Capturing variation impact on molecular interactions: the IMEx Consortium mutations data set

Authors

The IMEx Consortium Curators†, del Toro N[1], Duesbury M[1], Koch M[1,2], Perfetto L[1], Shrivastava A[1], Ochoa D[1], Wagih O[1,3], Piñero J[4], Kotlyar M[5], Pastrello C[5], Beltrao P[1], Furlong LI[4], Jurisica I[5], Hermjakob H[1], Orchard S[1], Porras P[1*].

Abstract

The current wealth of genomic variation data identified at the nucleotide level has provided us with the challenge of understanding by which mechanisms amino acid variation affects cellular processes. These effects may manifest as distinct phenotypic differences between individuals or result in the development of disease. Physical interactions between molecules are the linking steps underlying most, if not all, cellular processes. Understanding the effects that amino acid variation of a molecule's sequence has on its molecular interactions is a key step towards connecting a full mechanistic characterization of nonsynonymous variation to cellular phenotype. Here we present an open access resource created by IMEx database curators over 14 years, featuring 28,000 annotations fully describing the effect of individual point sequence changes on physical protein interactions. We describe how this resource was built, the formats in which the data content is provided and offer a descriptive analysis of the data set. The data set is publicly available through the IntAct website at www.ebi.ac.uk/intact/resources/datasets#mutationDs and is being enhanced with every 4-weekly release.

Main

Cells process information and respond to their environments through dynamic networks of molecular interactions, where the nodes are bio-molecules (e.g. proteins, genes, metabolites, miRNAs) and edges represent functional relationships, including physical protein-protein interactions, transcriptional regulation, genetic interactions and gene/protein modifications. Comprehensive and systematic characterization of these networks is essential to gain a full understanding of complex biological

processes, of how cells behave in response to specific cues, and of how individual components of the network contribute to the whole phenotype, in physiological, pathological or synthetic conditions.

Interactions between molecules may be inherently stable and essentially irreversible, resulting in the formation of stable macromolecular complexes, or weak transient interactions characterized by a dissociation constant (KD) in the micromolar range and a lifetime of seconds. A change to a single amino acid in a protein chain can be enough to disrupt a protein binding site and may then alter the composition of a sub-network of transient binders or the formation of a protein complex. A variant leading to the inactivation of a protein kinase molecule may result in widespread disruption of post-translation phosphorylation events and the rewiring of related signalling networks. Many diseases are caused by specific mutations, and prognosis or response to treatment is frequently mutation-specific. The study of how mutations affect molecular interactions is thus of extreme interest since it can help ascertain the role of specific protein residues on the universal function of molecular binding. Several studies[1–4] have explored the impact of disease-related variation in molecular interaction networks, using structural studies and computational predictions to attempt to both identify variation-affected interfaces and predict the effect of specific variants on interactions. These studies suggest that interaction interfaces contain a significantly higher rate of disease-related variants than the rest of the molecule and that variant location in these interfaces can determine disease specificity.

Despite available high-throughput interaction screening platforms, the experimental validation of these variation effect predictions on a systems-scale remains a major challenge. However, these data can be found, reported in the literature but difficult to search and concatenate. Researchers have for many years been examining the effect of single, or multiple, induced point mutations on both binary and n-ary interactions in small-scale experiments. Targeted changes to the amino acid sequence of a protein have been engineered, largely by site-directed mutagenesis, with the aim of mimicking known variants[5,6], removing known, or predicted, post-translational modifications[7,8], disrupting regions required for protein stability or altering the properties of protein binding domains[9,10], and their effects of the interaction of interest monitored. It has been the work of the IMEx Consortium[11] to capture such information into a single data

set and thus make it available for researchers to re-use and reanalyse. IMEx Consortium annotators follow a detailed curation model, capturing not only full details of the experiment (including interaction detection method, participant identification method and the host organism) but also a description of the constructs used. This may include the co-ordinates of deletion mutants used to derive a minimum binding domain and also the effect of point mutations. Databases in the Consortium perform detailed, archival curation of published literature and also receive pre-publication data through direct submissions. This close collaboration with data producers often entails access to unpublished details in the data, such as experiments reporting mutations that have no effect on interactions, which enables the capture of added value for the scientific community.

Here we describe the largest literature-derived data set, to our knowledge, capturing the effect of sequence changes over interaction outcome. We discuss how the data set was generated and how it is maintained by the EMBL-EBI IntAct team. We also provide an initial analysis of the data set, highlighting its overlap with genomic variation data, discussing possible biases and exploring its potential as a benchmarking tool for variant effect prediction tools.

The IMEx mutations data set: data curation and quality control

The IMEx Consortium databases have been collecting point mutation data for over 14 years, which has resulted in a sizeable data set of almost 28,000 fully annotated events (www.ebi.ac.uk/intact/resources/datasets#mutationDs). The IMEx resources curate interaction data into structured database fields, and from there into community standard interchange formats, and each observation is described using controlled vocabulary terms. Mutations are mapped to the underlying protein sequence in UniProtKB and updated in line with changes to that sequence, to ensure that they stay mapped to the correct amino acid residue with every proteome release.

In order to make the mutant data set more accessible to the biomedical scientist, the Consortium has released the mutation data set in a tab-delimited format (Box 1), which includes details of the position and the amino acid change of the mutation, the

molecules in the interaction and the effect of the mutation on the interaction, as well as additional fields containing contextual information.

Additionally, a data-update pipeline has been specifically developed to ensure the accuracy of the annotation of mutation events as interaction participant features (suppl. figure 1). The construction of this pipeline has been made possible by the creation of specific fields capturing sequence changes in our recently developed standard format PSI-MI XML3.0[12]. It is run in coordination with the IntAct database monthly protein update procedure, which ensures synchronization with UniProtKB[13] and automatically shifts feature positions if there are changes in referenced protein sequences. The pipeline has been applied to the entire data in the IntAct database (www.ebi.ac.uk/intact), in which all IMEx data, and also legacy data generated by the IntAct, MINT, DIP and UniProt curation groups is housed (see Supplementary Methods for details on re-annotation and data update procedures). The mutation data update pipeline will continue to be run in quality control mode with every release of IntAct to ensure the mutation data set is kept entirely up to date with UniProtKB.

Data set statistics

The full IMEx mutations data set contains 27,868 fully annotated events in which a sequence change has been experimentally tested in an interaction experiment. All this information has been manually curated, representing over 33,000 person-hours' worth of biocurators' work, and it is continuously growing with on-going IMEx curation activities. The 4,353 proteins annotated come from 297 different species, with over 60% of the events annotated in human proteins and roughly 90% annotated in seven main model organisms (see table I).

In total, 13,926 interaction evidences are annotated with differentially reported effects, using the PSI-MI controlled vocabulary. Most of the effects reported are of a 'deleterious' nature, either disrupting (10,976 annotations, 39.3%) or decreasing the interaction (8,553 annotations, 30.7%), but there is a significant number of interactions that are either strengthened (2,256 annotations, 8.1%) or caused (188 annotations, 0.7%) by the mutation when compared with the wild type sequence (figure 1a). The data set also includes those mutations that were experimentally tested but found to

have no effect over the interaction (3,057 annotations, 11%) and 'undefined' mutations that were present in constructs used in the experiment but where the comparison with the "wild type" reference is either absent or not possible (2,838 annotations, 10.2%). It is important to note that the 'causing' and 'no effect' mutation effect categories have been only recently adopted into the controlled vocabulary and captured by the biocurators, so they have a much lower number of annotations and are not directly comparable with the other categories.

Protein-protein interaction (PPI) experiments reporting this type of data have been steadily increasing in the last 20 years, with over 4,100 publications containing data pertaining to mutated proteins sequences curated by the IMEX Consortium. However, the fraction of PPIs in which a mutated version of a protein has been reported remains relatively low (figure 1b). The majority of the interactions where a mutated protein was involved were detected using either affinity chromatography-related methods (such as co-immunoprecipitations or pull-downs) or by complementation assays based on transcriptional reporters, mainly variations of the yeast two hybrid method (see figure 1c). Most of our data set comes from the curation of small-scale papers each reporting only a few mutations (figure 1d). 99% of the publications (4,173) contain less than 100 mutation annotations and represent 80% of the annotations (22,218). Only 8 publications contain over 100 annotations, with one of them describing over 4,000 events, a study in which the authors systematically tested large numbers of variants and their effect on interactions[5]. Recording large-scale data sets such as this one has been enabled by the development of the flexible PSI-MI XML 3.0 format cited above.

