Identification of Discriminative Gene-level and Protein-level Features Associated with Gain-of-Function and Loss-of-Function Mutations

Cigdem S. Bayrak[1], Aayushee Jain[1,2], David Stein[3], Kumardeep Chaudhary[1], Girish N. Nadkarni[1,4,5], Tielman Van Vleck[1], Stephanie Boisson-Dupuis[6,7,8], Peter D. Stenson[9], David N. Cooper[9], Avner Schlessinger[10,*], and Yuval Itan[1,5,*]

[1]Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, US

[2]Department of Neurology, Massachusetts General Hospital, Boston, MA, US

[3]Department of Genetics and Data Science, Icahn School of Medicine at Mount Sinai, New York, US

[4]Division of Nephrology, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, US

[5]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, US

[6]Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM 1163, Paris, France

[7]Imagine Institute, University of Paris, Paris, France

[8]St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, Rockefeller University, New York, US

[9]Institute of Medical Genetics, Cardiff University, Cardiff, UK

[10]Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, New York, US

[*] To whom correspondence should be addressed: Yuval Itan, Email: yuval.itan@mssm.edu; Avner Schlessinger, E-mail: avner.schlessinger@mssm.edu

Abstract

Gain-of-function (GOF) and loss-of-function (LOF) mutations in the same gene may result in markedly different clinical phenotypes and hence require different therapeutic treatments. Identifying the functional consequences of mutations is an important step toward understanding disease mechanisms. While there are numerous computational tools (e.g., CADD, SIFT, PolyPhen-2) to predict the pathogenicity of a variant, there are currently no methods to predict whether a given genetic mutation results in a gene product with increased (gain-of-function; GOF) or diminished (loss-of-function; LOF) activity. Here, we investigated various gene- and protein-level features of GOF and LOF mutations. We generated the first extensive database of all currently known germline GOF and LOF pathogenic mutations by employing natural language processing (NLP) on the available abstracts in the *Human Gene Mutation Database*. Machine learning and statistical analyses of gene- and protein-level features associated with GOF and LOF mutations indicated significant differences. For example, GOF mutations were enriched in essential genes, autosomal dominant inheritance, protein binding and interaction domains, whereas LOF mutations were enriched in singleton genes, protein-truncating variants, and protein core regions. These results are consistent with the notion that mutations underlying recessive and dominant disorders have significantly different structural and functional properties. These findings could ultimately improve our understanding of how mutations affect gene/protein function thereby guiding future treatment options.

Introduction

To understand the mechanisms underlying human genetic diseases, it is crucial to study the effects of genetic mutations on their protein products. Based on their effect on protein function, mutations can be categorized into two main types: gain-of-function (GOF) mutations, which enhance the activity of the mutated protein, and loss-of-function (LOF) mutations which cause partial or complete protein inactivation [1]. Over the past decades, it has been shown that GOF and LOF mutations in the same gene may result in completely different disease phenotypes [1-6]. For example, GOF mutations in the *STAT1* gene cause chronic mucocutaneous candidiasis (CMC), which is characterized by recurrent disease of the nails, skin, and oral and genital mucosae, whereas LOF mutations in *STAT1* cause Mendelian susceptibility to mycobacterial disease (MSMD), which is characterized by predisposition to clinical disease caused by weakly virulent mycobacteria [7, 8]. *WAS*, *MDA5*, *PCSK9*, and *CACNA1C* are additional known examples of genes in which GOF and LOF mutations impact different genetic pathways, leading to distinct clinical phenotypes that require different drug treatment regimens [2-6]. To determine the disease phenotype that can in turn guide the identification of drug targets, it is essential to understand the different functional effects of mutations, i.e., whether the protein enhances or diminishes activity.

Whether a specific mutation causes a loss- or gain-of-function is ultimately determined by its impact on the biological pathway(s) associated with the protein(s) encoded by the mutated gene [9]. Mutations can have a significant effect on protein stability, hydrogen bond network, conformational dynamics, as well as on binding to other molecules, and many other important biochemical and physiological properties of proteins [9-11]. Thus, to understand the genetic

mutation's effect on protein structure and function, it is necessary to accurately and reliably distinguish between GOF and LOF mutations.

Understanding the properties of genetic variants identified by of next-generation sequencing (NGS) is of great interest for human health and genetic research [12-14]. So far, various computational deleteriousness and pathogenicity prediction tools based on machine learning algorithms and DNA sequence information, such as CADD, PolyPhen-2, SIFT and REVEL, have been developed to facilitate the identification of disease-causing mutations [15-18]. For example, CADD integrates multiple annotations including functional information and conservation, whereas SIFT and PolyPhen-2 mainly consider sequence conservation and biochemical features, whilst REVEL integrates scores from multiple prediction tools. However, none of these tools have the capability to distinguish GOF from LOF mutations.

To date, there has only been a limited effort to identify and analyze the discriminative features of GOF and LOF mutations, and these studies have numerous limitations. Thus, no study to date has attempted to generate a comprehensive public germline GOF/LOF database that could prove invaluable for developing an accurate and generalizable GOF/LOF classifier and for annotating NGS variants. For example, in a recent study, a computational method to predict missense GOF and LOF variants in voltage-gated sodium and calcium channels, *SCNxA* and *CACNA1x* family genes, has been presented [19]. In a separate study, 129 GOF mutations from 59 genes and 258 LOF mutations from 109 genes were analyzed and six features were proposed to be discriminant [20]. Additionally, a computational method, HMMvar-func, that predicts functional outcome of a mutation has been reported [21]. HMMvar-func was however trained only on two genes, *TSHR* and *TP53*, and it defines GOF mutations as having acquired a new function, without including GOF mutations with enhanced activation, which accounts for the majority of GOF mutations

[22-24]. Other shortcomings of such existing studies include: (i) the GOF and LOF mutations were not shown to be disease-causing; (ii) the number of incorporated features was limited; and (iii) they lacked publicly accessible data. In particular, no study has generated a comprehensive dataset of GOF and LOF mutations.

In this study, we applied various machine learning and statistical approaches to construct a comprehensive dataset of currently known disease-causing GOF and LOF mutations from the Human Gene Mutation Database (HGMD) [25]. We first applied text-mining and natural language processing (NLP) to abstracts of published pathogenic mutations to extract GOF and LOF mutations from HGMD (Professional version 2019.1), and then annotated 12 protein-level, 11 gene-level, and 10 variant-level features that were broken down into 1,725 variables. We demonstrated that there are variant-, gene-, and protein-level features that serve to distinguish GOF and LOF mutations, and discuss the biology of these discriminant features (Figure 1). Finally, we generated and made publicly available the first extensive database of 9,541 pathogenic GOF and LOF mutations based on the HGMD Public release.

Throughout this paper we use the term 'mutation' to refer to pathogenic variants.

Results

Database of GOF and LOF mutations

Currently, there is no available genome-wide database of ascertained germline GOF and LOF disease-causing mutations in human. Such a database could serve as an invaluable resource with which to study features of mutations causing different diseases and to develop computational prediction tools. To fulfill this need, we developed an automated pipeline based on text-mining and NLP to extract all disease-causing GOF and LOF mutations from paper abstracts and

established a comprehensive dataset of GOF and LOF mutations. We searched 123,737 paper abstracts associated with disease-causing mutations in the HGMD database (Professional version 2019.1) for different possible terms describing GOF and LOF mutations [25, 26]. It should be noted that the terms defining GOF and LOF are inconsistent, with examples including "gain-of-function", "gain of function", "GoF", "GOF", making it challenging to analyze in an automated fashion. Therefore, we used NLP to search for different patterns and phrases matching the patterns we defined. The NLP framework helps to extract phrases with specific pattern types from a given text. We thus identified 1,237 mutations as GOF and 10,133 mutations as LOF. To the best of our knowledge, this is the largest GOF and LOF mutation dataset established to date. In Supplementary Table 1, we provide a subset of the predicted GOF/LOF dataset containing 949 GOF and 8,592 LOF mutations from the publicly available version of HGMD December 2019.

Although the majority of both GOF (89%) and LOF (55%) mutations in our dataset were missense mutations, the frameshift (23%) and stop-gains (19%) account for a significant percentage of LOF mutations. To address the question of whether the discriminative features we observed resulted from this difference, we compared the various features specifically in the proportion of missense GOF and LOF mutations. In our dataset, we have 1,101 missense GOF and 5,507 missense LOF mutations.

Discriminative variant-level features of GOF and LOF mutations

LOF mutations have a higher impact on protein sequence and function

To evaluate the significance of the mutational effects on protein function, we compared the consequence (e.g., missense, stop-gain, frameshift), the impact rating (e.g., high, moderate, low),

and the variant class (e.g., indel, SNV) of the GOF and LOF mutations and likely neutral variants from gnomAD by using the Ensembl Variant Effect Predictor (VEP) tool [27] (see Methods). GOF mutations were mostly predicted to be missense (89%), whereas LOF mutations were predicted to be missense, frameshift, and stop-gain (55%, 23%, and 19%, respectively) (Figure 2a). The variants with high impact (i.e., frameshift, splice acceptor/donor, start-lost, stop-gained, stop-lost) and moderate impact (i.e., inframe insertion/deletion, missense) were found to be highly enriched among both GOF and LOF mutations as compared to neutral variants, while variants with low and modifier impacts (i.e., synonymous, intron) were more enriched among the neutral variants (Figure 2b). Since we specifically selected the GOF and LOF mutations from the disease-causing mutations collected in the HGMD database, and the reference dataset comprises likely neutral variants from the gnomAD database, this result affirms the pathogenic effect of the GOF and LOF mutations in our dataset. More specifically, predicted functional consequences on protein sequence showed that LOF mutations have higher (disruptive) impact in the protein than GOF mutations as they often resulted from a frameshift or stop-gain. The missense, frameshift, and stop-gained mutations exhibited significant differences between the GOF and LOF mutation datasets (Fisher's exact 2-sided, Bonferroni corrected P-values were 6.26×10^{-135}, 2.92×10^{-71}, 4.83×10^{-44}, respectively) (Supplementary Table 2).

We also analyzed the differences in variant classification of GOF and LOF mutations, and neutral variants. Whereas the GOF dataset had similar proportions of insertion, deletion and single nucleotide variations (SNVs) to neutral variants, the LOF dataset had fewer SNVs and more insertions and deletions. Comparing the variant class differences between GOF and LOF mutations by Fisher's exact test indicates that GOF mutations are significantly enriched among SNVs (Bonferroni corrected 2-sided P-value=1.21×10^{-61}) whereas LOF mutations are enriched

among micro-deletions and micro-insertions (P-values=8.94×10^{-40} and 1.0×10^{-14}) (Figure 2c, Supplementary Table 2). Overall, the impact rate, consequence for protein sequence, and variant class features of GOF and LOF mutations indicate that LOF mutations have a greater effect on protein sequence.

