Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations

Benjamin J. Livesey[1] and Joseph A. Marsh[1]*

[1] MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK

*joseph.marsh@igmm.ed.ac.uk

Abstract

To deal with the huge number of novel protein-coding variants identified by genome and exome sequencing studies, many computational variant effect predictors (VEPs) have been developed. Such predictors are often trained and evaluated using different variant datasets, making a direct comparison between VEPs difficult. In this study, we use 31 previously published deep mutational scanning (DMS) experiments, which provide quantitative, independent phenotypic measurements for large numbers of single amino acid substitutions, in order to benchmark and compare 46 different VEPs. We also evaluate the ability of DMS measurements and VEPs to discriminate between pathogenic and benign missense variants. We find that DMS experiments tend to be superior to the top-ranking predictors, demonstrating the tremendous potential of DMS for identifying novel human disease mutations. Among the VEPs, DeepSequence clearly stood out, showing both the strongest correlations with DMS data and having the best ability to predict pathogenic mutations, which is especially remarkable given that it is an unsupervised method. We further recommend SNAP2, DEOGEN2, SNPs&GO and REVEL based upon their performance in these analyses.

Introduction

Many genetic disorders can be attributed to sequence changes in protein-coding regions of DNA, yet pathogenic mutations account for only a tiny fraction of the overall genetic variation seen in humans. A typical pair of unrelated individuals will differ by approximately one nonsynonymous single nucleotide variant (SNV) per protein-coding gene (Rauch *et al*, 2012), while *de novo* mutations lead to roughly one new nonsynonymous SNV per child not observed in either parent (de Ligt *et al*, 2012; Epi4K Consortium *et al*, 2013; Fitzgerald *et al*, 2015; Neale *et al*, 2012). The vast majority of mutations identified by sequencing are of unknown phenotypic consequence, *i.e.* we are unsure if they have significant phenotypic effects or are functionally neutral. Thus, the ability to distinguish damaging variants from those that are benign is of tremendous importance for the diagnosis and treatment of human genetic disease.

In order to prioritise potentially pathogenic variants, many different computational variant effect predictors (VEPs) have been developed. These predictors make use of various protein sequence, structural, evolutionary and biophysical features to produce an effect score for the variant. By far the most commonly used feature is evolutionary sequence conservation and known variation (Table 1). This is the only information used by several methods such as SIFT (Sim *et al*, 2012) and DeepSequence (Riesselman *et al*, 2018). Other predictors integrate additional features including biophysical properties of amino acids, protein functional annotations and epigenetic data (Rentzsch *et al*, 2019). Protein structural information, derived from experimentally determined models, is also used by several methods (Adzhubei *et al*, 2010; Capriotti & Altman, 2011), although there is conflicting information regarding whether its inclusion significantly improves predictor performance (Carraro *et al*, 2017).

While many of these approaches are able to make impressive predictions on test datasets, and are widely applied in both clinical and research environments, there remain a number of unresolved sources of biases and inaccuracies. For example, when employing a supervised machine-learning method, overfitting of the training set can become an issue. Instead of learning general rules, the predictor learns the niche peculiarities and noise of its training set (Srivastava *et al*, 2014). For this reason, machine-learning techniques are usually subject to out-of-sample validation, whereby data not present in the training set is used to verify that the predictor has learned how to classify the data. Furthermore, when benchmarking these predictors with alternative datasets, they should contain as few mutations used during training and validation as possible. Biased representation within these data sets will skew the reported accuracy of methods trained and benchmarked with them (Schaafsma & Vihinen, 2018).

Grimm et al. describe two types of data circularity that can bias the assessment of predictor accuracy (Grimm *et al*, 2015). Type 1 circularity occurs when the data from the training set is re-used for

assessing predictor performance. This can occur due to overlap between commonly used variant databases. The result is a better apparent performance than if a more appropriate validation set were used. Metapredictors (trained using the outputs of other predictors) amplify this issue, as the methods they are built from often use different overlapping training sets. Type 2 circularity results in the weighting of predictor output by biases in the training examples. This can come about in VEPs due to ascertainment biases in the training set (long-studied proteins will have more annotated mutations than recently analysed ones). Another source is the association of certain genes with pathogenicity (*e.g.* many mutations in P53 will be damaging, while other genes may have no pathogenic mutations). Tools that use this information to weight their predictions can achieve excellent results on proteins with annotated pathogenic or benign mutations, but perform poorly when faced with unannotated proteins.

An alternative to computational predictions is experimental characterisation of mutation phenotypes. While this can be extremely time consuming if a separate experiment is required for each mutation, in recent years, an assortment of approaches have been developed for the high-throughput characterisation of mutation phenotypes. Deep mutational scanning (DMS) experiments combine saturation mutagenesis of a protein with a high-throughput functional test and deep sequencing (Fowler & Fields, 2014). The result is a framework, allowing the design of experiments to quantify the functional impact of a huge number of mutations at the same time. DMS experiments could potentially be hugely valuable for variant prioritisation, allowing direct identification of damaging human variants on a large scale (Majithia *et al*, 2016; Matreyek *et al*, 2018). DMS experiments can also be tailored to the specific definition of protein fitness required - something which computational methods are not able to account for (Harris *et al*, 2016). Even the best performing predictors struggle with more complex biological concepts such as allosteric regulation (Xu *et al*, 2017).

In addition to directly identifying damaging variants, another major benefit of DMS experiments is that they produce large variant-effect datasets that can be used to benchmark and assess the performance of VEPs. These are fully independent from any training and testing data used by the phenotype predictors, with one exception (Gray *et al*, 2018). Previous studies have found that using DMS datasets to benchmark computational predictors resulted in reduced predictive power compared to other commonly used datasets, suggesting that these predictors may not be as accurate for human variants as previously reported (Mahmood *et al*, 2017). The Critical Assessment of Genome Interpretation (CAGI) experiment, which aims to drive innovations in VEPs frequently assesses predictors against novel unseen datasets (Hoskins *et al*, 2017) including those derived from DMS experiments.

In this study, we have taken advantage of the large number of DMS experiments that have now been published for a variety of diverse proteins from different organisms. First, we have used these datasets to perform an independent assessment and comparison of many different VEPs. Second, we have compared the ability of DMS experiments and VEPs to directly identify pathogenic human mutations.

Results

Overview of DMS datasets and variant effect predictors used in this study

To identify DMS datasets, we performed a literature search for papers presenting such experiments with available data. Using search terms such as ‘deep mutational scan’, ‘fitness landscape’, ‘massively parallel mutagenesis’ and ‘saturation mutagenesis’, we identified 31 viable DMS datasets (Table 2). As shown in Fig 1, human proteins were the most numerous targets for these DMS experiments. *Saccharomyces cerevisiae* and *Escherichia coli* were also highly represented as they endogenously produce a number of model proteins, are easy to culture and maintain, and are amenable to several effective assays for protein activity (*e.g.* growth rate and two hybrid). Proteins from viruses were also represented from studies investigating viral adaptation through massively parallel mutagenesis techniques.

There was considerable variation in functional assays applied between the DMS studies. Growth rate of yeast was a most common technique, and was applied to multiple human proteins by knocking out the yeast orthologue and replacing it with the human gene that is capable of rescuing the null strain (Weile *et al*, 2017). Viral replication assays, performed by quantitative sequencing after a certain time point, were applied to all of the viral proteins (Haddox *et al*, 2016; Doud & Bloom, 2016; Lee *et al*, 2018; Wu *et al*, 2015). Survival assays involved placing the organism in hostile conditions where the target protein confers an advantage such as antibiotic resistance (Dandage *et al*, 2018; Firnberg *et al*, 2014; Stiffler *et al*, 2015; Jacquier *et al*, 2013; Deng *et al*, 2012). Two-hybrid assays allow protein-protein interactions to be analysed, while fluorescence can be used to investigate enzyme activity, protein stability or transcriptional pathway activation (Starita *et al*, 2015; Bandaru *et al*, 2017; Kitzman *et al*, 2015). Phage-display assays allow a number of protein attributes to be tested *ex vivo* by externalising the protein of interest followed by selection based on its attributes (Starita *et al*, 2015). The *E. coli* toxin ccdB was assayed by reverse survival, investigating its ability to restrict cell growth (Adkar *et al*, 2012).

Each study also varied in the coverage of possible single amino acid substitutions across the entire protein (Fig 1). Many of the studies included only those mutations that were possible by introducing a single nucleotide change, reducing potential coverage of all possible amino acid substitutions by around 70%. Some studies focused on specific regions of the target protein. In addition, most studies excluded low confidence mutants from their data, *i.e.* those with exceptionally low sequencing counts. For inclusion in this analysis, we required at least 5% coverage of all possible mutations in order to prevent unrepresentative low coverage data from skewing the results.

The computational VEPs used in this study were found using a number of approaches, including the OMICtools database (Henry *et al*, 2014), identifying tools tagged with 'variant effect prediction' and searching for 'protein variant effect prediction' and 'protein phenotype predictor' using standard internet search engines. Priority was given to tools that featured either a web interface or an API that could be queried for thousands of mutations simultaneously. We also made use of the dbNSFP (Liu *et al*, 2016) database of pre-calculated predictions from multiple VEPs for the human genome (downloaded 2020-02-12). We split the predictors into four broad categories, based on the way in which they make predictions:

1. *Supervised predictors*. These predictors use a machine learning technique that relies on learning from labelled examples, in particular datasets of known or suspected pathogenic and benign variants. Different predictors make use of a variety of different machine learning approaches, *e.g.* support vector machines and random forest algorithms.

2. *Unsupervised predictors*. These predictors make use of an unsupervised machine learning technique, *i.e.* they are not trained using labelled pathogenic and benign variants. Instead, they rely mostly on evolutionary conservation from multiple sequence alignments. This includes unsupervised clustering techniques, hidden Markov models and generative models.

3. *Empirical predictors*. These predictors do not make use of any machine learning techniques, instead making an empirical calculation using the input data. This category also includes amino acid substitution matrices and many evolutionary conservation metrics. Along with the unsupervised predictors, they should be free from any training bias.

4. *Metapredictors*. These predictors integrate other VEP results as input features, although many also use additional features. The metapredictors used in our study are nearly all trained using a supervised learning approach, with one exception (Ionita-Laza *et al*, 2016). To qualify for this category in our classification a predictor must include at least two other VEPs as input features, not including substitution matrices or simple conservation metrics (such as GERP, PhyloP or SiPhy).

Among the DMS datasets, there are several instances of the same protein being investigated in different studies by different groups. Specifically, there are four independent datasets for β-lactamase (bla) and two each for UBI4, PTEN and BRCA1. There are also two datasets for the influenza protein

HA, but these were from different strains, so not directly comparable. To assess the reproducibility of DMS and its viability as a benchmark, we calculated the Spearman's correlation coefficient between the functional scores of each DMS set in the same protein. Our results (Table EV1) demonstrate a range of correlations from 0.94 (bla(a)/(b)) to 0.34 (PTEN(a)/(b)). The average correlations observed over all pairs of analyses was 0.66. Some level of variance is expected due to differences in experimental method, fitness assays and conditions between experiments. Overall the moderately high correlations suggest that DMS scores constitute a reasonably robust benchmark despite differing experimental conditions. We can also treat this correlation as a rough guide for how well we could expect a 'perfect' computational predictor to perform against DMS data from these experiments.

We also assessed the correlation between different DMS datasets generated by the same studies. The purposes of these assays varied and included controls with no selection pressure, biological replicates, incrementally differing conditions and different fitness assays (Table EV2). Incremental changes in conditions tended to result in high correlations while larger alterations to conditions, assay type or protein partners resulted in much lower correlations between the datasets. Nonselective controls produced low correlations while comparison of positive to negative selection assays produced a negative correlation. These results indicate that interpretation of DMS results depends to some extent on the exact fitness assay. However, certain mutations (*e.g.* those that destabilise the protein) are likely to always have an impact on fitness if a selection pressure is present.

Assessment of variant effect predictors using DMS data

Where possible, we applied every computational predictor to each protein in the DMS datasets, substituting every possible amino acid at all positions. Some predictors failed to generate results for some proteins; this can occur due to an insufficiently deep multiple sequence alignment, mapping errors or other causes depending on the predictor. In order to get a measure of relative performance for each predictor, we calculated the Spearman's rank correlation between the independent DMS scores for each protein and the predictions of every VEP (Fig 2). We also performed the same analysis using Kendall's Tau (Fig EV1) which produced only minor changes in predictor ranking and lower average correlations.

