classification of genetic variants as pathogenic or benign, which necessitates the estimation of impact for all single variants. Clinical associations and experimental data are too limited for characterizing all variants, since more than 98% of the variants in human exomes have frequency of less than 1% (Karcewski et al. 2020; Van Hout et al. 2020) and over 40% of the ClinVar entries are catalogued as variants of unknown significance (Henrie et al. 2018). Protein function effect prediction methods have shown strong correlations with established associations and may be cautiously used to start bridging the gap in genome interpretation. With the advent of less costly sequencing technologies, clinicians can read patient’s genomes and search for precise therapies tailored to the genetic etiology of the disease. The insights provided by variant impact prediction methods may assist clinicians in selecting beneficial treatments.

Acknowledgements This work was supported by the National Institutes of Health (Grants GM066699, AG061105, AG068214 and AG074009).

Availability of data and material Not applicable.

Code availability Not applicable.

Declarations

Conflicts of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Harles ME, McVean GA (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073. https://doi.org/10.1038/nature09534

Abugessaisa I, Ramiloslawski JA, Lizio M, Severin J, Hasegawa A, Harshbarger J, Kondo A, Noguchi S, Yip CW, Ooi JLC, Tagami M, Hori F, Agrawal S, Hon CC, Cardon M, Ikeeda S, Ono H, Bono H, Kato M, Hashimoto K, Bonetti A, Kato M, Kobayashi N, Shin J, de Hoon M, Hayashizaki Y, Carninci P, Kawaji H, Kasukawa T (2021) FANTOM enters 20th year: expansion of transcriptomic atlases and functional annotation of non-coding RNAs. Nucleic Acids Res 49:D892–D898. https://doi.org/10.1093/nar/gkaa1054

Adikesavan AK, Katsonis P, Marciano DC, Luu R, Herman C, Lichtarge O (2011) Separation of recombination and SOS response in Escherichia coli RecA suggests LexA interaction sites. PLoS Genet 7:e1002244. https://doi.org/10.1371/journal.pgen.1002244

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249. https://doi.org/10.1038/nmeth0410-248

Ahmad S, Gromiya M, Fawareh H, Sarai A (2004) ASAView: database and tool for solvent accessibility representation in proteins. BMC Bioinform 5:51. https://doi.org/10.1186/1471-2105-5-51

Ainscough BJ, Griffith M, Coffman AC, Wagner AH, Kinseki J, Choudhary MN, McMichael JF, Fulton RS, Wilson RK, Griffith OL, Mardis ER (2016) DoCM: a database of curated mutations in cancer. Nat Methods 13:806–807. https://doi.org/10.1038/nmeth.4000

Alirozaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD (2018) ClinPred: prediction tool to identify disease-relevant non-synonymous single-nucleotide variants. Am J Hum Genet 103:474–483. https://doi.org/10.1016/j.ajhg.2018.08.005

All of Us Research Program I, Denny JC, Rutter JL, Goldstein DB, Philippakis A, Smoller JW, Jenkins G, Dishman E (2019) The “All of Us” Research Program. N Engl J Med 381:668–676. https://doi.org/10.1056/NEJMsr1809937

Almnaani M, Wang J, Dai H, El-Hattab AW, Faqiib EA, Saleh MA, Al Asmari A, Alwaidei AH, Aljadhai YI, AlHashem A, Tabarki B, Lines MA, Grange DK, Benini R, Alsamasan AS, Mahmoud A, Katsonis P, Lichtarge O, Wong LC (2018) FARS2 deficiency, new cases, review of clinical, biochemical, and molecular spectra, and variants interpretation based on structural, functional, and evolutionary significance. Mol Genet Metab 125:281–291. https://doi.org/10.1016/j.ymgme.2018.07.014

Ameur A, Dahlberg J, Olsson P, Vezzi F, Karlsson R, Martin M, Viklund J, Kahari AK, Lundin P, Che H, Thutkawkorapin J, Eissfeldt J, Lampa S, Dahlberg M, Hagberg J, Jareborg N, Liljedahl U, Jonasson I, Johansson A, Feuk L, Lundeberg J, Syvanen AC, Lundin S, Nilsson D, Nystedt B, Magnusson PK, Gylensten U (2017) SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. Eur J Hum Genet 25:1253–1260. https://doi.org/10.1038/ejhg.2017.130

Andreoletti G, Pal LR, Moiul J, Brenner SE (2019) Reports from the fifth edition of CAGI: The critical assessment of genome interpretation. Hum Mutat 40: 1197–1201. https://doi.org/10.1002/humu.23876

Armon A, Graur D, Ben-Tal N (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 307:447–463. https://doi.org/10.1006/jmbi.2000.4474

Springer
Carcinoma. Cell 169(1327–1341):e23. https://doi.org/10.1016/j.cell.2017.05.046

Capriotti E, Fariselli P, Casadio R (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33:W306–W310. https://doi.org/10.1093/nar/gki375

Capriotti E, Calabrese R, Casadio R (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22:2729–2734. https://doi.org/10.1093/bioinformatics/btl423

Capriotti E, Fariselli P, Rossi I, Casadio R (2008) A three-state prediction of single point mutations on protein stability changes. BMC Bioinform 9(Suppl 2):S6. https://doi.org/10.1186/1471-2105-9-S2-S6

Capriotti E, Altman RB, Bromberg Y (2013) Collective judgment of single point mutations of non-coding disease-specific risk variants using multi-omics profiles. Genome Biol 17:252. https://doi.org/10.1186/s13059-016-1112-z

Chen S, Fragozo R, Klei L, Liu Y, Wang J, Roeder K, Devlin B, Yu H (2018) An interactome perturbation framework prioritizes damaging missense mutations for developmental disorders. Nat Genet 50:1032–1040. https://doi.org/10.1038/s41588-018-0130-z

Chen H, Li J, Wang Y, Ng PK, Tsang YH, Shaw KR, Mills GB, Liang H (2020) Comprehensive assessment of computational algorithms in predicting cancer driver mutations. Genome Biol 21:43. https://doi.org/10.1186/s13059-020-01954-z

Cheng J, Randall A, Baldi P (2006) Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 62:1125–1132. https://doi.org/10.1002/prot.20810

Chenchen K, Weber T, Lorange X, Kress A, Bohm J, Thompson J, Laporte J, Poch O (2020) MISTIC: A prediction tool to reveal disease-relevant deleterious missense variants. PLoS One 15:e0236962. https://doi.org/10.1371/journal.pone.0236962

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP (2012) Predicting the functional effect of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol 307:683–706. https://doi.org/10.1016/j.jmb.2001.4510