Currently, the only resources that represent the impact of amino acid substitutions on binding events are the SKEMPI database[14], UniProtKB and IMEx Consortium member databases through IntAct (see table II for a detailed comparison). Of these resources, IMEx is the biggest and the only one that can provide easily accessible, systematically described, up-to-date annotations. UniProtKB mutagenesis annotations record whether a change in sequence affects an interaction, but the experimental context is not captured and the effects are described in a semi free-text field that is difficult to parse. SKEMPI offers a detailed overview of sequence change effects on binding derived from in vitro experiments, recording changes in affinity and other kinetic

parameters. Only very specific interaction detection methods, using purified proteins, are considered, which limits its scope.

The IMEx Consortium is currently formed by 11 groups, each one with their own area of interest, that have agreed to use the same curation standards and data representation download formats. All members of the consortium[15–22] use the curation platform provided by the IntAct team at EMBL-EBI. Figure 1e shows the number of events annotated by each data resource. Large databases such as IntAct, DIP and MINT, with an exclusive focus on interaction data curation, have produced the majority of the annotations, but a sizeable part of the data set has been entered by other, domain-specific, members of the Consortium.

According to the IMEx schema and curation policy, interaction evidence, rather than interacting pairs of molecules, is the focus of the data representation. This results in the curation of multiple distinct pieces of evidence describing the same interacting pairs and offers a way to weight how well characterized is a given interacting group of molecules. It also enables us to capture separate experiments where different sequence variants are tested for their effect on an interaction. Most of the proteins in the data set have a low number of associated mutations, with most proteins having 5 or less sequence changes (suppl. figure 2a) and less than 15 annotations (suppl. figure 2b). There is a greater depth of information available for human proteins, since the relative amount of human data vs other species increases with the number of annotations per protein.

The IMEx evidence-centric curation model also makes it possible to check whether the same mutation has been tested on identical interacting molecules using different interaction detection methodologies (or by different research groups) and whether the outcome of the mutations has been consistent in all these experiments. In figure 1f we show that the majority of the mutations have only been annotated once (tested in one experiment only). In those cases where there have been multiple instances of evidence testing, the results appear to be highly consistent, with only a small number of cases identified for which conflicting results have been reported. For 7,212 cases where the effect of a mutation on an interface was tested 2 or more times, only 131 (1.8%) show different effects, and only 61 cases (0.8%) reported antagonistic effects.

One reason for these contradictory results may be differences in experimental methodologies used to measure the effect, since IMEx databases recognize a large variety of experimental approaches that provide molecular interaction evidence.

The vast majority of the data set refers to amino acid substitutions, with a marginal amount of insertions and deletions reported (only 65 deletion and 83 insertion annotations). Figure 2a shows that arginine, leucine and serine are the most frequently replaced residues, while histidine and methionine residues are mutated less often (see suppl. figure 3a for a more detailed view on specific replacements). Alanine is by far the most frequently used residue for replacement (figure 2b), which is probably reflective of the widespread use of alanine scanning[23] to identify residues critical for binding to other molecules, either because they are found on the interacting interface or at an allosteric binding site. When we checked the relative proportion of the different mutation effects per replacing residue (figure 2c, suppl. figure 3b), alanine replacements mostly associate with deleterious effect on interactions. The dominance of deleterious effects most probably reflects the authors of the original study using alanine scanning to locate binding-related residues.

Genomic variation and the IMEx mutations data set

In this era of deep-sequencing genomics, there is a wealth of data concerning nonsynonymous genomic variants. As discussed before, the motivation behind the design of these experiments varies, and only a fraction were specifically designed to systematically test known variants vs reference ("wild type") versions of the participant proteins[5,24]. Hence, we decided to explore how much of currently available information for natural or disease related variation can be linked to the data set. Because of the strong predominance of human data both in IMEx mutations and in variation data sets, we decided to focus on human proteins only.

We used the EMBL-EBI Proteins API[25] to access variation data both manually annotated by and mapped to UniProtKB from large-scale sequencing studies such as the 1000 Genomes[26], ExAC[27] and COSMIC[28] projects. We queried 8,820 sequence changes in 1,990 human proteins, corresponding to 16,765 IMEx mutation annotations (see table III and figure 3). 29% (4,804) of the mutation annotations (figure 3a) and

12% (1,073) of the sequence changes (figure 3b) were fully mapped to natural variants. We also checked cases in which there is a variant described in the same position as a mutation reported in the IMEx data set, but the amino acid change is different in the two datasets (positional matches), and also those where mutations span more than one residue and only some of the residue changes or positions are matched in UniProtKB (partial matches). 16% (2,671) of the mutation annotations (figure 3a) and 16% (1,415) of the sequence changes (figure 3b) are positional or partial matches. The biological significance of positional and partial mappings does not go beyond stating that the region or position in question is important for interaction and is variable. However, we believe this information might be useful for researchers interested in exploring specific regions in more detail.

We also checked how many of the mapped variant annotations have been linked to disease according to UniProtKB. Disease associations were complemented with data from the DisGeNET database[29]. There were disease-associated variants for 42% (840) of the proteins queried, with a median value of 4 disease variants mapped per protein. As seen in Table III, 20% (3,432) of IMEx mutation annotations have been tagged as related to disease, with over 900 known disease variants represented in the data set. UniProtKB derives disease annotations for variants from both manual curation[30] and imports of cross-referenced data from ClinVar[31] via Ensembl[32], while DisGeNET also includes variants from the GWAS Catalog[33], and from text-mining the scientific literature. Figure 3c provides an overview of the diseases with most mutations annotated as mapped by full match to disease-associated variants in UniProtKB.

We then checked if the proportion of disease-related annotations in IMEx varies depending on the reported effect on interaction. As seen in figures 3d-e, disease-related mutations tend to have mostly ‘deleterious’ effects on interaction outcome, but we could also map a considerable number of annotations where there was an increase or even gain of function in terms of binding (411 annotations representing 116 variants). When we look at mutation recurrence in different types of cancer as extracted from cBioPortal[34], mutations strengthening interaction seem to have both statistically higher recurrence values and a higher proportion of mutations with extremely high recurrence in cancer data sets (figures 3f and 3g).

Variant effect annotation: computational predictions and literature curation

There is currently a variety of computational tools used to annotate variation data sets[35]. These tools can report the effect of variation on protein function, folding or binding, usually based almost exclusively on sequence or structural data, or can also report genome-derived parameters such as allele frequencies or conservation scores. We wanted to study how variation annotations provided by these tools align with experimental effect over interaction as reported in the literature.

For this purpose, we used mutfunc (www.mutfunc.com)[36], a database reporting the effect of almost any possible mutation on protein stability, interaction interfaces, post-translational modifications, protein translation, conserved regions, and regulatory regions. It hosts pre-computed variation effect data derived from established resources such as SIFT[37], Interactome3D[38] or FoldX[39].

We first examined the predicted destabilization effect of mutations on structural models of protein-protein interfaces, dividing them by the literature-reported effect. As can be seen in figure 4a, mutations with a ‘decreasing’ and especially a ‘disrupting’ effect over interactions had a significantly higher predicted destabilization effect than those with no effect, a difference that was not seen in mutations that would strengthen or even cause an interaction. These “deleterious” groups also contained a significantly higher proportion of mutations predicted to be very destabilizing for interfaces (figure 4b).

We next studied genome-derived parameters that are useful to study variation, such as residue conservation or natural allele frequencies. The experimentally-observed impact on binding stability that we report in our data set may also be reflected on these parameters. This assumption was partially confirmed using two independent measurements. First, we used the ‘sorting intolerant from tolerant’ (SIFT) method[37] , observing that the proportion of variants with low tolerance scores was significantly higher in all groups where an effect was reported vs the ‘no effect’ reference (figure 4c). We also checked allele frequencies as derived from ExAC data. Again, mutations with a reported effect seemed to have significantly lower allele frequencies (figure 4d)

and a higher proportion of alleles with extremely low frequencies (figure 5e) than those reported to have no effect over interaction.

The interaction-perturbing effects reported in the IMEx data set can be caused by modifying overall protein structure or by alteration of binding interfaces. We can determine if the mutations reported fall within sequence regions associated with binding using both computational predictions and literature-reported experimental data. We obtained predicted interfaces, based on available structural data, from Interactome3D[38]. Literature-curated interfaces were inferred from IMEx records that contained participant features of the 'binding-associated region' (MI: 0117) branch. These represent experiments where the authors have tested fragment constructs in an attempt to find sequence regions that are critical for binding, although they may not necessarily represent the actual binding surface. As seen in figures 4f-g, most of the mutations fall within predicted or curated interfaces. The proportion of mutations having an effect over the interaction seems to be higher in binding interfaces, both predicted and inferred from IMEx curation. Disease-associated variants seem to show the same pattern (suppl. figure 4a-b). Thus, the majority of the variants reported to have effects on protein interactions (68%) can be linked to perturbations inside binding regions, with a smaller proportion of variants (32%) potentially representing systemic or allosteric effects influencing interactions.