Given the large number of predictors that are specific to humans, we split this analysis up into human (Fig 2A) and non-human (Fig 2B) proteins. The top-performing predictor for each protein is labelled on the plot, while the full set of correlations are provided in Tables EV3-4. Table 3 shows the relative ranking of each predictor using a rank score that combines the rankings for all proteins in the human, yeast, bacterial and viral datasets (Table EV5 shows the same using Kendall's Tau instead).

DeepSequence was the overall top performing method for predicting DMS results in the human proteins, showing the highest correlations out of all predictors for ADRB2, CALM1 and PTEN(b), and ranking within the top five predictors for 7 of the 13 DMS datasets. It also had by far the highest rank score. To assess the statistical significance of this, we used a bootstrapping approach and re-calculated the ranking by re-sampling all DMS datasets with replacement 1000 times. Strikingly, we found that DeepSequence always ranked the highest, showing that it is significantly better than all other predictors ($p<0.001$). DeepSequence also ranked best for bacterial proteins, being the top predictor for six proteins: infA, bla(b), bla(c), ccdB, haeIIIM and GmR. In contrast, DeepSequence produced only a moderate rank score for yeast proteins, with a high coefficient of variation. This was largely due to poor performance on the ubiquitin (UBI4) datasets, which reduced the overall rank score considerably. If the UBI4 datasets are excluded from the analysis, then DeepSequence becomes the second-highest ranked predictor for yeast proteins. Interestingly, DeepSequence performs poorly for viral proteins, ranking second to last out of all predictors tested. This is consistent with the original publication, where the creators report poor performance on viral proteins due to insufficient sequence diversity within the alignments used (Riesselman *et al*, 2018).

Among the other predictors, certain supervised approaches were particularly notable. SNPs&GO (Capriotti *et al*, 2013) ranked 2nd for human and 1st for yeast proteins, although its predictions were relatively poor for non-eukaryotic (bacterial and viral) proteins. SNAP2 (Hecht *et al*, 2015) also

performed well, ranking 3rd for human proteins and 2nd for yeast. DEOGEN2 (Raimondi *et al*, 2017), a human-specific predictor came 4th. SuSPect (Yates *et al*, 2014) showed good performance across most groups, ranking 5th for humans 3rd for bacteria and 2nd for viral proteins. REVEL was the only metapredictor to show notable performance, ranking 6th overall and having the highest correlation with the SUMO1 DMS data.

Certain predictors incorporate features derived from experimentally determined protein structures into their predictions. Specifically SNP&GOs3D and S3D-PROF (Capriotti & Altman, 2011) require a PDB structure to be provided in order to make their predictions and use features representing the 3D environment of the mutation. Other predictors such as PolyPhen-2 (Adzhubei *et al*, 2010), DEOGEN2 and MPC derive some features from experimental structures, but are still capable of making predictions without them. While these methods ranked average-to-high and achieved a number of top correlations with the DMS data, overall they do not perform better than the top-performing sequence-based methods. We also find that, in proteins with partial coverage of high-resolution structures, the difference in performance between areas of structural coverage and areas with no coverage is comparable between structural methods and pure sequence-based methods (Fig EV2). This may be due to regions without structures being more likely to be disordered and less conserved, and thus harder to characterise by conservation metrics.

Of all the predictors, FATHMM (Shihab *et al*, 2013) produced the most significant outlier, generating predictions with by far the highest correlation for P53, but having low correlations for all other proteins, resulting in an overall low rank score with a high coefficient of variation. The explanation for this is unclear, but it may be due to overfitting of the predictor for specific proteins, given the enrichment of P53 mutations in human disease databases compared to many of the other proteins in this study.

Different DMS datasets varied greatly in their correlations with the computational predictors. In particular, BRCA1, CALM1 and TPK1 among the human proteins, and ccdB, Cas9 and env among the non-human proteins showed low correlations, even from the best predictors. As far as we can tell, this effect appears to be unrelated to protein coverage, dataset size or experimental methodology. For example, UBE2I, SUMO1, TPK1 and CALM1 were all studied by the same group using the same approach (growth rate in yeast) (Weile *et al*, 2017), yet UBE2I and SUMO1 show markedly higher correlations with all predictors than the others. Viral proteins also showed low correlations, and in fact, the simple BLOSUM62 substitution matrix (Henikoff & Henikoff, 1992) was the most highly correlated with the env dataset when using Kendall's Tau (second highest when using Spearman's). This indicates that the inclusion of typical training features are of less use when predicting the fitness of viral proteins, likely due to lack of viral representation in training sets and lack of viral sequence diversity in many databases used to generate multiple sequence alignments. Viral proteins may also be more likely protein to undergo adaptive evolution, thus potentially confounding conservation-based approaches.

It is also interesting to note that, despite that fact that most of the predictors used in this study are human-specific, the top-ranking predictors for the human DMS datasets tend to be general predictors applicable to proteins from all species. For example, for the human DMS datasets, only one of the top five predictors is specific to humans, whereas many of the lowest ranked predictors are human-specific. An important contributing factor to this may be overfitting against human mutation datasets for some predictors, which causes them to perform poorly against independent experimental phenotype measurements. In addition, several of the worst predictors are also based upon nucleotide-level constraint (GERP++ (Davydov *et al*, 2010), SiPhy (Garber *et al*, 2009), phastCons (Siepel & Haussler, 2005) and fitCons (Gulko *et al*, 2015)). These predictors ranked even lower than the simple BLOSUM62 and Grantham (Grantham, 1974) substitution matrices, suggesting that such approaches are poorly suited to predicting the protein-level effects of mutations.

Many DMS experiments included amino acid substitutions that are not possible by single nucleotide changes, *i.e.* they are technically not missense variants. Some VEPs do not produce predictions for these mutations, particularly those that take nucleotide-level substitutions into consideration. Therefore, we repeated our analysis, limiting our predictions to only those amino acid substitutions

that are possible by single nucleotide changes (Table EV6). The rankings remain broadly similar and the top-ranking method did not change for any group.

Identification of pathogenic human mutations using DMS data and computational variant effect predictors

We next investigated the ability of both DMS experiments and VEPs to distinguish pathogenic human missense mutations, taken from the ClinVar database (Landrum *et al*, 2014), from missense variants observed in the human population, taken from gnomAD (Karczewski *et al*, 2019). While some gnomAD variants may be damaging under certain circumstances (*e.g.* if associated with recessive, late-onset or incomplete penetrance disease), we assume that the vast majority of them should be non-pathogenic, and therefore refer to them as "putatively benign". Of the 11 human proteins with DMS datasets, 7 have known pathogenic or likely pathogenic missense variants in ClinVar as of 2019-10-25 (93 for BRCA1, 31 for HRAS, 189 for P53, 108 for PTEN, 9 for CALM1, 5 for TPK1 and 2 for MAPK1). For CALM1 and TPK1, we identified additional pathogenic missense mutations in the literature (Banka *et al*, 2014; Crotti Lia *et al*, 2013; Jensen *et al*, 2018; Nomikos *et al*, 2018; Zhu *et al*, 2019), leading to a total of 19 for CALM1 and 8 for TPK1. MAPK1 has too few recorded pathogenic missense variants to include in this analysis.

For each predictor, we plotted a receiver operating characteristic (ROC) curve for classification performance on every protein, identifying pathogenic ClinVar mutations as true positives and the putatively benign gnomAD mutations as true negatives (removing any ClinVar mutations from the gnomAD set). We then calculated the area under the curve (AUC) for each plot as a measure of that predictor's performance in classifying the data (Fig 3). We also calculated the precision recall AUCs (Fig EV3). Descriptions of each DMS dataset displayed in Fig 3 are provided in Table EV7.

In the ROC analysis, an experimental DMS metric performed better than any of the 46 VEPs for four of the six human proteins, and ranked relatively high for the remaining two (CALM1 and PTEN). To determine the significance of the performance of the DMS data, we used a bootstrapping approach and individually re-sampled the gnomAD and ClinVar datasets with replacement 10,000 times, re-calculating the AUC scores with the new data. DMS ranked first in 9202 trials, while DeepSequence came top in 600 trials, REVEL in 154, MutPred in 34, SNPs&GO in 7, SIFT4G in 2 and PhD-SNP in 1. Thus, we cannot quite state at this point that DMS is significantly better than all computational predictors together ($p = 0.080$). However, it clearly ranks higher than all VEPs in our analysis, and is significantly better than all except DeepSequence. Very similar results are observed for the precision recall AUCs (Fig EV3), except that the TPK1 DMS dataset changed from ranking 1st to 6th.

The DMS results for CALM1 and TPK1 were generated by the same group using the same method, assessing the effects of mutations on growth rate in a yeast system (Weile *et al*, 2017). The data processing pipeline used in this study penalised 'hypercomplementing' variants (*i.e.* those with fitness greater than the wild type) by setting the fitness to the reciprocal of the measured value. These are labelled as 'flipped' in Fig 3. Interestingly we found that, while these 'flipped' DMS results show a better correlation with the outputs of VEPs than the raw DMS data (Tables EV8-9), the raw scores are better for directly identifying pathogenic variants (Fig 3). This suggests that VEPs in general tend to be predictive of a perturbation away from wild type activity (regardless of whether it is an increase or decrease), whereas only a decrease in activity is predictive of disease, at least for these two proteins. This is consistent with a recent observation that beneficial effects on protein function, as measured by DMS experiments, were predicted less well than detrimental effects for all four tested VEPs (Reeb *et al*, 2019).

While the primary objective of this study is to compare the DMS datasets to the VEPs, it is also interesting to observe the relative performances of the different computational predictors in terms of directly identifying pathogenic mutations for the six human proteins in Fig 3. This comparison is limited to some extent by the fact that there is likely some overlap between the mutations used to evaluate the predictors here, and the mutations originally used to train some of the supervised

predictors and metapredictors. In the regard, it is especially interesting to see that the unsupervised predictor DeepSequence again stood out amongst the VEPs, ranking 1st for TPK1 and CALM1, 2nd for HRAS, and within the top 11 predictors for all remaining proteins. This is considerably better and more consistent performance than any of the other computational predictors.

A few supervised predictors and metapredictors performed well. While this could in part be due to overfitting, for those predictors that showed strong correlations with the DMS data, we can assume that the effects of this should be fairly small. SNAP ranked 3rd in the DMS analyses, and also showed strong predictions of disease mutations in PTEN (1st) and CALM1 (3rd). DEOGEN2 ranked 4th in the DMS analyses and performed well for P53 (3rd), CALM1 (4th) and PTEN (4th). REVEL, which was the best performing metapredictor against the DMS data, showed consistently good predictions, ranking 2nd for P53 and TPK1, and in the top 9 for all but CALM1. VEST4 ranked 1st for P53, 2nd for CALM1 and 3rd for TPK1, but was not as highly ranked in the DMS analysis (9th overall). SNPs&GO, which ranked 2nd in the human DMS analysis, performs well for HRAS (3rd) and CALM1 (5th), and never ranked worse than 17th.

Some of these predictors did perform poorly on certain targets. For example, DEOGEN2 ranked last, by far, for BRCA1. Interestingly, however, the relative performance of DEOGEN2 improved markedly if only predictions of mutations with DMS measurements, which cover primarily just the RING domain of BRCA1, are considered (Fig EV4). This appears to be due to DEOGEN2 assigning extremely different weights to different domains in BRCA1, thus obscuring good predictions when analysing the entire protein. We also investigated other predictors with low AUCs for additional domain-specific effects. A further three datasets which showed a similar pattern were MPC on P53, VEST4 on BRCA1 and PROVEAN on BRCA1, which are all highlighted in Fig EV4.

Discussion

The number of available genome and protein sequences has increased tremendously in the last decade due to advances in next-generation sequencing technologies. In this wealth of new data, we have discovered a large number of previously unseen coding variants of unknown functional significance. To assist us in analysing this new data, computational predictors have been developed, but the training and evaluation of these predictors often suffer from biases. DMS experiments provide an ideal benchmark for testing predictors, ensuring that none of the training data is included in the evaluation. The availability of a large number of such experimental DMS datasets has facilitated this study.