Using our GOF/LOF mutations dataset, we evaluated the predictive power of various computational tools designed to predict different aspects of mutation effect on function. We calculated the functional effect scores of the mutations by polymorphism phenotyping (PolyPhen-2), sorting intolerant from tolerant (SIFT), Combined Annotation Dependent Depletion (CADD), consensus deleteriousness (Condel), and rare exome variant ensemble learner (REVEL) tools [15, 17, 18, 28, 29]. PolyPhen-2, SIFT, Condel, and REVEL are only available for missense mutations. Therefore, non-missense mutations could not to be scored by these tools. LOF mutations failed to receive a score more often compared to GOF mutations (P-values: 7.86×10^{-131}, 1.9×10^{-131}, 1.82×10^{-131}, 2.94×10^{-135}, respectively).

Remarkably, in the missense GOF/LOF dataset, missense LOF mutations were more often predicted to be "probably damaging" by PolyPhen-2, "deleterious" by Condel and SIFT, having high scores (≥20) by CADD, and "likely pathogenic" by REVEL (P-values: 1.13×10^{-15}, 1.06×10^{-7}, 3.77×10^{-7}, 2.3×10^{-4}, 0.001, respectively) as compared to missense GOF mutations. Overall, pathogenicity prediction methods tend to predict missense LOF mutations as more damaging than missense GOF mutations, whereas GOF mutations were more often predicted to be "benign" by PolyPhen-2, "neutral" by Condel, "tolerated" by SIFT, having medium scores (between 10-20) by CADD, and "benign" by REVEL (P-values: 6.28×10^{-8}, 4.14×10^{-8},1.4×10^{-4}, 0.003, 0.0007, respectively). Figure 3 shows the differences in score predictions for GOF, LOF and neutral mutations. A low specificity of SIFT and PolyPhen

for predicting GOF and LOF missense mutations in *ABCC8*, *KCNJ11*, and *GCK* genes has been previously reported [30]. Our results confirm that current bioinformatics tools are relatively successful at capturing LOF mutations (likely due to being trained mostly by LOF rather than GOF mutations), whilst GOF mutations were more often incorrectly predicted to be benign.

Discriminative gene-level features of GOF and LOF mutations

In our dataset, we identified 135 genes having both GOF and LOF mutations, 1,437 genes having only LOF mutations, and 154 genes having only GOF mutations. In the following analysis, we compared the 289 genes harboring GOF mutations to 1,572 genes harboring LOF mutations.

GOF mutations tend to occur in essential genes

A gene is defined as essential when it does not tolerate loss-of-function mutation. In other words, when an essential gene loses its function, it results in lethality or a profound loss of fitness [31]. The genes with GOF mutations in our dataset were predicted to be significantly more intolerant compared to the genes with LOF mutations by all five tools (Bonferroni-corrected 1-sided P-values are 6.78×10^{-38}, 1.83×10^{-24}, 3.18×10^{-34}, 2.3×10^{-4}, and 0.001 by Z-score (*de novo* excess rate), RVIS, LoFtool, IS (predicted indispensability score), and ExACpLI, respectively) (Figure S1, Supplementary Table 2). Z-score and RVIS evaluate functional variation; LoFtool and ExACpLI calculate ratio of protein truncating mutations; and indispensability score (IS) combines network and evolutionary properties (see Methods for details). In addition, when compared to neutral variants, LOF mutations occur significantly less frequently in intolerant genes. Figure 4 illustrates the differences in gene essentiality between GOF, LOF and neutral

variants based on different metrics. The essentiality of the genes also remained significantly different for only missense GOF and LOF mutations (Figure S1).

Paralogs are genes that have been duplicated (or even amplified) within the same organism during evolution, sharing a common ancestor. Paralogous genes are expected to be more tolerant to mutations since they may still exhibit some functional redundancy and hence can compensate to some extent for loss of function of one paralog [32-35]. The degree of protection afforded also depends upon divergence in tissue distribution and molecular functionality over evolutionary time. When paralogs maintain related functions, they can buffer each other's loss, as inactivation of one copy may not affect the other(s) [35-37]. However, when paralogs have different or interdependent functions, they tend to be more intolerant, i.e. less protective against deleterious mutations [37].

To relate paralogy to GOF and LOF mutations, we determined the number of paralogs of each gene by Ensembl BioMart [38]. This showed that GOF mutations occur more frequently in genes that have more than 4 paralogs (67%) as compared to LOF (41%) and neutral mutations (50%) (Figures 4f and 5a). Loss of function in the genes that have more than 4 paralogs appears to be buffered by the other paralogs. The genes with GOF mutations more often had more than four paralogs compared to the genes with LOF mutations (P-value=5.08×10^{-64}), and the genes with LOF mutations were found disproportionally in singleton genes compared to GOF mutations (P-value=1.02×10^{-60}) (Supplementary Table 2).

As genetic variant intolerance metrics are often correlated with coding sequence (CDS) length, we investigated the differences in CDS by defining 3 categories (long, medium and short) based on the distribution of sequence length in our dataset (see Methods) [31]. We identified significant enrichment for longer length (>1,747 base pairs) in the genes with GOF mutations (P-

value=5.19×10^{-22}) and medium length (592-1,747 base pairs) in the genes with LOF mutations (P-value=1.06×10^{-29}) suggesting that GOF mutations occur significantly more frequently in genes with longer CDS. The number of paralogs and the coding sequence length of genes were also significant features to distinguish missense GOF and LOF mutations. Missense GOF mutations occur more frequently in the genes that have paralogs (Figure 5a), whilst the CDS was significantly longer (>1,747 base pairs) for missense GOF mutations (P-value=1.03×10^{-44}), and having medium length (592-1,747 base pairs) for LOF mutations (P-value=1.68×10^{-51}).

GOF mutations occur mostly in genes underlying autosomal dominant conditions

Since loss-of-function mutations are often associated with autosomal recessive (AR) diseases, and autosomal dominant (AD) diseases are associated with either loss- or gain-of-function mutations, we analyzed the mode of inheritance (MOI) of the genes in our dataset [39-41]. We found that the mutations in genes with autosomal dominant conditions were significantly enriched for GOF mutations (P-value=4.55×10^{-43}), whilst the mutations in genes with autosomal recessive conditions were enriched for LOF mutations (P-value=6.74×10^{-42}). Genes underlying X-linked dominant or X-linked recessive conditions were enriched for LOF mutations (P-values: 9.23×10^{-20}, and 1.06×10^{-8}, respectively). Moreover, compared to neutral mutations, GOF mutations were enriched in genes underlying autosomal dominant conditions, whilst LOF mutations were enriched in genes underlying X-linked conditions (Figure 5b). Overall, these results reveal the significant differences in the MOI of the disease genes harboring GOF and LOF mutations, consistent with the known association between mode of inheritance and mutation type [42-44].

For the subgroup of missense GOF and LOF mutations, the MOI of the genes was the most significant difference; missense GOF mutations were significantly enriched in genes with AD inheritance (P-value=4.88×10^{-58}), whereas missense LOF mutations were significantly enriched in genes with AR, X-linked dominant, and X-linked recessive inheritance (P-values: 2.37×10^{-47}, 4.29×10^{-30}, and 8.07×10^{-8}, respectively) (Figure 5b, Supplementary Table 3).

Discriminative protein-level features of GOF and LOF mutations

GOF and LOF mutations are enriched in certain amino acid substitution types

Each amino acid has distinct physicochemical characteristics that affect its role in protein structure and function. We analyzed the amino acid substitution spectra of GOF and LOF mutations first in our full dataset, and neutral variants from gnomAD, and then in the subset of missense mutations. In line with predicted mutational consequences, LOF mutations were significantly enriched for premature termination of the protein and frameshift. Interestingly, hydrophobic amino acids Valine (V), Methionine (M), Phenylalanine (F), Glycine (G), and Isoleucine (I) (P-values: 4.32×10^{-5}, 7.4×10^{-4}, 6.5×10^{-3}, 0.001, 0.01, respectively), as well as charged amino acids Lysine (K) (P-value=1.87×10^{-5}) and Glutamic acid (E) (P-value=0.03) were found to be enriched in alternate (i.e., substituting) residues in GOF mutations (Supplementary Table 2, Figure 6a). Tryptophan (W) in the reference position was enriched for LOF mutations (P-value=5.2×10^{-4}).

Furthermore, compared to neutral mutations, we observe mutations to Tryptophan (W) more frequently in both GOF (P-value=6.19×10^{-6}) and LOF mutations (P-value=1.05×10^{-16}), whereas mutations to Alanine (A) are less frequent in both GOF (P-value=9.3×10^{-5}) and LOF (P-value=1.88×10^{-109}) (Figure 6a, Supplementary Tables 4-5). Mutations to Tryptophan are highly

associated with disease causation because Tryptophan is a hydrophobic and bulky amino acid typically found in the protein core; thus, mutation of a tryptophan residue, whether on the surface or buried, is likely to disrupt protein folding. Conversely, mutation to Alanine is less likely to be disease-causing since Alanine is a neutral, small amino acid and its introduction is less likely to have a dramatic effect on protein structure.

Missense GOF and LOF mutations also exhibited significant differences in terms of the types of amino acid involved (Figure 6b). Among the 20 amino acid types in alternate (i.e., substituting) and reference (i.e., substituted) positions, only Proline (Pro/P) in alternate position was significantly more enriched (P-value=0.02) in missense LOF mutations compared to missense GOF mutations. Proline, which consists of a rotationally constrained rigid ring structure, typically breaks or kinks a helical structure, causing dramatic effects on protein structure. However, an amino acid substitution matrix (20x20) showed significant differences for multiple substitution types (Figure 7a). Remarkably, amino acid changes in GOF usually involve only subtle changes in the physicochemical properties of the mutated amino acid, and hence are less likely to cause misfolding and complete loss of function. For example, we observe frequent transitions from Isoleucine, Leucine, and Methionine to Valine (P-values: 1.8×10^{-4}, 0.008, 0.02, respectively) (all hydrophobic). Other significant transitions in GOF mutations were from Alanine (small amino acid) to Glycine (small amino acid), from Lysine (charged) to Glutamic acid (charged), from Glutamine (neutral polar) to Arginine (neutral polar), and from Phenylalanine (aromatic) to Tyrosine (aromatic). A substitution by Glycine is less likely to cause a drastic change since it is the simplest amino acid that possesses a single hydrogen atom as its side chain. By contrast, LOF mutations were associated with significant changes in the physicochemical properties of the amino acid. For example, transitions from the neutral

Cysteine, Glycine and Leucine to the basic Arginine were more significantly enriched (P-values: 0.001, 0.014, 0.01, respectively). Cysteine is often involved in disulfide linkages that enhance protein stability. In addition, another substitution with significant structural effect in LOF mutations was from Leucine and Serine to Proline, which serves to break alpha helices.

