Mutation severity spectrum of rare alleles in the human genome is predictive of disease type

Jimin Pei[1], Lisa N. Kinch[1], Zbyszek Otwinowski[2] and Nick V. Grishin[1,2,*]

[1]Howard Hughes Medical Institute, and [2]Departments of Biophysics and Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9050, USA

[*]Contact: grishin@chop.swmed.edu

Abstract

The human genome harbors a variety of genetic variations. Single-nucleotide changes that alter amino acids in protein-coding regions are one of the major causes of human phenotypic variation and diseases. These single-amino acid variations (SAVs) are routinely found in whole genome and exome sequencing. Evaluating the functional impact of such genomic alterations is crucial for diagnosis of genetic disorders. We developed DeepSAV, a deep-learning convolutional neural network to differentiate disease-causing and benign SAVs based on a variety of protein sequence, structural and functional properties. Our method outperforms most stand-alone programs, and the version incorporating population and gene-level information (DeepSAV+PG) has similar predictive power as some of the best available. We transformed DeepSAV scores of rare SAVs observed in the general population into a mutation severity measure of protein-coding genes. This measure reflects a gene's tolerance to deleterious missense mutations and serves as a useful tool to study gene-disease associations. Genes implicated in cancer, autism, and viral interaction are found by this measure as intolerant to mutations, while genes associated with a number of other diseases are scored as tolerant. Among known disease-associated genes, those that are mutation-intolerant are likely to function in development and signal transduction pathways, while those that are mutation-tolerant tend to encode metabolic and mitochondrial proteins.

Author Summary

Human genetic variations in various forms are constantly found in whole genome and exome sequencing of general population and patients. It remains a challenging task to assess the functional impact of these variations. In this study, we performed comprehensive analysis of single-amino-acid variations (SAVs) in terms of their sequence, structure, and functional properties. We further developed a deep neural network-based method to predict the functional impact of SAVs. Our method is among the top performers compared to existing programs in differentiating pathogenic and benign SAVs. We designed a mutation severity measure for human protein-coding genes by aggregating the predicted scores of SAVs found in the human general population. Such a measure reflects a gene's tolerance to deleterious missense mutations and serves as a useful tool to study gene-disease associations. We found that genes implicated in cancer, autism, and viral interaction are more likely to be intolerant to mutations than genes with other diseases. Disease-associated genes with strong mutation intolerance tend to function in development and signal transduction pathways. On the other end of the mutation severity spectrum, mutation-tolerant genes often encode proteins functioning in mitochondria and metabolic pathways.

Introduction

Genetic variations are major determinants of human diseases and phenotypes [1]. Accelerating pace of large-scale sequencing projects on genomes and exomes has greatly expanded the landscape of human genetic variations. It remains a challenging task to assess the functional impact of these variations [2]. Comprehensive analysis of genetic variations, especially those found in and near the exons of protein-coding genes [3], may shed light on gene-disease relationships and provide insight into the mechanisms of diseases and variations in phenotypes [4]. The increasing number of sequenced human genomes and exomes from the general population would enhance the statistical power of such analyses [5].

Different types of genetic variations occur at a range of scales from large structural variations such as chromosomal rearrangements and copy number variations (CNVs), to insertions and deletions (indels) of up to hundreds of nucleotide positions, and to single-base-pair (single-nucleotide) variations (SNVs) [6]. Any type of genetic variation could cause human disease with a variety of mechanisms, including effects on chromatin organization, gene expression and regulation, protein function, and genetic instability [7-11]. The observed frequencies of genetic variations in the general population are tied to their fitness cost as well as the evolutionary history of the human species and its ancestors. While common variations, most notably SNVs, were first documented, more rare genetic variations (e.g., those with minor allele frequency (MAF) less than 0.0001) at the individual level have been identified in large-scale sequencing projects of the general population [5] as well as patients with certain diseases such as cancer [12] and intellectual disability [13]. Although some recurring variations have been identified to be the drivers of diseases, a significant number of rare mutations are persistently found, and their clinical significance are difficult to evaluate. Genome-wide association studies can pinpoint the genetic loci, mostly marked by common SNVs, with statistically significant disease or phenotype associations [14, 15]. Association of rare and de novo mutations to common and rare diseases could be unveiled through familial or trio studies that are facilitated by genome or exome sequencing nowadays [16, 17]. Coupled with pathway profiling, systematic analysis of genetic variations in patients could shed light on the biological processes underlying diseases [18]. However, disease gene prioritization and disease-causing variation discovery are still difficult [19, 20].

The identity change in a single base pair position is the most common type of genetic variation. In protein-coding regions, non-synonymous variations (missense mutations) result in the change of a single amino acid in the protein product [21]. Clinical consequences of these missense mutations, referred to as single amino acid variations (SAVs), are generally more difficult to evaluate than synonymous mutations (generally benign) and nonsense (stop codon) mutations (often resulting in loss of function). A number of computational methods [22] have been developed to assess the mutational effects of SAVs found in the human proteome encoded by around 20,000 protein-coding genes.

Essential genes compromise the viability of an individual when their function is lost. Such genes can be identified by observing intolerance to loss-of-function variants at the population level [23]. In genetic terms, essential genes tend to exhibit haploinsufficiency, where the loss of one of two gene alleles is detrimental. Genetic alterations of haploinsufficient genes are not only a major cause of dominant diseases [24], but also play key roles in developmental disorders [17]. On the one hand, haploinsufficient genes can function as tumor supressors [25]. On the other hand, essential genes tend to be expressed at higher levels in cancer cells than in normal cells [26]. Thus, knowledge about gene essentiality can help prioritize deleterious variants in genetic studies and could help prioritize therapeutic targets in cancer.

Given the role of essential genes in human disease, considerable efforts have gone into developing methods for haploinsufficiency prediction [5, 27-30].

In this study, we developed a deep convolutional neural network-based method for predicting the clinical impact of SAVs in the human proteome based on analysis of their sequence, structural and functional properties. The neural network prediction results of SAVs observed in the general population were used to calculate a mutation severity measure that estimates tolerance of each human protein-coding gene to deleterious missense mutations. This measure correlates with gene essentiality and specific disease classes such as cancer and autism. Finally, we observed a dichotomy of mutation severity for disease-associated genes: those that are mutation-intolerant tend to function in development and signal transduction pathways, while those that are mutation-tolerant tend to function in metabolism.

Results and discussion

Analysis of human disease-related genes and their variants

We obtained a set of likely pathogenic (disease-causing) genetic variants from two database resources: ClinVar [31] and UniProt [32]. ClinVar aggregates reported variant-disease associations from submissions of research studies. ClinVar variants annotated as "Pathogenic" or "Likely pathogenic" were found in ~4,200 protein-coding genes, about one fifth of the human proteome. SAVs were found in the majority (3,410) of these genes. Non-SAV variants were also found in most of them (~3,300 genes). Non-SAV variants include indel variants, single-nucleotide variations in noncoding regions (mostly at splice sites), nonsense single-nucleotide variations (to stop codons), and a small number of synonymous variants (Figure 1A). The 31,171 SAVs made up about 30% of all ClinVar variants (Figure 1B). UniProt is another curated resource for likely pathogenic SAVs. The number of proteins with UniProt SAVs annotated as disease-related is 2,755 (Figure 1C). Most of these genes (2,590) overlap with the ClinVar disease-associated gene set, with UniProt contributing only 165 disease-associated genes not found in the ClinVar set. On the other hand, more than half of the UniProt pathogenic variants (15,697 out of 29,300, Figure 1D) were not found in the set of ClinVar pathogenic variants. The total number of likely pathogenic variants in the unified ClinVar and UniProt set is ~47,000. We also obtained a set of benign variants (~45,000) by combining the ClinVar variants annotated as "Benign" or "Likely benign" and the UniProt variants in the category of "Polymorphism".

The number of likely pathogenic SAVs are not evenly distributed among the disease-associated genes. The three genes with the greatest number of SAVs encode long proteins: FBN1 (Fibrillin-1, 2,871 amino acids), LDLR (Low-density lipoprotein receptor, 860 amino acids), and SCN1A (Sodium channel protein type 1 subunit alpha, 2,009 amino acids), each of which has more than 500 pathogenic SAVs. In part, it may be due to the length of these proteins. 75 genes possess more than 100 pathogenic SAVs. More than half of the disease-associated genes with SAVs (2,003 out of 3,575) have less than 5 pathogenic SAVs, and 916 of them have only one pathogenic SAV. One cause of the uneven distribution of SAVs could be the bias in research studies of common diseases and certain genes (e.g., the *LDLR* gene involved in hypercholesterolemia).

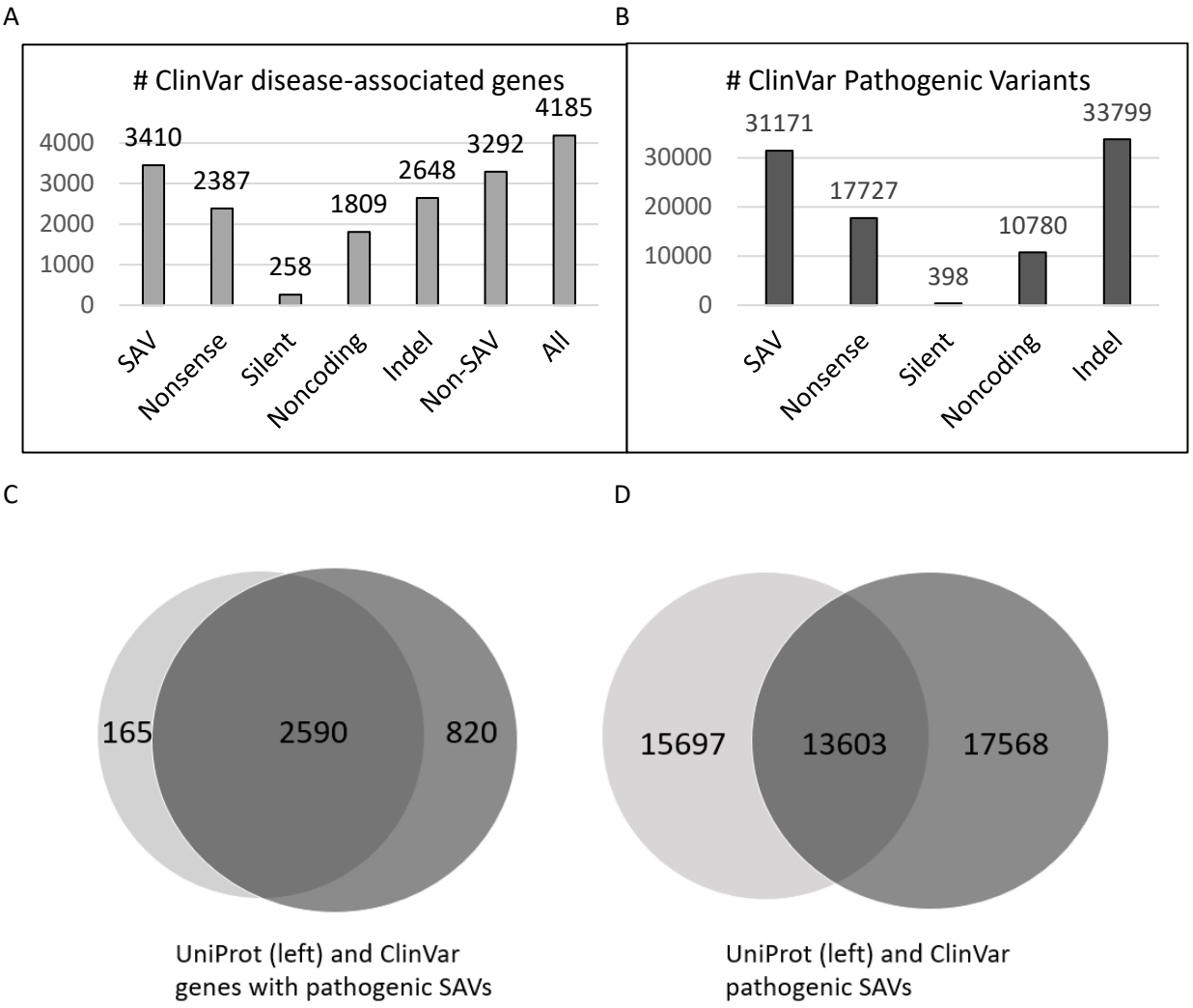


Figure 1. **Distribution of disease-associated genes and variants**. **A)** The number of ClinVar disease-associated genes with different types of variants. The non-SAV category combines the categories of nonsense, silent (synonymous), noncoding, and indel. **B)** The number of variants of different types in ClinVar disease-associated genes. **C)** Venn diagram of genes with pathogenic SAVs from UniProt and ClinVar. **D).** Venn diagram of pathogenic variants from UniProt and ClinVar.