We are aware that numerous technical and computational factors can impact the quality of data from DMS studies. These can stem from experimental procedure, and thus be assessed through reproducibility in biological replicates, or measurement uncertainty assessed by technical replicates. The largest source of error from DMS is encountered in the sequencing stage, where next-generation sequencing typically misreads between 1/100 and 1/1000 bases incorrectly (Ma *et al*, 2019). Many groups adopt a barcoding strategy to address this issue, so that a multi-base unique artificial sequence is associated with each variant. In addition, reads below a certain quality threshold are rejected and variants which are present at a rate below a given detection threshold are removed. Several groups provide both their full fitness scores and a filter for high quality results (Mighell *et al*, 2018; Starita *et al*, 2015). In these cases, we find that the filtered high quality results have a higher average correlation with the VEPs (Tables EV8-9), as well as superior predictive power for disease mutations (Fig 3).

Of the 46 different predictors evaluated in this study, we find that a single program, DeepSequence, clearly stands out from all of the others, both in terms of performance, and in terms of methodology. DeepSequence showed the strongest correlations with the DMS data in humans and bacteria and was the top predictor of human disease mutations. Most machine learning methods make use of several features, often including some measure of sequence conservation at the site of interest, and then learn the patterns of these features that result in a mutation being classified as damaging or benign. DeepSequence makes use of deep generative models to integrate factors from the entire sequence at once, rather than only one or a few sites. This type of problem is largely intractable for traditional machine learning, given the number of parameters involved; however, DeepSequence overcomes this

by learning the latent factors underlying the protein sequence. This approach also produces advantages in terms of the biases inherent in supervised methods. We can expect a machine learning method confronted with an example it was trained on to correctly classify it correctly most of the time, producing an unrepresentative assessment of its accuracy. DeepSequence makes use of multiple sequence alignments and never sees labelled protein data, resulting in scores that are not biased by training examples. This is not to say that DeepSequence is a completely unbiased method, however. The scores which are generated depend entirely upon the database from which multiple sequence alignments are drawn. If certain sequences are under-represented, then predictions for those will be lower quality, such as the results we observe for viral proteins drawn from the UniRef100 database. The success that DeepSequence has achieved in predicting mutations effects for human proteins show that deep generative models may well be the way forward in this field, removing the reliance on labelled datasets for making predictions.

Most predictors, supervised or otherwise, undergo hyperparameter optimisation, a process to tweak internal variables such as learning rate, network architecture or regularisation in order to obtain better performance. This process invariably involves repeatedly testing the predictor's performance against a certain 'test' dataset, and has potential to introduce another source of bias, even into unsupervised methods. For example, some of the DMS datasets used in this study were used to evaluate DeepSequence in the original study, and it is conceivable that some level of hyperparameter optimisation influenced its success here. Analysis of this bias is beyond the scope of this study, but we regard it as relatively minor compared to training biases of supervised methods.

One of the VEPs we assessed, Envision is trained with a supervised learning approach using DMS data rather than labelled pathogenic and benign variants. This method uses a number of the same DMS sets we used in this analysis for training (BRCA1(a), HSP82, UBI4(a and b), PAB1 and bla(a)). Thus, the ranking of this method in Table 3 is almost certainly subject to training bias. It is interesting, however, that despite this advantage Envision only produces moderate overall performance for human DMS datasets (although it does rank 1st for TPK1). In terms of predicting pathogenic missense mutations, Envision performs well for BRCA1 (ranking 3rd) and P53 (ranking 4th), but its performance is unremarkable for the other proteins. Notably, although Envision was not trained on a P53 dataset, it was evaluated using one (although not the same DMS dataset used in this study). While the approach used by Envision is innovative, assessing its performance with DMS has the same caveats as assessing performance of other supervised VEPs using pathogenic mutation databases. Thus it is notable that, despite this advantage, Envision showed only modest performance against the DMS data.

Certain DMS experiments appears to show outstanding performance at identifying disease mutations. It is interesting to compare performance with respect the experimental phenotypes used, as the utility of an experimental phenotype for identifying pathogenic mutations should be related to the mechanism by which mutations cause disease. We note that those DMS experiments based upon competitive growth assays appear to perform particularly well, ranking above all computational predictors for three of the four proteins where they are available. For BRCA1, where there are DMS datasets based upon three different experimental phenotypes, the growth-rate-based assay (Findlay *et al*, 2018) performs much better than those based upon yeast two-hybrid or E3 ubiquitin ligase activity (Starita *et al*, 2015). Growth rate is likely to be a very general experimental phenotype that will reflect any loss of function occurring at a molecular level. In contrast, if some of the pathogenic BRCA1 mutations acted by some mechanism other than perturbation of its interaction with specific binding partners (BARD1) or disrupting E3 activity, this could explain the underperformance of the DMS data based upon these alternate phenotypes. Interestingly, however, the HRAS DMS data, which is also superior to all computational predictors, is based upon a two-hybrid probe of its interaction with RasGAP (Bandaru *et al*, 2017), suggesting that disruption of this interaction is reflective of the molecular mechanisms underlying disease.

PTEN is also noteworthy, as it too has different DMS datasets available based upon different experimental phenotypes. The screen for the PTEN(b) dataset assesses the disruption of an artificial gene circuit in yeast, essentially probing phosphatase activity. This dataset is superior to all but four VEPs, suggesting it is reasonably reflective of molecular disease mechanisms. In contrast, the

phenotypic screen for PTEN(a) measures protein abundance in the cell by fluorescence of EGFP bound to the protein (Matreyek *et al*, 2018). This technique, called VAMP-seq, identifies thermodynamically unstable variants; however, this may fail to capture disease mechanisms acting through interaction disruption and loss or gain of function unrelated to destabilisation. Indeed, it was noted in this study that dominant-negative variants were not significantly different from wild type, consistent with our previous observation that dominant-negative mutations tend to be very mild at the protein structural level (McEntagart *et al*, 2016). Thus, great care must be taken when selecting an experimental phenotype. In the absence of a better phenotypic assay specifically related to a known disease mechanism, experiments based upon growth may be the most general way of probing loss of protein function, and thus the most useful for predicting disease.

Our results in analysing the predictive capability of DMS datasets largely recapitulate the results presented in the original studies. The CALM1 dataset (Weile *et al*, 2017) is reported to have superior precision-recall performance than PolyPhen-2 and PROVEAN, which we also find (for the raw scores rather than the flipped scores). The TPK1 dataset (Weile *et al*, 2017) allowed complete separation of the neutral and disease alleles as did PolyPhen-2 and PROVEAN, but only after additional filtering for recessive disease alleles, which we did not perform. The BRCA1(a) dataset (Starita *et al*, 2015) is used by the authors to train a model to predict homology-direct DNA repair rescue; however predictions are primarily made outside of the region of DMS coverage, which we are unable to assess. BRCA1(b) (Findlay *et al*, 2018) is reported by the authors to separate pathogenic and benign mutations in ClinVar almost perfectly, a result which we also see in our analysis. The PTEN(a) (Matreyek *et al*, 2018) dataset is stated to identify upwards of 90% of PTEN pathogenic variants, although no false positive rate is given since no PTEN variants were officially classified as benign. Again, our results are similar, given the high precision-recall AUC of the PTEN(a) dataset, but the considerably lower-ranked ROC AUC indicates a significant false positive rate. Finally the PTEN(b) authors (Mighell *et al*, 2018) employed an approach similar to us, using gnomAD variants to stand-in for benign substitutions. Their results indicate that their data has a superior positive predictive value than PROVEAN, SIFT and PolyPhen-2, which we also find.

The two most commonly used VEPs are probably PolyPhen-2 and SIFT, which are both still very widely utilised in variant prioritisation. Neither showed exceptional performance in this study, ranking 14th and 25th against the human DMS data (although SIFT4G, a genomic-conservation based implementation of the SIFT algorithm (Vaser *et al*, 2016) ranked 9th). They also both tended to rank in the middle compared to all other predictors for identifying human pathogenic mutations. Therefore, we recommend other VEPs based upon our analyses. Unfortunately, DeepSequence is very computationally intensive and could be quite difficult for a typical end user to run. It also does not have defined disease thresholds; these would need to be assessed on a protein-by-protein basis, likely by analysis of putatively benign (*e.g.* gnomAD) variants. We therefore highlight SNAP2, DEOGEN2 and SNPs&GO, which also tended to perform well against both the DMS and human mutation datasets, and have simple-to-use web interfaces. We further recommend REVEL - although it lacks a web interface, it has been pre-calculated for all human chromosomes and is available online to download. We suggest that these methods would make good choices for routine variant prioritisation. Importantly, however, they all showed large variation in their performance between different proteins, suggesting that one should still not rely too much on the results of any single predictor.

While evolutionary conservation is widely accepted to be the most predictive feature used in variant effect prediction, some VEPs also integrate features derived from experimentally determined protein structures (PolyPhen-2, S3D-PROF, SNP&GOs3D, DEOGEN2 and MPC). It is interesting that the inclusion of protein structural models did not appear to be particularly useful for the VEPs. In principle, since disease mechanisms can often be explained by protein structural effects (Steward *et al*, 2003), one might expect that protein structure should be useful. It may be that the value of evolutionary information simply dwarfs any contribution from the inclusion of structure, *i.e.* if a mutation is damaging at a structural level, this is likely to be reflected in the evolutionary conservation of that residue. Moreover, many pathogenic mutations are not highly damaging at a protein structural level, *e.g.* those associated with a dominant-negative effect in protein complexes (Bergendahl *et al*, 2019) or those that affect transcription factor binding specificity (Williamson *et al*,

2019). It is possible that future strategies that take into consideration the diverse molecular mechanisms underlying human genetic disease and the unique structural properties of individual proteins will be able to make better use of the huge amount of protein structural data now available.

The value of DMS data for directly identifying pathogenic mutations is especially exciting, based on the results we observed here. Given the proper choice of experimental phenotype, DMS experiments are likely to be better than (or at the very least competitive with) the best computational VEPs. The applicability of DMS data for direct variant prioritisation is currently limited by the small fraction of human protein residues for which DMS experiments have been performed. In the coming years, as more proteins are studied and experimental strategies are improved, we expect that the utilisation of such data for the identification of damaging variants will become routine.

Methods

Selecting DMS datasets for correlation analysis

Most of the DMS studies analysed provided multiple fitness maps for the protein of interest. Depending on the study, this was due to replicates in differing conditions (*e.g.* multiple antibiotic concentrations), different functional assays or quality filtering of the results. As our interest was to see how well VEPs could replicate the results of DMS experiments, for proteins with multiple datasets available from a single study, we selected the fitness map with the highest average Spearman correlation to all predictors to assess in Fig 2 and Table 3. Where a quality threshold was given, separating high and low quality results, we tested all results, and high quality filtered results. We did not investigate imputed results or those generated by predictive models trained on the DMS fitness maps.

Structure selection

The SNP&GOs3D VEP along with S3D-PROF require a protein structure to be provided in order to generate results. Where possible, we selected an X-ray crystallography structure with a resolution ≤2.5Å and selected the structure with greatest coverage of the DMS results for that protein. Otherwise we selected the highest resolution structure available. A full list of the structures and chains used for these predictors is provided in Table EV10.

Calculating rank scores

Rank score is defined as the mean, normalised correlation over all proteins, given by the following formula:

$$R = \frac{\sum_m \frac{c - c_{min}}{c_{max} - c_{min}}}{m_x}$$

where c is each correlation for a specific protein, c_{min} is the minimum correlation for each protein and c_{max} is the maximum correlation for each protein. This represents the correlation, normalised to a scale between 1 for the highest ranking method and 0 for the lowest. This is then summed across all proteins (m) for the same method, and divided by the number of proteins for which this method generated a result (m_x), in order to normalise for instances where a predictor failed to generate results for a certain protein. Where multiple DMS datasets are present for a single protein, we averaged the normalised correlations of each predictor between these datasets, and treated the resulting values as scores from a single protein.

Coefficient of variation is calculated from the normalised correlations, before the mean is taken. It is the standard deviation of these values across all proteins, divided by the mean. This represents the variation in predictor rank between different proteins.

It should be noted that rank scores are only comparable within the set of proteins that were used to calculate it, nor does it convey any information about predictor accuracy. The rank score metric can only be used for relative ranking within a set of proteins.

Human mutation datasets

Data was retrieved from gnomAD v2.1 by searching for each of the human genes at https://gnomad.broadinstitute.org. Because CALM1 had only 8 missense variants in gnomAD v2.1, we also included an additional 3 missense variants from gnomAD v3.0 (for CALM1 only). We did not filter for allele frequency. Each gene was also searched for in the ClinVar database at https://www.ncbi.nlm.nih.gov/clinvar. Data was filtered so that only missense mutations labelled as 'pathogenic' or 'likely pathogenic' were present.