Choudhuri TK, Paul S (2006) Protein-misfolding diseases and chaperone-based therapeutic approaches. FEBS J 273:1331–1349. https://doi.org/10.1111/j.1742-4658.2006.05181.x

Chuang T, Ollila S, Nystrom M, Grimm AJ, Christodoulou J, Clarke CN, Katsonis P, Hsu TK, Koire AM, Silva-Figueroa A, Christidis A, Tosatto SCE, Yin Y, Buckley BA (2017) Lessons from the CAGI-4 Hopkins clinical panel challenge. Hum Mutat 38:1155–1168. https://doi.org/10.1002/humu.23225

Chosman D, Adams RM (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol 307:683–706. https://doi.org/10.1002/jmb.2001.4510

Choudhuri TK, Paul S (2006) Protein-misfolding diseases and chaperone-based therapeutic approaches. FEBS J 273:1331–1349. https://doi.org/10.1111/j.1742-4658.2006.05181.x

Chu NS, Yasmin N, Lusitsch A, Agarwala R, Sidore C, Marth RJ, Karchin R, Bickel EE (2019) Machine learning models for analyzing oncogenic driver missense mutations for developmental disorders. Nat Genet 50:1032–1040. https://doi.org/10.1038/s41588-018-0130-z

Chung Y, Tseng DC, Xiao L, Aloia TA, Eng C, Kopetz SE, Lichtarge O, Vauthney JN (2019) Deletiouser effect of RAS and evolutionary high-risk TP53 double mutation in colorectal liver metastases. Ann Surg 269:917–923. https://doi.org/10.1097/SLA.0000000000002450

Chung Y, Tseng DC, Xiao L, Aloia TA, Eng C, Kopetz SE, Lichtarge O, Vauthney JN (2019) Deletiouser effect of RAS and evolutionary high-risk TP53 double mutation in colorectal liver metastases. Ann Surg 269:917–923. https://doi.org/10.1097/SLA.0000000000002450

Clark WT, Kusak L, Bakolitsa C, Hu Z, Andreoletti G, Babb 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, Radiovjac P, Savojardo C, Shi X, Zhou Y, Uppal A, Xu Q, Yin Y, Pejaver V, Wang M, Wei L, Moutt J, Yu GK, Brenner SE, LeBowitz JH (2019) Assessment of predicted enzymatic activity of alpha-N-acetylglucosaminidase coding disease-specific risk variants using multi-omics profiles. Brief Bioinform 14(Suppl 3):S3. https://doi.org/10.1093/bib/bbt013

Chen S, Fragozo R, Klei L, Liu Y, Wang J, Roeder K, Devlin B, Yu H (2018) An interactome perturbation framework prioritizes damaging missense mutations for developmental disorders. Nat Genet 50:1032–1040. https://doi.org/10.1038/s41588-018-0130-z

Cheng J, Randall A, Baldi P (2006) Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 62:1125–1132. https://doi.org/10.1002/prot.20810

Chennen K, Weber T, Lorange X, Kress A, Bohm J, Thompson J, Laporte J, Poch O (2020) MISTIC: A prediction tool to reveal disease-relevant deleterious missense variants. PLoS One 15:e0236962. https://doi.org/10.1371/journal.pone.0236962

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP (2012) Predicting the functional effect of amino acid substitutions and indels. PLoS One 7:e46688. https://doi.org/10.1371/journal.pone.0046688

Chun S, Fay JC (2009) Identification of deleterious mutations within three human genomes. Genome Res 19:1553–1561. https://doi.org/10.1101/gr.092619.109

Chun Y, Passot G, Yamashita S, Nusrat M, Katsonis P, Loree JM, Conrad C, Tzeng CD, Xiao L, Aloia TA, Eng C, Kopetz SE, Lichtarge O, Vauthney JN (2019) Deletiouser effect of RAS and evolutionary high-risk TP53 double mutation in colorectal liver metastases. Ann Surg 269:917–923. https://doi.org/10.1097/SLA.0000000000002450

Clarke WT, Kusak L, Bakolitsa C, Hu Z, Andreoletti G, Babb 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, Radiovjac P, Savojardo C, Shi X, Zhou Y, Uppal A, Xu Q, Yin Y, Pejaver V, Wang M, Wei L, Moutt J, Yu GK, Brenner SE, LeBowitz JH (2019) Assessment of predicted enzymatic activity of alpha-N-acetylglucosaminidase coding disease-specific risk variants using multi-omics profiles. Brief Bioinform 14(Suppl 3):S3. https://doi.org/10.1093/bib/bbt013

Cea-Rama I, Coscolin C, Katsonis P, Bargiela R, Golyshin PN, Lichtarge O, Ferrer M, Sanz-Aparicio J (2021) Structure and evolutionary trace-assisted screening of a residue swapping the evolutionary impact of SNVs. Bioinformatics 27:441–448. https://doi.org/10.1093/bioinformatics/btq695
Cooper GM, Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 12:628–640. https://doi.org/10.1038/nrg3046

Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H (2013) Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet 132:1077–1130. https://doi.org/10.1007/s00439-013-1331-2

Cline MS, Babbi G, Bonache S, Cao Y, Casadio R, de la Cruz X, Diez-Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6(4):D794–D801. https://doi.org/10.1093/nar/gkx1081

Davoli T, Xu AW, Mengwasser KE, Sack LM, Yoon JC, Park PJ, Elledge SJ (2013) Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell 155:948–962. https://doi.org/10.1016/j.cell.2013.10.011

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S (2010) Identifying a high fraction of the human genome to be under selective constraint using GERnP++. PLoS Comput Biol 6:e1001025. https://doi.org/10.1371/journal.pcbi.1001025

Dehouck Y, Kwasiogroch JM, Gilis D, Rooman M (2011) PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinform 12:151. https://doi.org/10.1186/1471-2105-12-151

Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM, Crawford DC (2010) PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 26:1205–1210. https://doi.org/10.1093/bioinformatics/btaq226

Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C (2009) Human splicing finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res 37:e67. https://doi.org/10.1093/nar/gkp215

Diehllein F, Weghorn D, Taylor-Weiner A, Richters A, Reardon B, Liu D, Lander ES, Van Allen EM, Sunyaev SR (2020) Identification of cancer driver genes based on nucleotide context. Nat Genet 52:208–218. https://doi.org/10.1038/s41588-019-0572-y

DiGiammarino EL, Lee AS, Cadwell C, Zhang W, Bothner B, Ribeiro RC, Zambetti G, Kriwacki RW (2002) A novel mechanism of tumorigenesis involving pH-dependent destabilization of a mutant p53 tetramer. Nat Struct Biol 9:12–16. https://doi.org/10.1038/nsb730