Enrichment analysis of sequence, structure and functional properties in likely pathogenic SAVs and gnomAD SAVs

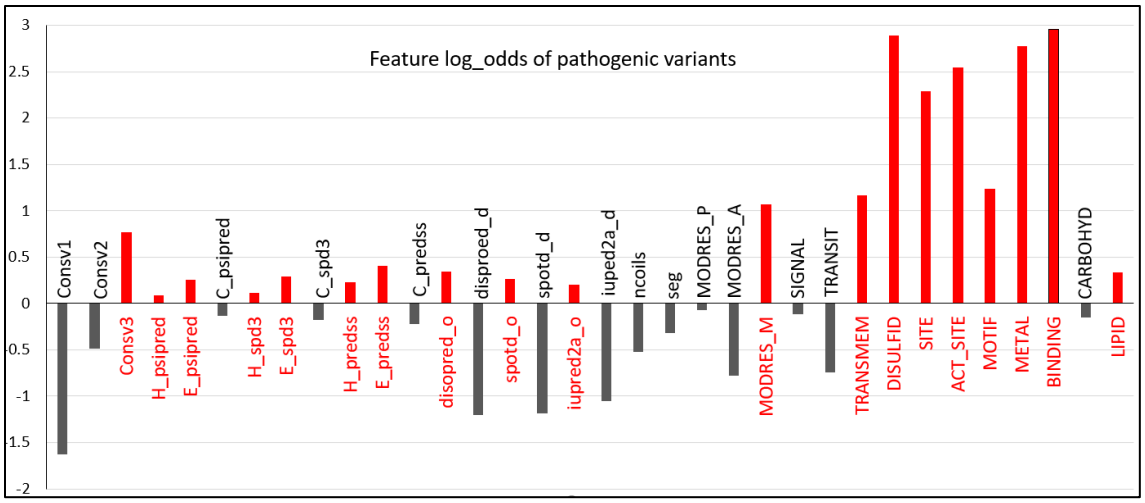
We compiled a set of protein sequence, structure, and functional properties (features) predicted by computer programs or retrieved from UniProt Feature fields (see Materials and methods). A log-odds score was used to determine if any feature is enriched or depleted in amino acid positions with pathogenic SAVs compared to the background frequency of that feature in all human proteins (see Materials and methods). We observed a 1.7-fold enrichment of conserved positions (Consv3 in Figure 2A) and more than 3-fold depletion of variable positions (Consv1 in Figure 2A) in pathogenic SAVs. Similarly, results of three disorder prediction programs (DISOPRED3 [33], SPOT-Disorder [34], and IUPred2A [35]) consistently show that ordered regions are enriched and disorder regions are depleted in pathogenic SAVs. Predicted β-

strands and α-helices are slightly preferred in pathogenic SAVs, while coil regions of secondary structure prediction, low complexity regions, and coiled coil regions are disfavored.

For regions with indications of subcellular localization, signal peptides and mitochondrial transit peptides are depleted in pathogenic SAVs, but transmembrane segments are enriched by more than 2 fold. Several UniProt features showing the strongest enrichments in pathogenic SAVs are related to protein stability (UniProt feature DISULFID: cysteine residues participating in disulfide bonds) or function (UniProt features: SITE, ACT_SITE, METAL, MOTIF, and BINDING, see their explanations in Materials and methods). Except the MOTIF feature, they exhibit more than 4-fold enrichment in pathogenic SAVs (log2-based odds score more than 2, Figure 2A).

We also analyzed SAVs found in more than 12,000 exomes (>24,000 alleles) in the gnomAD [5] database, which provides a comprehensive catalogue of natural variants from the general population. Common SAVs (MAF ≥ 0.01) should be mostly benign, and they only make up a small fraction of gnomAD SAVs (27,813 out of 4,885,239, about 0.57%). The gnomAD database possesses many more rare SAVs,

A



B

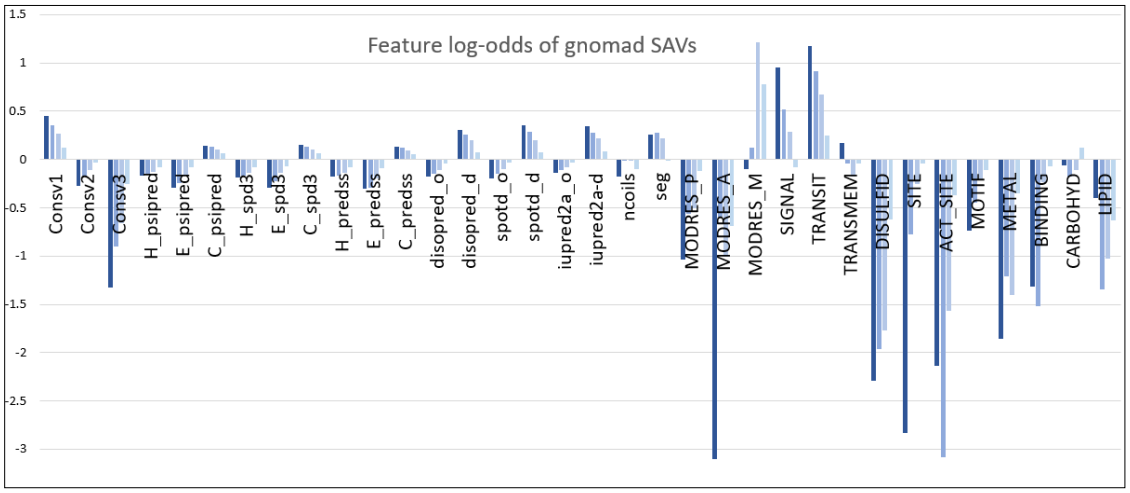


Figure 2. **Enrichment of SAVs among sequence, structure and functional properties**. **A)** Enrichment/depletion of features in pathogenic SAVs (y-axis shows log2 based log-odds scores). **B)** Enrichment/depletion of features in gnomAD SAVs with different MAF ranges (from light blue to dark blue: MAF < 0.0001, 0.0001 ≤ MAF < 0.001, 0.001 ≤ MAF < 0.01, 0.01 ≤ MAF).

with MAF less than 0.01, a significant portion of which are singletons (found only once in all exomes). We partition gnomAD SAVs according to their MAFs into four categories (MAF < 0.0001, 0.0001 ≤ MAF < 0.001, 0.001 ≤ MAF < 0.01, and MAF ≥ 0.01). The majority of SAVs (4,588,805 out of 4,885,239, about 94%) fall into the category of rare SAVs with MAF < 0.0001, while about 4.4% (27,813) and 1.1% (53,489) belong to the categories 0.0001 ≤ MAF < 0.001 and 0.001 ≤ MAF < 0.01, respectively. The population bottleneck events could be partially responsible for the depletion of common SAVs [36], and the explosive population growth in recent history can lead to excessive amount of rare SAVs [37].

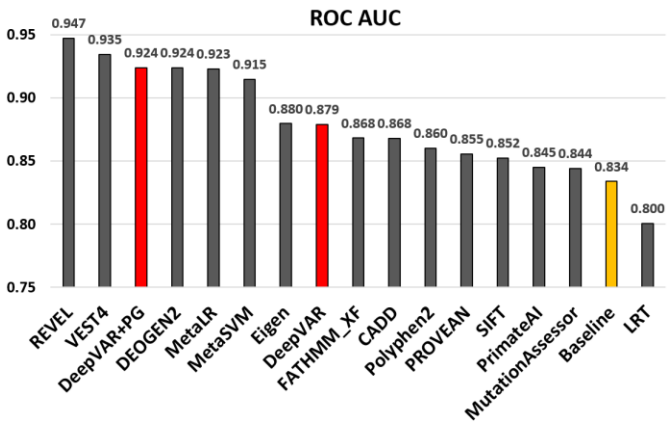
Enrichments of protein sequence, structure, and functional features in each gnomAD SAV category were analyzed in the same way as for pathogenic SAVs (Figure 2B). Common gnomAD SAVs (MAF ≥ 0.01) generally exhibit opposite enrichment/depletion trends compared to pathogenic SAVs. Features such as DISULFID, SITE, ACT_SITE, METAL, MOTIF, and BINDING exhibit the most prominent depletion in common gnomAD SAVs and the strongest enrichment in pathogenic SAVs. In contrast, features enriched in common SAVs include variable positions (Consv1), coil regions of secondary structure prediction, predicted disordered regions, low complexity regions, signal peptides, and mitochondrial transit peptides. The enrichment or depletion of features were gradually curtailed when moving from the category of common gnomAD SAVs to less frequent gnomAD SAV categories (Figure 2B). This behavior suggests that many low frequency SAVs, especially those with MAF less than 0.0001 in the general population could be deleterious, because functionally important residues (specified by UniProt features SITE, ACT_SITE, BINDING, METAL, and MOTIF) are found more frequently in these rare SAVs than in the common SAVs.

DeepSAV – a deep neural network-based method for SAV pathogenicity prediction

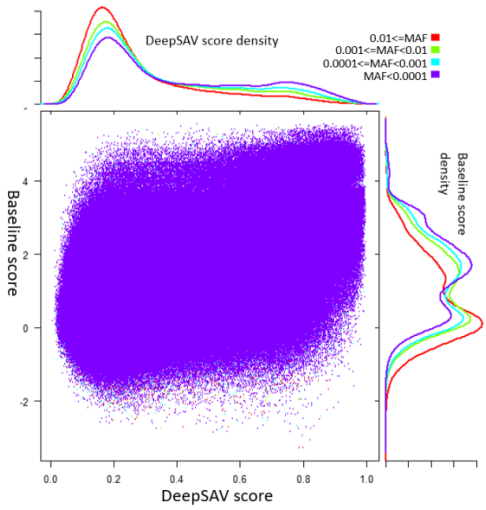
We developed an artificial intelligence method (DeepSAV) that uses a deep-learning convolutional neural network to predict SAV pathogenicity based on input features of sequence, structure, and functional information (see Materials and methods). The features include amino acid type, sequence profile, sequence conservation, secondary structure and disorder predictions, coiled coil and low complexity region predictions, sequence regions indicating subcellular localization (signal peptide, transit peptide, transmembrane segments), and functional and stability properties from the UniProt database such as post-translational modifications, disulfide bond, active site, and motifs. Features of a window of 21 amino acid positions centered around the mutated amino acid were encoded as input. The neural network has mainly convolutional layers and applies techniques such as max-pooling, residual network, and dropout (supplemental Figure S1). It is trained and tested on a large set (43,000 pathogenic and 43,000 benign) of SAVs from the ClinVar and UniProt database.

Cross validation test of DeepSAV showed that it yielded better performance (measured by area under the ROC (receiver operating characteristic) curve (AUC)) to differentiate pathogenic from benign SAVs than most stand-alone programs such as SIFT [38], PolyPhen-2 [39], FATHMM-XF [40], PROVEAN [41], CADD [42], LRT [43], MutationAssessor [44], PrimateAI [45], and a simple baseline fitness score (Baseline) we used before in the Critical Assessment of Genome Interpretation (CAGI) evaluations [46]

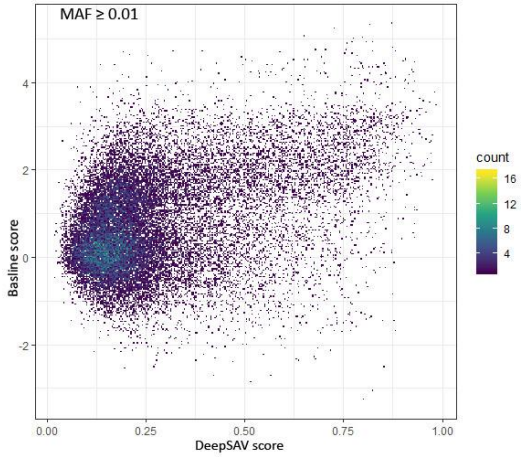
A


ROC AUC
0.95
0.90
0.85
0.80
0.75
0.947
0.935
0.924
0.924
0.923
0.915
0.880
0.879
0.868
0.868
0.860
0.855
0.852
0.845
0.844
0.834
0.800
REVEL
VEST4
DeepVAR+PG
DEOGEN2
MetaLR
MetaSVM
Eigen
DeepVAR
FATHMM_XF
CADD
Polyphen2
PROVEAN
SIFT
PrimateAI
MutationAssessor
Baseline
LRT

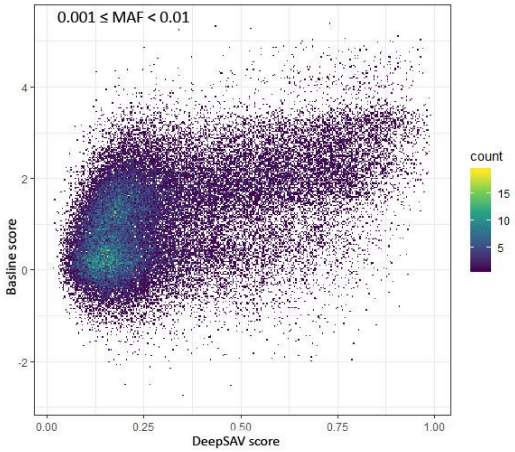
B


DeepSAV score density
0.01<=MAF
0.001<=MAF<0.01
0.0001<=MAF<0.001
MAF<0.0001
Baseline score
Baseline score density
DeepSAV score
0.0
0.2
0.4
0.6
0.8
1.0