Plotting ROC and precision recall curves

To plot the ROC curves, mutations present in the gnomAD dataset were taken as true negatives, while mutations present in the ClinVar dataset were taken as true positives. Mutations present in both sets were removed from the gnomAD set. The 'roc_curve' and 'auc' functions for the sklearn Python package were used to calculate the true positive rate (TPR) and false positive rate (FPR) and the AUC. As some predictors utilise inverse metrics and thus produce an AUC under 0.5, we multiplied the predictions of all such methods by -1 to bring the value above 0.5; this is equivalent to inverting the TPR and FPR. Precision-recall curves were calculated using the 'precision_recall_curve' and 'auc' functions from the sklearn package. A list of methods with inverted scores was retained from the ROC calculations. The scores from these methods were deducted from one to retain comparability. As precision-recall curves are sensitive to class balance, we removed methods with less-than complete coverage of the DMS mutations within the ClinVar and gnomAD datasets from the analysis. We also plotted individual curves for DMS assays in the same protein with differing coverage of the available ClinVar and gnomAD mutations.

Bootstrapping

To calculate statistical significance, we utilised a bootstrapping methodology and applied it to both the VEP ranking analysis using DMS data and the ROC curve calculation. For the ranking analysis, we re-sampled mutations from each protein with replacement 1000 times and re-calculated the rank scores. Our p-value for the top-ranking method was therefore the number of times it did not produce the top rank score, divided by 1000. The ROC curve bootstrapping was carried out using the same method with 10,000 replicates, except the ClinVar and gnomAD mutations were sampled individually to retain class balance and ensure that there was no chance one class could be lost from the analysis. The p-value for one method performing significantly better than another was the number of times it underperformed the second method, divided by 10,000.

Data availability

Datasets containing all variant effect predictions and DMS measurements used in this study are available at https://doi.org/10.7488/ds/2800 for the variants from all organisms, and https://doi.org/10.7488/ds/2799 for the human pathogenic and putatively benign variants.

Acknowledgements

JM is supported by an MRC Career Development Award (MR/M02122X/1) and is a Lister Institute Research Prize Fellow. BL is supported by the MRC Precision Medicine Doctoral Training Programme.

Author contributions

The study idea was conceived by JM. BL acquired all data for use in the study and carried out all data analysis. JM supervised and aided in interpreting the results. BL took a lead in writing the manuscript, which was finalised with considerable contributions from JM.

Conflict of interest

The authors declare that they have no conflicts of interest.

References

Adkar BV, Tripathi A, Sahoo A, Bajaj K, Goswami D, Chakrabarti P, Swarnkar MK, Gokhale RS & Varadarajan R (2012) Protein Model Discrimination Using Mutational Sensitivity Derived from Deep Sequencing. *Structure* **20:** 371–381

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

Bandaru P, Shah NH, Bhattacharyya M, Barton JP, Kondo Y, Cofsky JC, Gee CL, Chakraborty AK, Kortemme T, Ranganathan R & Kuriyan J (2017) Deconstruction of the Ras switching cycle through saturation mutagenesis. *eLife* **6:** e27810

Banka S, de Goede C, Yue WW, Morris AAM, von Bremen B, Chandler KE, Feichtinger RG, Hart C, Khan N, Lunzer V, Mataković L, Marquardt T, Makowski C, Prokisch H, Debus O, Nosaka K, Sonwalkar H, Zimmermann FA, Sperl W & Mayr JA (2014) Expanding the clinical and molecular spectrum of thiamine pyrophosphokinase deficiency: A treatable neurological disorder caused by TPK1 mutations. *Molecular Genetics and Metabolism* **113:** 301–306

Bergendahl LT, Gerasimavicius L, Miles J, Macdonald L, Wells JN, Welburn JPI & Marsh JA (2019) The role of protein complexes in human genetic disease. *Protein Science* **28:** 1400–1411

Brenan L, Andreev A, Cohen O, Pantel S, Kamburov A, Cacchiarelli D, Persky NS, Zhu C, Bagul M, Goetz EM, Burgin AB, Garraway LA, Getz G, Mikkelsen TS, Piccioni F, Root DE & Johannessen CM (2016) Phenotypic characterization of a comprehensive set of MAPK1/ERK2 missense mutants. *Cell Rep* **17:** 1171–1183

Capriotti E & Altman RB (2011) Improving the prediction of disease-related variants using protein three-dimensional structure. *BMC Bioinformatics* **12:** S3

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

Capriotti E, Calabrese R, Fariselli P, Martelli PL, Altman RB & Casadio R (2013) WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. *BMC Genomics* **14:** S6

Carraro M, Minervini G, Giollo M, Bromberg Y, Capriotti E, Casadio R, Dunbrack R, Elefanti L, Fariselli P, Ferrari C, Gough J, Katsonis P, Leonardi E, Lichtarge O, Menin C, Martelli PL, Niroula A, Pal LR, Repo S, Scaini MC, et al (2017)

Performance of in silico tools for the evaluation of p16INK4a (CDKN2A) variants in CAGI. *Human Mutation* **38:** 1042–1050

Choi Y, Sims GE, Murphy S, Miller JR & Chan AP (2012) Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS One* **7:** Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3466303/ [Accessed June 3, 2019]

Chun S & Fay JC (2009) Identification of deleterious mutations within three human genomes. *Genome Res* **19:** 1553–1561

Crotti Lia, Johnson Christopher N., Graf Elisabeth, De Ferrari Gaetano M., Cuneo Bettina F., Ovadia Marc, Papagiannis John, Feldkamp Michael D., Rathi Subodh G., Kunic Jennifer D., Pedrazzini Matteo, Wieland Thomas, Lichtner Peter, Beckmann Britt-Maria, Clark Travis, Shaffer Christian, Benson D. Woodrow, Kääb Stefan, Meitinger Thomas, Strom Tim M., et al (2013) Calmodulin Mutations Associated With Recurrent Cardiac Arrest in Infants. *Circulation* **127:** 1009–1017

Dandage R, Pandey R, Jayaraj G, Rai M, Berger D & Chakraborty K (2018) Differential strengths of molecular determinants guide environment specific mutational fates. *PLOS Genetics* **14:** e1007419

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 Computational Biology* **6:** e1001025

Deng Z, Huang W, Bakkalbasi E, Brown NG, Adamski CJ, Rice K, Muzny D, Gibbs RA & Palzkill T (2012) Deep Sequencing of Systematic Combinatorial Libraries Reveals β-lactamase Sequence Constraints at High Resolution. *J Mol Biol* **424:** 150–167

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

Doud MB & Bloom JD (2016) Accurate Measurement of the Effects of All Amino-Acid Mutations on Influenza Hemagglutinin. *Viruses* **8:** 155

Epi4K Consortium, Epilepsy Phenome/Genome Project, Allen AS, Berkovic SF, Cossette P, Delanty N, Dlugos D, Eichler EE, Epstein MP, Glauser T, Goldstein DB, Han Y, Heinzen EL, Hitomi Y, Howell KB, Johnson MR, Kuzniecky R, Lowenstein DH, Lu Y-F, Madou MRZ, et al (2013) De novo mutations in epileptic encephalopathies. *Nature* **501:** 217–221

Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, Gasperini M, Janizek JD, Huang X, Starita LM & Shendure J (2018) Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562:** 217–222

Firnberg E, Labonte JW, Gray JJ & Ostermeier M (2014) A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Mol Biol Evol* **31:** 1581–1592

Fitzgerald T, Gerety S, Jones W, van Kogelenberg M, King D, McRae J, Morley K, Parthiban V, Al-Turki S, Ambridge K, Barrett D, Bayzetinova T, Clayton S, Coomber E, Gribble S, Jones P, Krishnappa N, Mason L, Middleton A, Miller R, et al (2015)

Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519:** 223–228

Fowler DM & Fields S (2014) Deep mutational scanning: a new style of protein science. *Nat Methods* **11:** 801–807

Garber M, Guttman M, Clamp M, Zody MC, Friedman N & Xie X (2009) Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25:** i54–i62

Giacomelli AO, Yang X, Lintner RE, McFarland JM, Duby M, Kim J, Howard TP, Takeda DY, Ly SH, Kim E, Gannon HS, Hurhula B, Sharpe T, Goodale A, Fritchman B, Steelman S, Vazquez F, Tsherniak A, Aguirre AJ, Doench JG, et al (2018) Mutational processes shape the landscape of TP53 mutations in human cancer. *Nature Genetics* **50:** 1381

González-Pérez A & López-Bigas N (2011) Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. *Am J Hum Genet* **88:** 440–449

Grantham R (1974) Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* **185:** 862–864

Gray VE, Hause RJ, Luebeck J, Shendure J & Fowler DM (2018) Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell Syst* **6:** 116-124.e3

Grimm DG, Azencott C-A, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW, Duncan LE & Borgwardt 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

Gulko B, Hubisz MJ, Gronau I & Siepel A (2015) Probabilities of Fitness Consequences for Point Mutations Across the Human Genome. *Nat Genet* **47:** 276–283

Haddox HK, Dingens AS & Bloom JD (2016) Experimental Estimation of the Effects of All Amino-Acid Mutations to HIV's Envelope Protein on Viral Replication in Cell Culture. *PLOS Pathogens* **12:** e1006114

Harris DT, Wang N, Riley TP, Anderson SD, Singh NK, Procko E, Baker BM & Kranz DM (2016) Deep Mutational Scans as a Guide to Engineering High Affinity T Cell Receptor Interactions with Peptide-bound Major Histocompatibility Complex. *J. Biol. Chem.* **291:** 24566–24578

Hecht M, Bromberg Y & Rost B (2015) Better prediction of functional effects for sequence variants. *BMC Genomics* **16:** S1

Henikoff S & Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89:** 10915–10919

Henry VJ, Bandrowski AE, Pepin A-S, Gonzalez BJ & Desfeux A (2014) OMICtools: an informative directory for multi-omic data analysis. *Database (Oxford)* **2014:** bau069

Hoskins RA, Repo S, Barsky D, Andreoletti G, Moult J & Brenner SE (2017) Reports from CAGI: The Critical Assessment of Genome Interpretation. *Human Mutation* **38:** 1039–1041

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

Jacquier H, Birgy A, Nagard HL, Mechulam Y, Schmitt E, Glodt J, Bercot B, Petit E, Poulain J, Barnaud G, Gros P-A & Tenaillon O (2013) Capturing the mutational landscape of the beta-lactamase TEM-1. *PNAS* **110:** 13067–13072

Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, Bernstein JA & Bejerano G (2016) M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature Genetics* **48:** 1581–1586

Jensen HH, Brohus M, Nyegaard M & Overgaard MT (2018) Human Calmodulin Mutations. *Front. Mol. Neurosci.* **11:** 396

Johansen MB, Izarzugaza JMG, Brunak S, Petersen TN & Gupta R (2013) Prediction of Disease Causing Non-Synonymous SNPs by the Artificial Neural Network Predictor NetDiseaseSNP. *PLoS One* **8:** Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3723835/ [Accessed June 5, 2019]

Jones EM, Lubock NB, Venkatakrishnan AJ, Wang J, Tseng AM, Paggi JM, Latorraca NR, Cancilla D, Satyadi M, Davis JE, Babu MM, Dror RO & Kosuri S (2019) Structural and Functional Characterization of G Protein-Coupled Receptors with Deep Mutational Scanning. *bioRxiv***:** 623108

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, Seaby EG, Kosmicki JA, Walters RK, Tashman K, et al (2019) Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* doi: 10.1101/531210 [PREPRINT]

Kelsic ED, Chung H, Cohen N, Park J, Wang HH & Kishony R (2016) RNA structural determinants of optimal codons revealed by MAGE-seq. *Cell Syst* **3:** 563-571.e6

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM & Shendure J (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46:** 310–315

Kitzman JO, Starita LM, Lo RS, Fields S & Shendure J (2015) Massively Parallel Single Amino Acid Mutagenesis. *Nat Methods* **12:** 203–206

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM & Maglott DR (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42:** D980–D985

Lee JM, Huddleston J, Doud MB, Hooper KA, Wu NC, Bedford T & Bloom JD (2018) Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. *Proc Natl Acad Sci U S A* **115:** E8276–E8285

de Ligt J, Willemsen MH, van Bon BWM, Kleefstra T, Yntema HG, Kroes T, Vulto-van Silfhout AT, Koolen DA, de Vries P, Gilissen C, del Rosario M, Hoischen A, Scheffer H, de Vries BBA, Brunner HG, Veltman JA & Vissers LELM (2012) Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *N Engl J Med* **367:** 1921-1929