Literature bias in the IMEx mutations data set

IMEx databases have a wide scope when selecting publications for curation and it is reasonable to assume that the proteins in this data set are representative of the interaction data that has been explored in the literature. Socially-driven, literature bias is a well-known phenomenon previously reported for literature-curated data sets[24,41] so we decided to explore to what extent it affects the data set.

First, we checked whether the number of annotations and variants found in the dataset and the number of publications in which the affected protein is reported are correlated. As seen in figures 5a and 5b, the data set contains examples of both heavily-researched proteins with a low number of annotations and variants and vice versa. If we fit linear models between the number of annotations / variants and number of

publications in which a protein is reported we find a slight positive correlation, especially in the case of disease-related variants. This observation is compatible with socially-induced bias, with known disease-related proteins and variants being more often reported in the literature.

Then we set out to find if the proteins represented in the mutations data set are involved in distinctive pathways versus all proteins for which IMEx has interaction information. To avoid database-specific biases we performed annotation enrichment analysis using PathDIP (http://ophid.utoronto.ca/pathdip), an analysis tool that integrates information from 20 source databases[42]. Human proteins were divided in different sets depending on the effect reported for their mutations and their pathway annotation enrichment was calculated using all the human proteins in IMEx as background. Pathways obtained from these sets have substantial overlap (figure 5c, 885 pathways). These results suggest that the proteins whose mutation effect on interactions have been collected in this dataset may be biased, possibly due to specific interest of the researchers exploring variation influence on molecular interactions. Specifically, in the group of mutations that show an effect on interactions, pathways related to the immune system, signalling, disease and cell cycle control ranked on the top (suppl. figure 5, see suppl. table 2 for full details), with little difference between effect categories. There seems to be a predominance of cancer-related pathways, with representatives in both the 'disease' and the 'signaling' categories, which agrees with the observation reported in figure 5b that the literature is biased towards disease-related variants.

Discussion

Here we present a unique resource containing experimental, publicly available information about the impact of sequence changes on specific protein-protein interaction outcomes. This is a direct result of the IMEx Consortium full-detail curation policies and represents an example of how expert curation, resulting in structured and standardized representations, is required in order to make the most of published experimental results. In comparison to similar, pre-existing data sets recording variation influence over interactions, this resource represents a leap forward in depth, size and scope (table II). A previous, relatively small study[43] reported a curated list of

about 100 mutations influencing interactions. This was used as benchmark in a study investigating the link between disease-related variation and interaction interfaces[1], showing an application of this type of data, despite obvious limitations due to its size. The curation infrastructure and practices of the IMEx consortia will enable the capture of data from a growing number of deep-mutagenesis interaction studies, where hundreds if not thousands of single amino acid changes over the whole length of a protein sequence are explored for their influence on interactions[44].

We have also acknowledged the social biases inherent to any literature-based resource in our data set, although it is difficult to ascertain its extent. Alanine scanning features prominently as a commonly used technique (figure 2b) and may represent amino acid changes that will never be seen in nature due to evolutionary constraints or simply because they would require extensive sequence alteration at the DNA level, but remains an invaluable source of information, identifying key binding-related positions. For the human sub-section of the data set, disease-related variants and proteins are possibly over-represented (figures 3b and 3c, supplementary figure 5) and have been preferentially been selected for biocuration over non-disease related proteins (supplementary figure 2a). Interestingly, we report over 100 disease-related variants described in the literature to either cause or increase existing interactions (figures 3d and 3e), some of which are found to be highly recurrent in cancer according to cBioPortal[45]. This contrasts with the findings reported by Sahni *et al.*[5], where only two cases of gain-of-function mutations were found in a systematic screening for disease-related mutations and their effect on interactions using yeast two-hybrid technology. Although interaction decreasing/disrupting effects were much more frequently reported, this highlights how gain-of-interaction mechanisms could play a significant role in disease pathogenesis, especially in cancer.

Analysis of variation is a fundamental tool in basic and clinical research, with direct application in the clinic through translational genomics. Variation effects are explored mainly through statistical analysis of large population datasets, GWAS studies, or by quantitative analysis of its influence on expression via identification of eQTLs. However, in order to unravel the mechanisms behind detected effects, it is key to explore how molecular interactions are affected[46]. Currently, most of the mechanistic insight into variation effects is generated by computational annotation and predictions,

using tools that are based on relatively small reference sets, generally based on structural data. As an example, the widely used FoldX algorithm is generated from protein complex structures and has been tested against a library of 1,008 mutants[39]. Our current data set already provides interaction effects for over 10 times more individual variants and is not limited to structural data. The wide scope of experimental setups represented (figure 1c) allows the capture of effects on proteins and protein regions that might be intrinsically non-structured[47]. We show that the data set gives a currently unparalleled and representative overview about which residues are key for protein interactions, with the results being in good accordance with commonly used variant annotators (figure 4). IMEx curation practices originally did not enforce capturing sequence changes that had no effect over interaction outcome, but as a result of consultations with tool developers and data users this policy has been amended and the data set now features a growing number of mutations with no effect that can be used as a training negative set for the development of computational annotation tools.

The IMEx mutations data set represents both a reference source for direct, literature-based variant characterization and a unique benchmark that can be used to further refine computational variant effect annotators. We will continue to expand the data set and improve its accessibility for users, as a part of IMEx global mission of ensuring data representation and re-use.

Methods

Source data

All analysis was performed using the September 2017 version of the IMEx mutations data set, which can be directly downloaded from ftp://ftp.ebi.ac.uk/pub/databases/intact/2017-09-02/various/mutations.tsv.

Software and packages used

The quality control pipeline for mutation annotations was developed and integrated within the production code used in the IntAct database. The code is written in Java and makes use of the Hibernate and Spring frameworks for interaction with the core SQL database and application implementation. Specific implementation details are available upon request. Statistical analysis, plots, mutation re-annotation checks and mappings were performed using the R programming language[48] through the RStudio programming suite[49]. The following R packages were used in the study: data.table, dplyr, ggplot2, ggpubr, gridExtra, gsubfn, httr, jsonlite, plyr, RCurl, reshape2, scales, seqinr, splitstackshape, XML, Biostrings, biomaRt.

Curation practices

Data has been produced through manual literature curation following the IMEx Consortium curation guidelines[11], which can be explored in detail on the Consortium's website: http://www.imexconsortium.org/curation. Briefly, every publication reviewed was curated for the entirety of the interaction data it contained, representing each experimental piece of evidence as a separate record. Full details of constructs used were registered and every entry was reviewed by at least two independent curators for quality control.

Mutations re-annotation effort

After the development of PSI-XML3.0 and the 'resulting sequence' field in the IMEx schemas to capture amino acid change in participant features of the type 'mutation (MI:0118)' and children, it was necessary to populate the field with legacy data from the participant feature short label. This free-text, manually-entered field was prone to contain typographical errors and was difficult to keep updated. Curators used a set of

simple rules to depict amino acid substitution, deletions and insertions. As a first step towards populating the 'resulting-sequence' field, we wrote ad hoc parsing scripts to evaluate and extract the information stored in the short labels. Several rounds of corrections took place until the data set got to its current state. Of the 27,868 records of the data set, 20,161 had to be corrected, with around 2,000 of them manually corrected. There are still about 2,500 records for which no fix was possible without fully amending the original entry. These have been left out of the dataset until being revisited by an IMEx curator in due course. An automated quality control pipeline has been put in place to handle newly-created entries and future changes in UniProtKB (details in Supplementary materials). Finally, we have also adapted the participant feature short labels to the Human Genome Variation Society (HGVS) recommendations for variant annotation[50], which can be accessed at http://varnomen.hgvs.org/recommendations/protein/.