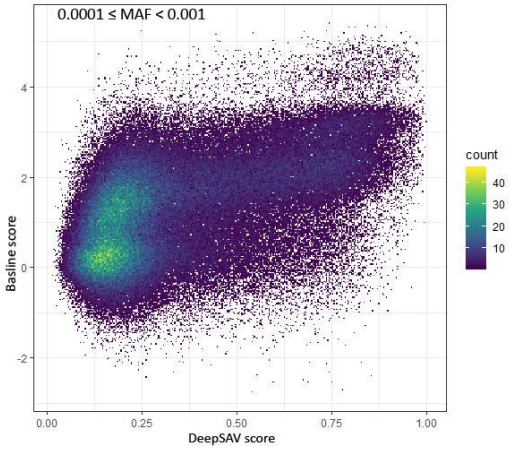
C


MAF ≥ 0.01
Basline score
DeepSAV score
count
16
12
8
4

D


0.001 ≤ MAF < 0.01
Basline score
DeepSAV score
count
15
10
5

E


0.0001 ≤ MAF < 0.001
Basline score
DeepSAV score
count
40
30
20
10

F

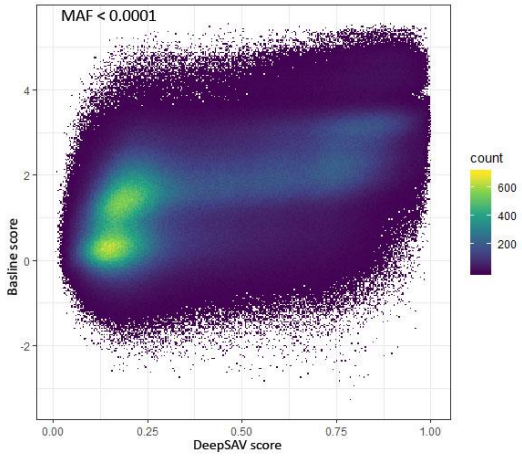

MAF < 0.0001
Basline score
DeepSAV score
count
600
400
200

Figure 3. **A)** Performance of variant pathogenicity prediction programs in terms of AUC (area under the ROC curve) measure. **B)** Scatter plot of DeepSAV scores and baseline fitness scores for SAVs observed in gnomAD. Datapoints for four different MAF categories are shown. Their density plots are shown by the axes above (DeepSAV) and right (baseline fitness). **C)** to **F)** Two-dimensional histograms (made with R ggplot2 package) of DeepSAV scores and baseline scores for gnomAD variants with MAF ≥ 0.01 (**C**), 0.001 ≤ MAF < 0.01 (**D**), 0.0001 ≤ MAF < 0.001 (**E**), and MAF < 0.0001 (**F**).

(Figure 3A). DeepSAV's performance is similar to (Eigen [47]) or worse than several meta-predictors (MetaSVM [48], MetaLR [48], and REVEL [49]) that use prediction results of a number of other stand-alone predictors (Figure 3A). DeepSAV trails the two best methods REVEL and VEST4 by about 0.06 in AUC. We were not able to find information about the algorithm of VEST4, an improved version of VEST [50], which shows the second best performance, and used VEST4 scores as given in dbNSFP for comparison [22].

The DeepSAV predictor is based on information derived from amino acid positions. Thus, the DeepSAV score reflects the level of deleterious effect to the target protein for any given variant based on its sequence, structural and functional properties. However, protein-level deleterious effects do not necessarily lead to human diseases, as a significant fraction of human protein-coding genes could be compromised without causing diseases. Variations in essential genes are more likely to cause diseases than in non-essential genes, and disease-associated genes are generally involved in more protein-protein interactions than genes not associated with diseases [36]. Adding information based on gene-level annotation or predictions such as gene essentiality and the number of protein-protein interactions have proven useful in improving the differentiation between disease-causing mutations and benign mutations [36]. Indeed, the performance of DEOGEN2 [51], which incorporates heterogeneous information such as the relevance of the gene and the number of protein interactions, is among the best (ROC AUC: 0.924) in our test. Another valuable source of information independent from the amino acid positions are the occurrence frequencies of the variants observed in the human general population, which is available as minor allele frequencies (MAFs) of the variants in the gnomAD database. We added gnomAD MAF and 17 gene-level features (extracted from dbNSFP, see Materials and methods) to the amino-acid-level features in DeepSAV. The resulting predictor DeepSAV+PG (DeepSAV with population and gene-level information) was able to boost the performance from ROC AUC 0.879 to 0.924, close to some of the best methods (Figure 3A).

We further calculated DeepSAV scores and baseline fitness scores for human protein SAVs observed in the gnomAD database. They show a positive correlation (correlation coefficient: 0.57), and both exhibit bimodal distributions for SAVs in each of the four different MAF categories (MAF ≥ 0.01 (common SAVs), 0.001 ≤ MAF < 0.01, 0.0001 ≤ MAF < 0.001, and MAF < 0.0001) (Figure 3B-3F). The range of DeepSAV scores is between 0 and 1, with higher scores suggesting an increasing likelihood of being deleterious (pathogenic). For common SAVs (MAF ≥ 0.01), the distribution of DeepSAV exhibits a high peak in the low score range, and a flat tail in the high score range, suggesting that the majority of common SAVs are predicted to be benign. With increasing stringencies of rare SAVs, the volume of the peak in the low-score range decreases while the tail in the high-score range increases, suggesting that pathogenic SAVs are more likely to occur in rarer SAVs. The baseline fitness scores display similar behavior for SAVs in different MAF categories, although the peaks in high and low scoring ranges appears to overlap more compared to the DeepSAV scores.

Mutation severity scores enrich for essential genes with potential disease associations

Deep sequencing of human exomes has highlighted the contribution of rare SAVs to gene function and complex diseases [52, 53]. Certain genes may be more tolerant to deleterious or partially deleterious SAVs due to their functional properties. To evaluate the mutation tolerance of genes, we chose to use the DeepSAV scores that reflect a variant's deleterious effect on the protein product, as the features used for training are based on protein sequence, structure, and functional properties. DeepSAV+PG scores are better than DeepSAV scores at discriminating pathogenic (disease-causing) variants from benign variants by adding gene-level information that correlates with the likelihood of a gene associated with diseases (e.g., gene essentiality and interaction numbers). However, to more objectively assess the mutation tolerance of genes, DeepSAV scores were used as they reflect the deleterious effects on the protein products regardless of whether the genes are disease-associated. We transformed DeepSAV predictions of SAVs present in the human population (from the gnomAD database [5]) into an average mutation severity measure for each gene (AvgAI scores, see Materials and methods). AvgAI scores based on our deep neural network predictor were calculated for SAVs with several filters for common variants (MAF less than 1, 0.01, 0.001, or 0.0001), and were compared to the same scores calculated using a simple baseline predictor [46] (AvgBF) (see Materials and methods). For rare SAVs (MAF < 0.0001) the baseline AvgBF score correlates well (R^2 = 0.78) with the AvgAI score (Figure 4A), suggesting that the deep neural network predictor reflects the profile score difference between amino acids of major and minor alleles that go into the baseline predictor.

Human genes have been classified by a measure (LOEUF) that reflects their tolerance to inactivation (loss-of-function) [5]. To see how our AvgAI score correlates with the LOEUF score, we ranked human genes from low AvgAI (mutation-intolerant) to high AvgAI (mutation-tolerant). The LOEUF distribution for top-ranking mutation-intolerant genes was compared to that of a set of known disease-associated genes (Figure 4B). The top-ranking mutation-intolerant genes selected by the mutation severity measure (lowest AvgAI scores) include progressively more loss-of-function constrained genes with increased filtering of common variants. In contrast, the disease-associated gene set displays a bimodal distribution of highly constrained genes at low LOEUF and less constrained genes at median LOEUF. Thus, the mutation severity measure for rare SAVs reiterates a gene's tolerance to inactivation, with top-ranking mutation-intolerant genes being more frequent in the percentile of lowest tolerance to inactivation.

A fraction of the disease-associated human gene set (17%) is annotated as essential by one or more CRISPR screens [54, 55]. Among all curated gene-disease associations in DisGeNET (May 2019 version) [56], the essential disease-associated genes contribute to 2,477 diseases or syndromes and 1,847 neoplastic processes. We originally reasoned that genes able to accumulate numerous detrimental SAVs (evaluated by high AvgAI scores) were less likely to contribute to disease phenotypes. However, the AvgAI scores do not discriminate disease-associated genes collectively, giving similar gene frequencies displayed across the AvgAI deciles (Figure 4C, gray bars). Instead, AvgAI scores tend to select for gene essentiality, with an increase in essential genes and a decrease in non-essential genes at the lowest AvgAI decile (Figure 4C). Similar to noted trends of both essential and disease-associated genes [57, 58], human genes with AvgAI scores in the lowest decile, regardless of their essentiality, exhibit increased numbers of protein interactions than those from higher AvgAI deciles (Figure 4D).

Although the loss-of-function constraint measure LOEUF and the mutation severity measure AvgAI display similar trends in reflecting gene essentiality, they define different gene sets that might be

used to evaluate potential new disease-associated genes. A comparison of essential genes with the lowest LOEUF scores, essential genes with the lowest AvgAI scores, and essential genes with pathogenic SAVs highlights the divide among these gene sets (Figure 4E). The overlap between low-AvgAI set and low-LOEUF set (126 genes, not including genes with pathogenic SAVs) provides a potential source of disease-associated genes. Indeed, despite the lack of documented pathogenic SAVs in the 126 mutation- and inactivation-intolerant genes, curated DisGeNET gene-disease associations annotate almost half (55 genes) as being involved in disease. The set includes almost all disease classes, with several being over-represented when compared to all gene-disease associations: including virus diseases, stomatognathic diseases, immune system diseases, and neoplasms (Figure 4F). Given their propensity to associate with disease, the essential genes selected by our AvgAI measure could provide insight into novel gene-disease associations.

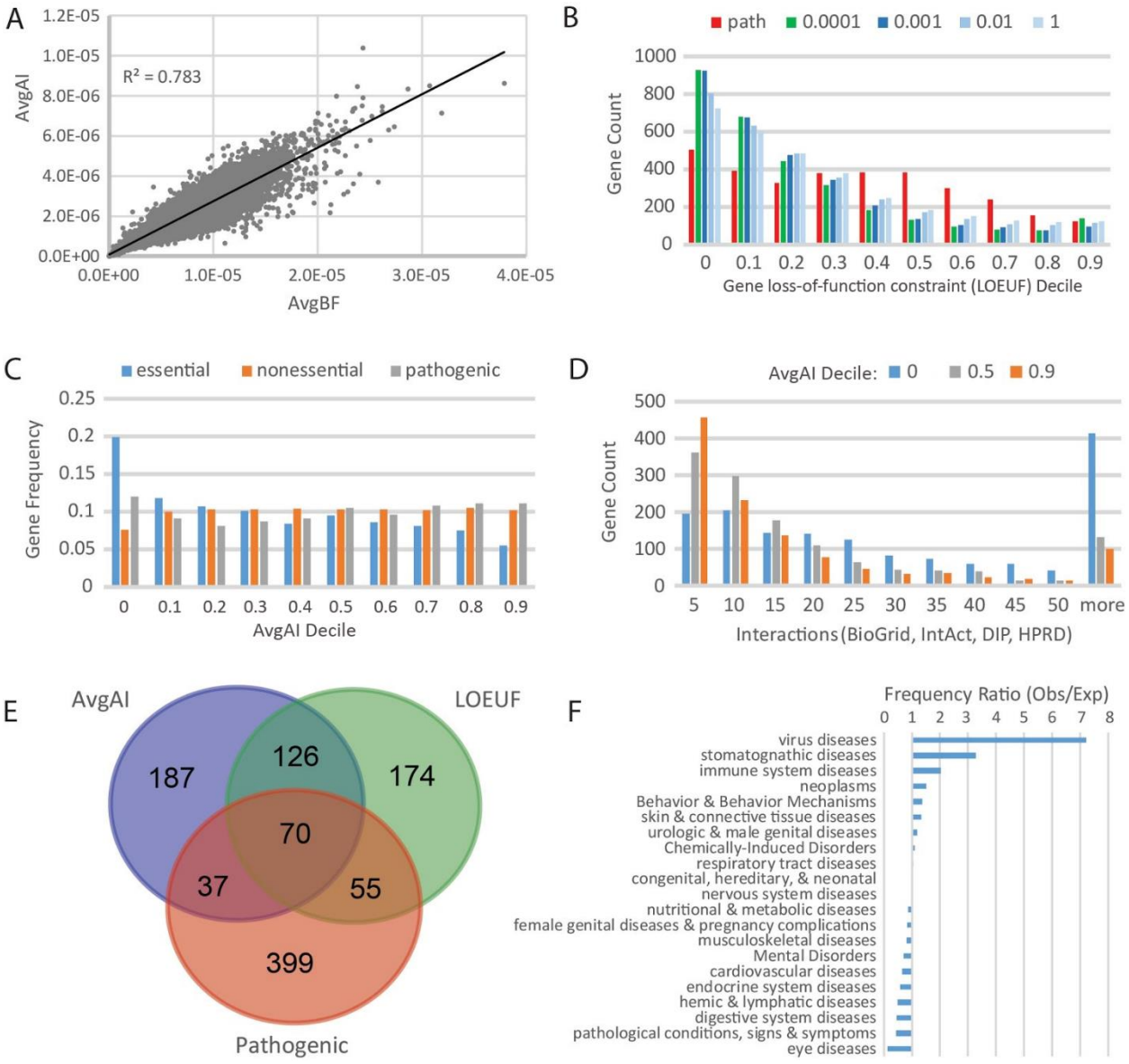


Figure 4. Mutation severity measures based on DeepSAV identify potential disease-associated genes. A) Mutation severity measure based on DeepSAV scores (AvgAI) correlates with the measure based on baseline fitness scores (AvgBF) for 17,480 human genes **B)** Distribution of gene count among decile bins of loss-of-function constraint measure (LOEUF) for a set of genes (>3,000) with pathogenic SAVs (red bars, labeled as "path") and for the gene sets (having the same number genes) with the lowest AvgAI scores computed at various cutoffs of minor allele frequencies (0.0001, 0.001, 0.01 and 1). On the x axis, 0 means the first LOEUF decile [0, 0.1] (the same extrapolation applies to other numbers). **C)** Distribution of gene frequency among AvgAI deciles (MAF cutoff 0.0001) for the same gene set with known pathogenic SAVs compared to essential and nonessential gene sets. **D)** Distribution of protein interactions from four databases (BioGrid [59], IntAct [60], DIP [61] and HPRD [62]) integrated in PICKLE [63] for gene sets within three different mutation severity AvgAI score deciles (0, 0.5 and 0.9). **E)** Venn diagram highlights overlap among essential genes with known pathogenic variants (labeled as "Pathogenic"), essential genes with lowest loss-of-function constraint scores (LOEUF), and essential genes with lowest mutation severity measure (AvgAI). **F)** Representation of disease class associated with genes from the overlapping set of top-ranked genes by LOEUF and AvgAI (126 genes, not including genes with known pathogenic SAVs).