Mutations at the beginning of sequences result in LOF

The relative position of the mutation in CDS, protein and cDNA provides information on the position of the mutation relative to the length of the sequence in question. LOF mutations were significantly more enriched in the first decile (1/10) of the CDS and protein sequences, and in the second decile (2/10) of the cDNA sequences compared to GOF mutations (P-values: 0.004, 0.004 and 0.02, respectively). This result is in line with the fact that frameshift and nonsense mutations that occur earlier in the sequence are more likely to result in misfolded protein which is unlikely to be functional.

Interestingly, missense LOF mutations were significantly enriched only in the last decile (10/10) of the cDNA sequences compared to missense GOF mutations (P-value=0.03).

GOF mutations occur in specific protein domains

The modularity of proteins is critical for protein function and may also have clinical relevance. For example, it has been previously suggested that GOF mutations often occur in specific protein regions including protein–protein interaction interfaces and allosteric regions [41]. To further assess the effect of GOF and LOF on protein function, we used Pfam domains to examine if GOF and LOF mutations occur in particular domains, which are associated with specific functions [45]. Interestingly, GOF mutations were found to be significantly enriched in various

Pfam domains including STAT_alpha, ion_trans, STAT_bind, sod_cu (superoxide dismutase, copper/zinc binding domain), tropomyosin, vWA (von Willebrand factor type A), and SH2, which often mediates protein-protein interactions in signaling (P-values: 3.49×10^{-37}, 4.62×10^{-33}, 1.47×10^{-23}, 9.78×10^{-15}, 5.98×10^{-12}, 2.46×10^{-11}, 3.3×10^{-10}, respectively). The only Pfam domains found to be enriched for LOF mutations were p450 that are associated with the oxidation-reduction process, BRCT (BRCA1 C terminus), and Melibiase (P-values: 3.16×10^{-5}, 0.006, 0.002, respectively). Missense GOF mutations were found significantly enriched in the same Pfam domains (i.e., STAT_alpha, ion_trans, STAT_bind, sod_cu, with P-values: 4.12×10^{-28}, 2.79×10^{-22}, 4.86×10^{-18}, 2.08×10^{-9}, respectively).

GOF mutations were found disproportionally in transmembrane regions (P-val=1.3×10^{-8}) compared to LOF mutations. This enrichment disappeared when we considered only missense GOF/LOF mutations, demonstrating that the difference between GOF and LOF mutations is highly likely to be driven by protein truncating LOF mutations (i.e., frameshift, stop-gain, stop-loss).

The proteins having GOF mutations were predicted to have lower Clark's distance (P-value=1.54×10^{-8}), a term of reference to proteins with an amino acid composition similar to that of the average human protein (unbiased composition) [46-48]. Low complexity regions are composed of a subset of amino acids (biased composition), and hence have greater distance from the average protein. Additionally, low complexity regions tend to be disordered. Our results are consistent with the notion that truncating LOF mutations are enriched in low-complexity disordered regions which often play a key role in signaling and regulation [40, 41, 49].

Missense LOF mutations are enriched in protein core regions

Predicted relative solvent accessibility (RSA) of proteins was found to be one of the most discriminatory features for missense GOF and LOF mutations although it did not show a significant difference when considering all mutations. Missense LOF mutations were enriched in buried regions, likely disrupting the stability and folding of the protein, whereas missense GOF mutations were found in regions more exposed to the solvent (P-value=5.33×10^{-19}), which likely have functional effect. Finally, missense GOF mutations were found to be enriched in disordered protein regions (P-value=3.88×10^{-5}). Disordered regions are considered to be evolutionarily less conserved and therefore might have higher tolerance to LOF mutations.

Feature importance

To evaluate the discriminatory predictive power of our features, we used two different types of feature selection methods: (i) Fisher's exact test, which is a statistical significance test calculating the association between two groups; and (ii) the Boruta algorithm, which is a wrapper built around the random forest machine learning classification algorithm [50]. Fisher's exact test compares two groups (GOF and LOF mutation datasets in our case) and assesses the correlation or dependence between the groups. Then, the important features can be selected by ranking the P-values based on a user-defined threshold. However, with the Fisher's exact test, the interactions and correlations between feature variables or feature and target variables are not considered. Wrapper methods on the other hand measure the importance of features based on the classifier performance by iteratively selecting different feature subsets and evaluating the importance of each feature. Therefore, to evaluate the discriminative power of features based on the random forest classifier, as well as Fisher's exact method, we also applied the Boruta algorithm.

We annotated 33 features (12 protein-level, 11 gene-level, and 10 variant-level). After binning the continuous features and binarizing the categorical features, 33 features were broken into 1,725 variables. We found 89 variables of 27 features to be significantly enriched by the Fisher's exact test (2-sided Bonferronni-corrected pvalue<0.05) (Figure 1a).

When using the Boruta algorithm, SIFT, PolyPhen-2, Condel and REVEL scores were excluded from the feature list since these are not available for non-missense mutations. The missing values in the remaining features were imputed (see Methods, Figure S2 and S3). The categorical features were encoded by one-hot-encoding and the continuous features were used without binning. A balanced subset of 2,400 GOF and LOF mutations was randomly selected by using an under-sampling method [51]. We determined 64 variables of 28 features as important (having a Z-score higher than the shadow features) with the Boruta algorithm (Figure 1b).

The most important feature identified by the Boruta algorithm was the number of paralogs of a gene which is a metric for essentiality (Figure 1b). Other important features include gene-level features associated with essentiality (*de novo* excess rate, IS essentiality score, gene damage index (GDI), CDS length, mutation significance cutoff (MSC), LoFtool, RVIS, selective pressure, ExACpLI) as well as MOI; protein-level features associated with protein sequence (Clark's distance, disordered binding region, disordered region, protein domain, secondary structure); and variant-level features including impact rank, consequence type, CADD score, and variant class. Most of these features were also identified as important by Fisher's exact test. Interestingly, the gene-level threshold values calculated by CADD scores (MSC-CADD) as well as GDI which measures of the mutational damage that has accumulated in the general population were also among the important features. Both GDI and MSC metrics correlate with gene tolerance/intolerance. The different results between Fisher's exact test and the Boruta algorithm

could be due to conversion of continuous variables into categorical features which leads to loss of information. Another important difference between these two approaches is that with Fisher's exact test the correlations are evaluated independently, whereas the Boruta algorithm considers the interactions among the feature variables.

Functions of genes significantly differ for GOF and LOF mutations

To analyze the functional properties of genes, we performed gene ontology and protein class overrepresentation analysis. The genes with GOF mutations were significantly enriched in metal ion, cation and ion transport among GO biological processes and in molecular functions related to ion channel and transmembrane transporter activity (Supplementary Table 6). Transporter protein class was enriched in both gene sets, whilst metabolite interconversion enzyme and defense/immunity protein classes were enriched in the genes harboring LOF mutations. The transporter protein class involves the ion channels in which an altered channel activity may cause gain or loss of ion channel function.

Discussion

We have generated and made available a comprehensive database of annotated GOF and LOF mutations. Our aim was to analyze gene-level and protein-level features of GOF and LOF mutations and to identify potential discriminative features, then to discuss the biological implications of these features in terms of how they contribute to protein gain- and loss-of function. We applied an NLP algorithm to the available abstracts of disease-causing mutations in the HGMD database followed by multiple annotation algorithms based on genomic and protein positions and sequence information. Our study yielded multiple discriminative features of GOF

and LOF mutations suggesting a computational method based on machine learning techniques to classify the functional consequences of mutations.

To minimize the numbers of potential true positives and false negatives in our automated GOF/LOF prediction process, we excluded the abstracts mentioning both GOF and LOF mutations, and applied an iterative search of all available abstracts for each mutation and used the common prediction as label (GOF or LOF). Another remaining limitation of our approach was searching only paper abstracts instead of full text which is a more complex undertaking when applying NLP and requires free access. Whilst our study provides a first comprehensive GOF/LOF database, the data should be interpreted with caution as the GOF and LOF predictions are not manually curated.

To be able to apply Fisher's exact test, we converted the continuous variables to categorical data. However, this approach might cause loss of information and variability in the data. To assess our findings, we also applied a machine learning based algorithm in which all variables were kept continuous. To compensate for missing values in our dataset, we applied data imputation which might itself introduce bias.

Among 33 tested features (1,725 variables), 27 features (89 variables) were identified as being significantly discriminative for GOF and LOF mutations with Fisher's exact test, and 28 features (64 variables) with random forest-based algorithm (Figure 1). Overall, 21 features were found to be significant with both approaches. For example, GOF mutations were enriched in some specific protein domains such as binding regions and domain interfaces where they may change the protein activation, allostery and interactions, whereas missense LOF mutations were enriched in protein core regions where they might disrupt the stability and folding of the protein. Overall, these results extend our knowledge of different properties of GOF and LOF mutations, and offer

a powerful resource for analyzing these unique features. The findings of this study should improve our understanding of how GOF and LOF mutations impact protein function and ultimately cause disease.

Methods

Construction of GOF and LOF database

Selection of GOF and LOF mutations using NLP

We used the HGMD database (Professional version, 2019.1 release) that contains 256,070 mutations of which 202,639 were from the disease-causing class (DM, highest confidence for being pathogenic) [25, 26]. For each disease-causing mutation, we searched all associated paper titles and abstracts for possible acronyms of GOF and LOF mutations including "gain of function(s)", "gain-of-function(s)", "GOF", "loss of function(s)", "loss-of-function(s)", "LOF" by using phrase-based matching in spaCy v2.1.3 open-source natural language processing (NLP) library in Python 3.7.3 (https://spacy.io). The NLP text processing pipeline includes several steps: (i) tokenization: breaking text into meaningful units, (ii) tagging: assigning a part of speech tag to each token (noun, adjective, verb, etc.) depending on its usage in the sentence, (iii) parsing: extracting the dependency relationship between the words, and (iv) entity recognition: locating named entities in text and classifying them into categories. We added a PhraseMatcher class before entity recognition step to search for all possible phrases we defined. We converted the text into lowercase to allow for case-sensitivity. Abstracts containing both GOF and LOF acronyms were filtered out. We applied an iterative search on all available abstracts for each mutation. When there was a disagreement, i.e. a mutation was found as GOF in one abstract, and LOF in another abstract, it was labeled as "Mixed", and excluded from the dataset. The

remaining mutations were labeled as “GOF” or “LOF”, if one of the keywords was found in at least one associated abstract.