Mapping IMEx mutations to UniProtKB and the genome

UniProtKB accessions for human proteins were extracted from the IMEx mutations data set, retaining isoform identifiers, and used to query the EMBL-EBI Protein API [25]. The API's 'variation' method was used to extract large-scale variation annotation from UniProtKB, regardless of its origin. Annotations extracted through this method were then mapped to the IMEx mutations data set using UniProtKB accession, sequence position and resulting amino acid for 'full' mappings and only UniProtKB accession and position for 'positional' mappings. Cases where the IMEx-reported mutation spans more than one amino acid position were split into individual substitutions and only labelled as 'full' matches if every individual position matches an annotation in UniProtKB. Otherwise, they were considered 'partial' mappings. Disease annotations were extracted from the API's output, along with rsIDs. These rsIDs were then used in DisGeNET to search for additional disease annotations that were brought in as well.

Predicting impact on protein interaction interfaces

Experimental and homology modelled structures for protein interactions were obtained from the Interactome3D database[38]. Relative solvent accessibility (RSA) for all residue atoms was computed using NACCESS[51] for proteins individually and in the interaction complex. Interface residues were defined as those with any change in RSA. The

impact of variant on interface stability was computed using FoldX v.4.0. All binary interface structures were repaired using the RepairPDB command, with default parameters. The Pssm command is then used to predict ΔG with numberOfRuns=5. This performs the mutation multiple times with variable rotamer configurations, to ensure the algorithm has achieved convergence. The average ΔG of all runs is computed and the ΔΔG is computed as the difference between the wildtype and mutant and provides a predictive estimate of how destabilising the mutant is to the interaction interface.

Predicting the functional impact of variants using conservation

All protein alignments were built against UniRef50[52], using the seqs_chosen_via_median_info.csh script in SIFT 5.1.1[40]. The siftr R package (https://github.com/omarwagih/siftr) was used to generate SIFT scores with parameters ic_thresh=3.25 and residue_thresh=2.

Allele frequencies

A total of 3,198,692 coding variants in H. sapiens for over 65,000 individuals was collected from the ExAC Consortium[27] in the ANNOVAR[53] output format along with corresponding adjusted allele frequencies. Ensembl transcript positions were mapped to UniProt by performing Needleman-Wunsch global alignment of translated Ensembl transcript sequences against the UniProt sequence using the pairwiseAlignment function in the Biostrings R package. The mapping between Ensembl transcript IDs (v81) and UniProt accessions was obtained from the biomaRt R package. In the case that multiple alleles mapped to the sample single amino acid substitution, the one with the highest adjusted allele frequency was retained.

Recurrence

Recurrence data for 1,183,665 variants was obtained from the cBioPortal MAF file (06/11/2015) containing data from 100 cancer genomics studies.

Mapping variants to interaction interfaces

Predicted interface and accessibility coordinates were obtained from Interactome3D. Curated interfaces were extracted from IntAct by selecting participant features under

the PSI-MI term 'binding-associated region' (MI:0117). Only human proteins for which accessibilities were calculated directly from structural data in Interactome3D were selected for this analysis, modelled structures were excluded.

Estimating literature bias

We used the NCBI 'geneID2pubmed' table, accessible at ftp://ftp.ncbi.nih.gov/gene/DATA/gene2pubmed.gz, to estimate how many papers were associated to individual proteins in the IMEx mutations data set. Only human proteins were considered. Entrez GeneIDs were mapped to UniProtKB accessions using UniProt's website REST API mapping service as described at https://www.uniprot.org/help/api_idmapping.

Pathway enrichment analysis using PathDIP

Pathway enrichment was performed using mutated PPIs (i.e., mutated protein + partner) of a given mutation type (causing, disrupting, etc.) and pathDIP 2.5 pathways (considering only core pathway, http://ophid.utoronto.ca/pathDIP/[42]). We considered whole IntAct human PPIs as a background for enrichment analysis (downloaded March 24, 2018). For pathways overlap Venny 2.1.0 (http://bioinfogp.cnb.csic.es/tools/venny/) was used and Wordle (http://www.wordle.net/) was used to prepare word clouds from enriched pathway titles.

Author information

Affiliations

1. European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Genome Campus, Hinxton, CB10 1SD, UK

2. Novartis Institutes for BioMedical Research (NIBR), Basel, Canton of Basel-Stadt, Switzerland

3. Deep Genomics, MaRS Centre, 661 University Ave, Suite 480, Toronto, Ontario, M5G 1M1, Canada

4. Research Programme on Biomedical Informatics (GRIB), Department of Experimental and Health Sciences (DCEXS), Hospital del Mar Medical Research Institute (IMIM), Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain

5. Krembil Research Institute, Data Science Discovery Centre for Chronic Diseases, Krembil Discovery Tower, 5KD-407, 60 Leonard Avenue, Toronto, Ontario, M5T 0S8, Canada

† See acknowledgements for full list of contributing IMEx Consortium curators

Contributions

S.O. and P.P. designed this study and wrote the manuscript. The IMEx Consortium curators generated the mutation annotations. M.D., S.O., M. Koch, N.dT., A.S. and P.P. re-curated the data set and implemented semi-automated quality control procedures. L.P., D.O., O.W., C.P., M. Kotlyar, J.P. and P.P. analysed the data. S.O., H.H., P.B., L.I.F., I.J. and P.P. interpreted the results and revised the manuscript.

Corresponding authors

Correspondence to Pablo Porras (pporras@ebi.ac.uk).

Acknowledgements

The IMEx Consortium curators that produced the annotations used for the IMEx mutations data set are listed here, along with their IMEx database affiliation: Sara Abbani (DIP), Mais G Ammari (HPIDB), Alan Bridge (UniProt), Nancy H Campbell (BHF-UCL), Carol Chen (InnateDB), Marta Iannuccelli (MINT), Sruthi Jagannathan (MBInfo), Jyoti Khadake (IntAct), Luana Licata (MINT), Ruth C Lovering (BHF-UCL), Usha Mahadevan (Molecular Connections), Anna N Melidoni (BHF-UCL, IntAct), Simona Panni (IntAct), Arathi Raghunath (Molecular Connections), Sylvie Ricard-Blun (MatrixDB), Milagros Rodriguez-Lopez (BHF-UCL, IntAct), Bernd Roechert (UniProt), Lukasz Salwinski (DIP), David Thorneycroft (IntAct) and Kim van Roey (IntAct). Principal Investigators that supported their work are Gianni Cesareni (MINT), David Lynn (InnateDB) and Fiona M McCarthy (HPIDB).

The IntAct database and EMBL-EBI-based authors received funding from EMBL core funding and Open Targets (grant agreement OTAR-044). The DIP database is funded by NIH grant R01GM123126. MINT received support from ERC grant "DEPTH project of the European Research Council (grant agreement 322749)". UniProt curation activities at EMBL-EBI and the Swiss Institute of Bioinformatics are funded by NIH grants U41HG007822 and U24HG007822. The British Heart Foundation-University College of London (BHF-UCL) curation team is funded with the British Heart Foundation grant RG/13/5/30112. DisGeNET is supported with EU-FP7 funds from ISCIII-FEDER (CP10/00524, CPII16/00026), IMI-JU (grant agreement no. 116030, TransQST) and EFPIA companies in kind contribution, and the EU H2020 Programme 2014-2020 (grant agreements no. 634143, MedBioinformatics and no. 676559, Elixir-Excelerate). The Research Programme on Biomedical Informatics (GRIB) is a member of the Spanish National Bioinformatics Institute (INB), PRB2-ISCIII and is supported by grant PT13/0001/0023, of the PE I+D+i 2013-2016, funded by ISCIII and FEDER. The DCEXS is a “Unidad de Excelencia María de Maeztu”, funded by the MINECO (ref: MDM-2014-0370).

The authors would like to especially thank Marco Galardini, Luz García-Alonso, Denes Turei and Martin Krallinger for valuable discussions when designing the data set output format.

Bibliography

1. Wang, X. *et al.* Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* **30,** 159–164 (2012).
2. Mosca, R. *et al.* dSysMap: exploring the edgetic role of disease mutations. *Nat. Methods* **12,** 167–168 (2015).
3. Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J. & Godzik, A. A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces., A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. *PLoS Comput. Biol. PLoS Comput. Biol.* **11, 11,** e1004518–e1004518 (2015).
4. Buljan, M., Blattmann, P., Aebersold, R. & Boutros, M. Systematic characterization of pan-cancer mutation clusters. *Mol. Syst. Biol.* **14,** e7974–e7974 (2018).
5. Sahni, N. *et al.* Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders. *Cell* **161,** 647–660 (2015).
6. Chen, S. *et al.* An interactome perturbation framework prioritizes damaging missense mutations for developmental disorders. *Nat. Genet.* 1 (2018). doi:10.1038/s41588-018-0130-z
7. Burén, S. *et al.* Regulation of OGT by URI in Response to Glucose Confers c-MYC-Dependent Survival Mechanisms. *Cancer Cell* **30,** 290–307 (2016).
8. Liu, X. *et al.* Bidirectional regulation of neutrophil migration by mitogen-activated protein kinases. *Nat. Immunol.* **13,** 457–464 (2012).
9. Maio, N., Kim, K. S., Singh, A. & Rouault, T. A. A Single Adaptable Cochaperone-Scaffold Complex Delivers Nascent Iron-Sulfur Clusters to Mammalian Respiratory Chain Complexes I–III. *Cell Metab.* **25,** 945-953.e6 (2017).