Mutation-intolerant essential genes cluster with disease-associated genes and contribute to diseases

The essential genes with pathogenic SAVs (70 genes, Figure 4E) that overlap with both the loss-of-function-constrained genes (lowest LOEUF) and the low mutation severity genes (lowest AvgAI) set a standard to prioritize other potential disease-associated genes (126 overlapping genes shared by the low LOEUF and the low AvgAI sets, but without pathogenic SAVs, Figure 4E). Clustering these two sets of genes (70+126) together using complete linkage of correlated distances over six scores (see Materials and methods) places potential disease-associated genes among those that are known to be associated with diseases (Supplemental Figure S2). Two clusters (40 genes) with the highest proportions of disease-associated genes exhibit lower AvgAI scores than other clusters, indicating their intolerance to detrimental missense mutations (red labels in Supplemental Figure S2). Inspection of gene-disease associations for genes in these two clusters reveals that 68% are linked to curated diseases.

Enriched GO biological process terms are similar for each identified gene cluster, and annotation clustering of terms from the combined set (40 genes from two clusters, 17 with pathogenic SAVs) highlights their function in RNA splicing (enrichment score 9.02, 13 genes), gene expression (enrichment score 6.81, 31 genes), and chromosome segregation (enrichment score 4.34, 9 genes). Two disease-associated genes and eleven others belong to the most enriched cluster and function in RNA splicing, including pre-mRNA processing factor 3 (*PRPF3*) having variants associated with Retinitis pigmentosa, and splicing factor 3b subunit 1 (*SF3B1*) having variants associated with acute myeloid leukemia, among other neoplastic processes. Five of the potential disease-associated genes involved in RNA splicing (*DHX15*, *HNRNPH1*, *SRSF1*, *PCBP2*, and *DHX9*) are reported to be associated with myelodysplasias in DisGeNET [56], and the spliceosome has become a therapeutic target for myeloid malignancies [64, 65].

The third most enriched functional cluster includes six disease-associated genes and three others that function in chromosome segregation. Three of the disease-associated genes (*RAD21*, *SMC3*, and *SMC1A*) functioning in chromosome segregation have genetic variants causing Cornelia de Lange syndrome (CdLS), which manifests developmentally as intellectual and growth retardations. The protein-coding products of these genes comprise three of the four subunits of the mitotic cohesion complex responsible for chromosome segregation. Mutations in this complex are known to cause a number of diseases termed cohesinopathies, of which CdLS is the best characterized [66]. One additional chromosome segregation gene from this set, PDS5 cohesin associated factor A (*PDS5A*), is associated with

CdLS in DisGeNET literature [67]. Gene dosage appears to be an important component of CdLS severity, which is consistent with the essential nature of our selected gene set [23].

Mutation-intolerant disease-associated genes function in development and signaling pathways

Over half of the top 1,000 human genes ranked by low AvgAI (571 genes) are associated with 1,618 diseases, 262 phenotypes and 184 disease groups such as “Intellectual Disability” that encompass multiple similar diseases or phenotypes. To understand the functional context of mutation-intolerant genes that are associated with disease, we assigned them to pathways in Reactome [68] (467 genes). Functional enrichment of these pathways highlight involvement in axon guidance (P-value < 1.78E-15), development (P-value < 9.64E-14), and neurotransmitter receptors and postsynaptic signal transmission (P-value < 1.12E-13), among others. Those genes in the enriched category of "developmental biology" describe early steps in development that give rise to diverse tissues in the body and thus represent critical processes that should contribute to fitness. In fact, 25% of this gene set participate in development, and many are annotated as essential (47 genes) or conditionally essential (17 genes). However, a significant portion of the developmental genes are not considered essential (52 genes). Many of them encode protein kinases (18 genes), homeobox transcription regulators (4 genes) or proteins with other signaling domains that are expanded in the genome like rho-binding domains (3 genes), pleckstrin homology domains (4 genes), or SH3 domains (3 genes).

While a relatively small core set of essential genes exists in eukaryotes whose loss of function results in lethality, a larger subset of genes exhibits conditional lethality that also affects fitness [69]. For example, deleterious mutation of immune system genes might not necessarily result in a lethal phenotype. However, their contribution to survival under specific conditions like being challenged with an infectious agent could be considered as essential. This spectrum of gene essentiality is indeed reflected in the disease-associated genes functioning in development, as they exhibit essential and conditionally essential responses in CRISPR screens [54, 55]. Furthermore, many of the mutation-intolerant and disease-associated genes not considered as essential belong to families like protein kinases that have expanded in the human genome and could be functionally redundant [70]. Thus, the concept of gene essentiality alone does not necessarily suggest the ability to cause disease.

The mitogen-activated protein kinases ERK1 and ERK2 function in development and signal transduction pathways. They represent a duplication that is thought to be functionally redundant [71]. However, ERK2 includes two known pathogenic variants that are associated with various neoplastic diseases (E322K) as well as with inborn genetic disease (R135T). The ERK2 structure (Figure 5A) includes a relatively small set (9 positions) of DeepSAV-predicted deleterious SAVs from the gnomAD database (DeepSAV score >0.75). One of these SAVs (D106G) lines the ATP-binding pocket, and four are buried in the structure core (D44Y, G136E, R148H, and R194T), with R148 belonging to the HRD motif that controls kinase activation. The rest are in a C-terminal extension to the catalytic domain that lines the surface of the kinase in between the N-lobe and the C-lobe. The known pathogenic variants cluster together with many of the predicted deleterious mutations. Thus, while this kinase is thought to be functionally redundant, some variants have been reported as pathogenic, several others are predicted as detrimental, and the gene is intolerant to deleterious mutation (AvgAI score 3.24E-7 and ranked 142 out of more than 17,000 genes). Accordingly, the ERK2 gene was shown to be conditionally essential in a CRISPR screen [54], suggesting conditions exist where the functional redundancy of the two kinases breaks down.

Mutation intolerance appears to be a quality exhibited not only by genes associated with developmental disorders [17], but also by genes contributing to other various disease types such as cancer (COSMIC) [12], autism [72] (https://gene.sfari.org), and viral interacting proteins [73] (Figure 5C). However, mutation severity does not select for collective disease-associated gene sets (PathVar (ClinVar and UniProt genes with pathogenic SAVs) and DisGeNET, Figure 5D), with nearly uniform distributions of the number of genes among AvgAI deciles. Genes associated with X-linked diseases (from the Clinical Genomic Database) [74] exhibit pronounced preference for high mutation intolerance, with ~70% falling into the two lowest mutation severity deciles (Figure 5D). Genes associated with autosomal dominant diseases and genes associated with autosomal recessive diseases [74] show opposite trends in mutation intolerance (Figure 5D). The preference for mutation-intolerance in selected disease types suggests that the AvgAI score can be particularly useful for prioritizing disease genes when combined with additional considerations, such a disease type or functional pathways contributing to the disease state.

Mutation-tolerant genes function in metabolic pathways and mitochondria

The concept of functional redundancy from gene duplication extends not only to critical components of developmental and signal transduction pathways, but also to those of metabolic pathways [59]. Enriched functional pathways of mutation-intolerant genes that are associated with disease highlight repeated involvement of core genetic information processing (e.g., transcription and RNA processing) and signal transduction components, but they tend to exclude those of metabolism. In fact, mutation-tolerant genes (a numeric matched set of genes with the highest AvgAI scores) are significantly enriched in metabolism (P-value < 2.05E-13) in the Reactome pathway database [68].

An example of a mutation-tolerant gene product is platelet glycoprotein 4 (CD36), which functions in cell adhesion by serving as a receptor for thrombospondin in platelets as well as in the metabolism of lipids through binding long chain fatty acids. CD36 represents one of the most mutation-tolerant genes in the diseases-associated set, with 155 DeepSAV-predicted detrimental mutations (DeepSAV score>0.75) in 98 positions covering the structure, including seven lining the fatty acid binding site (Figure 5B). Although this gene is tolerant to mutation, known pathogenic variants (I413L, R386W, P90S, and F254L), with I413L lining the fatty acid binding pocket, cause platelet glycoprotein deficiency, a congenital disease of the hemic and lymphatic class. Furthermore, DisGeNET associates this gene with metabolic phenotypes of impaired glucose tolerance, insulin resistance, and insulin sensitivity. While mutations in CD36 can still lead to disease, the mutation tolerance of the gene might be explained by the recessive nature of the associated disease, by the ability of two paralogs, SCARB1 and SCARB2, to serve as functional replacements, or by the tissue-specific nature of the disease [75].

Therefore, a dichotomy seems to exist for disease-associated genes, where those that are mutation-intolerant tend to function in development and signal transduction pathways, while those that are mutation-tolerant tend to function in metabolism. These trends imply a greater overall fitness cost of mutations in developmental and signal transduction genes when compared to metabolic genes. However, extreme functional redundancy in some signal transduction proteins may lead to their tolerance to mutations. The mutation severity spectrum of signal transduction proteins with numerous paralogs that could exhibit functional redundancy are shown in Figure 5E. Paralogous olfactory receptors (OR), which represent a specialized set of G protein-coupled receptors (GPCRs) that detect odors, are more mutation-tolerant than other GPCRs. In fact, human OR paralogs include more pseudogenes [76] (464, not included


A
B
C
COSMIC
Virus Interact
Autism
Development
Gene Frequency
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8
9
10
Mutation Severity (AvgAI Percentile)
D
PathVar
DisGeNET
CGD_XL
CGD_AD
CGD_AR
Gene Frequency
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8
9
10
Mutation Severity (AvgAI Percentile)
E
Metabolic
Mitochondrial
Olfactory
GPCRs
Protein Kinases
Gene Frequency
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8
9
10
Mutation Severity (AvgAI Percentile)

Figure 5. Mutation-intolerant and mutation-tolerant genes prefer different pathways and disease types. A) Top ranked AvgAI genes like ERK2 kinase (PDB 4fmq) have relatively few DeepSAV predicted deleterious variant positions (DeepSAV score > 0.75, red spheres). One of these (black sphere) is near (< 4Å) the active site (ANP substrate analog in black stick). **B)** Bottom ranked AvgAI genes like CD36 (PDB 5lgd) are tolerant to predicted deleterious mutations (DeepSAV score > 0.75, red spheres), including several positions (black spheres) lining the fatty acid (black stick) binding sites or with known pathogenic variation in platelet glycoprotein deficiency (blue spheres). **C)** Mutation severity spectrum of disease-associated genes measured by their frequencies in AvgAI deciles. Associated disease for each gene set is labeled above. **D)** Mutation severity spectrum of disease-associated genes in general, measured by their frequencies in AvgAI deciles. (PathVar – genes with pathogenic SAVs in ClinVar and UniProt, DisGeNET – genes with diseases in DisGeNET database, CGD_XL, CGD_AR, and CGD_AD correspond to sets of genes associated with X-linked, autosomal recessive, and autosomal dominant diseases in the Clinical Genome Database, respectively) **E)** Mutation severity spectrum of pathway gene sets and large paralogous gene sets measured by their frequencies in AvgAI deciles.

in Figure 5E) that have accumulated enough mutations to render them inactive than functional genes (361, included in Figure 5E), and this well-known OR variability likely contributes to an individual's sense of smell [77]. Both non-OR GPCRs (Figure 5E, gray bars) and protein kinases (Figure 5E, orange bars) shift in the spectrum towards mutation intolerance when compared to either metabolic enzymes [78] (Figure 5E, dark green bars) or nucleus-encoded proteins functioning in the mitochondria [79] (Figure 5E, medium green bars), organelles that provide energy from nutrients using metabolic processes [80].

Metabolic enzymes exhibit similar tendency towards mutation-tolerance as the ORs (Figure 5E). One explanation for the greater tolerance of metabolic genes to mutations might be the redundancy not only in gene duplications, but also in non-homologous proteins that can serve as functional analogs of the same reactions [81]. Metabolites can also be acquired through transport mechanisms, relieving the evolutionary constraints on certain metabolic enzymes. Finally, metabolic pathways exhibit both redundancy and plasticity, allowing for multiple ways to arrive at the same metabolite [82].