Selection of neutral variants

We used variants selected from the Genome Aggregation Database (gnomAD release 2.0.2) to compare GOF and LOF (function affecting) mutations which were very likely to be neutral variants [52]. The gnomAD database provides harmonized exome and genome sequencing data from unrelated individuals sequenced as part of various disease-specific and population genetic studies. The individuals affected by severe disease and their first-degree relatives were not included in the dataset, allowing it to be used as a reference dataset that represents the general population. We randomly selected 10^6 variants and extracted 134,450 likely neutral variants that occur in the same genes as our GOF and LOF mutations.

Feature annotations

We used several tools and resources to annotate an extensive set of features (Supplementary Table 7). This subsection is organized as (i) variant-level features, (ii) gene-level features, and (iii) protein-level features.

Annotation of variant-level features

All variants were first annotated with Ensembl Variant Effect Predictor (VEP, release 93) based on the Human GRCh37 assembly. Variant annotation tools provide a prediction for each transcript that a variant overlaps with. To filter one prediction result per variant, we selected the annotations of the Ensembl canonical transcripts [27]. The canonical transcript for a human gene is defined by Ensembl based on following: (1) Longest CCDS translation with no stop codons;

(2) If no (1), choose the longest Ensembl/Havana merged translation with no stop codons; (3) If no (2), choose the longest translation with no stop codons; (4) If no translation, choose the longest non-protein-coding transcript.

Consequence. The predicted variant consequence by Ensembl (i.e., missense, frameshift, stop_gain, splice, etc.) and the impact rating (i.e., high, moderate, low and modifier) indicating the severity of the consequence were used as two separate features [27].
The variant classes were defined by Ensembl as: (i) single nucleotide polymorphism (SNP), (b) insertion of one or several nucleotides (insertion), and (iii) deletion of one or several nucleotides (deletion).

Impact. The predicted functional impact of amino acid substitutions by SIFT and PolyPhen-2 were calculated by VEP [16, 17, 27]. Both predicted terms and scores were used as features.
The highest allele frequency (max_af) observed in any population from 1000 Genomes Project, Exome Sequencing Project (ESP) or gnomAD was calculated by VEP and used as a feature [52, 53].
We used several available plugins with VEP (https://github.com/Ensembl/VEP_plugins) to extend functional impact annotations including CADD, Condel and REVEL. Combined Annotation-Dependent Depletion (CADD) scores have been widely used for predicting the deleteriousness of single nucleotide variants and insertion/deletion variants in the human genome [15]. Condel is a predicted deleteriousness score of missense variants calculated based on the pre-calculated SIFT and PolyPhen-2 scores from the Ensembl API [16, 17, 29, 54].
REVEL is a score predicting the pathogenicity of missense variants [18].

Position. The relative position of the mutation in the protein sequence, cDNA, and CDS were calculated by dividing the mutation position by the sequence length. The feature indicates the percentile interval in which the mutation occurs: (1): 0-10%, (2): 10-20%, (3): 20-30%, (4): 30-40%, (5): 40-50%, (6): 50-60%, (7): 60-70%, (8): 70-80%, (9): 80-90%, (10): 90-100%.

Annotation of gene-level features

Essentiality. Genic intolerance was evaluated using different tools. We calculated LoFtool and ExACpLI scores using VEP plugins. LoFtool is a gene intolerance ranking score based on the ratio of LoF (stop-gain, splice site disrupting, frameshift) mutations to synonymous variants from the Exome Aggregation Consortium (ExAC) dataset (http://exac.broadinstitute.org/), but adjusted for the *de novo* mutation rate and protein evolutionary conservation [55]. A lower LoFtool score indicates more intolerance to functional variation. The ExACpLI score represents the probability of a gene being loss-of-function intolerant (pLI) computed by the ExAC consortium based on observed and expected protein-truncating variant counts in exome sequence data for 60,706 individuals of diverse ethnicities [56].

The Residual Variation Intolerance Score (RVIS) is a gene-based score calculated based on common functional variation compared to neutral variation from whole exome sequence data from large datasets from the human population [57]. While a positive RVIS score points to more common functional variation, a negative score indicates intolerance. We used the latest RVIS release (v4), which is based on the ExAC v2 (http://gnomad.broadinstitute.org/), the consensus coding sequence (CCDS) release 20 (http://www.ncbi.nlm.nih.gov/CCDS), and Ensembl release 87 [57].

The *de novo* excess rate (Z-score) estimates the rate of *de novo* mutation per gene and calculates if a given gene contains more *de novo* mutations than expected by chance alone [58].

The predicted functional indispensability score (IS) estimates the essentiality based on a gene's network and evolutionary properties [59].

The gene damage index (GDI) provides an accumulated mutational damage estimate (based on CADD scores) for each gene in the healthy human population from the 1,000 Genomes Project database [15, 48, 60].

Selective evolutionary pressure acting on each gene was estimated by the McDonald–Kreitman neutrality index at the population level [61]. The neutrality index compares the numbers of non-silent and silent substitutions to the numbers of non-silent and silent polymorphisms, and measures the direction and degree of departure from neutrality.

The length of the coding sequence (CDS) of the canonical transcript and number of paralogs were extracted from Ensembl BioMart release 75 [62, 63].

Inheritance. Mode of inheritance (MOI) of each gene was extracted from the HGMD database and sub-categorized as autosomal recessive (AR), autosomal dominant (AD), autosomal recessive/dominant (ADAR), X-linked dominant (XLD), X-linked recessive (XLR), and unknown [26].

Pathogenicity. The mutation significance cutoff (MSC) introduces gene-level thresholds for predicted mutation impact scores such as CADD, PolyPhen-2 and SIFT [15-17, 64]. In this study, we included MSC-CADD cutoffs with 95% confidence interval.

Annotation of protein-level features

Each protein can generate a number of alternative protein sequences (isoforms) through different biological events such as alternative splicing, alternative promoter or polyadenylation site usage, etc. The canonical protein isoform is defined as the most relevant sequence and/or the most similar sequence. To annotate the correct protein isoform, we used Ensembl protein IDs instead of UniProt IDs of canonical isoforms. As a result, in this study, we considered protein sequences that were translated from canonical transcripts but not necessarily canonical isoform protein sequences.

Amino acids. The reference and alternative amino acid types, as well as their positions were extracted by Ensembl VEP and used as features [63]. When there was an insertion or deletion, only the first amino acid type and position were considered.

Protein domain. Pfam is a database of protein families where families are represented by multiple sequence alignments and hidden Markov models [45]. Pfam protein domain information of all genes in our dataset were downloaded from the Ensembl BioMart (release 99) using Ensembl protein IDs with version corresponding to the canonical transcripts as annotated by Ensembl VEP [38, 62]. We also searched the InterPro database for the proteins lacking Pfam domain information in Ensembl BioMart. When a protein exists in a protein database but the mutation position did not fall into a known protein domain, it was considered as “outside” of the domain, and when a protein was absent from both the Pfam and InterPro databases, it was considered as “unknown” domain.

Transmembrane helix regions were derived from Ensembl VEP, which uses predictions by TMHMM [65].

Post-translational modification (PTM) regions for phosphorylation, acetylation, ubiquitination, methylation, N-linked glycosylation, and O-linked glycosylation were downloaded from the dbPTM. The dbPTM is a database that includes experimentally validated PTM sites from available databases and manually curated PTMs from the literature [66].

Disordered region. Disordered regions are defined as regions that lack a stable well-defined tertiary structure under native conditions. IUPred2A was used to predict intrinsically disordered protein regions (IDPRs), and disordered binding regions [67]. IUPred2A is based on pairwise energies estimated from observed amino acid compositions and returns probabilities of each residue being part of a disordered region and being part of a disordered binding region. We used the default (long) disorder prediction type predicting global structural disorder that encompasses at least 30 consecutive amino acid residues of the protein. Clark's distance was used to estimate the amino acid composition complexity of each protein [46, 47] from a previously published study [48].

Solvent accessible area and secondary structure. NetsurfP-1.1 was used to predict the solvent accessible surface area and secondary structure for each protein [68]. This method is based on neural networks that had been trained on experimentally determined protein structures. NetsurfP-1.1 returns solvent accessible surface area (ASA) given in $Å^2$, relative surface area (RSA), RSA class ((buried or exposed i.e., below or above 25% exposure), reliability of each prediction (Z-

fit). In addition, NetsurfP-1.1 provides the probability of each residue being in an alpha-helix, beta-strand, or coil region.

Feature importance

To evaluate the discriminatory power of our features, we used (i) Fisher's exact test, (ii) Boruta algorithm.

Fisher's exact test

We used SciPy v1.1.0 python library to perform Fisher's exact test on 2x2 contingency tables. Each category of a feature was evaluated separately based on its presence/absence in GOF and LOF datasets. All continuous features were converted to categorical features to be able to calculate p-values with Fisher's exact test. When available, we used the original definitions of categorical features instead of continuous values. When there were no defined cutoff values, we identified three classes based on distributions of the continuous values. First, the interquartile range (IQR) of non-normally distributed data which is defined as $IQR=Q_3-Q_1$, where Q_1 is the first quartile and Q_3 is the third quartile, was calculated [69]. Then, the boundaries for categories were identified as median±0.25*IQR. A drawback of converting continuous features into categorical variables is that it might result in loss of information and power. A full list of categorical features can be found in Supplementary Table 7. All p-values were corrected using the Bonferroni correction method.

For the significant p-values showing associations between GOF versus LOF mutations, GOF versus neutral variants, and LOF versus neutral variants, we also calculated the logarithm of the odds ratio (OR) to compare features of GOF and LOF mutations by using neutral variants as

background. That is, we determined log-odds ratios indicating enrichment for GOF mutations and LOF mutations separately, compared to neutral variants.

Boruta algorithm

We used the Boruta feature selection package implemented in R, which is a wrapper algorithm built around the random forest classifier [50]. Whereas traditional algorithms return a subset of features that gives the best possible classification result (minimal-optimal approach), the Boruta algorithm identifies all features that are somewhat relevant (weakly or strongly) for classification (all-relevant approach).

In random forest, the importance of a feature is calculated as the loss of accuracy of classification caused by the random permutation of feature values. It is computed separately for all trees in the forest. Then the Z-score is computed by dividing the average accuracy loss by its standard deviation. The Z-score cannot be directly used to measure feature importance as a reference value is needed. For this reason, the Boruta algorithm creates shadow features which are copies of the original features with randomly shuffled values. Then, the maximum importance of shadow features can be used as a reference to decide if original features are important. The features, which significantly outperform the best shadow feature, are interpreted as confirmed.