DNA sequencing costs: Data (2020) DNA Sequencing Costs: Data

Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X (2015) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet 24:2125–2137. https://doi.org/10.1093/hmg/ddu733

Dousse A, Junier T, Zdobnov EM (2016) CEGA—a catalog of conserved elements from genomic alignments. Nucleic Acids Res 44:D96-100. https://doi.org/10.1093/nar/gkj1163

Douvillie C, Masica DL, Stenson PD, Cooper DN, Gygax DM, Kim R, Ryan M, Karchin R (2016) Assessing the pathogenicity of insertion and deletion variants with the variant effect scoring tool (VEST-Indel). Hum Mutat 37:28–35. https://doi.org/10.1002/humu.22911

Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z (2002) Intrinsic disorder and protein function. Biochemistry 41:6573–6582. https://doi.org/10.1021/bi012159+4

Encode Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74. https://doi.org/10.1038/nature11247

Fariselli P, Martelli PL, Savojardo C, Casadio R (2015) INPS: predicting phenotypes from exomes in the critical assessment of genome interpretation (CAGI) challenges. Hum Mutat 38:1182–1192. https://doi.org/10.1002/humu.23280

Fischbach GD, Lord C (2010) The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron 68:192–195. https://doi.org/10.1016/j.neuron.2010.10.006

Flanagan SE, Patch AM, Ellard S (2010) Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. Genet Test Mol Biomarkers 14:533–537. https://doi.org/10.1089/gtmb.2010.0036

Fokkema IF, den Dunnen JT, Taschner PE (2005) LOVD: easy creation of a locus-specific sequence variation database using an “LSDB-in-a-box” approach. Hum Mutat 26:63–68. https://doi.org/10.1002/humu.20201

Folkman L, Yang Y, Li Z, Stantic B, Sattar A, Mort M, Cooper DN, Liu Y, Zhou Y (2015) DDIG-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels. Bioinformatics 31:1599–1606. https://doi.org/10.1093/bioinformatics/btv862

French JD, Edwards SL (2020) The role of noncoding variants in heritable disease. Trends Genet 36:880–891. https://doi.org/10.1016/j.tig.2020.07.004
Fu W, Akey JM (2013) Selection and adaptation in the human genome. Annu Rev Genom Hum Genet 14:467–489. https://doi.org/10.1146/annurev-genom-091212-153509

Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, Khurana E, Gerstein M (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol 15:480. https://doi.org/10.1186/s13059-014-0480-5

Gallagher MD, Chen-Plotkin AS (2018) The post-GWAS Era: from association to function. Am J Hum Genet 102:717–730. https://doi.org/10.1016/j.ajhg.2018.04.002

Gallion J, Koire A, Katsonis P, Schoenegg AM, Bouvier M, Lichtarge O (2017) Predicting phenotype from genotype: Improving accuracy through more robust experimental and computational modeling. Hum Mutat 38:569–580. https://doi.org/10.1002/humu.23193

Ganesan K, Kulandaiaasamy A, Binny Priya S, Gromiha MM (2019) HuVarBase: A human variant database with comprehensive information at gene and protein levels. PLoS One 14:e0210475. https://doi.org/10.1371/journal.pone.0210475

Garber M, Guttmann M, Clamp M, Zody MC, Friedman N, Xie X (2009) Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics 25:i54–62. https://doi.org/10.1093/bioinformatics/btp190

Gelfman S, Wang Q, McSweeney KM, Ren Z, La Carpia F, Halvorsen GB, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N, Schlager M, Gisel A, Ballestar E, Bongcam-Rudloff E, Conesa O’Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew YE, DeFazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenexiv-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung SY, Woooster R, Futreal PA, Stratton MR (2007) Patterns of somatic mutation in human cancer genomes. Nature 446:153–158. https://doi.org/10.1038/nature05610

Grimm DG, Azencott CA, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW, Duncan LE, Gorbard KM (2015) The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat 36:513–523. https://doi.org/10.1002/humu.23199

Gutierrez-Sacristan A, De Niz C, Kothari C, Kong SW, Mandl KD, Aviñach P (2021) GenoPheno: cataloging large-scale phenotypic and next-generation sequencing data within human datasets. Brief Bioinform 22:55–65. https://doi.org/10.1093/bib/bbab033

Gutierrez-Sacristan A, De Niz C, Kothari C, Kong SW, Mandl KD, Aviñach P (2021) GenoPheno: cataloging large-scale phenotypic and next-generation sequencing data within human datasets. Brief Bioinform 22:55–65. https://doi.org/10.1093/bib/bbab033

Hassan MS, Shaalan AA, Dessouky MI, Abdelnaiem AE, ElHefnawi M (2019) A review study: Computational techniques for expecting the impact of non-synonymous single nucleotide variants in human diseases. Gene 680:20–33. https://doi.org/10.1016/j.gene.2018.09.028

A review study: Computational techniques for expecting the impact of non-synonymous single nucleotide variants in human diseases. Gene 680:20–33. https://doi.org/10.1016/j.gene.2018.09.028

This page has been reviewed by experts to ensure the accuracy and completeness of the text. It includes a variety of topics related to human genetics, including the impact of non-synonymous single nucleotide variants in human diseases. The text is written in a clear and concise manner, making it accessible to a wide audience.
Loeb DD, Swanstrom R, Chandesakhar P, Gibson G, Kumar S (2019) Biological relevance of computationally predicted pathogenicity of noncoding variants. Nat Commun 10:330. https://doi.org/10.1038/s41467-019-08270-y

Livesey BJ, Marsh JA (2020) Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. Mol Syst Biol 16:e9380. https://doi.org/10.15252/msb.20199380

Livingstone M, Folkman Y, Zhang P, Mort M, Cooper DN, Livesey BJ, Marsh JA (2018) Autism spectrum disorder. Nat Rev Dis Primers 6:5. https://doi.org/10.1038/s41572-019-0138-4

Lord C, Brugha TS, Charman T, Cusack J, Dumas G, Frazier T, Jones CA 3rd (1989) Complete mutagenesis of the HIV-1 protease. Mol Biol 336:1265–1282. https://doi.org/10.1016/j.jmb.2003.1678-4685-2018-0148

Liu L, Sanderford MD, Patel R, Chandrashekar P, Gibson G, Kumar S, Lord C, Elsabbagh M, Baird G, Veenstra-Vanderweele J (2018) Autism spectrum disorder. Nat Rev Dis Primers 4:1038/18964