The mutation-tolerance observed for nucleus-encoded mitochondrial proteins might reflect their roles in metabolic processes [80]. However, this tendency is also exhibited by the ribosomal proteins that function in mitochondria compared to ribosomal proteins functioning in cytoplasm (Figure 6A): the majority of mitochondrial ribosomal proteins have high AvgAI scores while the majority of cytoplasmic ribosomal proteins have low AvgAI scores. As an example, side-by-side comparison of ribosomal L14P/L23E-like proteins functioning in the cytoplasm (RPL23, Figure 6B and 6C) and the mitochondria (MRPL14, Figure 6B and 6D) highlights their different levels of mutation-intolerance. Both proteins adopt similar small 5-stranded meander barrel folds with relatively long loops that interact with RNA in the assembled ribosome. Cytoplasmic RPL23 and mitochondrial MRPL14 have 34 and 89 gnomAD SAVs respectively, and exhibit quite different DeepSAV distributions (Figure 6B). The cytoplasmic RPL23 includes only a single predicted pathogenic variant (I40F, DeepSAV score = 0.788, Figure 6C) (protein length: 140 amino acid residues), while MRPL14 includes 34 predicted pathogenic variants (DeepSAV score > 0.75, all but one are rare with MAF < 0.0001) covering 28 positions (Figure 6D) (protein length: 145 residues). Neither of these examples possess known pathogenic variants, and only the cytoplasmic version is associated with a neoplastic process in DisGeNET. This marked difference in rare allele mutation severity cannot be explained by either domain or pathway redundancy. The main function of mitochondria is to supply energy, which can be partly salvaged by increasing nutrient intake and decreasing energy-demanding activities. In addition, mitochondria might be able to overcome lowered fitness of mutations in the ribosome through their processes of fusion and fission that help maintain both

functional properties and the integrity of the mitochondrial genome that harbors the mitochondrial ribosomal RNA genes [80].

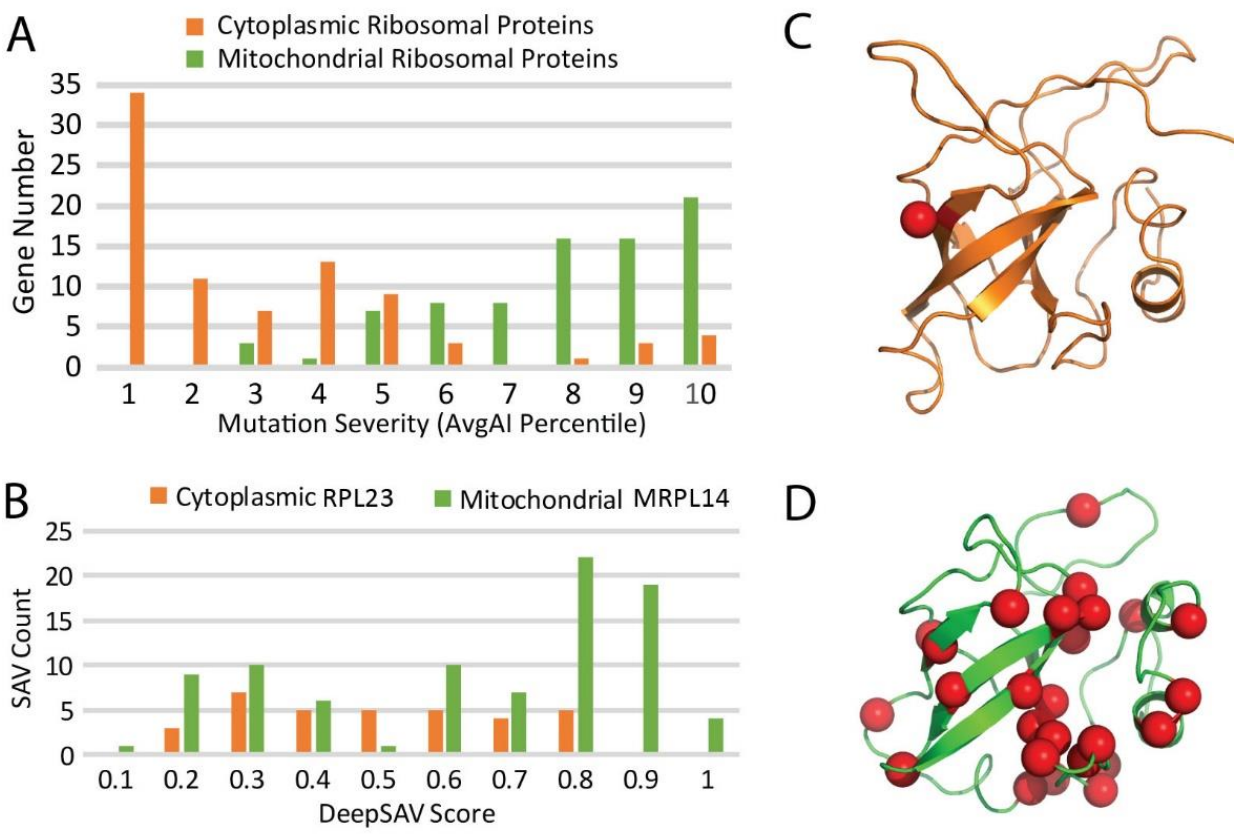


Figure 6. A) Mutation severity spectrum of ribosomal proteins functioning in the cytoplasm (orange bars) and in the mitochondria (green bars) measured by their numbers in AvgAI deciles. B). DeepSAV score distribution for 34 gnomAD SAVs of cytoplasmic ribosomal protein RPL23 (orange) and 89 gnomAD SAVs of mitochondrial ribosomal protein MRPL14. **C)** 60S ribosomal protein RPL23 from cytoplasm (PDB: 6ek0, chain LV) in orange cartoon has a single detrimental predicted SAV (red sphere). D**)** Mitochondrial 39S ribosomal protein MRPL14 (PDB 5oom, chain L) in green cartoon has multiple predicted detrimental SAVs.

Mutation-intolerant and mutation-tolerant genes function in different disease classes

The set of mutation-intolerant genes define several over-represented disease classes, including virus diseases, behavior & behavior mechanisms, stomatognathic diseases, hemic & lymphatic diseases, immune system diseases, musculoskeletal diseases, nervous system diseases, neoplasms, pathological conditions, signs & symptoms, and mental disorders (Figure 7A). Development and signal transduction are enriched among the mutation-intolerant genes associated with these specific disease classes. Furthermore, top mutation-intolerant genes tend to participate in relevant functional pathways. For example, mutation-intolerant genes associated with behavior diseases are enriched in neurotransmitter receptors and postsynaptic signal transmission (P-value < 1.11E-16), and those involved in immune system diseases are enriched in cytokine signaling of the immune system (P-value < 1.4E-14).

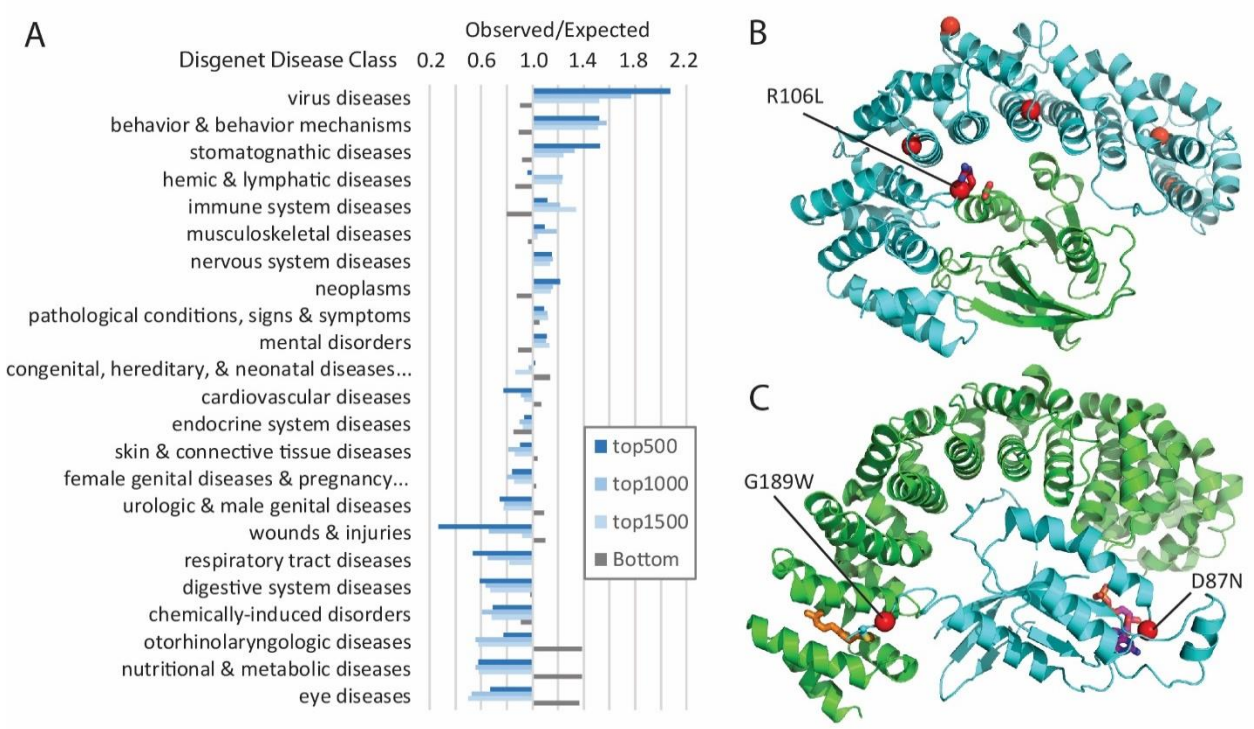


Figure 7. Mutation-intolerant genes exhibit pathway preferences and are exploited by viruses. A) Genes are ranked from low to high by mutation severity measure, AvgAI. The top ranked genes are mutation-intolerant, and the bottom ranked are mutation-tolerant. Ratios of observed/expected frequencies of disease class associations for sets of mutation-intolerant (top-) and mutation-tolerant (Bottom) genes are shown. Diseases are ordered by the exp/obs frequency ratios in the top1000 set (top 1000 genes with the lowest AvgAI score). **B)** Ribbon diagram of KPNB1 (cyan) bound to Ran GTPase (green) with DeepSAV-predicted detrimental variants (red spheres), including R106L (stick) at the interaction interface (from PDB 1ibr). **C)** Ribbon diagram of GEF (green) bound to RHOA GTPase (cyan, PDB 5zhx), with labeled DeepSAV-predicted detrimental variants (red spheres) adjacent to a farnesylation site (orange stick) and near the active site (stick colored by atom, from superimposed GTPase 1tx4).

The disease classes that are the most under-represented in mutation-intolerant (and over-represented in mutation-tolerant) genes include eye diseases, nutritional and metabolic diseases, otorhinolaryngologic diseases, chemically induced disorders, digestive system diseases, respiratory tract diseases, and wounds and injuries. These disease classes tend to be either tissue-specific or related to metabolism. For example, eye diseases involve genes functioning in visual perception, in cilium morphogenesis, as structural constituents of the eye lens, and in phototransduction. Alternatively, nutritional diseases involve genes of respiratory electron transport, steroid metabolism, TCA cycle, and fatty acid metabolism, among others. The nutritional diseases associated with mutation-intolerant genes tend to be dominated by a clinically heterogeneous group of disorders that arise as a result of dysfunction of the mitochondrial respiratory chain (mitochondrial diseases), as well as by obesity and diabetes that display a range of severity in affected individuals and can develop in adolescence or later in life. The relatively modest impact of these diseases on survival may be a reason for the genes associated with such diseases to tolerate mutations.

Viruses exploit disease-causing mutation-intolerant genes for infection

Viral diseases are the highest over-represented disease class among mutation-intolerant genes. Potentially, viral strategies for successful replication and evasion of host immunity could benefit from targeting essential genes that accumulate fewer mutations. In fact, similar observations of viral proteins interacting with more evolutionarily constrained host genes suggest that viruses have driven close to 30% of adaptive amino acid changes in the human proteome, with HIV infection causing a statistically significant increase in adaptation [73]. These evolutionarily constrained viral-interacting host proteins tend to be mutation-intolerant (Figure 5C), while a set of similarly highly adaptive proteins that interact with *Plasmodium* [83] do not have the same degree of preference for mutation-intolerance (data not shown).

Over a third of the virus disease gene set is involved in HIV coinfection, which describes simultaneous infection of a single host cell by two or more virus particles. Identified HIV coinfection-associated gene products function in pathways such as signaling by interleukins/cytokines, regulation of RUNX3, stabilization of P53, and host interactions with HIV factors, among others (ordered by Reactome enrichment). Cytokines, including interleukins, play a critical role in immunity. Because HIV infects immune CD4 T cells, the connection to interleukin/cytokine signaling molecules that regulate T cell growth and differentiation (i.e. through IL2 or CCL2) is known [84, 85], and two of the interleukin signaling examples (PSME3 and PSMC3) represent biomarkers for the disorder [86]. A significant portion of the HIV-related genes (19 out of 23 or 82%) are annotated as essential, including all host interaction factors (*KPNB1, RAN, PSMC3, PSME3, PSMA5, PSMA6, PSMC5, and PSMA4*), supporting the notion that infection strategies involving essential proteins are utilized by HIV, and potentially other viruses.