10. Rebsamen, M. *et al.* SLC38A9 is a component of the lysosomal amino acid sensing machinery that controls mTORC1. *Nature* **519,** 477–481 (2015).
11. Orchard, S. *et al.* Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods* **9,** 345–350 (2012).
12. Sivade Dumousseau, M. *et al.* Encompassing new use cases - level 3.0 of the HUPO-PSI format for molecular interactions. *BMC Bioinformatics* **19,** 134 (2018).
13. UniProt Consortium, U. C. U. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (2017).
14. Moal, I. H. & Fernández-Recio, J. SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics* **28,** 2600–2607 (2012).
15. Orchard, S. *et al.* The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42,** D358-363 (2014).
16. Licata, L. *et al.* MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* **40,** D857–D861 (2012).
17. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43,** D204-212 (2015).
18. Kotlyar, M., Pastrello, C., Sheahan, N. & Jurisica, I. Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res.* **44,** D536-41 (2016).
19. Ammari, M. G., Gresham, C. R., McCarthy, F. M. & Nanduri, B. HPIDB 2.0: a curated database for host-pathogen interactions. *Database J. Biol. Databases Curation* **2016,** (2016).
20. Lynn, D. J. *et al.* Curating the innate immunity interactome. *BMC Syst. Biol.* **4,** 117 (2010).

21. Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32,** D449–D451 (2004).
22. Launay, G., Salza, R., Multedo, D., Thierry-Mieg, N. & Ricard-Blum, S. MatrixDB, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities. *Nucleic Acids Res.* **43,** D321-327 (2015).
23. Morrison, K. L. & Weiss, G. A. Combinatorial alanine-scanning. *Curr. Opin. Chem. Biol.* **5,** 302–307 (2001).
24. Rolland, T. *et al.* A Proteome-Scale Map of the Human Interactome Network. *Cell* **159,** 1212–1226 (2014).
25. Nightingale, A. *et al.* The Proteins API: accessing key integrated protein and genome information. *Nucleic Acids Res.* doi:10.1093/nar/gkx237
26. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
27. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
28. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43,** D805-811 (2015).
29. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45,** D833–D839 (2017).
30. Famiglietti, M. L. *et al.* Genetic variations and diseases in UniProtKB/Swiss-Prot: the ins and outs of expert manual curation. *Hum. Mutat.* **35,** 927–935 (2014).
31. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42,** D980-985 (2014).
32. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43,** D662-669 (2015).

33. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45,** D896–D901 (2017).

34. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6,** pl1 (2013).

35. Verma, R., Schwaneberg, U. & Roccatano, D. Computer-Aided Protein Directed Evolution: a Review of Web Servers, Databases and other Computational Tools for Protein Engineering. *Comput. Struct. Biotechnol. J.* **2,** e201209008–e201209008 (2012).

36. Wagih, O. *et al.* Comprehensive variant effect predictions of single nucleotide variants in model organisms. *bioRxiv* 313031 (2018). doi:10.1101/313031

37. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–1081 (2009).

38. Mosca, R., Céol, A. & Aloy, P. Interactome3D: adding structural details to protein networks. *Nat. Methods* **10,** 47–53 (2013).

39. Van Durme, J. *et al.* A graphical interface for the FoldX forcefield. *Bioinforma. Oxf. Engl.* **27,** 1711–1712 (2011).

40. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11,** 1–9 (2016).

41. Schaefer, M. H., Serrano, L. & Andrade-Navarro, M. A. Correcting for the study bias associated with protein-protein interaction measurements reveals differences between protein degree distributions from different cancer types., Correcting for the study bias associated with protein–protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Front. Genet. Front. Genet.* **6, 6,** 260–260 (2015).

42. Rahmati, S., Abovsky, M., Pastrello, C. & Jurisica, I. pathDIP: an annotated resource for known and predicted human gene-pathway associations and pathway enrichment analysis. *Nucleic Acids Res.* **45,** D419–D426 (2017).
43. Schuster-Böckler, B. & Bateman, A. Protein interactions in human genetic diseases., Protein interactions in human genetic diseases. *Genome Biol. Genome Biol.* **9, 9,** R9, R9–R9 (2008).
44. Woodsmith, J. *et al.* Protein interaction perturbation profiling at amino-acid resolution. *Nat. Methods* **14,** 1213–1221 (2017).
45. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov.* **2,** 401–404 (2012).
46. Sahni, N. *et al.* Edgotype: a fundamental link between genotype and phenotype. *Curr. Opin. Genet. Dev.* **23,** 649–657 (2013).
47. Babu, M. M. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem. Soc. Trans.* **44,** 1185–1200 (2016).
48. Ihaka, R. & Gentleman, R. R: A Language for Data Analysis and Graphics. *J. Comput. Graph. Stat.* **5,** 299–314 (1996).
49. RStudio Team. RStudio: Integrated Development for R. (2015).
50. Dunnen, J. T. den *et al.* HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum. Mutat.* **37,** 564–569 (2016).
51. Hubbard, S. & Thornton, J. *NACCESS*. (Department of Biochemistry and Molecular Biology, University College London, 1993).

52. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31,** 926–932 (2015).

53. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164–e164 (2010).

Figure legends

Figure 1. IMEx mutations data set overview

a: Number of annotations by effect type; b: Increase of reported protein interactions involving wild type and mutated proteins over time; c: Distribution of the number of mutation annotations by interaction detection method; d: Distribution of the number of mutation annotations captured per publication. The number of annotations per publication is shown on a log scale; e: Number of mutation annotations per database of origin; f: Internal consistency of repeatedly reported mutations. 'conflict' cases are those in which the effects reported are antagonistic (e.g. 'disrupting' vs 'increasing'). 'mild conflict' cases are those in which the mutation is sometimes reported as having some effect vs others in which there is no detectable effect.

Figure 2. Amino acid replacement frequencies in the full data set

a: Replacement frequencies by original residue; b: Replacement frequencies by resulting residue; c: Normalized frequencies of resulting sequences by mutation effect over the interaction. Substitutions with non-standard amino acids and deletions are not shown for simplicity.

Figure 3. Genomic variation and disease annotations in the IMEx mutations data set

a: Mapping IMEx mutation annotations to UniProtKB human variants; b: Mapping UniProtKB human variants to IMEx reported sequence changes; c: Top 15 most represented diseases in fully mapped variants according to UniProtKB disease associations; d: IMEx mutation annotations by effect type and their relation to disease; e: IMEx mutation sequence changes by effect type and their relation to disease; f: cBioPortal recurrence scores for mutations grouped by effect type. P-values calculated with one-sided Wilcoxon test are indicated; g: Proportion of highly-recurrent cancer variants according to cBioPortal by effect type. p-values calculated with Fisher exact test are indicated.

Figure 4. Computational annotations and the IMEx mutations data set

a: Interaction interface disruption as predicted with FoldX, by mutation effect type; b: Proportion of highly disruptive variants by mutation effect type; c: Proportion of low

tolerance residue positions according to the SIFT, by mutation effect type; d: ExAC-extracted allele frequencies for mutations represented in the IMEx data set, by mutation effect type; e: Low frequency variants, by mutation effect type; f: Number of mutation annotations located in binding interfaces (curated and predicted), by effect; g: Normalized frequencies of mutation annotations reporting effects over interactions or not and their localization in binding interfaces. p-values from figures a and d calculated with Wilcoxon test. p-values indicated in figures b, c and e were calculated with Fisher exact test.

Figure 5. Literature biases in IMEx mutations data set

a and b: Scatter plot of number of publications (in a logarithmic scale) in which a protein is reported vs a: the number of annotations and b: the number of variants reported in the IMEx mutations data set; c: Overlap of significantly enriched pathways ($q<0.01$) across different sets of proteins and word enrichment analysis (using Wordle on enriched pathway names) for the overlapping set (any mutational effect), the set of proteins annotated with no effect and the remaining proteins in IMEx (non-mutated). In “no effect” word enrichment analysis, the words “pathway” and “action” have been removed to make remaining words more visible (original wordle available as supplementary figure 6a), while in “common mutated” wordle the words “pathway” and “signalling” have been removed (original wordle available as supplementary figure 6b). The analysis in this figure was performed taking into account human proteins only.