First, we omitted the features SIFT, PolyPhen-2, Condel and REVEL scores which are only available for missense mutations to deal with high missing rates in our dataset. We treated the missing categorical data as another category, “unknown”. The continuous features were used without binning. A full list of features can be found in Supplementary Table 7. The remaining missing values in continuous features were imputed using the MICE R package with the

predictive mean matching (pmm) imputation method [70]. We also performed scaling on data to normalize the features.

To deal with imbalanced GOF and LOF dataset sizes, we used the ROSE R package with under-sampling option [51]. The probability of the minority class (GOF) was chosen as p=0.4. In the resulting balanced dataset of 2,400 mutations, there were 1,254 GOF and 1,146 LOF mutations.

Overrepresentation analysis

The Gene Ontology (GO) and protein class overrepresentation analysis were performed using PANTHER for gene sets having GOF mutations and LOF mutations separately compared to all human genes in the PANTHER database [71-73]. Fisher's exact test has been performed with Bonferroni correction.

Acknowledgements

Avner Schlessinger is a Co-founder of AIchemy.

References

1. Kagawa, R., et al., *Alanine-scanning mutagenesis of human signal transducer and activator of transcription 1 to estimate loss- or gain-of-function variants.* J Allergy Clin Immunol, 2017. **140**(1): p. 232-241.
2. Miner, J.J. and M.S. Diamond, *MDA5 and autoimmune disease.* Nat Genet, 2014. **46**(5): p. 418-9.
3. Boisson, B., P. Quartier, and J.-L. Casanova, *Immunological loss-of-function due to genetic gain-of-function in humans: autosomal dominance of the third kind.* Current opinion in immunology, 2015. **32**: p. 90-105.
4. Farnier, M., *PCSK9: From discovery to therapeutic applications.* Arch Cardiovasc Dis, 2014. **107**(1): p. 58-66.
5. Wemhöner, K., et al., *Gain-of-function mutations in the calcium channel CACNA1C (Cav1.2) cause non-syndromic long-QT but not Timothy syndrome.* J Mol Cell Cardiol, 2015. **80**: p. 186-95.
6. Templin, C., et al., *Identification of a novel loss-of-function calcium channel gene mutation in short QT syndrome (SQTS6).* Eur Heart J, 2011. **32**(9): p. 1077-88.
7. Lanternier, F., et al., *Primary immunodeficiencies underlying fungal infections.* Curr Opin Pediatr, 2013. **25**(6): p. 736-47.
8. Bustamante, J., et al., *Mendelian susceptibility to mycobacterial disease: Genetic, immunological, and clinical features of inborn errors of IFN-γ immunity.* Seminars in Immunology, 2014. **26**(6): p. 454-470.
9. Wang, Z. and J. Moult, *SNPs, protein structure, and disease.* Hum Mutat, 2001. **17**(4): p. 263-70.
10. Peng, Y. and E. Alexov, *Investigating the linkage between disease-causing amino acid variants and their effect on protein stability and binding.* Proteins, 2016. **84**(2): p. 232-9.
11. Peng, Y., E. Alexov, and S. Basu, *Structural perspective on revealing and altering molecular functions of genetic variants linked with diseases.* Int J Mol Sci, 2019. **20**(3).
12. Zhang, J., et al., *The impact of next-generation sequencing on genomics.* J Genet Genomics, 2011. **38**(3): p. 95-109.
13. Pavlopoulos, G.A., et al., *Unraveling genomic variation from next generation sequencing data.* BioData Min, 2013. **6**(1): p. 13.
14. Bamshad, M.J., et al., *Exome sequencing as a tool for Mendelian disease gene discovery.* Nat Rev Genet, 2011. **12**(11): p. 745-55.
15. Kircher, M., et al., *A general framework for estimating the relative pathogenicity of human genetic variants.* Nat Genet, 2014. **46**(3): p. 310-5.
16. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. **7**(4): p. 248-9.
17. Kumar, P., S. Henikoff, and P.C. Ng, *Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.* Nat Protoc, 2009. **4**(7): p. 1073-81.
18. Ioannidis, N.M., et al., *REVEL: An Ensemble method for predicting the pathogenicity of rare missense variants.* Am J Hum Genet, 2016. **99**(4): p. 877-885.
19. Heyne, H.O., et al., *Predicting functional effects of missense variants in voltage-gated sodium and calcium channels.* Sci Transl Med, 2020. **12**(556): p. eaay6848.

20. Jung, S., et al., *Identification of genomic features in the classification of loss- and gain-of-function mutation.* BMC Med Inform Decis Mak, 2015. **15 Suppl 1**(Suppl 1): p. S6.
21. Liu, M., L.T. Watson, and L. Zhang, *HMMvar-func: a new method for predicting the functional outcome of genetic variants.* BMC Bioinformatics, 2015. **16**: p. 351.
22. Eyre-Walker, A. and P.D. Keightley, *The distribution of fitness effects of new mutations.* Nature Reviews Genetics, 2007. **8**(8): p. 610-618.
23. Bridgham, J.T., et al., *Evolution of a new function by degenerative mutation in cephalochordate steroid receptors.* PLOS Genetics, 2008. **4**(9): p. e1000191.
24. Lynch, M. and J.S. Conery, *The evolutionary fate and consequences of duplicate genes.* Science, 2000. **290**(5494): p. 1151.
25. Stenson, P.D., et al., *The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting.* Human Genetics, 2020. **139**(10): p. 1197-1207.
26. Stenson, P.D., et al., *The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies.* Hum Genet, 2017. **136**(6): p. 665-677.
27. McLaren, W., et al., *The Ensembl Variant Effect Predictor.* Genome Biology, 2016. **17**(1): p. 122.
28. Adzhubei, I., D.M. Jordan, and S.R. Sunyaev, *Predicting functional effect of human missense mutations using PolyPhen-2.* Curr Protoc Hum Genet, 2013. **Chapter 7**: p. Unit7.20.
29. González-Pérez, A. and N. López-Bigas, *Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel.* Am J Hum Genet, 2011. **88**(4): p. 440-9.
30. Flanagan, S.E., A.M. Patch, and S. Ellard, *Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations.* Genet Test Mol Biomarkers, 2010. **14**(4): p. 533-7.
31. Bartha, I., et al., *Human gene essentiality.* Nat Rev Genet, 2018. **19**(1): p. 51-62.
32. Gu, Z., et al., *Role of duplicate genes in genetic robustness against null mutations.* Nature, 2003. **421**(6918): p. 63-6.
33. Kamath, R.S., et al., *Systematic functional analysis of the Caenorhabditis elegans genome using RNAi.* Nature, 2003. **421**(6920): p. 231-7.
34. White, J.K., et al., *Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes.* Cell, 2013. **154**(2): p. 452-64.
35. De Kegel, B. and C.J. Ryan, *Paralog buffering contributes to the variable essentiality of genes in cancer cell lines.* PLoS Genet, 2019. **15**(10): p. e1008466.
36. Conant, G.C. and K.H. Wolfe, *Turning a hobby into a job: how duplicated genes find new functions.* Nat Rev Genet, 2008. **9**(12): p. 938-50.
37. Dandage, R. and C.R. Landry, *Paralog dependency indirectly affects the robustness of human cells.* Molecular systems biology, 2019. **15**(9): p. e8871-e8871.
38. Cunningham, F., et al., *Ensembl 2019.* Nucleic Acids Res, 2019. **47**(D1): p. D745-d751.
39. Turner, T.N., et al., *Proteins linked to autosomal dominant and autosomal recessive disorders harbor characteristic rare missense mutation distribution patterns.* Hum Mol Genet, 2015. **24**(21): p. 5995-6002.
40. Zhong, Q., et al., *Edgetic perturbation models of human inherited disorders.* Mol Syst Biol, 2009. **5**: p. 321.

41. Taipale, M., *Disruption of protein function by pathogenic mutations: common and uncommon mechanisms (1).* Biochem Cell Biol, 2019. **97**(1): p. 46-57.
42. Coban-Akdemir, Z., et al., *Identifying Genes Whose Mutant Transcripts Cause Dominant Disease Traits by Potential Gain-of-Function Alleles.* Am J Hum Genet, 2018. **103**(2): p. 171-187.
43. Meyts, I., et al., *Exome and genome sequencing for inborn errors of immunity.* J Allergy Clin Immunol, 2016. **138**(4): p. 957-969.
44. Wilkie, A.O., *The molecular basis of genetic dominance.* Journal of medical genetics, 1994. **31**(2): p. 89-98.
45. Punta, M., et al., *The Pfam protein families database.* Nucleic Acids Res, 2012. **40**(Database issue): p. D290-301.
46. Wootton, J.C. and S. Federhen, *Statistics of local complexity in amino acid sequences and sequence databases.* Computers & Chemistry, 1993. **17**(2): p. 149-163.
47. Cha, S. *Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions*. 2007.
48. Itan, Y., et al., *The human gene damage index as a gene-level approach to prioritizing exome variants.* Proceedings of the National Academy of Sciences, 2015. **112**(44): p. 13615.
49. Guo, Y., et al., *Dissecting disease inheritance modes in a three-dimensional protein network challenges the "guilt-by-association" principle.* Am J Hum Genet, 2013. **93**(1): p. 78-89.
50. Kursa, M.B. and W.R. Rudnicki, *Feature Selection with the Boruta Package.* 2010, 2010. **36**(11): p. 13.
51. Lunardon, N., G. Menardi, and N. Torelli, *ROSE: A Package for Binary Imbalanced Learning.* R journal, 2014. **6**(1).
52. Karczewski, K.J., et al., *The mutational constraint spectrum quantified from variation in 141,456 humans.* Nature, 2020. **581**(7809): p. 434-443.
53. Auton, A., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
54. Flicek, P., et al., *Ensembl 2012.* Nucleic Acids Res, 2012. **40**(Database issue): p. D84-90.
55. Fadista, J., et al., *LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals.* Bioinformatics, 2017. **33**(4): p. 471-474.
56. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706 humans.* Nature, 2016. **536**(7616): p. 285-91.
57. Petrovski, S., et al., *Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes.* PLOS Genetics, 2013. **9**(8): p. e1003709.
58. Samocha, K.E., et al., *A framework for the interpretation of de novo mutation in human disease.* Nat Genet, 2014. **46**(9): p. 944-50.
59. Khurana, E., et al., *Interpretation of Genomic Variants Using a Unified Biological Network Approach.* PLOS Computational Biology, 2013. **9**(3): p. e1002886.
60. Abecasis, G.R., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.
61. Stoletzki, N. and A. Eyre-Walker, *Estimation of the neutrality index.* Mol Biol Evol, 2011. **28**(1): p. 63-70.
62. Haider, S., et al., *BioMart Central Portal--unified access to biological data.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W23-7.