The essential HIV host factor importin subunit beta-1 (KPNB1) includes several gnomAD SAV positions predicted as detrimental using both DeepSAV and baseline fitness scores. However, the gene does not belong to our disease-causing set and has no disease associations in DisGeNET. KPNB1 mediates nuclear import of ribosomal proteins [87], and also works together with the RAN GTP-binding protein to bind and import HIV Rev into the nucleus where it exports viral mRNAs for translation [88]. In a structure of KPNB1 bound to RAN (Figure 7B) [89], these positions are buried (T150P, L238S, and A389V) or partially buried (R234G, C436Y) in the hydrophobic core of the KPNB1 repetitive α-hairpins. Such variations could result in local structure instability and loss of function. One surface SAV, R105L, interacts with a nearby E in RAN. Replacement of R by L, which removes a potential interaction of charges, could lower the key interaction of KPNB1 with RAN that drives nuclear import of HIV Rev.

Another example of a GTP-binding protein, RHOA, contributes to viral diseases such as Burkitt Lymphoma, which is a cancer of the lymphatic system with a subtype linked to Epstein-Barr virus (EBV) [90]. RHOA variants are deemed likely pathogenic for several other neoplastic disorders, and several missense variants are listed in DisGeNET, although not in association with Burkitt Lymphoma. Despite the apparent tumor-promoting effects of RHOA in various cancers, previous studies suggest mutations of the gene in the case of Burkitt lymohoma and other neoplastic processes are inhibitory [91, 92]. There are only two predicted detrimental SAVs in RHOA in gnomAD. One of them (G189W) maps to the disordered C-terminus adjacent to a residue that gets farnesylated. The disordered and modified C-terminus adopts a coil structure when bound to the RAP1GDS1 guanine nucleotide exchange factor (GEF) (Figure 7C), and the replacement of a small G to W with the larger sidechain would incur steric problems in the GEF-bound conformation. Similarly, a larger sidechain adjacent to the farnesylation site might reduce the

modification and influence RHOA localization. While the second SAV (D87N) is relatively conservative, its position near the GTP binding pocket adjacent to a K sidechain that mediates Guanine nucleotide binding might influence enzymatic activity.

Materials and methods

Human proteome, sequence alignment, and baseline fitness score

The human proteome was obtained from the UniProt database (version 2018.12) [93]. The orthologous groups of human proteins were obtained by OrthoFinder [94] applied to a set of representative vertebrate proteomes. For human proteins in large orthologous groups, we replaced their orthologous groups by the ones retrieved from the OMA database [95] that are usually much smaller and thus exhibit better alignment quality. Multiple sequence alignments of orthologs were obtained by MAFFT [96]. Sequence profile of each position of an alignment, represented as the estimated amino acid frequencies, was calculated as described before [97]. For any amino acid change, we used a previously devised baseline fitness score to represent the severity of the mutation, based on the log-odds ratio between original amino acid and mutated amino acid [46, 98].

Positional features used in impact predictions of deep convolutional neural network

For each human protein position, we deduced features reflecting amino acid type, sequence profile, sequence conservation, structure properties, and available functional annotations. The type of 20 amino acids is used as one feature with one-hot encoding. Both the original amino acid and the variant amino acid are encoded in this way, resulting in 40 features. The estimated amino acid frequencies of each position in the multiple sequence alignment of orthologs were used as 20 features. Sequence conservation scores of the multiple sequence alignment of orthologs were calculated by AL2CO [99] and used as one feature. Prediction of 3-state secondary structures (helix, strand, and coil) were made by three programs (PSIPRED [100], SPIDER [101], and PSSpred [102]), resulting in nine features. Three features are based on disorder propensities predicted by three programs (DISOPRED3 [33], SPOT-Disorder [34], and IUPred2A [35]). In addition, low complexity region predictions by SEG [103] and coiled coil predictions by NCOILS [104] were encoded as two features. We also used features reflecting protein-targeting or functional regions or positions from the UniProt sequence annotations. Regions of N-terminal signal peptide (indication of proteins going through secretory pathway), transit peptide (indication of mitochondrion targeting), and transmembrane segments were obtained from UniProt feature records SIGNAL, TRANSIT, and TRANSMEM, respectively. Three post-translational modifications (phosphorylation, acetylation, and methylation) were extracted from the UniProt MODRES records. Other UniProt Features includes DISULFID (cysteines participating in disulfide bonds), CARBOHYD (site with covalently attached glycan group), METAL (binding site for a metal ion), BINDING (binding site for any chemical group (co-enzyme, prosthetic group, etc.)), ACT_SITE (amino acid directly involved in the activity of an enzyme), SITE (any single amino acid site that could be functionally relevant), LIPID (site with covalently attached lipid group(s)), and MOTIF (short, i.e. up to 20 amino acids, sequence motif of biological interest). For 1-dimensional convolutional network, the above 89 features from a window of 21 amino acids (the target position and 10 neighboring positions on each side) were used as input. Features in neighboring positions

beyond the first or last residues were zero-filled (zero-padding). One additional feature encodes the indicator of zero-padding for such positions (1 for positions beyond the first or last residues, and zero for normal amino acid positions within the protein length). The number of features for each position is 90. By using a window of 21 positions, a total of 90 x 21=1890 values serve as the input of the convolutional neural network for each training and testing data point.

Architecture and hyperparameters of the deep-learning convolutional neural network

We used a deep-learning artificial neural network for prediction of SAV pathogenicity. The diagram of neural network structure is shown in supplemental Figure S1. It consists of seven 1-dimensional convolutional (conv1d) layers, two max-pooling layers, and two dense layers before the output. The residual network architecture is implemented twice by combining the input of a conv1d layer with the output after several layers of that input (thick arrows, supplemental Figure S1). The initial input has a window size 21 and 90 channels corresponding to 90 features encoding protein sequence, structure and functional properties (described above). The number of filters and the kernel size of other conv1d layer are 200 and 3, respectively. Each of the two dense layers has 100 nodes and has a following dropout layer with the dropout rate of 0.5. The ReLU activation function is used in all layers except the output layer that uses the softmax function. The batch size is set to 128 in the training process. The neural network was written in python with the TensorFlow package. The prediction score of any SAV, ranging from zero to one, reflects the likelihood of the SAV being pathogenic, and is termed DeepSAV score.

The DeepSAV neural network predictor

We obtained SAVs that were classified as likely pathogenic from the ClinVar [31] and UniProt database. For the ClinVar database, these SAVs are classified as “Pathogenic” or “Likely Pathogenic”. For the UniProt database, these SAVs are classified as “Disease” in the SwissVar database [105]. Benign SAVs are those classified as “Benign” or “Likely Benign” in the Clinvar database and those classified as “Polymorphism” by SwissVar in the UniProt database. To evaluate the performance of our neural network predictor, we performed 4-fold cross validation tests. The sets of 43,000 pathogenic variants and 43,000 benign variants were divided into to 4 subsets of equal size. Three subsets of pathogenic variants and three subsets of the benign variants were used to train the neural network and the remaining variants are used for testing. This process is repeated four times with each of the four subsets serving as the validation set. We also obtained scores of various prediction methods from the dbNSFP database [22] and evaluated their performance on the 43,000 pathogenic variants and 43,000 benign variants.

DeepSAV+PG: variant pathogenicity prediction incorporating population-level and gene-level information

We combined amino acid-level features used in DeepSAV with information from human general population (minor allele frequency of any variant from the gnomAD database) and gene-level information (17 features from dbNSFP) in a deep neural network predictor called DeepSAV+PG. The 17 gene-level features include three numbers of protein-protein interactions (IntAct, BioGrid, and ConsensusPathDB) [106-108], four experimental measures of gene essentiality [54, 55, 109, 110], four scores of estimated haploinsufficiency (P(HI), HIPred_score and GHIS) [24, 27, 111] or gene essentiality (Gene_indispensability_score) [112], estimated probability of the gene involved in recessive diseases

(P(rec)) [113], gene damage index score (GDI-Phred)[29], a loss-of-function intolerance score (LoFtool_score) [114], and three measures of loss-of-function (lof) intolerance/tolerance (gnomAD_pLI (the probability of being loss-of-function intolerant), gnomAD_pRec (the probability of being intolerant of homozygous, but not heterozygous lof variants), and gnomAD_pNull (the probability of being tolerant of both heterozygous and homozygous lof variants)) [5]. We applied the same four-fold cross-validation test to evaluate the performance of DeepSAV+PG on the same set consisting of 43,000 pathogenic variants and 43,000 benign variants.

Enrichment analysis of features in likely pathogenic SAVs and gnomAD SAVs

he enrichment score is defined as the logarithm of the ratio between two probabilities. This ratio is the probability of observing a feature among a subset of amino acid positions (e.g., positions with pathogenic SAVs, or positions with gnomAD SAVs with MAF in a certain range) divided by the probability of observing that feature among all amino acid positions in the human proteome. It reflects enrichment (if the log-odds score is above zero) or depletion (log-odds score less than zero) of the feature in the subset compared to the background (the whole proteome).

Quantification of mutation severity of gnomAD SAVs at the gene level

We applied our deep neural network method to the prediction of mutational impact of gnomAD [5] SAVs obtained from the dbNSFP database [22] (version 4.0). Rare SAVs were defined as those with MAF less than a certain cutoff (e.g., 0.01, 0.001, 0.0001). For any given MAF cutoff, the cumulative mutation severity measure according to our predictions is calculated as

$$\text{CumuAI} = \text{sum}(\text{DeepSAV\_score}(k) * \text{MAF}(k)),$$

where DeepSAV_score(k) is the DeepSAV score of any rare SAV k, and MAF(k) is its minor allele frequency. The average mutation severity score of rare SAVs is defined as

$$\text{AvgAI} = \text{CumuAI}/\text{protein\_len},$$

where the normalization factor is the protein length (protein_len) of the gene. Similarly, the average mutation severity measure according to baseline fitness predictions is calculated as

$$\text{AvgBF} = \text{CumuBF}/\text{protein\_len} = \text{sum}(\text{BF\_score}(k) * \text{MAF}(k))/\text{protein\_len},$$

where BF_score(k) is the baseline fitness score of any rare SAV k in the gene.

Analysis of mutation severity measures for potential disease-associated genes

Average mutation severity scores calculated using baseline fitness (AvgBF) [98] or our deep neural network (AvgAI) predictors were transformed into percentiles (using Excel percentrank) for 17,480 human protein-coding genes. For comparison of our average mutation severity scores to constrained genes that are more likely to be detrimental when inactivated (LOEUF score [5]), we transformed LOEUF scores by percentrank (for 16,670 human genes with LOEUF score). The resulting gene count distributions among LOEUF deciles were plotted for a set of genes with pathogenic SAVs or the top 3,284 genes ranked by four sets of AvgAI scores from lowest (unlikely to acquire damaging mutations) to highest (tolerates acquired mutations). We chose the AvgAI mutation severity measure filtered at MAF<0.0001 for further evaluation

and transformed the score into decile rank. For the disease-related gene set, we compared decile rank of AvgAI score to those of human genes annotated as essential by either of two large-scale CRISPR experiments [54, 55] (2,108 genes) or annotated as non-essential (11,589 genes) in both.

To identify potential disease-associated genes, we compared essential genes having the lowest AvgAI and LOEUF scores with essential genes with pathogenic SAVs using a venn diagram. The overlap between the AvgAI and LOEUF sets is considered to be enriched for potential disease-associated genes. The overlapping set (126 genes) was assigned to disease classes using the DisGeNET curated gene-disease associations (GDAs) [56]. We removed group and phenotype associations from the GDAs. MeSH (Medical Subject Headings) disease class frequencies for the set (observed frequencies) were compared to disease class frequencies assigned to all curated genes (expected frequencies) to evaluate over- and under-representation (observed/expected frequency ratios).

To further select among potential disease-related genes, we clustered the gene set (126 genes) together with the genes with pathogenic SAVs (70 genes) using ClustVis [115] with six measures for each gene (AvgAI, LOEUF [5], lnGDI [29] , number of rare gnomAD mutations (MAF filter: 0.0001), HIPred [27], and P(HI) [24]). The raw scores for each measure were converted to *Z*-scores and were pre-processed with row centering and no scaling. Principal component analysis using the SVD with imputation option indicated the first and second components explain 40.8 % and 25.9% of the data variance, respectively. Scores were plotted as a heatmap from high (red) to low (blue) *Z*-score, and both genes and measures were clustered using complete linkage of correlation distances. The genes were split into three large groups for visualization (supplemental Figure S2), with the top 20 clusters separated by space in the resulting heatmaps. Functional analysis for potential disease-related genes were performed using DAVID clustering (medium stringency with 0.001 ease) of GOfat biological process terms [116] and GO enrichment of PANTHER classification [117].