Limongelli I, Marini S & Bellazzi R (2015) PaPI: pseudo amino acid composition to score human protein-coding variants. *BMC Bioinformatics* **16:** 123

Liu X, Wu C, Li C & Boerwinkle E (2016) dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human Mutation* **37:** 235–241

Lu Q, Hu Y, Sun J, Cheng Y, Cheung K-H & Zhao H (2015) A Statistical Framework to Predict Functional Non-Coding Regions in the Human Genome Through Integrated Analysis of Annotation Data. *Scientific Reports* **5:** 1–13

Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, Liu Y, Chen X, Newman S, Nakitandwe J, Li Y, Li B, Shen S, Wang Z, Shurtleff S, Robison LL, Levy S, Easton J & Zhang J (2019) Analysis of error profiles in deep next-generation sequencing data. *Genome Biology* **20:** 50

Mahmood K, Jung C, Philip G, Georgeson P, Chung J, Pope BJ & Park DJ (2017) Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. *Human Genomics* **11:** 10

Majithia AR, Tsuda B, Agostini M, Gnanapradeepan K, Rice R, Peloso G, Patel KA, Zhang X, Broekema MF, Patterson N, Duby M, Sharpe T, Kalkhoven E, Rosen ED, Barroso I, Ellard S, Kathiresan S, O'Rahilly S, Chatterjee K, Florez JC, et al (2016) Prospective functional classification of all possible missense variants in PPARG. *Nat Genet* **48:** 1570–1575

Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, Kircher M, Khechaduri A, Dines JN, Hause RJ, Bhatia S, Evans WE, Relling MV, Yang W, Shendure J & Fowler DM (2018) Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature Genetics* **50:** 874

McEntagart M, Williamson KA, Rainger JK, Wheeler A, Seawright A, De Baere E, Verdin H, Bergendahl LT, Quigley A, Rainger J, Dixit A, Sarkar A, López Laso E, Sanchez-Carpintero R, Barrio J, Bitoun P, Prescott T, Riise R, McKee S, Cook J, et al (2016) A Restricted Repertoire of De Novo Mutations in ITPR1 Cause Gillespie Syndrome with Evidence for Dominant-Negative Effect. *Am. J. Hum. Genet.* **98:** 981–992

Melamed D, Young DL, Gamble CE, Miller CR & Fields S (2013) Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA* **19:** 1537–1551

Mighell TL, Evans-Dutson S & O'Roak BJ (2018) A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. *Am J Hum Genet* **102:** 943–955

Mishra P, Flynn JM, Starr TN & Bolon DNA (2016) Systematic mutant analyses elucidate general and client-specific aspects of Hsp90 function. *Cell Rep* **15:** 588–598

Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin C-F, Stevens C, Wang L-S, Makarov V, Polak P, Yoon S, Maguire J, Crawford EL, Campbell NG, Geller ET, Valladares O, Shafer C, Liu H, Zhao T, et al (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485:** 242–245

Niroula A, Urolagin S & Vihinen M (2015) PON-P2: Prediction Method for Fast and Reliable Identification of Harmful Variants. *PLoS One* **10:** Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4315405/ [Accessed June 5, 2019]

Niroula A & Vihinen M (2017) Predicting Severity of Disease-Causing Variants. Available at: https://onlinelibrary.wiley.com/doi/full/10.1002/humu.23173 [Accessed June 5, 2019]

Nomikos M, Thanassoulas A, Calver BL, Beck K, Vassilakopoulou V, Buntwal L, Konotgianni I, Smith A, Safieh-Garabedian B, Livaniou E, Toft ES, Nounesis G & Lai FA (2018) Calmodulin Mutations Associated with Congenital Cardiac Disease Display Novel Biophysical and Biochemical Characteristics. *Biophysical Journal* **114:** 467a

Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam H-J, Mort M, Cooper DN, Sebat J, Iakoucheva LM, Mooney SD & Radivojac P (2017) MutPred2: inferring the molecular and phenotypic impact of amino acid variants. *bioRxiv***:** 134981

Pollard KS, Hubisz MJ, Rosenbloom KR & Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20:** 110–121

Qi H, Chen C, Zhang H, Long JJ, Chung WK, Guan Y & Shen Y (2018) MVP: predicting pathogenicity of missense variants by deep learning. *bioRxiv***:** 259390

Quang D, Chen Y & Xie X (2015) DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31:** 761–763

Raimondi D, Tanyalcin I, Ferté J, Gazzo A, Orlando G, Lenaerts T, Rooman M & Vranken W (2017) DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res* **45:** W201–W206

Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, Albrecht B, Bartholdi D, Beygo J, Di Donato N, Dufke A, Cremer K, Hempel M, Horn D, Hoyer J, Joset P, Röpke A, Moog U, Riess A, Thiel CT, et al (2012) Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *The Lancet* **380:** 1674–1682

Reeb J, Wirth T & Rost B (2019) Variant effect predictions capture some aspects of deep mutational scanning experiments. *bioRxiv***:** 859603

Rentzsch P, Witten D, Cooper GM, Shendure J & Kircher M (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47:** D886–D894

Reva B, Antipin Y & Sander C (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* **39:** e118

Riesselman AJ, Ingraham JB & Marks DS (2018) Deep generative models of genetic variation capture the effects of mutations. *Nature Methods* **15:** 816

Rockah-Shmuel L, Tóth-Petróczy Á & Tawfik DS (2015) Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations. *PLOS Computational Biology* **11:** e1004421

Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR & Campbell C (2018) FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* **34:** 511–513

Roscoe BP & Bolon DNA (2014) Systematic exploration of ubiquitin sequence, E1 activation efficiency, and experimental fitness in yeast. *J Mol Biol* **426:** 2854–2870

Roscoe BP, Thayer KM, Zeldovich KB, Fushman D & Bolon DNA (2013) Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *J Mol Biol* **425:** 1363–1377

Samocha KE, Kosmicki JA, Karczewski KJ, O'Donnell-Luria AH, Pierce-Hoffman E, MacArthur DG, Neale BM & Daly MJ (2017) Regional missense constraint improves variant deleteriousness prediction. *bioRxiv***:** 148353

Schaafsma GCP & Vihinen M (2018) Representativeness of variation benchmark datasets. *BMC Bioinformatics* **19:** 461

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

Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, Day INM & Gaunt TR (2013) Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Hum Mutat* **34:** 57–65

Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, Gaunt TR & Campbell C (2015) An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31:** 1536–1543

Siepel A & Haussler D (2005) Phylogenetic Hidden Markov Models. In: *Statistical Methods in Molecular Evolution. Statistics for Biology and Health*, Rasmus N (ed) pp 325-351 Springer, New York, NY

Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G & Ng PC (2012) SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* **40:** W452–W457

Spencer JM & Zhang X (2017) Deep mutational scanning of S. pyogenes Cas9 reveals important functional domains. *Scientific Reports* **7:** 16836

Srivastava N, Hinton G, Krizhevsky A, Sutskever I & Salakhutdinov R (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res* **15:** 30

Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, Fowler DM, Parvin JD, Shendure J & Fields S (2015) Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* **200:** 413–422

Steward RE, MacArthur MW, Laskowski RA & Thornton JM (2003) Molecular basis of inherited diseases: a structural perspective. *Trends in Genetics* **19:** 505–513

Stiffler MA, Hekstra DR & Ranganathan R (2015) Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. *Cell* **160:** 882–892

Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, Fritzilas N, Hakenberg J, Dutta A, Shon J, Xu J, Batzoglou S, Li X & Farh KK-H (2018) Predicting the clinical impact of human mutation with deep neural networks. *Nature Genetics* **50:** 1161–1170

Thomas PD & Kejariwal A (2004) Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: Evolutionary evidence for differences in molecular effects. *PNAS* **101:** 15398–15403

Vaser R, Adusumalli S, Leng SN, Sikic M & Ng PC (2016) SIFT missense predictions for genomes. *Nat Protoc* **11:** 1–9

Weile J, Sun S, Cote AG, Knapp J, Verby M, Mellor JC, Wu Y, Pons C, Wong C, van Lieshout N, Yang F, Tasan M, Tan G, Yang S, Fowler DM, Nussbaum R, Bloom JD, Vidal M, Hill DE, Aloy P, et al (2017) A framework for exhaustively mapping functional missense variants. *Mol Syst Biol* **13:** 957

Williamson KA, Hall HN, Owen LJ, Livesey BJ, Hanson IM, Adams GGW, Bodek S, Calvas P, Castle B, Clarke M, Deng AT, Edery P, Fisher R, Gillessen-Kaesbach G, Heon E, Hurst J, Josifova D, Lorenz B, McKee S, Meire F, et al (2019) Recurrent heterozygous PAX6 missense variants cause severe bilateral microphthalmia via predictable effects on DNA–protein interaction. *Genetics in Medicine***:** 1–12

Wu NC, Olson CA, Du Y, Le S, Tran K, Remenyi R, Gong D, Al-Mawsawi LQ, Qi H, Wu T-T & Sun R (2015) Functional Constraint Profiling of a Viral Protein Reveals Discordance of Evolutionary Conservation and Functionality. *PLOS Genetics* **11:** e1005310

Xu Q, Tang Q, Katsonis P, Lichtarge O, Jones D, Bovo S, Babbi G, Martelli PL, Casadio R, Lee GR, Seok C, Fenton AW & Dunbrack RL (2017) Benchmarking predictions of allostery in liver pyruvate kinase in CAGI4. *Human Mutation* **38:** 1123–1131

Yates CM, Filippis I, Kelley LA & Sternberg MJE (2014) SuSPect: Enhanced Prediction of Single Amino Acid Variant (SAV) Phenotype Using Network Features. *J Mol Biol* **426:** 2692–2701

Zhu L, Wu R, Ye Z, Gu R, Wang Y, Hou Y, Feng Z & Ma X (2019) Identification of two novel TPK1 gene mutations in a Chinese patient with thiamine pyrophosphokinase

deficiency undergoing whole exome sequencing. *Journal of Pediatric Endocrinology and Metabolism* **32:** 295–300

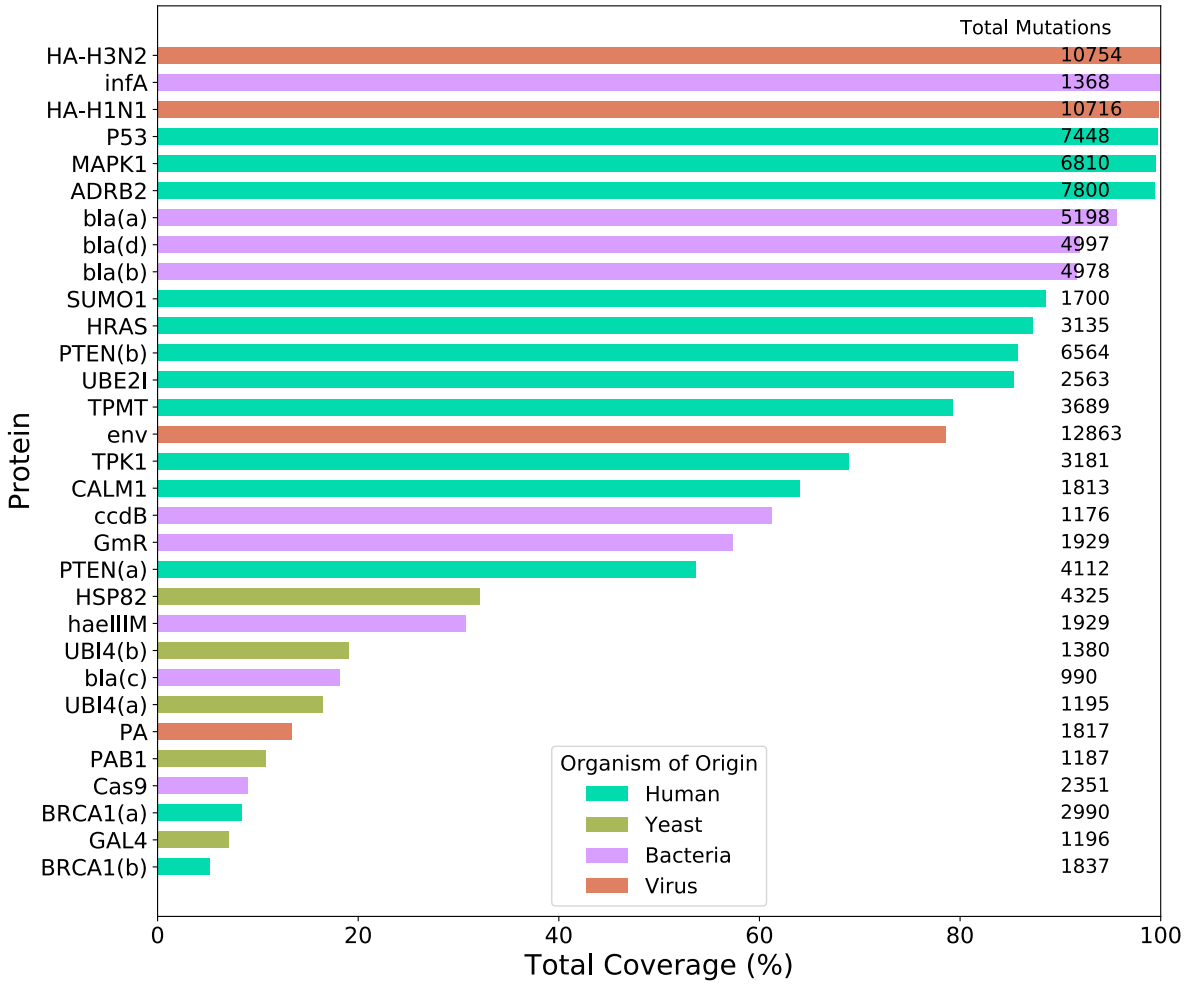


Figure 1. The percentage of all amino acid substitutions covered by each DMS experiment. The total number of mutations assessed by each DMS experiment is indicated on the right. Where multiple datasets exist for a single protein, sequential letters are used to distinguish them.