Marciano DC, Liu RC, Katsonis P, Amin SR, Herman C, Lichtarge O (2014) Negative feedback in genetic circuits confers evolutionary resilience and capacitance. Cell 1753:1789–1795. https://doi.org/10.1016/j.celrep.2014.05.018

Markiewicz W, Amikam S, Roguin N, Riss E (1975) Changing haemodynamics in patients with papillary muscle dysfunction. Br Heart J 37:445–448. https://doi.org/10.1136/hrt.37.4.445

Markiewicz P, Klein LA, Cruz C, Ehret S, Miller JH (1994) Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as “spacers” which do not require a specific sequence. J Mol Biol 240:421–433. https://doi.org/10.1006/jmb.1994.1458

Mathe E, Olivier M, Kato S, Ishioka C, Hainaut P, Tavtigian SV (2006) Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. Nucleic Acids Res 34:1317–1325. https://doi.org/10.1093/nar/gkj518

Matimba A, Del-Favero J, Van Broeckhoven C, Masimirembwa C (2009) Novel variants of major drug-metabolising enzyme genes in diverse African populations and their predicted functional effects. Hum Genomics 3:169–190. https://doi.org/10.1186/1678-4685-3-2-169

McInnes G, Daneshjou R, Katsonis P, Lichtarge O, Srinivasan R, Rana S, Radiovic P, Money SD, Pagel KA, Stamboulian M, Yang J, Capriotti E, Wang Y, Bromberg Y, Bosso S, Savojardo C, Martelli PL, Casadio R, Pal LR, Moulit J, Brenner SE, Altman R (2019) Predicting venous thromboembolism risk from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. Hum Mutat 40:1314–1320. https://doi.org/10.1002/humu.23825

Michels M, Matte U, Fraga LR, Mancuso ACB, Ligabue-Braun R, Berner EA, Siebert M, Sanseverino MTV (2019) Determining the pathogenicity of CFTR missense variants: Multiple comparisons of in silico predictors and variant annotation databases. Genet Mol Biol 42:560–570. https://doi.org/10.1590/1678-4685-GMB-2018-0148

Mihalek I, Res I, Lichtarge O (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. J Mol Biol 336:1265–1282. https://doi.org/10.1016/j.jmb.2003.12.078

Miller DT, Adam MP, Aradhyta S, Biesecker LG, Brothman AR, Carter NP, Church DM, Crolla JA, Eichler EE, Epstein JC, Faucett WA, Feuk L, Friedman JM, Hamosh A, Jackson L, Kaminsky EB, Kok K, Krantz ID, Kuhn RM, Lee C, Ostell JM, Rozenberg C, Scherer SW, Spinelli NB, Statoupolos DJ, Tepperberg JH, Thoraland EC, Vermeesch JR, Waggoner DJ, Watson MS, Martin CL,
Ledbetter DH (2010) Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet 86:749–764. https://doi.org/10.1016/j.ajhg.2010.04.006

Miller M, Wang Y, Bromberg Y (2019) What went wrong with variant effect predictor performance for the PM1 challenge. Hum Mutat 40:1486–1494. https://doi.org/10.1002/humu.23832

Miosge LA, Field MA, Sontani Y, Cho V, Johnson S, Palkova A, Balakishnan B, Li B, Ball EV, Cooper DN, Radivojac P, San-Munplaisir N, Merault G, Poyart C, Rhoda MD, Craescu C, Vidaud M, Miosge LA, Field MA, Sontani Y, Cho V, Johnson S, Palkova A, Balakishnan B, Li B, Ball EV, Cooper DN, Radivojac P, San-Monzon AM, Carraro M, Chiricosta L, Reggiani F, Han J, Ozturk K, Mullaney JM, Mills RE, Pittard WS, Devine SE (2010) Small inser -

Monplaisir N, Merault G, Poyart C, Rhoda MD, Craescu C, Vidaud M, Galacteros F, Blouquit Y, Rasia J (1986) Hemoglobin S Antilles: a variant with lower solubility than hemoglobin S and producing sickle cell disease in heterozygotes. Proc Natl Acad Sci U S A 83:9363–9367. https://doi.org/10.1073/pnas.83.24.9363

Montanucci L, Capirotti E, Frank Y, Ben-Tal N, Fariselli P (2019) DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. BMC Bioinformatics 20:335. https://doi.org/10.1186/s12859-019-2923-1

Monzon AM, Carraro M, Chiricosta L, Reggiani F, Han J, Ozturk K, Wang Y, Miller M, Bromberg Y, Capirotti E, Savojardo C, Babi G, Martelli PL, Casadiso R, Katsanis P, Lichtarge O, Carter H, Kousi M, Katsanis N, Andreoliotti G, Moul J, Brenner SE, Ferrari C, Leonardo E, Tosatto SCE (2019) Performance of computational methods for the evaluation of pericentriolar material 1 missense variants in CAGI-5. Hum Mutat 40:1474–1485. https://doi.org/10.1002/humu.23856

Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A 108:E1293–E1301. https://doi.org/10.1073/pnas.111471108

Mort M, Sterne-Weiler T, Li B, Ball EV, Cooper DN, Radivojac P, Sanford JR, Moorey SD (2014) MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. Genome Biol 15:R19. https://doi.org/10.1186/gb-2014-15-1-r19

Mullaney KM, Mills RE, Pittard WS, Devine SE (2010) Small insertions and deletions (INDELs) in human genomes. Hum Mol Genet 19:R131–R136. https://doi.org/10.1093/hmg/ddq400

Mullaney KM, Wong KK, Marciano DC, Katsanis P, King-Crane ER, Ren YA, Lichtarge O, Richards JS (2015) Specific TP53 mutants overrepresented in ovarian cancer impact CNV, TP53 Activity, Responses to nutlin-3a, and cell survival. Neoplasia 17:789–803. https://doi.org/10.1016/j.neo.2015.11.003

Munro D, Singh M (2020) DeMaSk: a deep mutational scanning substitution matrix and its use for variant impact prediction. Bioinformatics. https://doi.org/10.1093/bioinformatics/btaa1030

Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Koijima K, Kawai Y, Yamaguchi-Kabata Y, Yokozawa J, Danjoh I, Saito S, Sato Y, Mimori T, Tsuda K, Saito R, Pan X, Nishikawa S, Ito S, Kuroki Y, Tanabe O, Fuse N, Kuriyama S, Kiyomoto H, Hozawa A, Minegishi N, Douglas J, Kinosita K, Kure S, Yaegashi N, To MJRPP, Yamamoto M (2015) Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat Commun 6:8018. https://doi.org/10.1038/ncomms9018