63. Flicek, P., et al., *Ensembl 2013.* Nucleic Acids Res, 2013. **41**(Database issue): p. D48-55.
64. Itan, Y., et al., *The mutation significance cutoff: gene-level thresholds for variant predictions.* Nature Methods, 2016. **13**(2): p. 109-110.
65. Krogh, A., et al., *Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.* J Mol Biol, 2001. **305**(3): p. 567-80.
66. Huang, K.Y., et al., *dbPTM in 2019: exploring disease association and cross-talk of post-translational modifications.* Nucleic Acids Res, 2019. **47**(D1): p. D298-d308.
67. Mészáros, B., G. Erdos, and Z. Dosztányi, *IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding.* Nucleic Acids Res, 2018. **46**(W1): p. W329-w337.
68. Petersen, B., et al., *A generic method for assignment of reliability scores applied to solvent accessibility predictions.* BMC Struct Biol, 2009. **9**: p. 51.
69. Madadizadeh, F., M. Ezati Asar, and M. Hosseini, *Common Statistical Mistakes in Descriptive Statistics Reports of Normal and Non-Normal Variables in Biomedical Sciences Research.* Iran J Public Health, 2015. **44**(11): p. 1557-8.
70. van Buuren, S. and K. Groothuis-Oudshoorn, *mice: Multivariate Imputation by Chained Equations in R.* 2011, 2011. **45**(3): p. 67.
71. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
72. *The Gene Ontology Resource: 20 years and still GOing strong.* Nucleic Acids Res, 2019. **47**(D1): p. D330-d338.
73. Mi, H., et al., *PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools.* Nucleic Acids Res, 2019. **47**(D1): p. D419-d426.

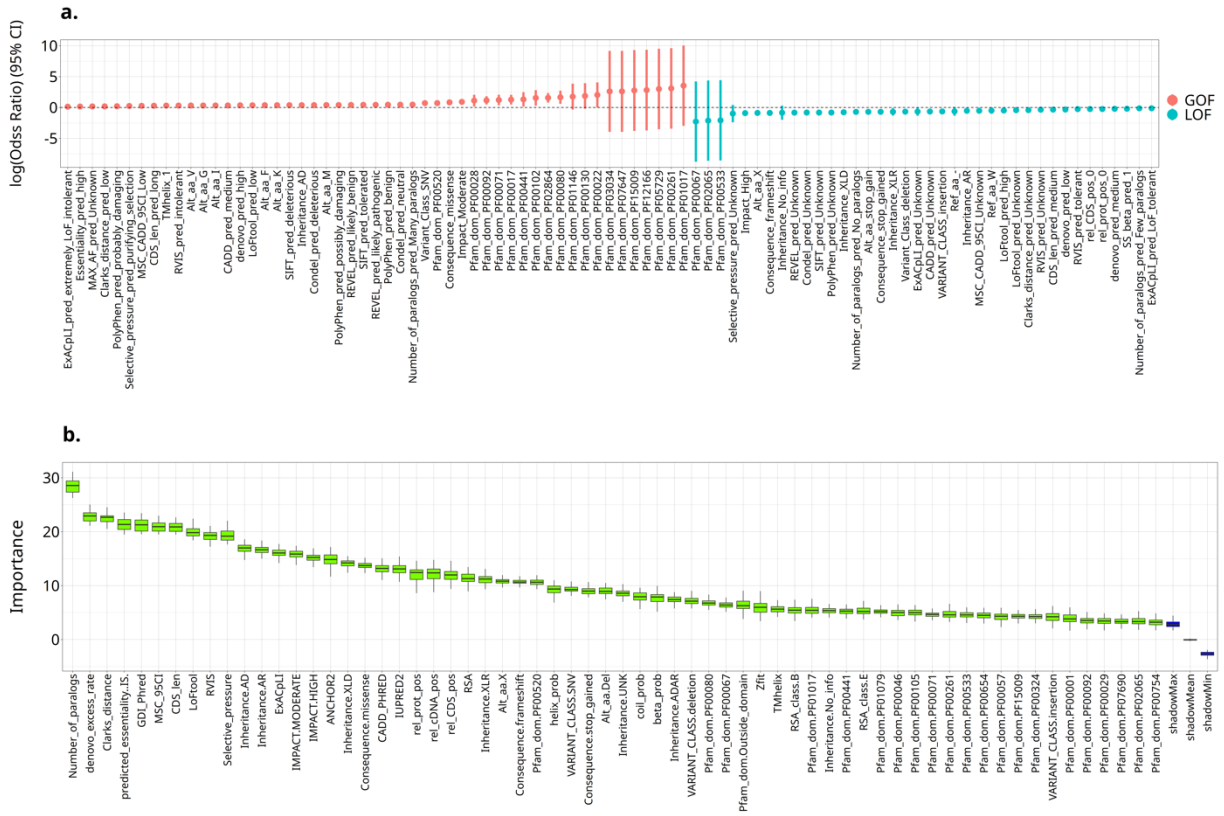

Figure 1. Enrichment of GOF and LOF mutations with gene- and protein- level features. **a)** Significant associations (Bonferroni-corrected 2-sided p-value < 0.05) calculated by Fisher's exact test are shown. The y-axis corresponds to log10 odds ratio. Positive values indicate a feature's enrichment for GOF mutations (shown in red) and negative values indicate a feature's enrichment in LOF mutations (shown in blue). See Supplementary Table 7 for details of the features. **b)** Feature importance ranked by the Boruta algorithm. The y-axis represents the Z-score of every feature in the shuffled dataset. The blue boxplots correspond to minimal, average and maximum Z-scores of shadow attributes (permuted copies of original features). The green boxplots represent Z-scores of confirmed attributes. A feature is rejected if it has a lower Z-score than that of max Z-score (if it under-performs the best shadow feature), and confirmed if it outperforms the best shadow feature.

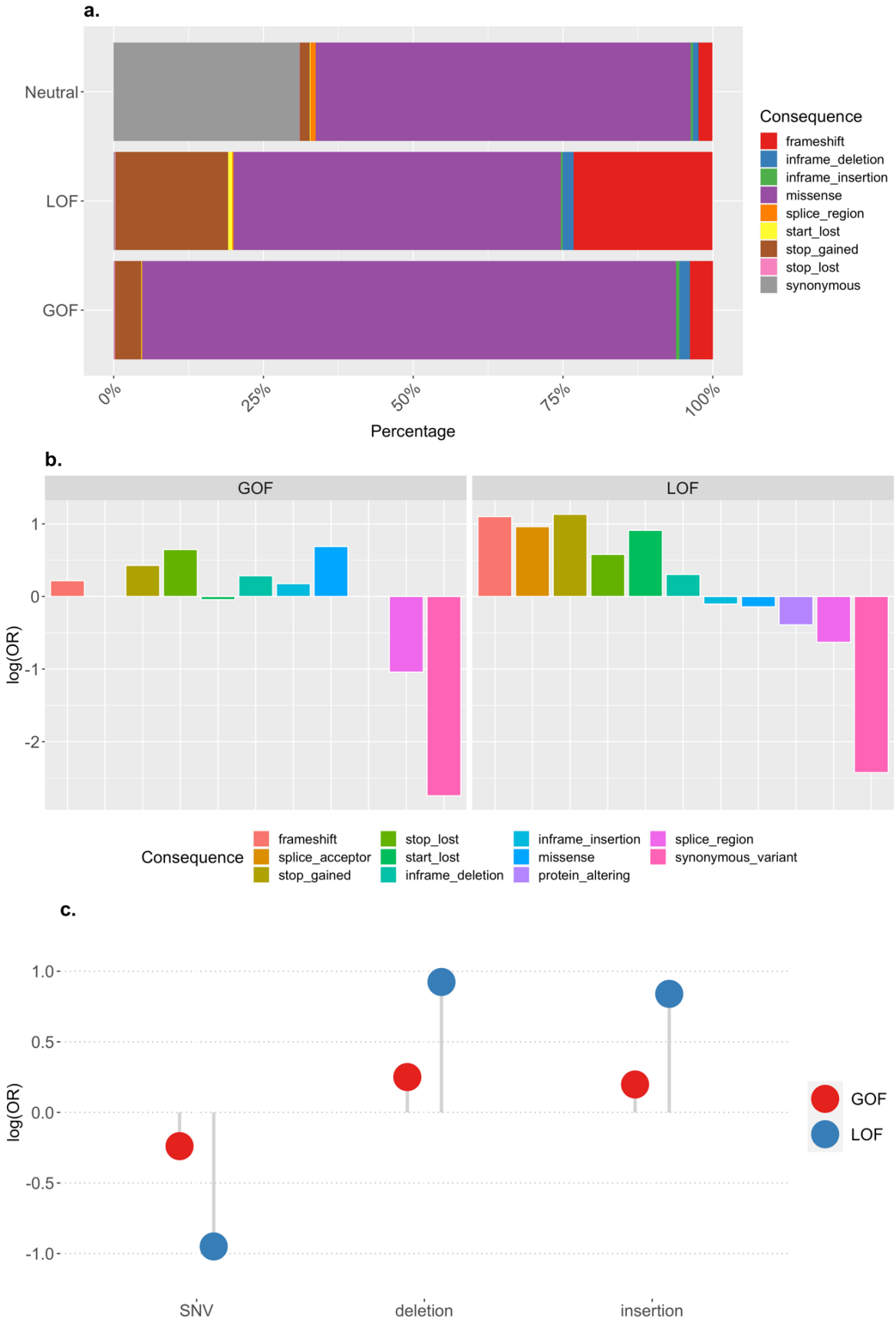
a.
Neutral
LOF
GOF
0%
25%
50%
75%
100%
Percentage
Consequence
frameshift
inframe_deletion
inframe_insertion
missense
splice_region
start_lost
stop_gained
stop_lost
synonymous
b.
GOF
LOF
log(OR)
1
0
-1
-2
Consequence
frameshift
splice_acceptor
stop_gained
stop_lost
start_lost
inframe_deletion
inframe_insertion
missense
protein_altering
splice_region
synonymous_variant
c.
log(OR)
1.0
0.5
0.0
-0.5
-1.0
SNV
deletion
insertion
GOF
LOF

Figure 2. Variant consequence and class types. a) Predicted mutational consequence percentages for each mutation type, including GOF, LOF and likely neutral variants. **b)** Predicted mutational consequences of GOF/LOF mutations. The y-axis corresponds to the log10 of the odds ratio of the enrichment of predicted consequences in GOF and LOF mutations compared to likely neutral variants. **c)** Predicted variant classes. The y-axis corresponds to the log10 of the odds ratio of the enrichment of variant classes in GOF and LOF mutations compared to likely neutral variants.