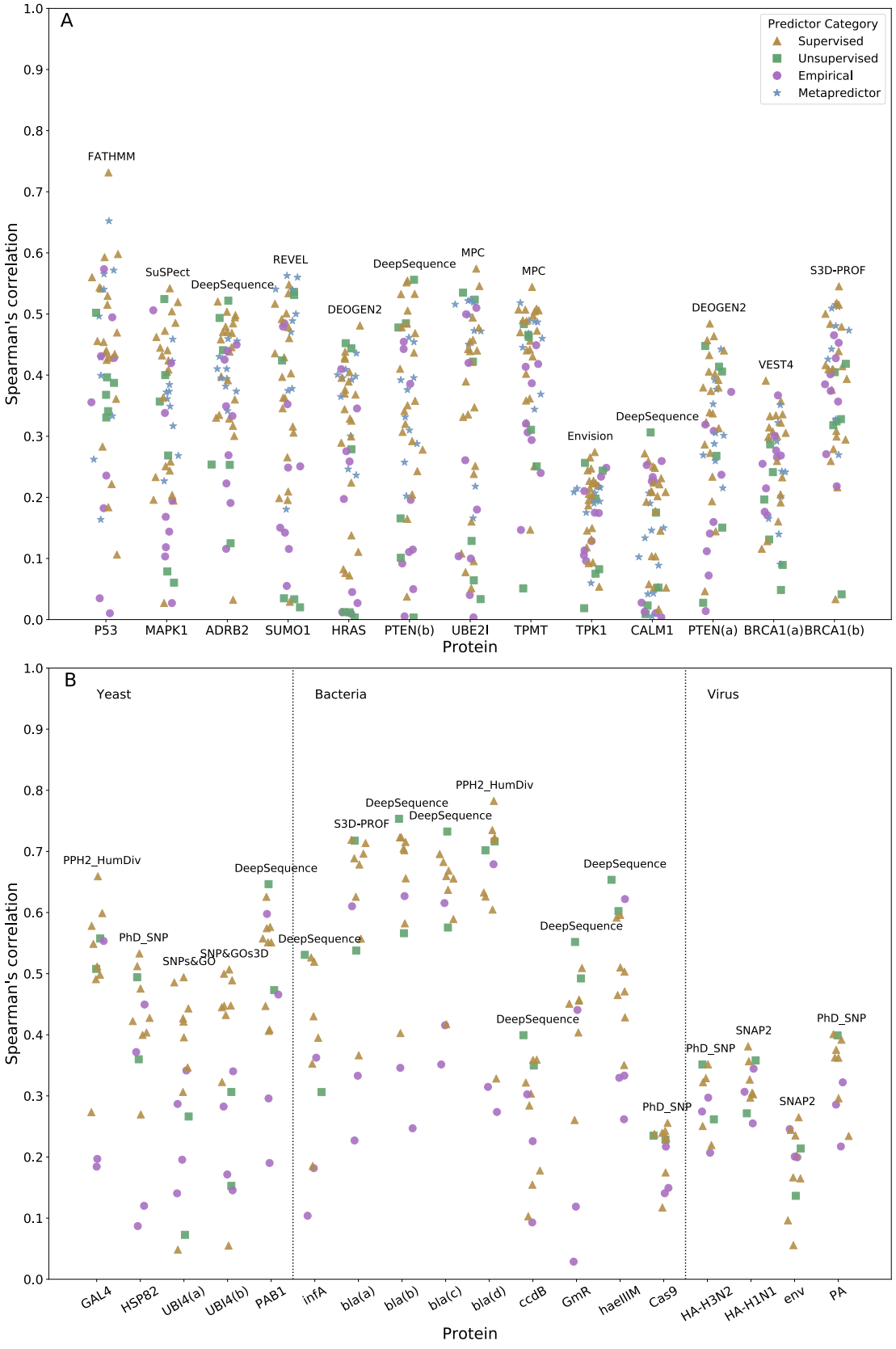


Figure 2. Spearman's correlation calculated between all VEP predictions and DMS datasets. The top-performing predictor for each protein is labelled on the plot. This analysis is split into **(A)** human and **(B)** non-human proteins.

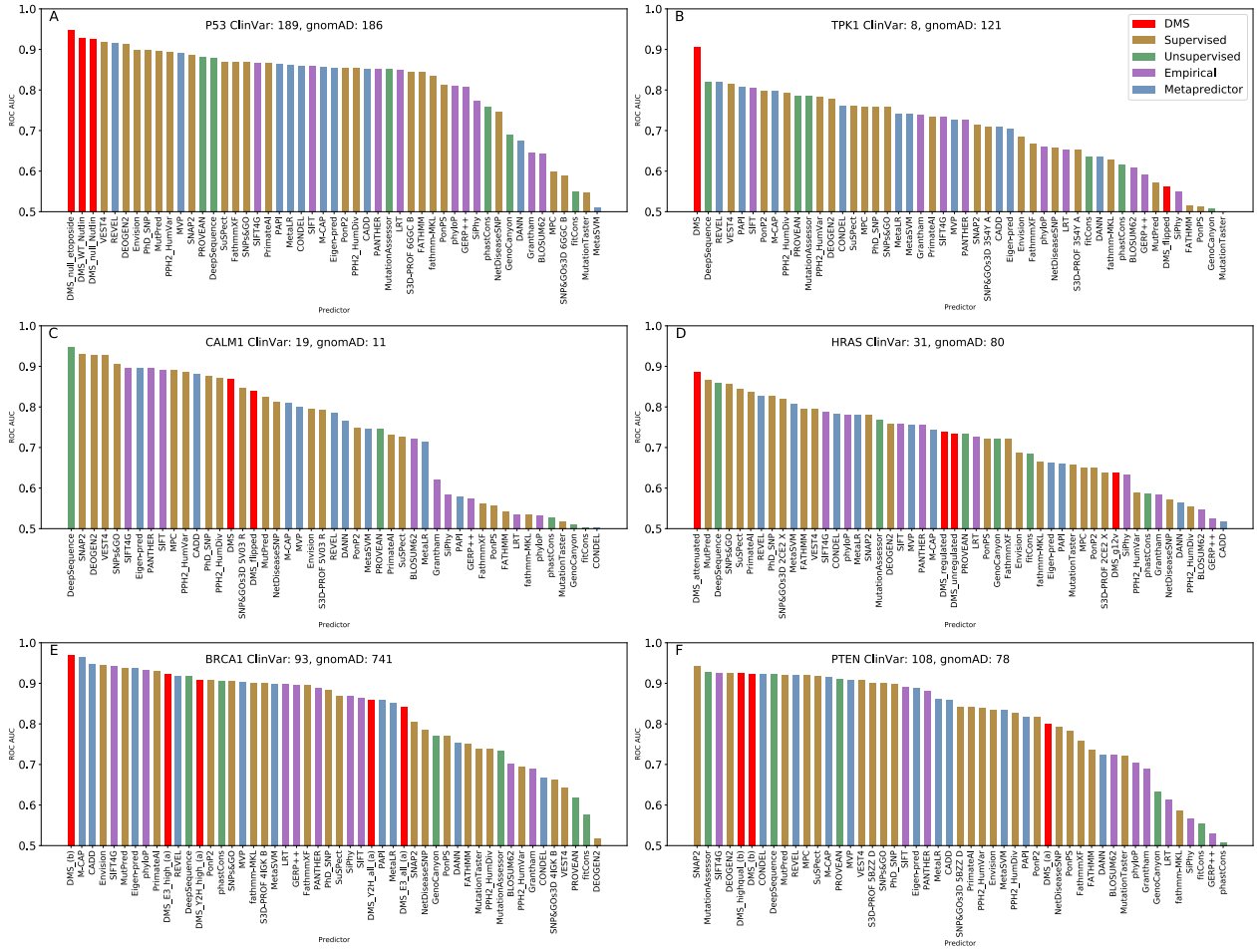


Figure 3. ROC AUC values for DMS datasets and VEPs in distinguishing between pathogenic missense variants from ClinVar and putatively benign missense variants from gnomAD for six human disease genes. The numbers of variants in each class are indicated on the plot. The different DMS datasets for each protein are described in Table EV7.

Table 1. Summary of the computational VEPs used in this analysis, along with a description of utilised predictive features and the total number of DMS datasets each method generated predictions for in the format human/yeast/bacterial/viral.

Predictor	Category	Features	Data Source, method source or online predictor.	Number of prediction sets (h/y/b/v)	Reference
DEOGEN2	Supervised	PROVEAN, sequence conservation, pathway features, early folding predictions, interface annotations from 3D structures	https://deogen2.mutaframe.com/	13/0/0/0	(Raimondi *et al*, 2017)
Envision	Supervised[1]	DMS measurements	https://envision.gs.washington.edu/shiny/envision_new/	13/3/0/0	(Gray *et al*, 2018)
FATHMM (weighted)	Supervised	HMM alignments, per-domain mutation consequences	http://fathmm.biocompute.org.uk/inherited.html	13/0/0/0	(Shihab *et al*, 2013)
Fathmm-MKL	Supervised	Sequence conservation, epigenetic features, genome site features, DNA footprints	dbNSFP database	13/0/0/0	(Shihab *et al*, 2015)
FathmmXF	Supervised	Sequence conservation, residue features, gene expression, RNA interactions, segmentation features	http://fathmm.biocompute.org.uk/fathmm-xf/	13/0/0/0	(Rogers *et al*, 2018)
MPC	Supervised	PolyPhen-2, 'missense badness', sequence conservation	dbNSFP database	11/0/0/0	(Samocha *et al*, 2017)
MutationTaster	Supervised	Regulatory features, PhyloP, phastCons,	dbNSFP database	13/0/0/0	(Schwarz *et al*, 2014)

		splice sites, sequence conservation, functional and domain annotations			
MutPred	Supervised	Sequence conservation, biophysical features, and sequence-based features	dbNSFP database	13/0/0/0	(Pejaver *et al*, 2017)
NetDiseaseSNP	Supervised	Predicted structural features, SIFT, sequence conservation	http://www.cbs.dtu.dk/services/NetDiseaseSNP/	13/5/9/4	(Johansen *et al*, 2013)
PhD_SNP	Supervised	Sequence conservation, sequence features	http://snps.biofold.org/snps-and-go/snps-and-go.html	13/5/9/4	(Capriotti *et al*, 2006)
PolyPhen-2 (HumDiv)	Supervised	Sequence conservation, sequence features, residue-level structural features	http://genetics.bwh.harvard.edu/pph2/	13/5/9/2	(Adzhubei *et al*, 2010)
PolyPhen-2 (HumVar)	Supervised	Sequence conservation, sequence features, residue-level structural features	http://genetics.bwh.harvard.edu/pph2/	13/0/0/0	(Adzhubei *et al*, 2010)
PonPS	Supervised	PonP2, sequence conservation, biophysical features, sequence features, co-evolution, predicted stability	http://structure.bmc.lu.se/PON-PS/	13/0/0/0	(Niroula & Vihinen, 2017)
PonP2	Supervised	Biophysical features, GO terms, sequence conservation	http://structure.bmc.lu.se/PON-P2/	13/0/0/0	(Niroula *et al*, 2015)
PrimateAI	Supervised	Sequence-based features, predicted structural features	dbNSFP database	13/0/0/0	(Sundaram *et al*, 2018)