DisGeNET gene-disease mapping

We mapped all human genes with AvgAI scores to curated DisGeNET diseases (using UniProt to GeneID provided in the Downloads section from the DisGeNET website [56]). Of the 9,414 total GeneIDs with curated GDAs, we mapped AvgAI scores and ranks from the complete AvgAI dataset to 8,426 genes with curated GDAs. Over-representation (enrichment) and under-representation (depletion) of disease classes from MeSH were calculated over various sets of genes as the ratio of the observed frequency of each class to the expected frequency of each class calculated from disease class frequencies in the entire curated gene-disease database. We chose sets of genes for plotting the distributions of overrepresented disease classes over all AvgAI ranks, where genes ranked up to 1,000 (top1000 set) tend to include increased frequencies, and genes ranked higher than 12,000 (bottom set) tend to have stable frequencies. We included additional sets surrounding the top1000 (top500 and top1500) to observe trends. We excluded disease classes with few representatives, including F02: psychological phenomena & processes, C22: animal diseases, C03: parasitic diseases, C01: bacterial infections & mycoses, C24: occupational diseases, and C21: disorders of environmental origin. We related genes from various disease classes to function using Reactome or DAVID enrichment analysis tools [68, 116].

Supporting information

S1 Figure. DeepSAV neural network structure.

S2 Figure. ClistVis heatmap of potential disease-causing genes (UniProt label on the right) and genes with pathogenic SAVs (UniProt label on right with _P). Scores (labeled below) for each gene are colored from blue (low) to red (high) and clustered (20 clusters delimited by spaces) according to complete linkage of correlation distances. Two clusters with low AvgAI scores (mutation resistant) have a relatively high proportion of genes with known pathogenic variants and could help identify new disease-associated genes (red labels).

S1 Table. **Table S1. Information of human protein coding genes (GeneSymbol, UniProt, AvgAI, LOEUF, and interaction_count) and catalogues of gene sets based on their essentiality (essential1 and essential2), involvement in diseases (autism, DDD (deciphering developmental disorders), cosmic, PathVar, disgenet_path, virus_interacting), and functional categories (kinase, metabolic_enzymes, olfactory_receptors, otherGPCR, ribosomal_protein_mitochondrial, ribosomal_protein_cytoplasmic, mitochondrial).**

Acknowledgements

We thank Alex Treacher and the Grishin lab members, Jing Zhang in particular, for helpful discussions. The study is supported in part by the grants (to NVG) from the National Institutes of Health (GM127390) and the Welch Foundation (I-1505).

Conflicts of Interest: non declared.

References

1. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet. 2015;97(2):199-215. doi: 10.1016/j.ajhg.2015.06.009. PubMed PMID: 26166479.
2. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-24. doi: 10.1038/gim.2015.30. PubMed PMID: 25741868.
3. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91. doi: 10.1038/nature19057. PubMed PMID: 27535533.
4. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011;12(11):745-55. doi: 10.1038/nrg3031. PubMed PMID: 21946919.
5. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv 2019:1-44. doi: 10.1101/531210.

6. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet. 2009;10(4):241-51. doi: 10.1038/nrg2554. PubMed PMID: 19293820.
7. Cheung VG, Spielman RS. Genetics of human gene expression: mapping DNA variants that influence gene expression. Nat Rev Genet. 2009;10(9):595-604. doi: 10.1038/nrg2630. PubMed PMID: 19636342.
8. Parker SC, Stitzel ML, Taylor DL, Orozco JM, Erdos MR, Akiyama JA, et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc Natl Acad Sci U S A. 2013;110(44):17921-6. doi: 10.1073/pnas.1317023110. PubMed PMID: 24127591.
9. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med. 2010;61:437-55. doi: 10.1146/annurev-med-100708-204735. PubMed PMID: 20059347.
10. Stefl S, Nishi H, Petukh M, Panchenko AR, Alexov E. Molecular mechanisms of disease-causing missense mutations. J Mol Biol. 2013;425(21):3919-36. doi: 10.1016/j.jmb.2013.07.014. PubMed PMID: 23871686.
11. Coelho MC, Pinto RM, Murray AW. Heterozygous mutations cause genetic instability in a yeast model of cancer evolution. Nature. 2019;566(7743):275-8. doi: 10.1038/s41586-019-0887-y. PubMed PMID: 30700905.
12. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 2011;39(Database issue):D945-50. doi: 10.1093/nar/gkq929. PubMed PMID: 20952405.
13. Harripaul R, Vasli N, Mikhailov A, Rafiq MA, Mittal K, Windpassinger C, et al. Mapping autosomal recessive intellectual disability: combined microarray and exome sequencing identifies 26 novel candidate genes in 192 consanguineous families. Mol Psychiatry. 2018;23(4):973-84. doi: 10.1038/mp.2017.60. PubMed PMID: 28397838.
14. Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet. 2005;6(2):109-18. doi: 10.1038/nrg1522. PubMed PMID: 15716907.
15. Alves MM, Sribudiani Y, Brouwer RW, Amiel J, Antinolo G, Borrego S, et al. Contribution of rare and common variants determine complex diseases-Hirschsprung disease as a model. Dev Biol. 2013;382(1):320-9. doi: 10.1016/j.ydbio.2013.05.019. PubMed PMID: 23707863.
16. Vissers LE, Gilissen C, Veltman JA. Genetic studies in intellectual disability and related disorders. Nat Rev Genet. 2016;17(1):9-18. doi: 10.1038/nrg3999. PubMed PMID: 26503795.
17. Deciphering Developmental Disorders S. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017;542(7642):433-8. doi: 10.1038/nature21062. PubMed PMID: 28135719.
18. Piccolo SR, Hoffman LM, Conner T, Shrestha G, Cohen AL, Marks JR, et al. Integrative analyses reveal signaling pathways underlying familial breast cancer susceptibility. Mol Syst Biol. 2016;12(3):860. doi: 10.15252/msb.20156506. PubMed PMID: 26969729.
19. Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian disease. Nat Rev Genet. 2017;18(10):599-612. doi: 10.1038/nrg.2017.52. PubMed PMID: 28804138.
20. Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. Nat Rev Genet. 2017;18(9):551-62. doi: 10.1038/nrg.2017.38. PubMed PMID: 28607512.
21. Vitkup D, Sander C, Church GM. The amino-acid mutational spectrum of human genetic disease. Genome Biol. 2003;4(11):R72. doi: 10.1186/gb-2003-4-11-r72. PubMed PMID: 14611658.
22. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat. 2016;37(3):235-41. doi: 10.1002/humu.22932. PubMed PMID: 26555599.
23. Bartha I, di Iulio J, Venter JC, Telenti A. Human gene essentiality. Nat Rev Genet. 2018;19(1):51-62. doi: 10.1038/nrg.2017.75. PubMed PMID: 29082913.

24. Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and predicting haploinsufficiency in the human genome. PLoS Genet. 2010;6(10):e1001154. doi: 10.1371/journal.pgen.1001154. PubMed PMID: 20976243.
25. Inoue K, Fry EA. Haploinsufficient tumor suppressor genes. Adv Med Biol. 2017;118:83-122. PubMed PMID: 28680740.
26. Chen H, Zhang Z, Jiang S, Li R, Li W, Zhao C, et al. New insights on human essential genes based on integrated analysis and the construction of the HEGIAP web-based platform. Brief Bioinform. 2019. doi: 10.1093/bib/bbz072. PubMed PMID: 31504171.
27. Shihab HA, Rogers MF, Campbell C, Gaunt TR. HIPred: an integrative approach to predicting haploinsufficient genes. Bioinformatics. 2017;33(12):1751-7. doi: 10.1093/bioinformatics/btx028. PubMed PMID: 28137713.
28. Bartha I, Rausell A, McLaren PJ, Mohammadi P, Tardaguila M, Chaturvedi N, et al. The Characteristics of Heterozygous Protein Truncating Variants in the Human Genome. PLoS Comput Biol. 2015;11(12):e1004647. doi: 10.1371/journal.pcbi.1004647. PubMed PMID: 26642228.
29. Itan Y, Shang L, Boisson B, Patin E, Bolze A, Moncada-Velez M, et al. The human gene damage index as a gene-level approach to prioritizing exome variants. Proc Natl Acad Sci U S A. 2015;112(44):13615-20. doi: 10.1073/pnas.1518646112. PubMed PMID: 26483451.
30. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014;46(9):944-50. doi: 10.1038/ng.3050. PubMed PMID: 25086666.
31. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062-D7. doi: 10.1093/nar/gkx1153. PubMed PMID: 29165669.
32. The UniProt C. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45(D1):D158-D69. doi: 10.1093/nar/gkw1099. PubMed PMID: 27899622.
33. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics. 2015;31(6):857-63. doi: 10.1093/bioinformatics/btu744. PubMed PMID: 25391399.
34. Hanson J, Yang Y, Paliwal K, Zhou Y. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. Bioinformatics. 2017;33(5):685-92. doi: 10.1093/bioinformatics/btw678. PubMed PMID: 28011771.
35. Meszaros B, Erdos G, Dosztanyi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. 2018;46(W1):W329-W37. doi: 10.1093/nar/gky384. PubMed PMID: 29860432.
36. Raimondi D, Gazzo AM, Rooman M, Lenaerts T, Vranken WF. Multilevel biological characterization of exomic variants at the protein level significantly improves the identification of their deleterious effects. Bioinformatics. 2016;32(12):1797-804. doi: 10.1093/bioinformatics/btw094. PubMed PMID: 27153718.
37. Gao F, Keinan A. High burden of private mutations due to explosive human population growth and purifying selection. BMC Genomics. 2014;15 Suppl 4:S3. doi: 10.1186/1471-2164-15-S4-S3. PubMed PMID: 25056720.
38. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001;11(5):863-74. doi: 10.1101/gr.176601. PubMed PMID: 11337480.
39. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7 20. doi: 10.1002/0471142905.hg0720s76. PubMed PMID: 23315928.

40. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics. 2018;34(3):511-3. doi: 10.1093/bioinformatics/btx536. PubMed PMID: 28968714.
41. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015;31(16):2745-7. doi: 10.1093/bioinformatics/btv195. PubMed PMID: 25851949.
42. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310-5. doi: 10.1038/ng.2892. PubMed PMID: 24487276.
43. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19(9):1553-61. doi: 10.1101/gr.092619.109. PubMed PMID: 19602639.
44. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39(17):e118. doi: 10.1093/nar/gkr407. PubMed PMID: 21727090.
45. Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet. 2018;50(8):1161-70. doi: 10.1038/s41588-018-0167-z. PubMed PMID: 30038395.
46. Zhang J, Kinch LN, Cong Q, Weile J, Sun S, Cote AG, et al. Assessing predictions of fitness effects of missense mutations in SUMO-conjugating enzyme UBE2I. Hum Mutat. 2017;38(9):1051-63. doi: 10.1002/humu.23293. PubMed PMID: 28817247.
47. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet. 2016;48(2):214-20. doi: 10.1038/ng.3477. PubMed PMID: 26727659.
48. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24(8):2125-37. doi: 10.1093/hmg/ddu733. PubMed PMID: 25552646.
49. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016;99(4):877-85. doi: 10.1016/j.ajhg.2016.08.016. PubMed PMID: 27666373.
50. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14 Suppl 3:S3. doi: 10.1186/1471-2164-14-S3-S3. PubMed PMID: 23819870.
51. Raimondi D, Tanyalcin I, Ferte J, Gazzo A, Orlando G, Lenaerts T, et al. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. Nucleic Acids Res. 2017;45(W1):W201-W6. doi: 10.1093/nar/gkx390. PubMed PMID: 28498993.
52. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012;337(6090):64-9. doi: 10.1126/science.1219240. PubMed PMID: 22604720.
53. Povysil G, Petrovski S, Hostyk J, Aggarwal V, Allen AS, Goldstein DB. Rare-variant collapsing analyses for complex traits: guidelines and applications. Nat Rev Genet. 2019. doi: 10.1038/s41576-019-0177-4. PubMed PMID: 31605095.
54. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, et al. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell. 2015;163(6):1515-26. doi: 10.1016/j.cell.2015.11.015. PubMed PMID: 26627737.
55. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, et al. Identification and characterization of essential genes in the human genome. Science. 2015;350(6264):1096-101. doi: 10.1126/science.aac7041. PubMed PMID: 26472758.