Nair A, Chung HC, Sun T, Tyagi S, Dobrolecki LE, Dominguez-Vidana R, Kurley SJ, Orellana M, Renwick A, Henke DM, Katsanis P, Schmitt E, Chan DW, Li H, Mao S, Petrovic I, Creighton CJ, Gutierrez C, Jostis J, Tyner JW, Lichtarge O, Lin CY, Zhang B, Scott KL, Hilsenbeck SG, Sun J, Yu X, Osborne CK, Schirf R, Christensen JG, Shields DJ, Rimawi MF, Ellis MJ, Shaw CA, Lewis MT, Westbrook TF (2018) Combinatorial inhibition of PTPN12-regulated receptors leads to a broadly effective therapeutic strategy in triple-negative breast cancer. Nat Med 24:505–511. https://doi.org/10.1038/nm.4507

Neskey DM, Osman AA, Ow TJ, Katsanis P, McDonald T, Hicks SC, Hsu TK, Pickering CR, Ward A, Patel A, Yardy JS, Skinner HD, Giri U, Sano D, Story MD, Beadle BM, El-Naggar AK, Kies MS, William WN, Caulin C, Frederick M, Kimmel M, Myers JN, Lichtarge O (2015) Evolutionary Action Score of TP53 Identifies High-Risk Mutations Associated with Increased Survival and Increased Distant Metastases in Head and Neck Cancer. Cancer Res 75:1527–1536. https://doi.org/10.1158/0008-5472.CAN-14-2735

Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11:863–874. https://doi.org/10.1101/gr.71766

Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814. https://doi.org/10.1093/nar/gkg509

NHLBI Exome Sequencing Project (2011) NHLBI Exome Sequencing Project (2011) NHLBI Exome Sequencing Project

Niroulia A, Urolagin S, Vihinen M (2015) PON-2P: prediction method for fast and reliable identification of harmful variants. PLoS ONE 10:e0117380. https://doi.org/10.1371/journal.pone.0117380

Olatubosun A, Valiatho J, Harkonen J, Thusberg J, Vihinen M (2012) PON-P: integrated predictor for pathogenicity of missense variants. Hum Mutat 33:1166–1174. https://doi.org/10.1002/humu.22102

Oluwole OG, Kuivaniemi H, Abrahams S, Haylett WL, Vorster AA, van Heyden CJ, Kenyon CP, Tabb DL, Favale MW, Summou TA, Ajose A, Olagunmo AO, Rossouw AC, van Hillelengordsberg LS, Carr J, Ross OA, Komolafe MA, Tromp G, Bardien S (2020) Targeted next-generation sequencing identifies novel variants in candidate genes for Parkinson’s disease in Black South African and Nigerian patients. BMC Med Genet 21:23. https://doi.org/10.1186/s12881-020-00953-1

Orenigo CA, Todd AE, Thornton JM (1999) From protein structure to function. Curr Opin Struct Biol 9:374–382. https://doi.org/10.1016/S0959-440X(99)80051-7

Osman AA, Monroe MM, Ortega Alves MV, Patel AA, Katsanis P, Fitzgerald AL, Neskey DM, Frederick MJ, W0 SH, Caulin C, Hsu TK, McDonald TO, Kimmel M, Meyn RE, Lichtarge O, Myers JN (2015a) Wee-1 kinase inhibition overcomes cisplatin resistance associated with high-risk TP53 mutations in head and neck cancer through mitotic arrest followed by senescence. Mol Cancer Ther 14:608–619. https://doi.org/10.1158/1535-7163.MCT-14-0735-T

Osman AA, Neskey DM, Katsanis P, Patel AA, Ward AM, Hsu TK, Hicks SC, McDonald TO, Ow TJ, Alves MO, Pickering CR, Skinner HD, Zhao M, Sturgis EM, Kies MS, El-Naggar A, Perrone F, Licitra L, Bossi P, Kimmel M, Frederick MJ, Lichtarge O, Myers JN (2015b) Evolutionary action score of TP53 coding variants is predictive of platinum response in head and neck cancer patients. Cancer Res 75:1205–1215. https://doi.org/10.1158/0005-8227.CAN-14-2729

Oufay GA, Whyte MP, Gottesman GS, McAlister WH, Eric Gordon J, Harkonen J, Iakoucheva LM, Mooney SD, Radivojac P (2017) When

Pachkov M, Balwierz PJ, Arnold P, Ozonov E, van Nimwegen E (2013) SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. Nucleic Acids Res 41:D214–D220. https://doi.org/10.1093/nar/gks1145

Pagel KA, Pejaver V, Lin GN, Nam HJ, Mort M, Cooper DN, Sebat J, Iakoucheva LM, Mooney SD, Radivojac P (2017) When
loss-of-function is loss of function: assessing mutational signatures and impact of loss-of-function genetic variants. Bioinformatics 33:1389–1398. https://doi.org/10.1093/bioinformatics/btx272

Pagel KA, Antaki D, Lian A, Mort M, Cooper DN, Sebat J, Iakoucheva LM, Mooney SD, Radivojac P (2019) Pathogenicity and functional impact of non-frameshifting insertion/deletion variation in the human genome. PLoS Comput Biol 15:e1007112. https://doi.org/10.1371/journal.pcbi.1007112

Pal LR, Kundu K, Yin Y, Moul J (2017) CAGI4 Crohn’s exome challenge: Marker SNP versus exome variant models for assigning risk of Crohn disease. Hum Mutat 38:1225–1234. https://doi.org/10.1002/humu.23256

Pal LR, Kundu K, Yin Y, Moul J (2020) 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 41:347–362. https://doi.org/10.1002/humu.23933

Parthiban V, Gromiya MM, Schomburg D (2006) CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res 34:W239–W242. https://doi.org/10.1093/nar/gkl190

Parvandeh S, Donehower LA, Katsonis P, Hsu TK, Asmussen JK, Lee K, Lichtarge O (2022) EPMUTESTR: a nearest neighbor machine learning approach to predict cancer driver genes from the evolutionary action of coding variants. Nucleic Acids Res. https://doi.org/10.1093/nar/gka215

Patwa Z, Wahl LM (2008) The fixation probability of beneficial mutations. J R Soc Interface 5:1279–1289. https://doi.org/10.1098/rsif.2008.0248

Pei J, Kinch LN, Otwinowski Z, Grishin NV (2020) Mutation severity spectrum of rare alleles in the human genome is predictive of disease type. PLoS Comput Biol 16(5):e1007775. https://doi.org/10.1371/journal.pcbi.1007775