S3D-PROF	Supervised	Structural environment, sequence conservation	http://snps.biofold.org/snps-and-go/snps-and-go-3d.html	12/5/7/2	(Capriotti & Altman, 2011)
SNAP2	Supervised	Biophysical features, sequence conservation, predicted structural features, co-evolution, residue annotations	https://www.rostlab.org/services/snap/	13/5/9/4	(Hecht *et al*, 2015)
SNPs&GO	Supervised	Sequence conservation, sequence features, PANTHER output, GO terms	http://snps.biofold.org/snps-and-go/snps-and-go.html	13/5/9/4	(Capriotti *et al*, 2013)
SNP&GOs3D	Supervised	PANTHER, GO terms, structural environment, sequence conservation	http://snps.biofold.org/snps-and-go/snps-and-go-3d.html	12/5/7/2	(Capriotti & Altman, 2011)
SuSPect	Supervised	Network centrality, UniProt annotations, sequence conservation, predicted surface accessibility	http://www.sbg.bio.ic.ac.uk/suspect/about.html	13/5/9/4	(Yates *et al*, 2014)
VEST4	Supervised	86 pre-calculated features from SNVBox	https://www.cravat.us/CRAVAT/	13/0/0/0	(Carter *et al*, 2013)
DeepSequence	Unsupervised	Sequence conservation	https://github.com/debbiemarkslab/DeepSequence	13/5/9/3	(Riesselman *et al*, 2018)
fitCons	Unsupervised	DNA accessibility, transcription, epigenetic features	dbNSFP database	13/0/0/0	(Gulko *et al*, 2015)
GenoCanyon	Unsupervised	Sequence conservation, biochemical signals	dbNSFP database	13/0/0/0	(Lu *et al*, 2015)
PANTHER	Unsupervised	HMM alignment scores from protein subfamilies	http://snps.biofold.org/snps-and-go/snps-and-go.html	13/4/1/0	(Thomas & Kejariwal)

phastCons	Unsupervised	Sequence conservation	dbNSFP database	13/0/0/0	(Siepel & Haussler, 2005)
PROVEAN	Unsupervised	Sequence conservation	http://provean.jcvi.org/index.php	13/5/9/4	(Choi *et al*, 2012)
MutationAssessor	Unsupervised	Sequence conservation, conservation between subfamilies	dbNSFP database	12/0/0/0	(Reva *et al*, 2011)
BLOSUM62	Empirical (substitution matrix)	Sequence conservation	https://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt	13/5/9/4	(Henikoff & Henikoff, 1992)
GERP++	Empirical	Sequence conservation	dbNSFP database	13/0/0/0	(Davydov *et al*, 2010)
Grantham	Empirical (substitution matrix)	Amino acid property differences	Matrix available in reference.	13/5/9/4	(Grantham, 1974)
LRT	Empirical	Sequence conservation	dbNSFP database	13/0/0/0	(Chun & Fay, 2009)
phyloP	Empirical	Sequence conservation	http://papi.unipv.it/	13/0/0/0	(Pollard *et al*, 2010)
SIFT	Empirical	Sequence conservation	https://sift.bii.a-star.edu.sg/www/code.html	13/5/9/4	(Sim *et al*, 2012)
SIFT4G	Empirical	Sequence conservation	dbNSFP database	13/0/0/0	(Vaser *et al*, 2016)
SiPhy	Empirical	Sequence conservation	dbNSFP database	13/0/0/0	(Garber *et al*, 2009)
CADD	Metapredictor	949 features including numerous other predictors; uses the same feature set as DANN	https://cadd.gs.washington.edu/snv	13/0/0/0	(Kircher *et al*, 2014)
CONDEL	Metapredictor	Mutation Assessor, SIFT, PolyPhen-2,	http://bbglab.irbbarcelona.org/fannsdb/	13/0/0/0	(González-Pérez & López-

		MAPP, Pfam E-value, Fathmm			Bigas, 2011)
DANN	Metapredictor	949 features including numerous other predictors. Uses the same feature set as CADD	dbNSFP database	13/0/0/0	(Quang *et al*, 2015)
Eigen-pred	Metapredictor[2]	SIFT, PolyPhen-2, MutationAssessor, conservation metrics	dbNSFP database	13/0/0/0	(Ionita-Laza *et al*, 2016)
M-CAP	Metapredictor	SIFT, PolyPhen2, CADD, MetaLR, MutationTaster, MutationAssessor, Fathmm, LRT, evolutionary conservation metrics, substitution matrices	http://bejerano.stanford.edu/mcap/	13/0/0/0	(Jagadeesh *et al*, 2016)
MetaLR	Metapredictor	SIFT, PolyPhen2, GERP++, MutationTaster, MutationAssessor, Fathmm, LRT, SiPhy, PhyloP	dbNSFP database	13/0/0/0	(Dong *et al*, 2015)
MetaSVM	Metapredictor	SIFT, PolyPhen2, GERP++, MutationTaster, MutationAssessor, Fathmm, LRT, SiPhy, PhyloP	dbNSFP database	13/0/0/0	(Dong *et al*, 2015)
MVP	Metapredictor	Sequence context, sequence conservation, conservation metrics, predicted structural features, mutational	dbNSFP database	13/0/0/0	(Qi *et al*, 2018)

		tolerance, Eigen, VEST3, MutationTaster, PolyPhen2, SIFT, PROVEAN, fathmm-MKL, FATHMM, MutationAssessor, LRT			
PAPI	Metapredictor	PolyPhen2, SIFT, PseAA RF model (evolutionary conservation metrics, sequence environment)	http://papi.unipv.it/	13/0/0/0	(Limongelli *et al*, 2015)
REVEL	Metapredictor	MutPred, PROVEAN, SIFT, PolyPhen2, LRT, MutationTaster, MutationAssessor, Fathmm, VEST3, GERP++, SiPhy, PhyloP	https://sites.google.com/site/revelgenomics/	13/0/0/0	(Ioannidis *et al*, 2016)

[1]Envision is a supervised method, but trained on DMS data, thus not used for calculating the RankScore in table 3.

[2]Unlike the other metapredictors, Eigen-pred uses an unsupervised machine learning method.

Table 2. Summary of the DMS datasets used in this analysis, including functional assay and method of mutagenesis. We also note the specific DMS assay from each study we use for calculating correlation with the VEP predictions.

Protein(s) (Uniprot ID)	**Organism**	**Functional Assay**	**Mutagenesis Method**	**Utilised assay**	**Access date**	**Reference**
UBE2I (P63279) SUMO1 (P63165) TPK1 (Q9H3S3) CALM1 (P0DP23)	Human	POPCode, a variant of multiple-site directed mutagenesis	Competitive growth assay in yeast	(screen.score column) UBE2I_flipped_scores SUMO1_flipped_scores TPK1_flipped_scores CALM1_flipped_scores	2018-10-12	(Weile *et al*, 2017)
BRCA1(a) (P38398)	Human	Systematic site-directed mutagenesis	Yeast two-hybrid assay and phage display	E3_score_800_filter_pass	2018-10-12	(Starita *et al*, 2015)
BRCA1(b) (P38398)	Human	Systematic site-directed mutagenesis	Competitive growth assay in HAP1 cells	function.score.mean	2020-02-14	(Findlay *et al*, 2018)
P53 (P04637)	Human	Systematic site-directed mutagenesis	Competitive growth assay in the presence of P53 agonists	A549_p53WT_Nutlin-3_Z-score	2018-10-12	(Giacomelli *et al*, 2018)
HRas (P01112)	Human	Systematic site-directed mutagenesis	Two-hybrid assay	Ras-G12V	2018-10-12	(Bandaru *et al*, 2017)
MAPK1 (P28482)	Human	Systematic site-directed mutagenesis	Competitive growth assay	Doxycycline	2018-10-12	(Brenan *et al*, 2016)
PTEN(a) (P60484) TPMT (P51580)	Human	Systematic site-directed mutagenesis	Fluorescence of a GFP fusion protein	score score	2018-10-12	(Matreyek *et al*, 2018)
PTEN(b) (P60484)	Human	Systematic site-directed mutagenesis	Disruption of an artificial genetic circuit in yeast	Cum_score_high_conf	2020-02-14	(Mighell *et al*, 2018)

ADRB2 (P07550)	Human	Systematic site-directed mutagenesis	Pathway-specific reporter gene transcription assessed by RNA-seq	condition_0.625	2019-04-31	(Jones *et al*, 2019)
HSP82 (P02829)	Yeast	Systematic site-directed mutagenesis	Competitive growth assay	norm_ratiochange	2018-10-12	(Mishra *et al*, 2016)
UBI4(a) (P0CG63)	Yeast	Systematic site-directed mutagenesis	Competitive growth assay	selection_coefficient	2018-10-12	(Roscoe *et al*, 2013)
UBI4(b) (P0CG63)	Yeast	Site directed mutagenesis by cassette ligation	Fluorescence activated cell sorting	Relative_E1-activity_limiting	2018-10-12	(Roscoe & Bolon, 2014)
PAB1 (P04147)	Yeast	Random mutagenesis by error-prone PCR	Competitive growth assay	Linear	2018-10-12	(Melamed *et al*, 2013)
GAL4 (P04386)	Yeast	Systematic site-directed mutagenesis	Two-hybrid assay	SEL_A_24h	2018-10-12	(Kitzman *et al*, 2015)
infA (P69222)	E. coli	Systematic site-directed mutagenesis	Competitive growth assay	fitness_min	2018-10-12	(Kelsic *et al*, 2016)
GmR (N/A)[1]	E. coli	Systematic site-directed mutagenesis	Antibiotic resistance	37C	2018-10-12	(Dandage *et al*, 2018)
bla(a) (P62593)	E. coli	Systematic site-directed mutagenesis	Antibiotic resistance	Fitness	2018-10-12	(Firnberg *et al*, 2014)
bla(b) (P62593)	E. coli	Systematic site-directed mutagenesis	Antibiotic resistance	Ampicillin_2500	2018-10-12	(Stiffler *et al*, 2015)
bla(c) (P62593)	E. coli	Random mutagenesis	Antibiotic resistance	MIC_Score_WT	2018-10-12	(Jacquier *et al*, 2013)

bla(d) (P62593)	E. coli	Random and site-directed mutagenesis	Antibiotic resistance	DMS	2018-10-12	(Deng *et al*, 2012)
ccdB (P62554)	E. coli	Systematic site-directed mutagenesis	Reverse survival assay (for toxin activity)	MSseq	2018-10-12	(Adkar *et al*, 2012)
haeIIIM (P20589)	H. aegyptius	Random mutagenesis	Competitive growth assay	W_rel_G17	2018-10-12	(Rockah-Shmuel *et al*, 2015)
Cas9 (Q99ZW2)	S. pyrogenes	Random mutagenesis by error-prone PCR	Survival assay	Log2_Fold_Change_after_Positive_Selection	2018-10-12	(Spencer & Zhang, 2017)
env (P03377)	HIV virus	Systematic site-directed mutagenesis	Competitive replication assay	Site_preferences	2018-10-12	(Haddox *et al*, 2016)
HA-H1N1 (A0A2Z5U3Z0)	Influenza virus	Helper virus	Competitive replication assay	Site_preferences	2018-10-12	(Doud & Bloom, 2016)
HA-H3N2 (A0A097PF60)	Influenza virus	Helper virus	Competitive replication assay	avg_prefs	2018-10-12	(Lee *et al*, 2018)
PA (P15659)	Influenza virus	Helper virus	Competitive replication assay	RF_index	2018-10-12	(Wu *et al*, 2015)

[1]The GmR sequence used in the study does not correspond to any Uniprot ID. We used the sequence in the paper for predictor input.