56. Pinero J, Bravo A, Queralt-Rosinach N, Gutierrez-Sacristan A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017;45(D1):D833-D9. doi: 10.1093/nar/gkw943. PubMed PMID: 27924018.
57. Fragoza R, Das J, Wierbowski SD, Liang J, Tran TN, Liang S, et al. Extensive disruption of protein interactions by genetic variants across the allele frequency spectrum in human populations. Nat Commun. 2019;10(1):4141. doi: 10.1038/s41467-019-11959-3. PubMed PMID: 31515488.
58. Dickerson JE, Zhu A, Robertson DL, Hentges KE. Defining the role of essential genes in human disease. PLoS One. 2011;6(11):e27368. doi: 10.1371/journal.pone.0027368. PubMed PMID: 22096564.
59. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34(Database issue):D535-9. doi: 10.1093/nar/gkj109. PubMed PMID: 16381927.
60. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014;42(Database issue):D358-63. doi: 10.1093/nar/gkt1115. PubMed PMID: 24234451.
61. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. DIP: the database of interacting proteins. Nucleic Acids Res. 2000;28(1):289-91. doi: 10.1093/nar/28.1.289. PubMed PMID: 10592249.
62. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database--2009 update. Nucleic Acids Res. 2009;37(Database issue):D767-72. doi: 10.1093/nar/gkn892. PubMed PMID: 18988627.
63. Gioutlakis A, Klapa MI, Moschonas NK. PICKLE 2.0: A human protein-protein interaction meta-database employing data integration via genetic information ontology. PLoS One. 2017;12(10):e0186039. doi: 10.1371/journal.pone.0186039. PubMed PMID: 29023571.
64. Jenkins JL, Kielkopf CL. Splicing Factor Mutations in Myelodysplasias: Insights from Spliceosome Structures. Trends Genet. 2017;33(5):336-48. doi: 10.1016/j.tig.2017.03.001. PubMed PMID: 28372848.
65. Seiler M, Yoshimi A, Darman R, Chan B, Keaney G, Thomas M, et al. H3B-8800, an orally available small-molecule splicing modulator, induces lethality in spliceosome-mutant cancers. Nat Med. 2018;24(4):497-504. doi: 10.1038/nm.4493. PubMed PMID: 29457796.
66. Piche J, Van Vliet PP, Puceat M, Andelfinger G. The expanding phenotypes of cohesinopathies: one ring to rule them all! Cell Cycle. 2019;18(21):2828-48. doi: 10.1080/15384101.2019.1658476. PubMed PMID: 31516082.
67. Zhang B, Chang J, Fu M, Huang J, Kashyap R, Salavaggione E, et al. Dosage effects of cohesin regulatory factor PDS5 on mammalian development: implications for cohesinopathies. PLoS One. 2009;4(5):e5232. doi: 10.1371/journal.pone.0005232. PubMed PMID: 19412548.
68. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018;46(D1):D649-D55. doi: 10.1093/nar/gkx1132. PubMed PMID: 29145629.
69. Zhan T, Boutros M. Towards a compendium of essential genes - From model organisms to synthetic lethality in cancer cells. Crit Rev Biochem Mol Biol. 2016;51(2):74-85. doi: 10.3109/10409238.2015.1117053. PubMed PMID: 26627871.
70. Kafri R, Springer M, Pilpel Y. Genetic redundancy: new tricks for old genes. Cell. 2009;136(3):389-92. doi: 10.1016/j.cell.2009.01.027. PubMed PMID: 19203571.
71. Busca R, Pouyssegur J, Lenormand P. ERK1 and ERK2 Map Kinases: Specific Roles or Functional Redundancy? Front Cell Dev Biol. 2016;4:53. doi: 10.3389/fcell.2016.00053. PubMed PMID: 27376062.
72. Weiss LA, Arking DE, Gene Discovery Project of Johns H, the Autism C, Daly MJ, Chakravarti A. A genome-wide linkage and association scan reveals novel loci for autism. Nature. 2009;461(7265):802-8. doi: 10.1038/nature08490. PubMed PMID: 19812673.

73. Enard D, Cai L, Gwennap C, Petrov DA. Viruses are a dominant driver of protein adaptation in mammals. Elife. 2016;5. doi: 10.7554/eLife.12469. PubMed PMID: 27187613.
74. Solomon BD, Nguyen AD, Bear KA, Wolfsberg TG. Clinical genomic database. Proc Natl Acad Sci U S A. 2013;110(24):9851-5. doi: 10.1073/pnas.1302575110. PubMed PMID: 23696674.
75. Yamamoto N, Akamatsu N, Sakuraba H, Yamazaki H, Tanoue K. Platelet glycoprotein IV (CD36) deficiency is associated with the absence (type I) or the presence (type II) of glycoprotein IV on monocytes. Blood. 1994;83(2):392-7. PubMed PMID: 7506948.
76. Hoover KC. Evolution of olfactory receptors. Methods Mol Biol. 2013;1003:241-9. doi: 10.1007/978-1-62703-377-0_18. PubMed PMID: 23585047.
77. Menashe I, Man O, Lancet D, Gilad Y. Different noses for different people. Nat Genet. 2003;34(2):143-4. doi: 10.1038/ng1160. PubMed PMID: 12730696.
78. Corcoran CC, Grady CR, Pisitkun T, Parulekar J, Knepper MA. From 20th century metabolic wall charts to 21st century systems biology: database of mammalian metabolic enzymes. Am J Physiol Renal Physiol. 2017;312(3):F533-F42. doi: 10.1152/ajprenal.00601.2016. PubMed PMID: 27974320.
79. Calvo SE, Clauser KR, Mootha VK. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. Nucleic Acids Res. 2016;44(D1):D1251-7. doi: 10.1093/nar/gkv1003. PubMed PMID: 26450961.
80. Lasserre JP, Dautant A, Aiyar RS, Kucharczyk R, Glatigny A, Tribouillard-Tanvier D, et al. Yeast as a system for modeling mitochondrial disease mechanisms and discovering therapies. Dis Model Mech. 2015;8(6):509-26. doi: 10.1242/dmm.020438. PubMed PMID: 26035862.
81. Piergiorge RM, de Miranda AB, Guimaraes AC, Catanho M. Functional Analogy in Human Metabolism: Enzymes with Different Biological Roles or Functional Redundancy? Genome Biol Evol. 2017;9(6):1624-36. doi: 10.1093/gbe/evx119. PubMed PMID: 28854631.
82. Guell O, Sagues F, Serrano MA. Essential plasticity and redundancy of metabolism unveiled by synthetic lethality analysis. PLoS Comput Biol. 2014;10(5):e1003637. doi: 10.1371/journal.pcbi.1003637. PubMed PMID: 24854166.
83. Ebel ER, Telis N, Venkataram S, Petrov DA, Enard D. High rate of adaptation of mammalian proteins that interact with Plasmodium and related parasites. PLoS Genet. 2017;13(9):e1007023. doi: 10.1371/journal.pgen.1007023. PubMed PMID: 28957326.
84. Ansari AW, Heiken H, Moenkemeyer M, Schmidt RE. Dichotomous effects of C-C chemokines in HIV-1 pathogenesis. Immunol Lett. 2007;110(1):1-5. doi: 10.1016/j.imlet.2007.02.012. PubMed PMID: 17434211.
85. Kreisberg JF, Yonemoto W, Greene WC. Endogenous factors enhance HIV infection of tissue naive CD4 T cells by stimulating high molecular mass APOBEC3G complex formation. J Exp Med. 2006;203(4):865-70. doi: 10.1084/jem.20051856. PubMed PMID: 16606671.
86. Krishnan V, Zeichner SL. Host cell gene expression during human immunodeficiency virus type 1 latency and reactivation and effects of targeting genes that are differentially expressed in viral latency. J Virol. 2004;78(17):9458-73. doi: 10.1128/JVI.78.17.9458-9473.2004. PubMed PMID: 15308739.
87. Jakel S, Gorlich D. Importin beta, transportin, RanBP5 and RanBP7 mediate nuclear import of ribosomal proteins in mammalian cells. EMBO J. 1998;17(15):4491-502. doi: 10.1093/emboj/17.15.4491. PubMed PMID: 9687515.
88. Henderson BR, Percipalle P. Interactions between HIV Rev and nuclear import and export factors: the Rev nuclear localisation signal mediates specific binding to human importin-beta. J Mol Biol. 1997;274(5):693-707. doi: 10.1006/jmbi.1997.1420. PubMed PMID: 9405152.
89. Vetter IR, Arndt A, Kutay U, Gorlich D, Wittinghofer A. Structural view of the Ran-Importin beta interaction at 2.3 A resolution. Cell. 1999;97(5):635-46. doi: 10.1016/s0092-8674(00)80774-6. PubMed PMID: 10367892.

90. Nagata Y, Kontani K, Enami T, Kataoka K, Ishii R, Totoki Y, et al. Variegated RHOA mutations in adult T-cell leukemia/lymphoma. Blood. 2016;127(5):596-604. doi: 10.1182/blood-2015-06-644948. PubMed PMID: 26574607.
91. O'Hayre M, Inoue A, Kufareva I, Wang Z, Mikelis CM, Drummond RA, et al. Inactivating mutations in GNA13 and RHOA in Burkitt's lymphoma and diffuse large B-cell lymphoma: a tumor suppressor function for the Galpha13/RhoA axis in B cells. Oncogene. 2016;35(29):3771-80. doi: 10.1038/onc.2015.442. PubMed PMID: 26616858.
92. Svensmark JH, Brakebusch C. Rho GTPases in cancer: friend or foe? Oncogene. 2019. doi: 10.1038/s41388-019-0963-7. PubMed PMID: 31427738.
93. UniProt Consortium T. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2018;46(5):2699. doi: 10.1093/nar/gky092. PubMed PMID: 29425356.
94. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157. doi: 10.1186/s13059-015-0721-2. PubMed PMID: 26243257.
95. Altenhoff AM, Levy J, Zarowiecki M, Tomiczek B, Warwick Vesztrocy A, Dalquen DA, et al. OMA standalone: orthology inference among public and custom genomes and transcriptomes. Genome Res. 2019;29(7):1152-63. doi: 10.1101/gr.243212.118. PubMed PMID: 31235654.
96. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform. 2008;9(4):286-98. doi: 10.1093/bib/bbn013. PubMed PMID: 18372315.
97. Pei J, Grishin NV. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics. 2007;23(7):802-8. doi: 10.1093/bioinformatics/btm017. PubMed PMID: 17267437.
98. Zhang J, Kinch LN, Cong Q, Katsonis P, Lichtarge O, Savojardo C, et al. Assessing predictions on fitness effects of missense variants in calmodulin. Hum Mutat. 2019;40(9):1463-73. doi: 10.1002/humu.23857. PubMed PMID: 31283071.
99. Pei J, Grishin NV. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics. 2001;17(8):700-12. doi: 10.1093/bioinformatics/17.8.700. PubMed PMID: 11524371.
100. Buchan DWA, Jones DT. The PSIPRED Protein Analysis Workbench: 20 years on. Nucleic Acids Res. 2019;47(W1):W402-W7. doi: 10.1093/nar/gkz297. PubMed PMID: 31251384.
101. Yang Y, Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, et al. SPIDER2: A Package to Predict Secondary Structure, Accessible Surface Area, and Main-Chain Torsional Angles by Deep Neural Networks. Methods Mol Biol. 2017;1484:55-63. doi: 10.1007/978-1-4939-6406-2_6. PubMed PMID: 27787820.
102. Yan R, Xu D, Yang J, Walker S, Zhang Y. A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. Sci Rep. 2013;3:2619. doi: 10.1038/srep02619. PubMed PMID: 24018415.
103. Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. Computers & Chemistry. 1993;17(2):149-63. doi: 10.1016/0097-8485(93)85006-X.
104. Lupas A. Prediction and analysis of coiled-coil structures. Methods Enzymol. 1996;266:513-25. doi: 10.1016/s0076-6879(96)66032-7. PubMed PMID: 8743703.
105. Mottaz A, David FP, Veuthey AL, Yip YL. Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. Bioinformatics. 2010;26(6):851-2. doi: 10.1093/bioinformatics/btq028. PubMed PMID: 20106818.
106. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, et al. The IntAct molecular interaction database in 2012. Nucleic Acids Res. 2012;40(Database issue):D841-6. doi: 10.1093/nar/gkr1088. PubMed PMID: 22121220.

107. Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, et al. The BioGRID interaction database: 2019 update. Nucleic Acids Res. 2019;47(D1):D529-D41. doi: 10.1093/nar/gky1079. PubMed PMID: 30476227.
108. Herwig R, Hardt C, Lienhard M, Kamburov A. Analyzing and interpreting genome data at the network level with ConsensusPathDB. Nat Protoc. 2016;11(10):1889-907. doi: 10.1038/nprot.2016.117. PubMed PMID: 27606777.
109. Georgi B, Voight BF, Bucan M. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. PLoS Genet. 2013;9(5):e1003484. doi: 10.1371/journal.pgen.1003484. PubMed PMID: 23675308.
110. Blomen VA, Majek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, et al. Gene essentiality and synthetic lethality in haploid human cells. Science. 2015;350(6264):1092-6. doi: 10.1126/science.aac7557. PubMed PMID: 26472760.
111. Steinberg J, Honti F, Meader S, Webber C. Haploinsufficiency predictions without study bias. Nucleic Acids Res. 2015;43(15):e101. doi: 10.1093/nar/gkv474. PubMed PMID: 26001969.
112. Khurana E, Fu Y, Chen J, Gerstein M. Interpretation of genomic variants using a unified biological network approach. PLoS Comput Biol. 2013;9(3):e1002886. doi: 10.1371/journal.pcbi.1002886. PubMed PMID: 23505346.
113. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012;335(6070):823-8. doi: 10.1126/science.1215040. PubMed PMID: 22344438.
114. Fadista J, Oskolkov N, Hansson O, Groop L. LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. Bioinformatics. 2017;33(4):471-4. doi: 10.1093/bioinformatics/btv602. PubMed PMID: 27563026.
115. Metsalu T, Vilo J. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Res. 2015;43(W1):W566-70. doi: 10.1093/nar/gkv468. PubMed PMID: 25969447.
116. Jiao X, Sherman BT, Huang da W, Stephens R, Baseler MW, Lane HC, et al. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. Bioinformatics. 2012;28(13):1805-6. doi: 10.1093/bioinformatics/bts251. PubMed PMID: 22543366.
117. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nat Protoc. 2013;8(8):1551-66. doi: 10.1038/nprot.2013.092. PubMed PMID: 23868073.