Pejaver V, Mooney SD, Radivojac P (2017) Missense variant pathogenicity predictors generalize well across a range of function-specific prediction challenges. Hum Mutat 38:1092–1108. https://doi.org/10.1002/humu.23258

Pejaver V, Urresti J, Pagel KA, Lin GN, Nam HJ, Mort M, Cooper DN, Sebat J, Lakoucheva LM, Mooney SD, Radivojac P (2020) Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun 11:5918. https://doi.org/10.1038/s41467-020-19669-x

Peltomaki P, Vasen H (2004) Mutations associated with HNPCC predisposition – Update of ICG-HNPCC/INSiGHT mutation database. Dis Markers 20:269–276. https://doi.org/10.1155/2004/105058

Pereira R, Oliveira J, Sousa M (2020) Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. J Clin Med. https://doi.org/10.3390/jcm9010132

Petersen RC, Aisen PS, Beckett LA, Donohue MC, Gamst AC, Haver DJ, Jack CR Jr, Jagust WJ, Shaw LM, Toga AW, Trojanowski JQ, Weiner MW (2010) Alzheimer’s disease neuroimaging initiative (ADNI): clinical characterization. Neurology 74:201–209. https://doi.org/10.1212/WNL.0b013e3181cbe525

Peterson SM, Pack TF, Wilkins AD, Urm NS, Urban DJ, Bass CE, Lichtarge O, Caron MG (2015) Elucidation of G-protein and beta-arrestin functional selectivity at the dopamine D2 receptor. Proc Natl Acad Sci U S A 112:7097–7102. https://doi.org/10.1073/pnas.1502742112

Petitjean A, Mathe E, Kato S, Ishioka C, Tavtigian SV, Hainaut P, Olivier M (2007) Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. Hum Mutat 28:622–629. https://doi.org/10.1002/humu.20495

Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB (2013) Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet 9:e1003709. https://doi.org/10.1371/journal.pgen.1003709

Pey AL, Stricher F, Serrano L, Martinez A (2007) Predicted effects of missense mutations on native-state stability account for phenotype outcome in phenylketonuria, a paradigm of misfolding diseases. Am J Hum Genet 81:1006–1024. https://doi.org/10.1086/521879

Pires DE, Ascher DB, Blundell TL (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res 42:W314–W319. https://doi.org/10.1093/nar/gku411

Plotkin JB, Kadla G (2011) Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 12:32–42. https://doi.org/10.1038/nrg2899

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20:110–121. https://doi.org/10.1101/gr.097857.109

Ponzoni L, Peñaherrera DA, Oltvai ZN, Bahar I, Ponty Y (2020) Rhapsody: predicting the pathogenicity of human missense variants. Bioinformatics 36(10):3084–3092. https://doi.org/10.1093/bioinformatics/btaa127

Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69:1–14. https://doi.org/10.1086/321275

Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18(Suppl 1):S71–S77. https://doi.org/10.1093/bioinformatics/18.suppl_1.s71

Qi H, Zhang H, Zhao Y, Chen C, Long JJ, Chung WK, Guan Y, Shen Y (2021) MVP predicts the pathogenicity of missense variants by deep learning. Nat Commun 12:510. https://doi.org/10.1038/s41467-020-20847-0

Quan L, Lv Q, Zhang Y (2016) STRUM: structure-based prediction of protein stability changes upon single-point mutation. Bioinformatics 32:2936–2946. https://doi.org/10.1093/bioinformatics/btw361

Quang D, Chen Y, Xie X (2015) DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 31:761–763. https://doi.org/10.1093/bioinformatics/btu703

Quinodoz M, Peter VG, Cisarova K, Royer-Bertrand B, Stenson PD, Cooper DN, Unger S, Superti-Furga A, Rivolta C (2022) Analysis of missense variants in the human genome reveals widespread gene-specific clustering and improves prediction of pathogenicity. Am J Hum Genet 109(3):457–470. https://doi.org/10.1016/j.ajhg.2022.01.006

Rababa’h A, Craft JW Jr, Wijaya CS, Atroooz F, Fan Q, Singh S, Guil- lory AN, Katsonis P, Lichtarge O, McConnell BK (2013) Protein kinase A and phosphodiesterase-4D3 binding to coding polymorphisms of cardiac muscle anchoring protein (mAKAP). J Mol Biol 425:3277–3288. https://doi.org/10.1016/j.jmb.2013.06.014

Raimondi D, Gazzo AM, Roosman T, Lemaerts T, Vranken WF (2016) Multilevel biological characterization of exomic variants at the protein level significantly improves the identification of their deleterious effects. Bioinformatics 32:1797–1804. https://doi.org/10.1093/bioinformatics/btw094

Raimondi D, Tanyalcin I, Ferte J, Gazzo A, Orlando G, Lemaerts T, Roosman M, Vranken W (2017) DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. Nucleic Acids Res 45:W201–W206. https://doi.org/10.1093/nar/gkx390