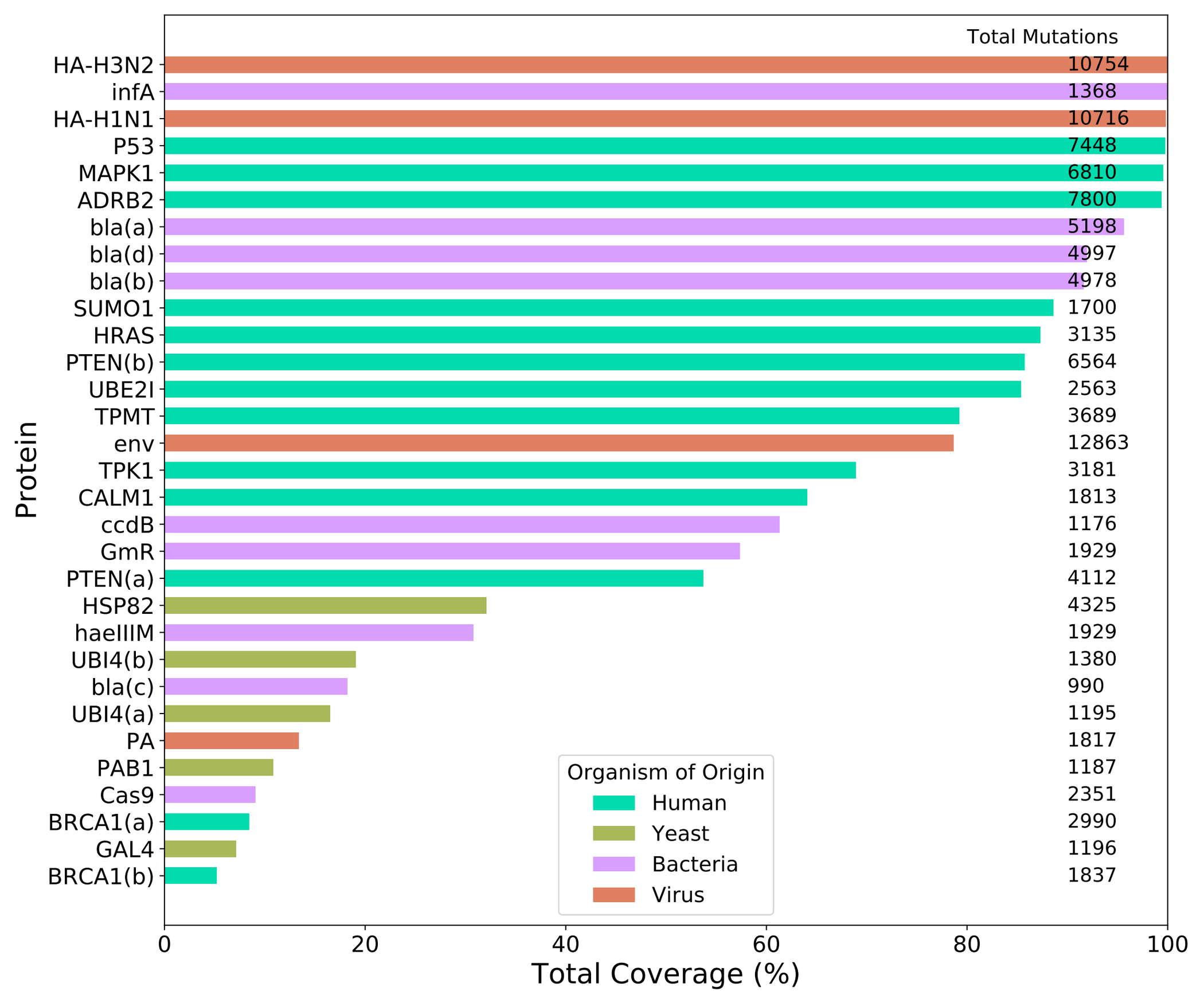
Table 3. Calculated rank scores of VEPs based on the mean, normalised correlation of DMS data for human, yeast, bacterial and viral proteins. Standard deviation and coefficient of variation are based on the normalised correlations of each predictor between different proteins.

Human proteins					
Predictor	**RankScore**	**Cofficient of variation**	**Standard deviation**	**Number of proteins**	**Category**
DeepSequence	0.884493	0.150842	0.133419	11	Unsupervised
SNPs&GO	0.832998	0.128118	0.106722	11	Supervised
SNAP2	0.821118	0.15786	0.129622	11	Supervised
DEOGEN2	0.810199	0.206995	0.167707	11	Supervised
SuSPect	0.806086	0.234057	0.18867	11	Supervised
REVEL	0.804425	0.219104	0.176253	11	Metapredictor
PhD_SNP	0.800309	0.091052	0.07287	11	Supervised
S3D-PROF	0.798352	0.102889	0.082142	10	Supervised
VEST4	0.779683	0.165271	0.128859	11	Supervised
SIFT4G	0.777466	0.127501	0.099128	11	Empirical
PANTHER	0.773401	0.14865	0.114966	11	Empirical
Eigen-pred	0.77282	0.143674	0.111034	11	Metapredictor
MPC	0.770669	0.26447	0.203819	10	Supervised
MutationAssessor	0.769645	0.208029	0.160109	10	Unsupervised
PPH2_HumVar	0.752304	0.203427	0.153039	11	Supervised
MetaLR	0.737236	0.230105	0.169642	11	Metapredictor
Envision	0.734916	0.292611	0.215045	11	Supervised
PROVEAN	0.734866	0.188153	0.138267	11	Unsupervised
SNP&GOs3D	0.732097	0.23776	0.174063	10	Supervised
MutPred	0.730093	0.244265	0.178336	11	Supervised
PPH2_HumDiv	0.717007	0.220618	0.158185	11	Supervised
M-CAP	0.714093	0.266564	0.190351	11	Metapredictor
MetaSVM	0.701199	0.30728	0.215465	11	Metapredictor
CONDEL	0.696421	0.327479	0.228063	11	Metapredictor
SIFT	0.689208	0.152913	0.105389	11	Empirical

CADD	0.67234	0.1284	0.086329	11	Metapredictor
MVP	0.645229	0.187356	0.120888	11	Metapredictor
PAPI	0.627417	0.377005	0.236539	11	Metapredictor
PonPS	0.594731	0.128064	0.076164	11	Supervised
NetDiseaseSNP	0.583933	0.221029	0.129066	11	Supervised
PonP2	0.540467	0.461565	0.249461	11	Supervised
PrimateAI	0.473828	0.41569	0.196966	11	Supervised
FathmmXF	0.455187	0.570612	0.259735	11	Supervised
BLOSUM62	0.44217	0.55099	0.243631	11	Empirical
fathmm-MKL	0.38756	0.557953	0.21624	11	Supervised
DANN	0.38604	0.41349	0.159624	11	Metapredictor
LRT	0.366185	0.70642	0.258681	11	Empirical
phyloP	0.354074	0.713528	0.252642	11	Empirical
FATHMM	0.33821	0.764497	0.258561	11	Supervised
Grantham	0.298924	0.723971	0.216412	11	Empirical
GERP++	0.267974	0.654054	0.175269	11	Empirical
MutationTaster	0.265101	0.627194	0.16627	11	Supervised
SiPhy	0.243562	0.891019	0.217018	11	Empirical
GenoCanyon	0.2307	0.70624	0.162929	11	Unsupervised
phastCons	0.221789	0.995113	0.220706	11	Unsupervised
fitCons	0.166508	1.090584	0.181591	11	Unsupervised
Yeast proteins					
Predictor	**RankScore**	**Cofficient of variation**	**Standard deviation**	**Number of proteins**	**Category**
SNPs&GO	0.858236	0.156899	0.134656	4	Supervised
SNAP2	0.852111	0.032678	0.027845	4	Supervised
PPH2_HumDiv	0.833507	0.127008	0.105862	4	Supervised
SNP&GOs3D	0.805788	0.139875	0.112709	4	Supervised
PANTHER	0.783467	0.132181	0.10356	3	Empirical
PhD_SNP	0.770593	0.244619	0.188501	4	Supervised
S3D-PROF	0.71706	0.137327	0.098472	4	Supervised

DeepSequence	0.682324	0.493253	0.336558	4	Unsupervised
PROVEAN	0.635131	0.149621	0.095029	4	Unsupervised
SIFT	0.578382	0.305186	0.176514	4	Empirical
SuSPect	0.538334	0.583936	0.314353	4	Supervised
NetDiseaseSNP	0.439179	0.404522	0.177658	4	Supervised
BLOSUM62	0.206271	0.965246	0.199102	4	Empirical
Grantham	0.057549	1.476692	0.084982	4	Empirical
Bacterial protein					
Predictor	**RankScore**	**Cofficient of variation**	**Standard deviation**	**Number of proteins**	**Category**
DeepSequence	0.961651	0.074785	0.071917	6	Unsupervised
PPH2_HumDiv	0.855432	0.094009	0.080418	6	Supervised
SuSPect	0.787043	0.104549	0.082285	6	Supervised
S3D-PROF	0.781873	0.157896	0.123455	4	Supervised
PROVEAN	0.764829	0.19325	0.147804	6	Unsupervised
SNAP2	0.75292	0.451042	0.339598	6	Supervised
SIFT	0.745189	0.129724	0.096669	6	Empirical
SNP&GOs3D	0.719106	0.199922	0.143765	4	Supervised
PhD_SNP	0.655059	0.393602	0.257833	6	Supervised
SNPs&GO	0.611539	0.334995	0.204862	6	Supervised
NetDiseaseSNP	0.25414	0.548375	0.139364	6	Supervised
BLOSUM62	0.227926	0.41492	0.094571	6	Empirical
Grantham	0.028098	2.236068	0.062829	6	Empirical
Viral proteins					
Predictor	**RankScore**	**Cofficient of variation**	**Standard deviation**	**Number of proteins**	**Category**
PROVEAN	0.890462	0.118196	0.105249	4	Unsupervised
SuSPect	0.853037	0.039862	0.034003	4	Supervised
PhD_SNP	0.774817	0.291176	0.225608	4	Sueprvised
SNAP2	0.759121	0.512382	0.38896	4	Supervised
BLOSUM62	0.614111	0.341185	0.209525	4	Empirical
SNPs&GO	0.610413	0.319614	0.195097	4	Supervised
NetDiseaseSNP	0.587275	0.410319	0.24097	4	Supervised

SIFT	0.574467	0.180874	0.103906	4	Empirical
DeepSequence	0.298481	0.396879	0.118461	3	Unsupervised
Grantham	0.171763	1.732051	0.297502	4	Empirical


Total Mutations
HA-H3N2 10754
infA 1368
HA-H1N1 10716
P53 7448
MAPK1 6810
ADRB2 7800
bla(a) 5198
bla(d) 4997
bla(b) 4978
SUMO1 1700
HRAS 3135
PTEN(b) 6564
UBE2I 2563
TPMT 3689
env 12863
TPK1 3181
CALM1 1813
ccdB 1176
GmR 1929
PTEN(a) 4112
HSP82 4325
haeIIIM 1929
UBI4(b) 1380
bla(c) 990
UBI4(a) 1195
PA 1817
PAB1 1187
Cas9 2351
BRCA1(a) 2990
GAL4 1196
BRCA1(b) 1837
Protein
Organism of Origin
Human
Yeast
Bacteria
Virus
0
20
40
60
80
100
Total Coverage (%)

A

Predictor Category
Supervised
Unsupervised
Empirical
Metapredictor

FATHMM
SuSPect
DeepSequence
REVEL
DEOGEN2
DeepSequence
MPC
MPC
Envision
DeepSequence
DEOGEN2
VEST4
S3D-PROF

Spearman's correlation

P53 MAPK1 ADRB2 SUMO1 HRAS PTEN(b) UBE2I TPMT TPK1 CALM1 PTEN(a) BRCA1(a) BRCA1(b)

Protein

B

Yeast
Bacteria
Virus

PPH2_HumDiv
PhD_SNP
SNPs&GO
SNP&GOs3D
DeepSequence
DeepSequence
S3D-PROF
DeepSequence
PPH2_HumDiv
DeepSequence
DeepSequence
PhD_SNP
SNAP2
SNAP2
PhD_SNP

Spearman's correlation

GAL4 HSP82 UBI4(a) UBI4(b) PAB1 infA bla(a) bla(b) bla(c) bla(d) ccdB GmR haeIIIM Cas9 HA-H3N2 HA-H1N1 env PA

Protein