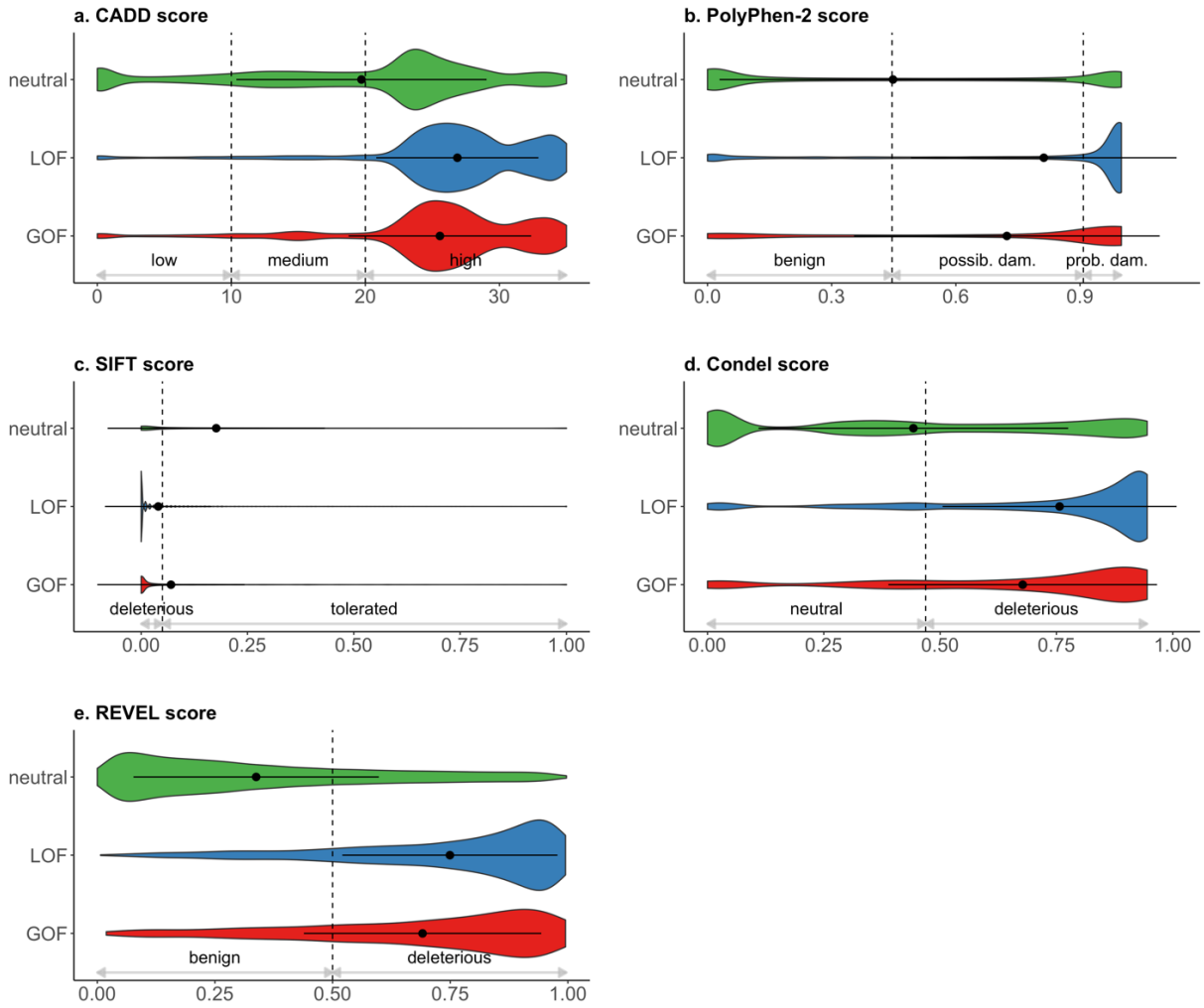

Figure 3. Pathogenicity prediction scores for GOF/LOF mutations by various methods. The x-axis corresponds to the predicted scores of GOF and LOF mutations, and likely neutral variants. The dashed lines represent the cutoff values we used for categorizing the scores (see Methods). **a)** The variants with higher CADD scores are predicted to be more deleterious. A CADD score of 10 indicates a variant to be among the 10% most deleterious mutations in the human genome, 20 indicates it to be among the top 1%, 30 indicates it to be among the top 0.1%, etc. **b)** The PolyPhen-2 score ranges from 0.0 (tolerated) to 1.0 (deleterious). The variants with less than a 0.446 score are predicted to be "benign", whereas variants with more than a 0.908 score are predicted to be "probably damaging". **c)** The SIFT score ranges from 0.0 (deleterious) to 1.0 (tolerated) where substitutions with a SIFT score < 0.05 are predicted to be "deleterious" and all others "tolerated". **d)** The Condel score ranges from 0.0 (tolerated) to 1.0 (deleterious) where a substitution is predicted to be "deleterious" if the Condel score is greater than 0.469, and "neutral" otherwise. **e)** REVEL score ranges from 0 to 1 and variants with higher scores are predicted to be more likely to be pathogenic. REVEL scores above 0.5 indicate "likely disease causing" and scores below 0.5 indicate "likely benign" variants.

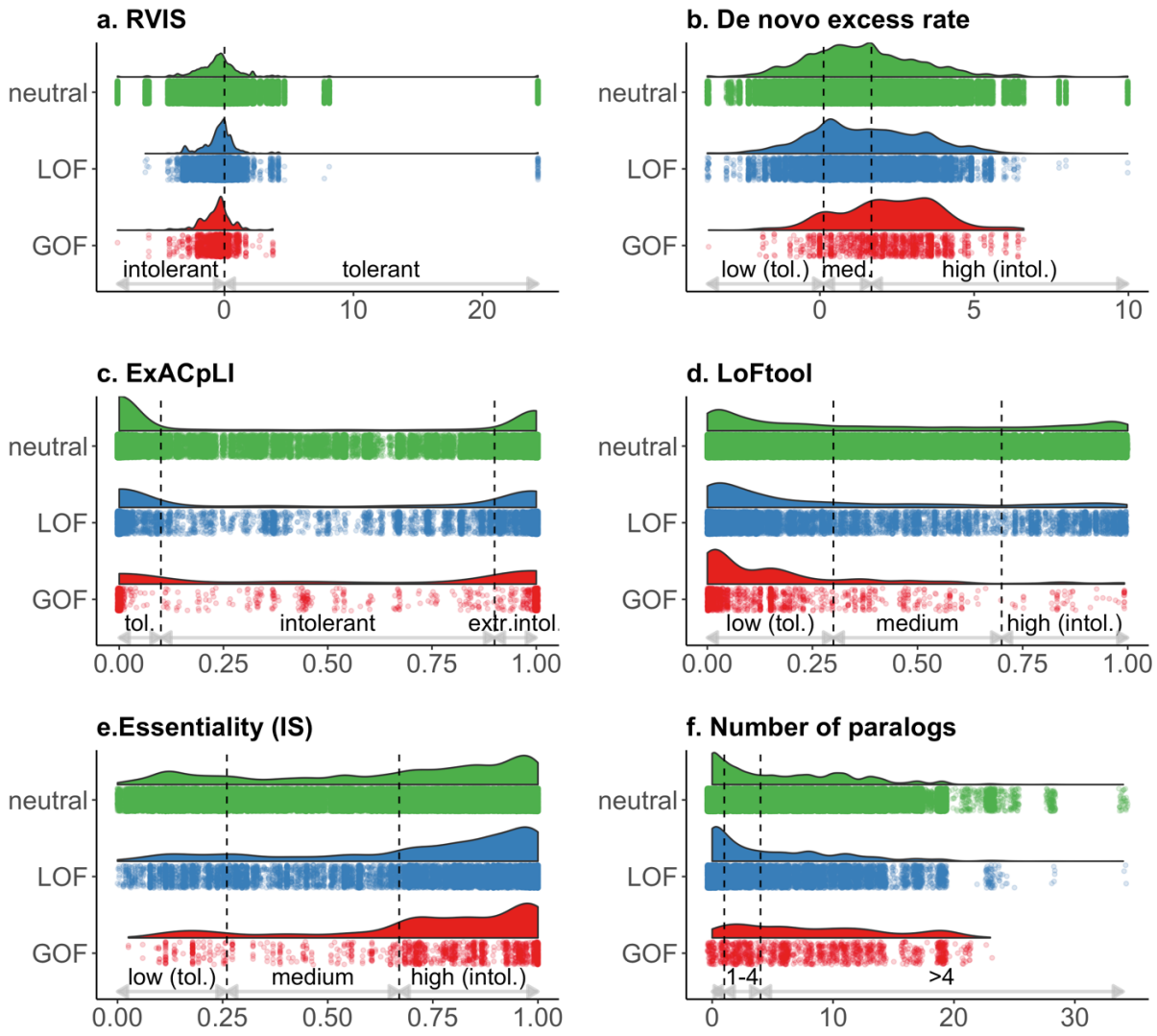

Figure 4. Essentiality predictions of the genes with GOF and LOF mutations, and likely neutral variants by various methods. In panels a-e, the x-axis corresponds to the predicted essentiality scores of the genes in each dataset, whereas in panel f the x-axis corresponds to the number of paralogs of the genes. The dashed lines represent the cutoff values we used for categorizing the features. When there were no defined cutoff values, three classes were identified based on distributions of the values (see Methods). **a)** A positive RVIS score is indicative of more common functional variation, a negative score implies intolerance. **b)** *de novo* excess rate (Z-core) is an estimation of the *de novo* mutation rate and a higher rate indicates intolerance. **c)** ExACpLI score represents the probability of a gene being LOF intolerant. **d)** a lower LoFtool score indicates more intolerance to functional variation. **e)** Predicted indispensability score indicates the essentiality based on a gene's network and evolutionary properties. **f)** Number of paralogs were shown in three categories: having more than 4 paralogs, 1-4 paralogs, and being a singleton.