Ramseney V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30:3894–3900. https://doi.org/10.1093/nar/gkf493
clinical genetics. Clin Chem 50:1974–1978. https://doi.org/10.1373/clinchem.2004.036053
Teng S, Madej T, Panchenko A, Alexov E (2009) Modeling effects of human single nucleotide polymorphisms on protein-protein interactions. Biophys J 96:2178–2188. https://doi.org/10.1016/j.bpj.2008.12.3904
Teng S, Sobitan A, Rhoades R, Liu D, Tang Q (2021) Systemic effects of missense mutations on SARS-CoV-2 spike glycoprotein stability and receptor-binding affinity. Brief Bioinform 22:1239–1253. https://doi.org/10.1093/bib/bbaa233
The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68–74. https://doi.org/10.1038/nature15393
The International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796. https://doi.org/10.1038/nature02168
Thomas PD, Campbell MJ, Kejarival A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13:2129–2141. https://doi.org/10.1101/gr.772403
Tian Y, Pesaran T, Chamberlin A, Fenwick RB, Li S, Gau CL, Chao EY, Lu HM, Black MH, Qian D (2019) REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification. Sci Rep 9:12752. https://doi.org/10.1038/s41598-019-49224-8
Uversky VN, Oldfield CJ, Dunker AK (2008) Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu Rev Biophys 37:215–246. https://doi.org/10.1146/annurev.bioph.37.030207.125924
van Dam S, Vosa U, van der Graaf A, Franke L, de Magalhaes JP (2018) Gene co-expression analysis for functional classification and gene-disease predictions. Brief Bioinform 19:575–592. https://doi.org/10.1093/bib/bbw139
Van Hout CV, Tachmazidou I, Backman JD, Hoffman JD, Liu D, Pandey AK, Gonzaga-Jauregui C, Khalid S, Ye B, Banerjee N, Li AH, O’Dushalaine C, Marketta A, Staples J, Schurmann C, Hawes A, Maxwell E, Barnard L, Lopez A, Penn J, Habegger L, Blumenfeld AL, Bai X, O’Keefe S, Yadav A, Praveen K, Jones ML, Salerno WJ, Chung WK, Surakka I, Willer CJ, Yu J, Garvie CW, Won J, Russell J, Ludwig S, Murray AR, Moresco EM, Beutler B (2015) Rare, damaging DNA variants in CORIN and risk of coronary artery disease: Insights from functional genomics and large-scale sequencing analyses. Circ Genom Precis Med 14:e003399. https://doi.org/10.1161/CIRCGENE.121.003399
Wei P, Liu X, Fu YX (2011) Incorporating predicted functions of nonsynonymous variants into gene-based analysis of exome sequencing data: a comparative study. BMC Proc 5(Suppl 9):S20. https://doi.org/10.1186/1753-6561-5-S9-S20
Vasicek R, Rosenthal P, Lichtarge O, Pejaver V, Radiovic P, Mooney SD, Capriottı E, Bromberg Y, Wang Y, Miller M, Martelli PL, Saveojardo C, Babbı G, Casadiso R, Cao Y, Sun Y, Shen Y, Garg A, Pal D, Yu Y, Huff CD, Tartaglino SV, Young E, Neuhausen SL, Ziv E, Pal LR, Andreoletti G, Brenner SE, Kann MG (2019) 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 40:1612–1622. https://doi.org/10.1002/humu.23849
Waalen J, Beutler E (2009) Genetic screening for low-penetrance variants in protein-coding genes. Annu Rev Genomics Hum Genet 10:431–450. https://doi.org/10.1146/annurev.genom.9.081307.164255
Walters-Sen LC, Hashimoto S, Thrush DL, Reshmi S, Gastier-Foster JM, Axbury C, Pyatt RE (2015) Variability in pathogenicity prediction programs: impact on clinical diagnostics. Mol Genet Genomic Med 3:99–110. https://doi.org/10.1002/mgg3.116
Wang Y, Bromberg Y (2019) Identifying mutation-driven changes in gene functionality that lead to venous thromboembolism. Hum Mutat 40:1321–1329. https://doi.org/10.1002/humu.23824
Wang Z, Mourt J (2001) SNPs, protein structure, and disease. Hum Mutat 17:263–270. https://doi.org/10.1002/humu.22
Wang T, Zhan X, Bu CH, Lyon S, Pratt D, Hildebrand S, Choi JH, Zhang Z, Zeng M, Wang KW, Turer E, Chen Z, Zhang D, Yue T, Wang Y, Shi H, Wang J, Sun L, SoRelle J, McAlpine W, Hutchins N, Zhan X, Fina M, Gobert R, Quan J, Kreutzer M, Arnett S, Hawkins K, Leach A, Tate C, Daniel C, Reyna C, Prince L, Davis S, Purrington J, Bearden R, Weatherly J, White D, Russell J, Sun Q, Tang M, Li X, Scott L, Moreeso EM, McNemey GM, Karlsson Hedestam GB, Xie Y, Beutler B (2015) Real-time resolution of point mutations that cause phenovariance in mice. Proc Natl Acad Sci U S A 112:E440–E449. https://doi.org/10.1073/pnas.1423216112
Wang T, Bu CH, Hildebrand S, Jia G, Siggs OM, Lyon S, Pratt D, Scott L, Russell J, Ludwig S, Murray AR, Moresco EM, Beutler B (2018) Probability of phenotypically detectable protein damage by ENU-induced mutations in the Mutagenetix database. Nat Commun 9:441. https://doi.org/10.1038/s41467-017-02806-4
Wang W, Corominas R, Lin GN (2019) De novo mutations from whole exome sequencing in neurodevelopmental and psychiatric disorders: from discovery to application. Front Genet 10:258. https://doi.org/10.3389/fgene.2019.00258
Wang C, Konecki DM, Marciano DC, Govindarajan H, Williams AM, Wastuwiidyaniytyas B, Bourquard T, Katsonis P, Lichtarge O (2021a) Identification of evolutionarily stable functional and immunogenic sites across the SARS-CoV-2 proteome and greater coronavirus family. Bioinformatics. https://doi.org/10.1093/bioinformatics/btab406
Wang M, Lee-Kim VS, Atri DS, Elowe NH, Yu J, Garvie CW, Won HH, Hadaya JE, MacDonald BT, Trindade K, Melander O, Rader DJ, Natarajan P, Kathiresan S, Kasturi VK, Khera AV, Gupta KMT (2021b) Rare, damaging DNA variants in CORIN and risk of coronary artery disease: Insights from functional genomics and large-scale sequencing analyses. Circ Genom Precis Med 14:e003399. https://doi.org/10.1161/CIRCGENE.121.003399
Wei P, Liu X, Fu YX (2011) Incorporating predicted functions of nonsynonymous variants into gene-based analysis of exome sequencing data: a comparative study. BMC Proc 5(Suppl 9):S20. https://doi.org/10.1186/1753-6561-5-5-S9-S20
Wichmann HE, Kuhn KA, Waldenberger M, Schmelcher D, Schuffenhauer S, Meitinger T, Wurst SH, Lamla G, Fortier I, Burton PR, Peltonen L, Perola M, Metspalu A, Riegman P, Landegren U, Taussig MJ, Litton JE, Frasson MN, Eder J, Cambon-Thomsen A, Bovenberg J, Dagher G, van Ommen GJ, Griffith M, Yuille M, Zatloukal K (2011) Comprehensive catalog of European biobanks. Nat Biotechnol 29:795–797. https://doi.org/10.1038/nbt.1958
Willsey AJ, Fernandez TV, Yu D, King RA, Dietrich A, Xing J, Sanders SJ, Mandell JD, Huang AY, Richer P, Smith L, Dong S, Samocha KE, Tourette International Collaborative G, Tourette Syndrome Association International Consortium for G, Neale BM, Coppola G, Mathews CA, Tischfield JA, Scharf JM, State MW, Heiman GA (2017) De novo coding variants are State MW, Heiman GA (2017) De novo coding variants are
Woolfe A, Mullikin JC, El精致fski L (2010) Genomic features defining exonic variants that modulate splicing. Genome Biol 11:R20. https://doi.org/10.1186/gb-2010-11-2-r20

Worth CL, Preisnser R, Blundell TL (2011) SDM—a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Res 39:W215–W222. https://doi.org/10.1093/nar/gkr363

Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O’Donovan C, Redaschi N, Suzek B (2006) The Universal protein resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 34:D187–D191. https://doi.org/10.1093/nar/gkj161

Wu S, Tian C, Liu P, Guo D, Zhong W, Huang X, Zhang Y, Liu L. (2021) Effects of SARS-CoV-2 mutations on protein structures and intraviral protein-protein interactions. J Med Virol 93:2132–2140. https://doi.org/10.1002/jmv.26597

Xu Q, Tang Q, Katsonis P, Lichtarge O, Jones D, Bovo S, Bajgierni Martelli PL, Casadio R, Lee GR, Seok C, Fenton AW, Dunbrack RL Jr (2017) Benchmarking predictions of allostery in liver pyruvate kinase in CAGI4. Hum Mutat 38:1123–1131. https://doi.org/10.1002/hum.23222

Yadegari F, Majdjudzeh K (2019) In silico analysis for determining the deleterious non-synonymous single nucleotide polymorphisms of BRCA genes. Mol Biol Res Commun 8:141–150. https://doi.org/10.22099/mbrc.2019.34198.1420

Yang J, Roy A, Zhang Y (2013) BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucleic Acids Res 41:D1096–D1103. https://doi.org/10.1093/nar/gkt966

Yang Y, Peng X, Ying P, Tian J, Li J, Ke J, Zhu Y, Gong Y, Zou D, Yang N, Wang X, Mei S, Zhong R, Gong J, Chang J, Miao X (2019) AWESOME: a database of SNPs that affect protein post-translational modifications. Nucleic Acids Res 47:D874–D880. https://doi.org/10.1093/nar/gky821

Yazar M, Ozbek P (2021) In Silico Tools and Approaches for the prediction of functional and structural effects of single-nucleotide polymorphisms on proteins: an expert review. OMICS 25:23–37. https://doi.org/10.1089/omi.2020.0141

Yin Y, Kundu K, Pal LR, Moult J (2017) Ensemble variant interpretation methods to predict enzyme activity and assign pathogenicity in the CAGI4 NAGLU (Human N-acetyl-glucosaminidase) and UBE21 (Human SUMO-ligase) challenges. Hum Mutat 38:1109–1122. https://doi.org/10.1002/hum.23267

Yue P, Moult J (2006) Identification and analysis of deleterious human SNPs. J Mol Biol 356:1263–1274. https://doi.org/10.1016/j.jmb.2005.12.025

Yue P, Li Z, Moult J (2005) Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol 353:459–473. https://doi.org/10.1016/j.jmb.2005.08.020

Yue P, Melamud E, Moult J (2006) SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics 7:166. https://doi.org/10.1186/1471-2105-7-166

Yue WW, Froese DS, Brennan PE (2014) The role of protein structural analysis in the next generation sequencing era. Top Curr Chem 336:67–98. https://doi.org/10.1007/128_2012_326

Zeng Z, Bromberg Y (2019) Predicting functional effects of synonymous variants: a systematic review and perspectives. Front Genet 10:914. https://doi.org/10.3389/fgene.2019.00914

Zeng S, Yang J, Chung BH, Lau YL, Yang W (2014) EFIN: predicting the functional impact of nonsynonymous single nucleotide polymorphisms in human genome. BMC Genomics 15:455. https://doi.org/10.1186/1471-2164-15-455

Zeng Z, Aptekmann AA, Bromberg Y (2021) Decoding the effects of synonymous variants. Nucleic Acids Res 49:12673–12691. https://doi.org/10.1093/nar/gkab1159

Zhang J, Kinch LN, Cong Q, Weile J, Sun S, Cote AG, Roth FP, Grishin NV (2017a) Assessing predictions of fitness effects of missense mutations in SUMO-conjugating enzyme UBE21. Hum Mutat 38:1051–1063. https://doi.org/10.1002/humu.23293

Zhang X, Li M, Lin H, Rao X, Feng W, Yang Y, Mort M, Cooper DN, Wang Y, Wang Y, Wells C, Zhou Y, Liu Y (2017b) regSNPs-splincing: a tool for prioritizing synonymous single-nucleotide substitution. Hum Genet 136:1279–1289. https://doi.org/10.1007/s00439-017-1783-x

Zhang T, Hou L, Chen DT, McMahon PJ, Wang JC, Rice JP (2018) Exome sequencing of a large family identifies potential candidate genes contributing risk to bipolar disorder. Gene 645:119–123. https://doi.org/10.1016/j.gene.2017.12.025

Zhang J, Kinch LN, Cong Q, Katsonis P, Lichtarge O, Savojardo C, Babi B, Martelli PL, Capriotti E, Casadio R, Garg A, Pal D, Weile J, Sun S, Verby M, Roth FP, Grishin NV (2019) Assessing predictions on fitness effects of missense variants in calmodulin. Hum Mutat 40:1463–1473. https://doi.org/10.1002/humu.23857

Zhou J, Fu BQ (2018) The research on gene-disease association based on text-mining of PubMed. BMC Bioinform 19:37. https://doi.org/10.1186/s12859-018-0204-y

Zhou L, Zhao F (2018) Prioritization and functional assessment of noncoding variants associated with complex diseases. Genome Med 10:53. https://doi.org/10.1186/s13073-018-0565-y

Zhou H, Zhou Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11:2714–2726. https://doi.org/10.1101/ps.0217002

Zhou H, Gao M, Skolnick J (2016) ENTPRISE: An Algorithm for Predicting Human Disease-Associated Amino Acid Substitutions from Sequence Entropy and Predicted Protein Structures. PLoS ONE 11:e0150965. https://doi.org/10.1371/journal.pone.0150965

Zhou H, Gao M, Skolnick J (2018a) ENTPRISE-X: Predicting disease-associated frameshift and nonsense mutations. PLoS ONE 13:e0196849. https://doi.org/10.1371/journal.pone.0196849

Zhou Y, Fujiwara K, Mrkrtchian S, Lauscher VM (2018b) Computational Methods for the Pharmacogenetic Interpretation of Next Generation Sequencing Data. Front Pharmacol 9:1437. https://doi.org/10.3389/fphar.2018.01437

Zlotogora J (2003) Penetrance and expressivity in the molecular age. Genet Med 5:347–352. https://doi.org/10.1097/01.gim.000086478.87623.69

Zou J, Yin J, Fang L, Yang M, Wang T, Wu W, Bellucci MA, Zhang P (2020) Computational prediction of mutational effects on SARS-CoV-2 binding by relative free energy calculations. J Chem Inf Model 60:5794–5802. https://doi.org/10.1021/acs.jcim.0c00679

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.