Box 1: Overview of the IMEx mutations data set downloadable flat file.

Figure 1

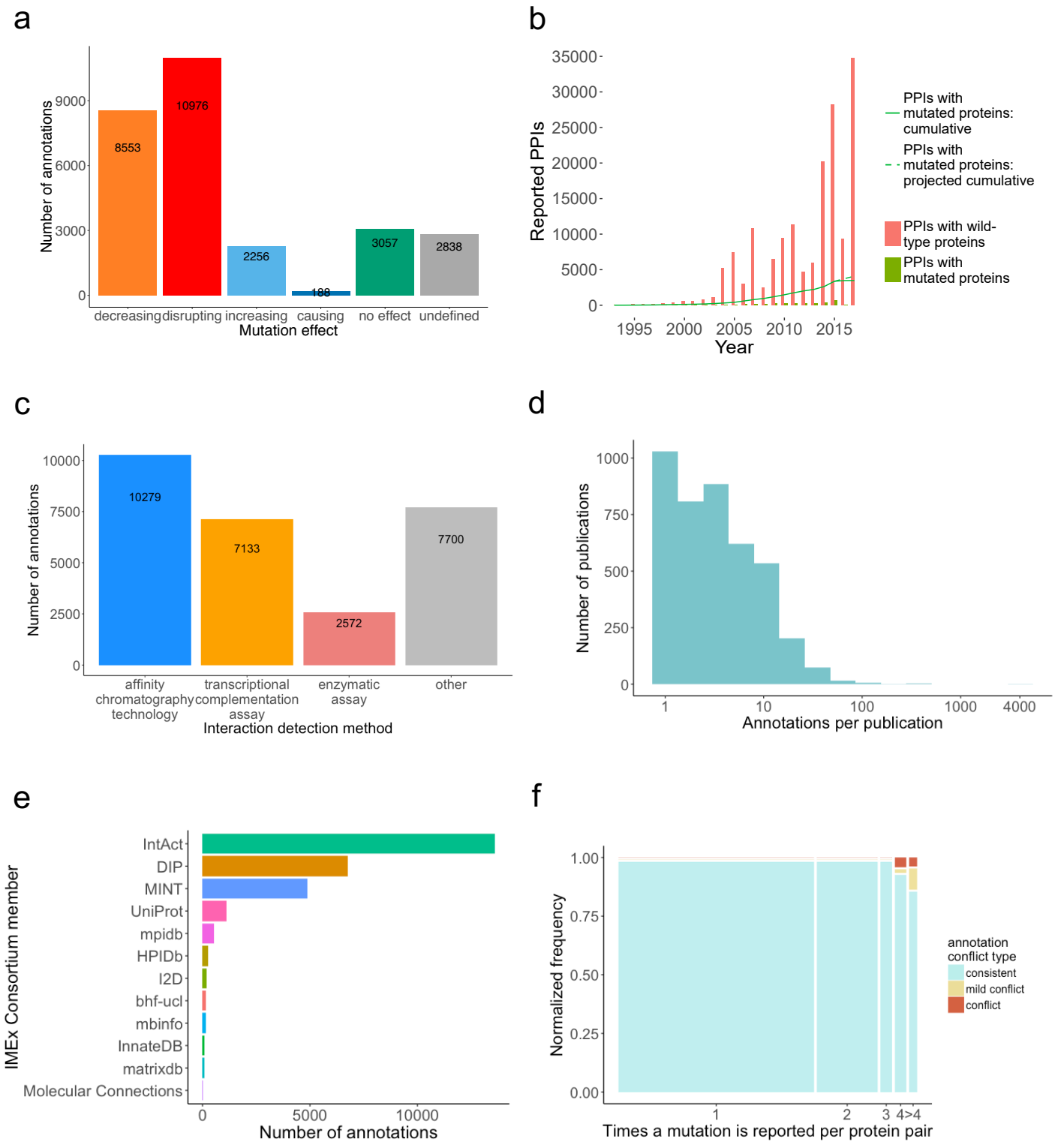


Figure 2

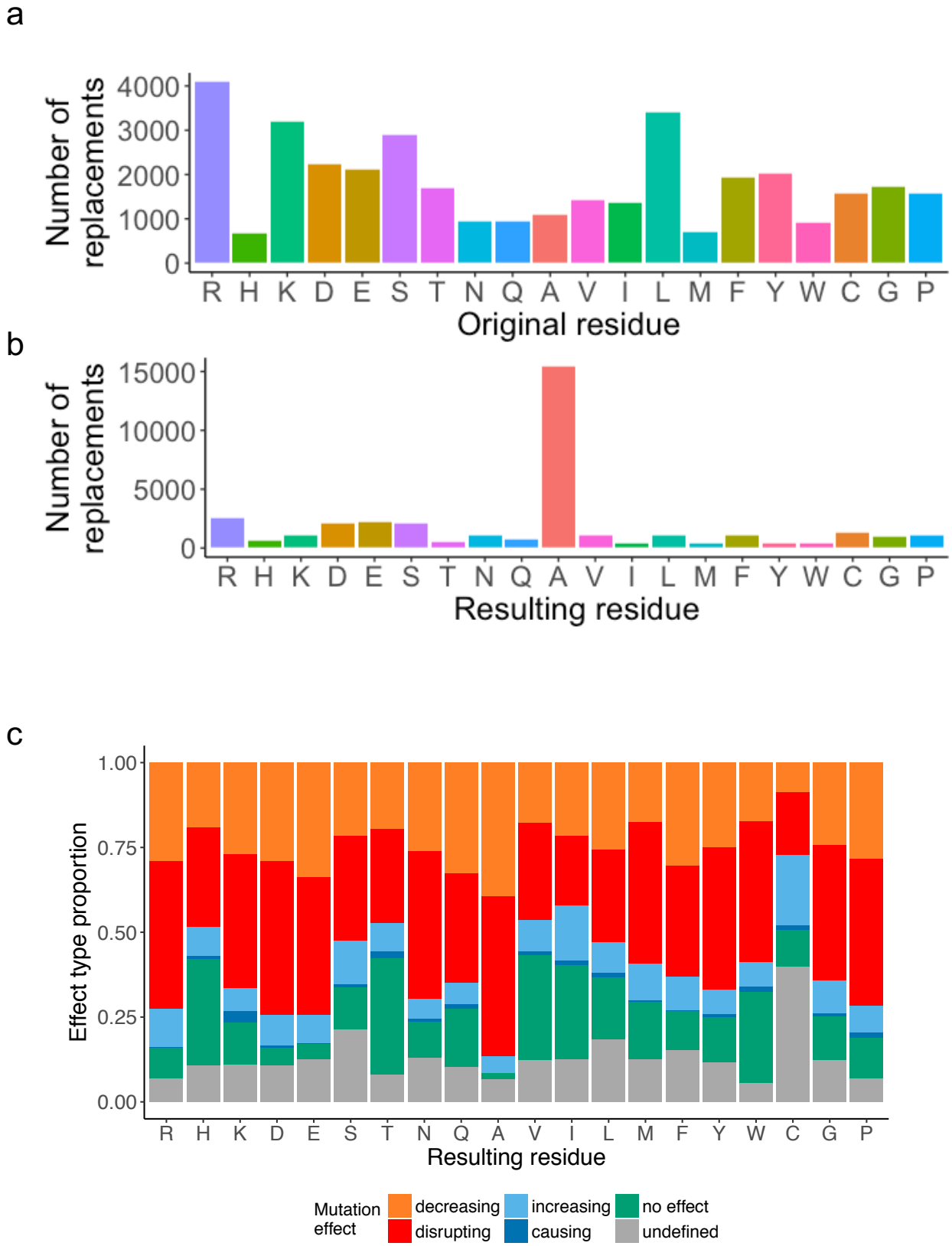


Figure 3

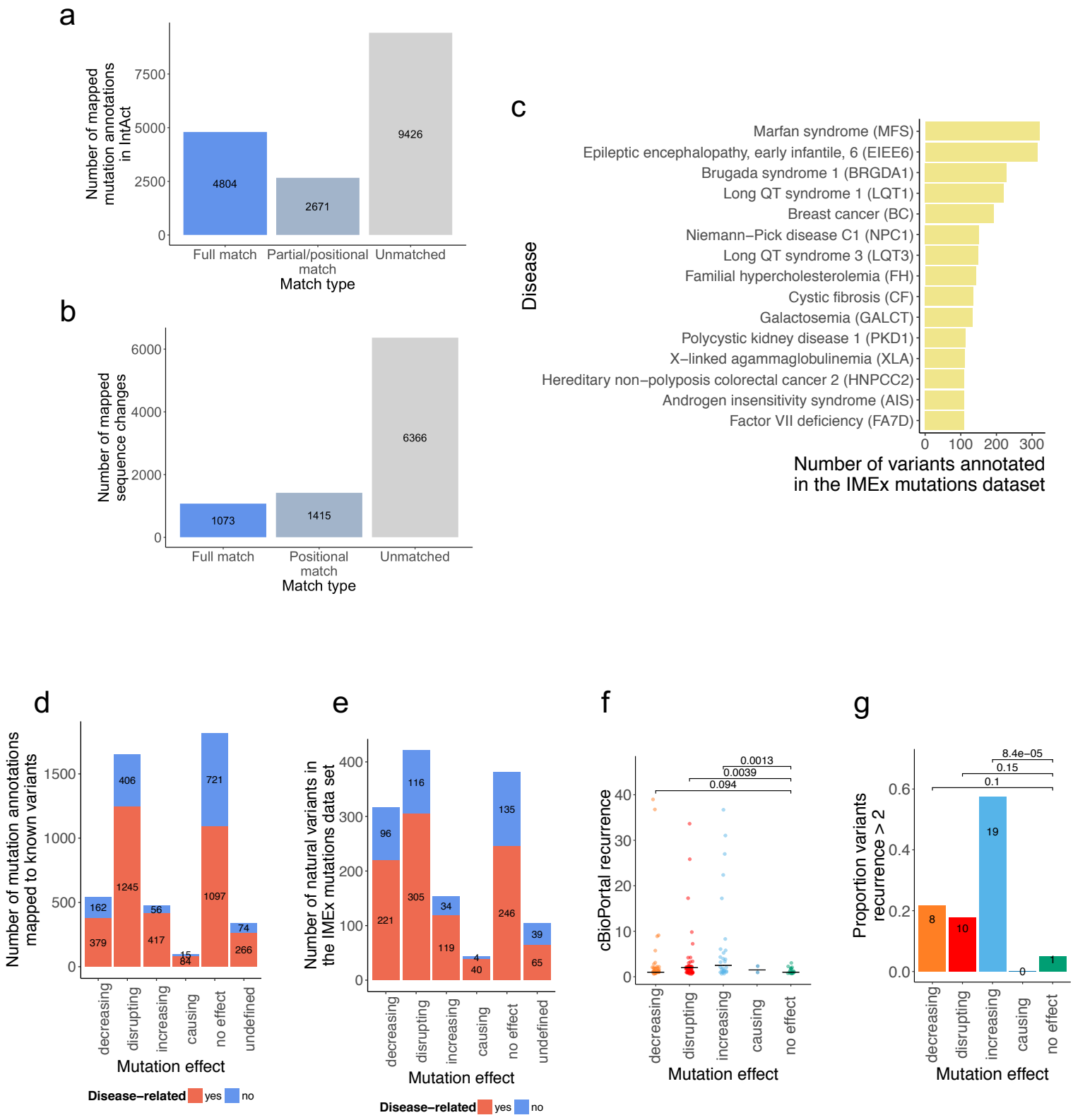


Figure 4

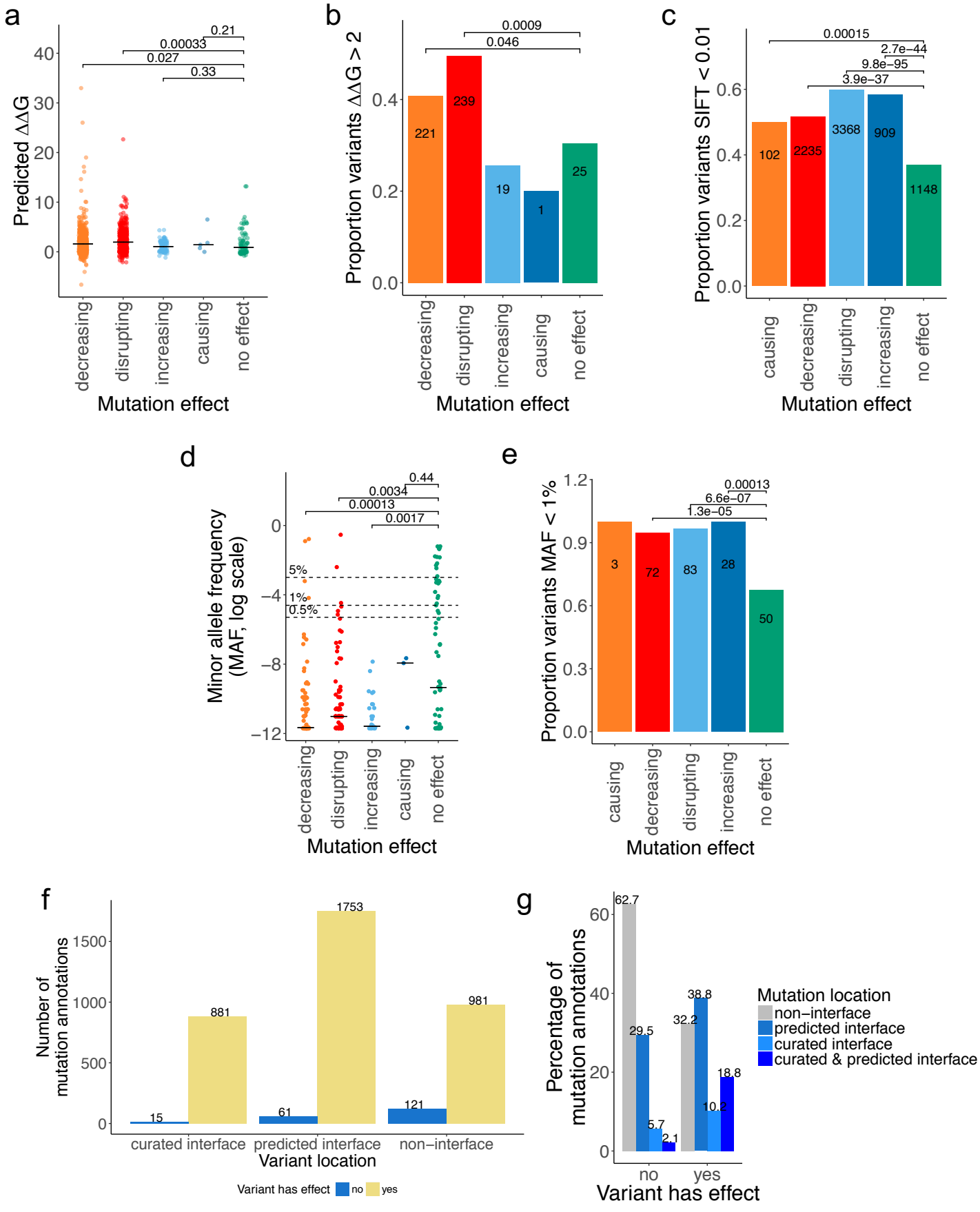
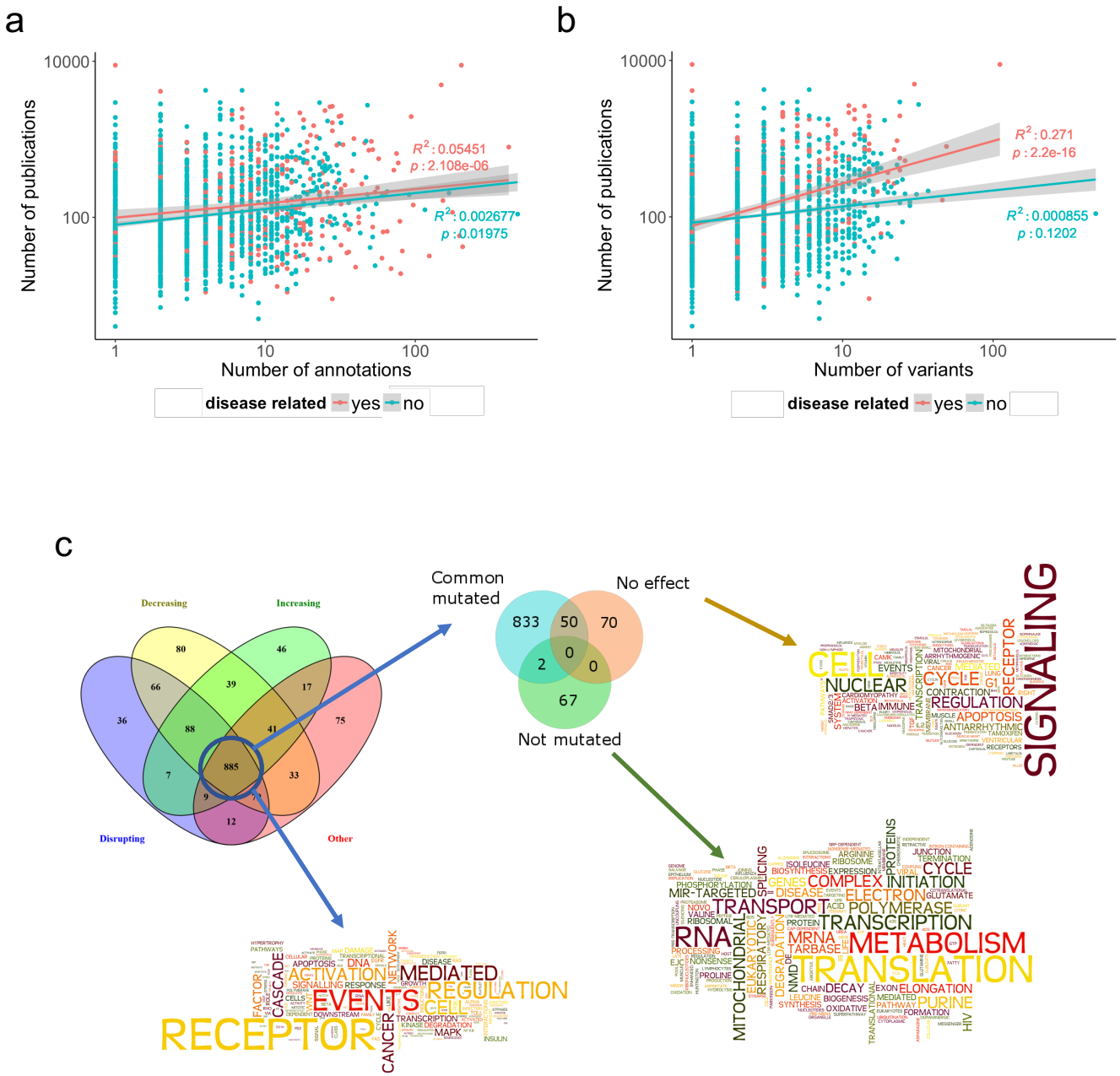
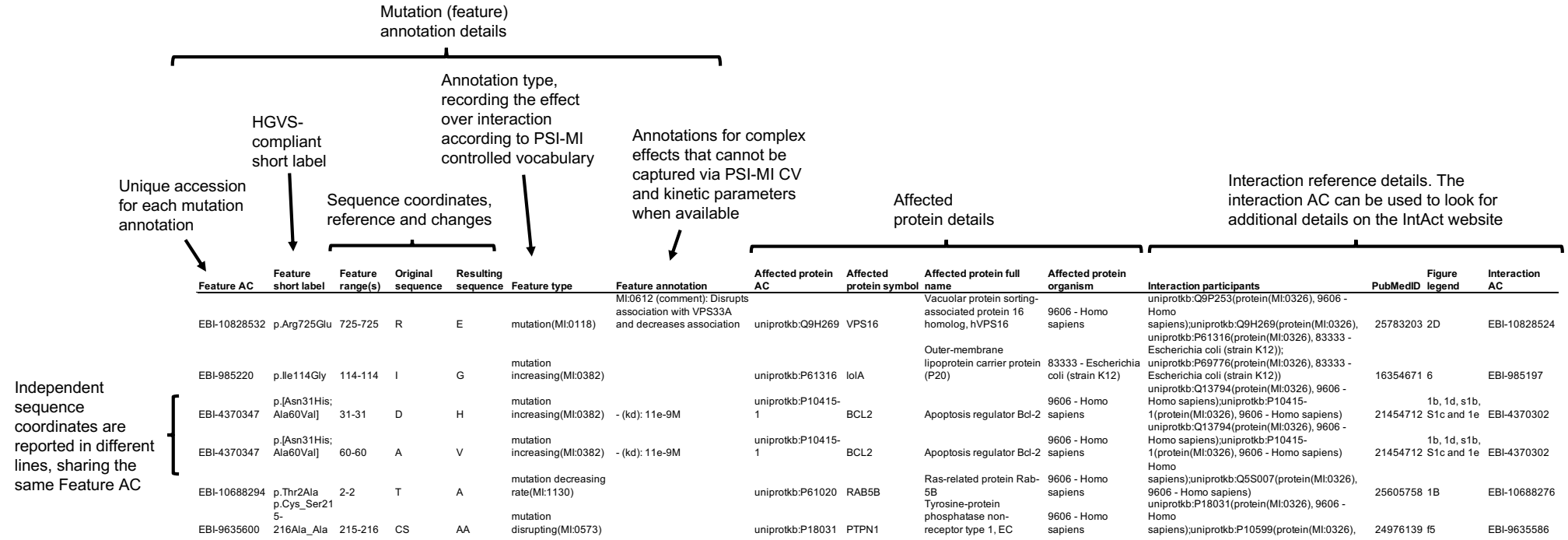


Figure 5



Box 1



Feature AC	Feature short label	Feature range(s)	Original sequence	Resulting sequence	Feature type	Feature annotation	Affected protein AC	Affected protein symbol	Affected protein full name	Affected protein organism	Interaction participants	PubMedID	Figure legend	Interaction AC
EBI-10828532	p.Arg725Glu	725-725	R	E	mutation(MI:0118)	MI:0612 (comment): Disrupts association with VPS33A and decreases association	uniprotkb:Q9H269	VPS16	Vacuolar protein