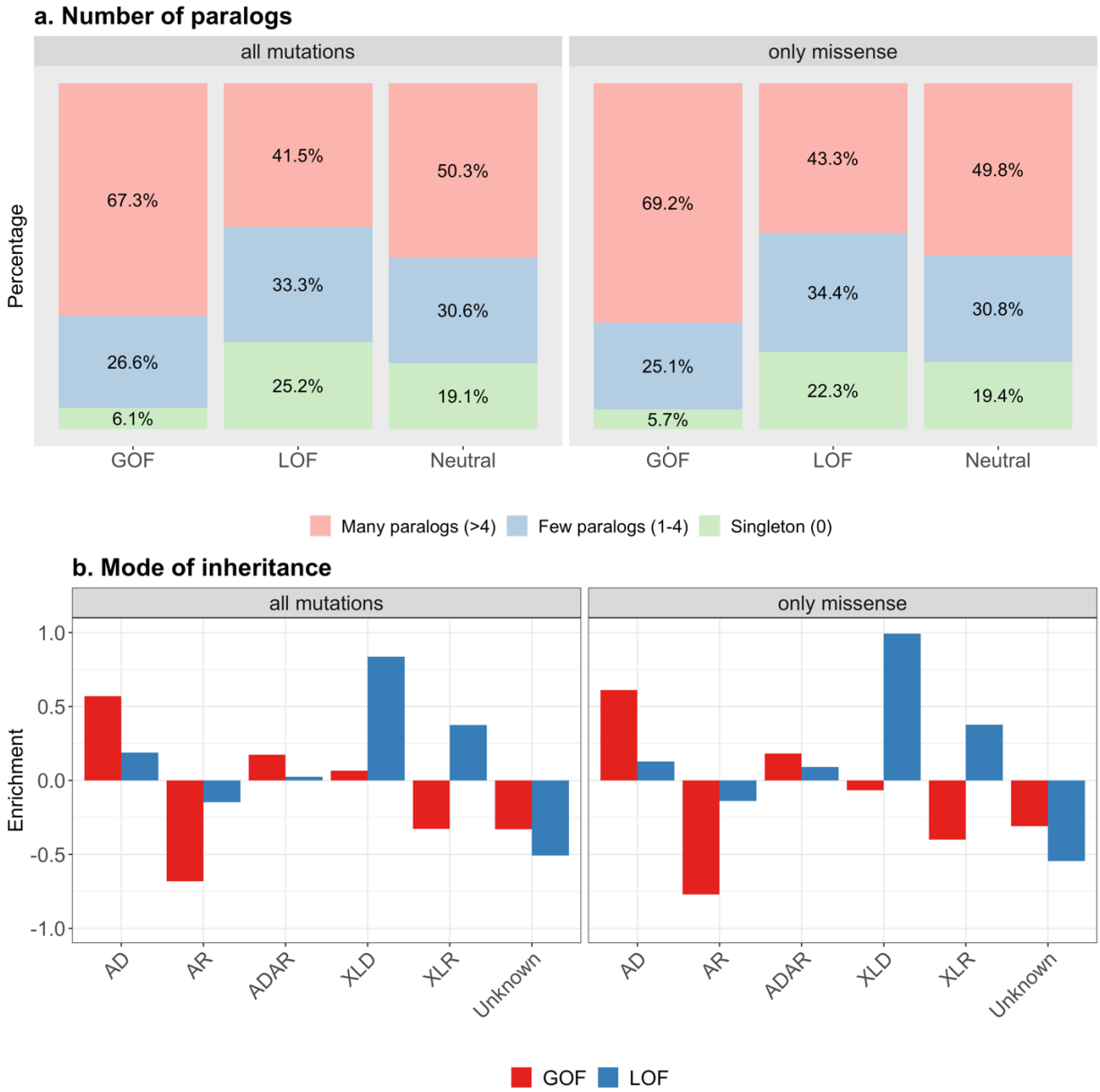

Figure 5. Number of paralogs and mode of inheritance. **a)** Fraction of GOF and LOF mutations, and likely neutral variants in genes with different numbers of paralogs. The y-axis corresponds to having more than 4 paralogs (many), having 1-4 paralogs (few), and being a singleton (none). The left-hand panel shows the fractions in the whole dataset, the right-hand panel shows the fractions in the subset of missense mutations. **b)** Mode of inheritance of the genes with GOF and LOF mutations. The y-axis corresponds to the log10 of the odds ratio of the enrichment of inheritance type in GOF and LOF mutations compared to likely neutral variants. The x-axis corresponds to autosomal dominant (AD), autosomal recessive (AR), autosomal dominant and recessive (ADAR), X-linked dominant (XLD), X-linked recessive (XLR), and not known (Unknown). Positive and negative values correspond to overrepresentation and underrepresentation in GOF/LOF mutations as compared to neutral variants, respectively. The

left-hand panel shows the fractions in the whole dataset, the right-hand panel shows the fractions in the subset of missense mutations.

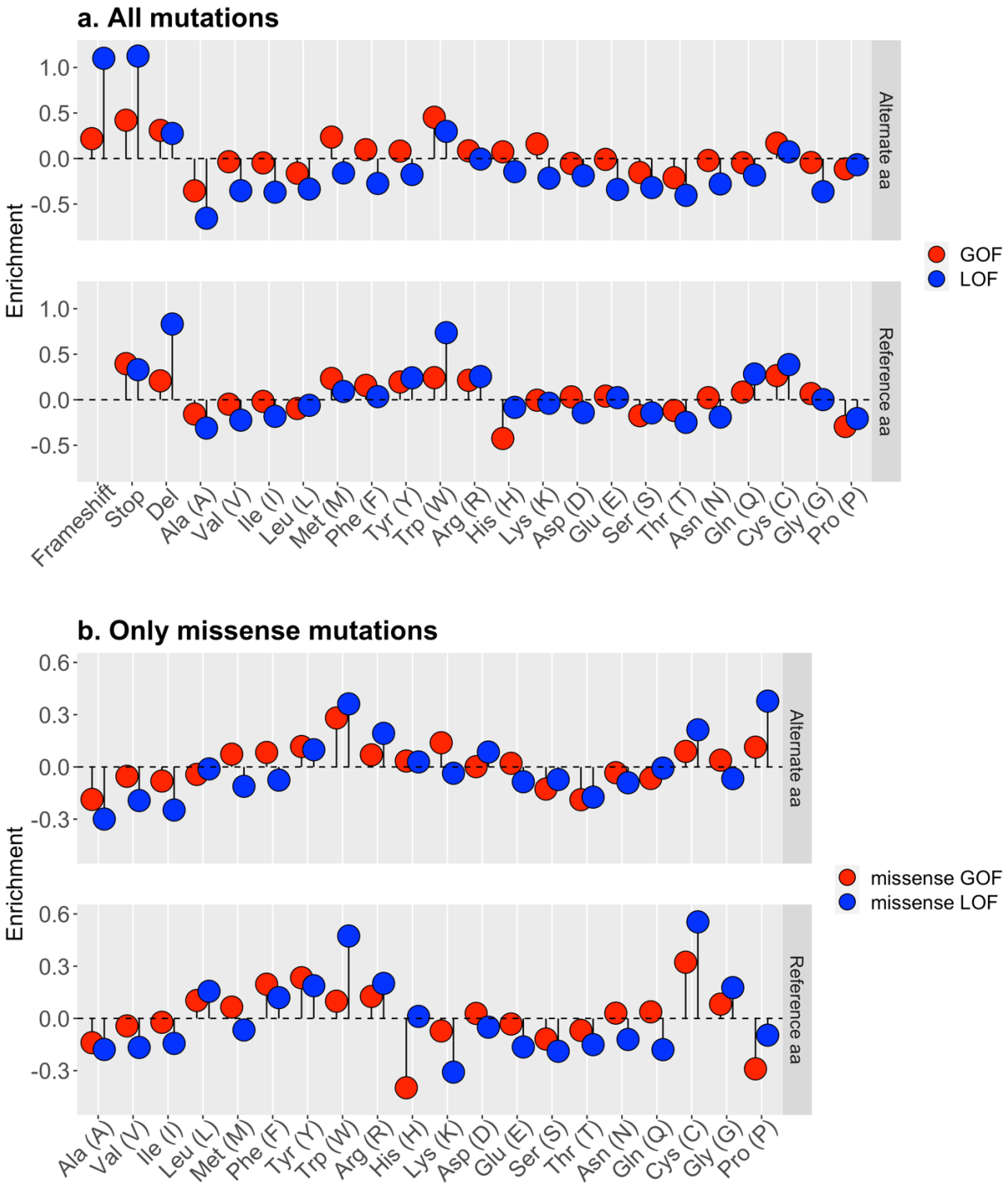

Figure 6. Amino acid changes in GOF and LOF mutations. **a)** Reference and alternate amino acid type enrichment of GOF and LOF mutations compared to likely neutral variants. The y-axis corresponds to the log10 of the odds ratio of the enrichment of reference amino acid type. The x-axis corresponds to 20 amino acid types plus frameshift, stop-gain (stop), and deletion (del). **b)** Reference and alternate amino acid type enrichment of missense GOF and missense LOF

mutations compared to likely neutral missense variants. The y-axis corresponds to the log10 of odds ratio of the enrichment of reference amino acid type. The x-axis corresponds to 20 amino acid types. Amino acid types are arranged based on their physicochemical properties (e.g., hydrophobicity, charge).

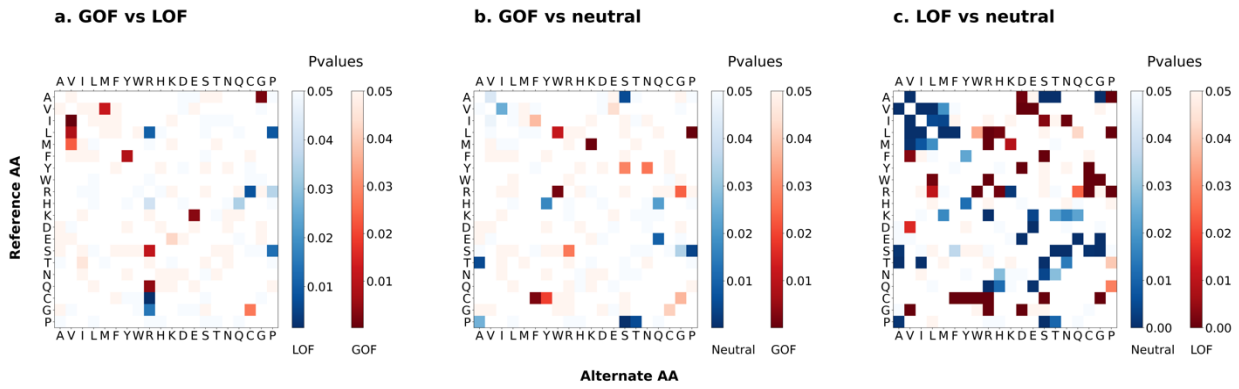

Figure 7. Enrichment of amino acid substitutions of missense mutations. **a)** Enrichment of amino acid substitutions between missense GOF and LOF mutations. Red color represents the substitutions more frequently seen in missense GOF mutations and blue color represents the substitutions more frequently seen in missense LOF mutations. **b)** Enrichment of amino acid substitutions between missense GOF mutations and likely neutral missense variants. **c)** Enrichment of amino acid substitutions between missense LOF mutations and likely neutral missense variants. Darker shade indicates more significant difference (i.e., lower P-values). The y-axis corresponds to 20 amino acid types in the reference protein sequence (i.e., substituted amino acid), and x-axis corresponds to 20 amino acid types in the alternate protein sequence (i.e., substituting amino acid). The p-values were calculated based on Fisher's exact test.