sorting-associated protein 16 homolog, hVPS16	9606 - Homo sapiens	uniprotkb:Q9P253(protein(MI:0326), 9606 - Homo sapiens);uniprotkb:Q9H269(protein(MI:0326),	25783203	2D	EBI-10828524
EBI-985220	p.Ile114Gly	114-114	I	G	mutation increasing(MI:0382)		uniprotkb:P61316	IolA	Outer-membrane lipoprotein carrier protein (P20)	83333 - Escherichia coli (strain K12)	uniprotkb:P61316(protein(MI:0326), 83333 - Escherichia coli (strain K12)); uniprotkb:P69776(protein(MI:0326), 83333 - Escherichia coli (strain K12))	16354671	6	EBI-985197
EBI-4370347	p.[Asn31His; Ala60Val]	31-31	D	H	mutation increasing(MI:0382)	- (kd): 11e-9M	uniprotkb:P10415-1	BCL2	Apoptosis regulator Bcl-2	9606 - Homo sapiens	uniprotkb:Q13794(protein(MI:0326), 9606 - Homo sapiens);uniprotkb:P10415-1(protein(MI:0326), 9606 - Homo sapiens)	21454712	1b, 1d, s1b, S1c and 1e	EBI-4370302
EBI-4370347	p.[Asn31His; Ala60Val]	60-60	A	V	mutation increasing(MI:0382)	- (kd): 11e-9M	uniprotkb:P10415-1	BCL2	Apoptosis regulator Bcl-2	9606 - Homo sapiens	uniprotkb:Q13794(protein(MI:0326), 9606 - Homo sapiens);uniprotkb:P10415-1(protein(MI:0326), 9606 - Homo sapiens)	21454712	1b, 1d, s1b, S1c and 1e	EBI-4370302
EBI-10688294	p.Thr2Ala	2-2	T	A	mutation decreasing rate(MI:1130)		uniprotkb:P61020	RAB5B	Ras-related protein Rab-5B	9606 - Homo sapiens	Homo sapiens);uniprotkb:Q5S007(protein(MI:0326), 9606 - Homo sapiens)	25605758	1B	EBI-10688276
EBI-9635600	p.Cys_Ser215-216Ala_Ala	215-216	CS	AA	mutation disrupting(MI:0573)		uniprotkb:P18031	PTPN1	Tyrosine-protein phosphatase non-receptor type 1, EC	9606 - Homo sapiens	uniprotkb:P18031(protein(MI:0326), 9606 - Homo sapiens);uniprotkb:P10599(protein(MI:0326),	24976139	f5	EBI-9635586