bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Missense variants in health and disease target distinct functional pathways and proteomics features

Anna Laddach[1], Joseph Chi-Fung Ng[1], and Franca Fraternali[1,*]

[1] *Randall Centre for Cell and Molecular Biophysics, King's College London, UK*
[*] *Corresponding author: franca.fraternali@kcl.ac.uk*

1 Abstract

Genetic variants are associated with a number of human diseases, but they also occur in healthy individuals, contributing to inter-person and ethnic differences. A subset of DNA variants alter the sequences of the proteins encoded, but differences in the nature of protein variants in health and disease, and the cellular processes they may affect are not fully understood. Because of this, distinguishing missense variants associated with "healthy" and "diseased" states remains a challenge.

To understand the molecular principles which underlie these differences, we quantify variant enrichment at multiple levels, from 3D structure defined regions to full-length proteins, and integrate this with available transcriptomic (gene expression) and proteomic data (half-life, thermal stability, abundance). We show a clear separation between population and disease-associated variants.

In comparison to population variants, we find that disease-associated variants preferentially target proliferative and nucleotide processing functions, localise to protein cores and interaction interfaces, and are enriched in more abundant proteins. In terms of their molecular properties, we find that common population variants and disease-associated variants show the greatest contrast. Additionally, we find that rare population variants display features closer to common than disease-associated variants.

We highlight that a multidimensional, integrative approach is essential to obtain a better understanding of the molecular features which separate these studied datasets. Ultimately, this understanding will contribute to the prediction of variant deleteriousness, and will help in prioritising variant-enriched proteins and protein domains for therapeutic targeting and development.

The ZoomVar database, which underlies our analysis, is available at `http://fraternalilab.kcl.ac.uk/ZoomVar`, and allows users to structurally annotate SAVs and calculate variant enrichments in protein structural regions.

2 Introduction

The genomic revolution has brought about large advances in the identification of disease-associated variants. However, despite the recent explosion of genetic data, the problem of "missing heritability" still persists [1], where the genetic component of a phenotype remains unidentified. This phenomenon can most likely be attributed to variants where a causal link is difficult to establish. Prime examples being variants with low penetrance, and/or those with higher penetrance, however unique to single/few individuals, such as *de novo* variants implicated in developmental disorders [2]. Somatic cancer variants pose a similar problem, as driver mutations can be difficult to segregate from passenger mutations; moreover this classification may vary from case to case [3]. A common feature of all such variants where disease associations are difficult to establish is that they defy detection by the use of statistical methods alone [4]; therefore one must understand the impact of variants at the molecular level on protein structure and function, to correctly classify them. Moreover, this knowledge is essential for a number of strategies associated with the design of therapeutic interventions, e.g. structure-based drug design and/or drug repurposing [5].

A number of recent studies highlight that the characteristics of genetic variants, particularly those found in nominally healthy individuals, are still poorly understood. As a consequence, the boundary separating disease-causing from neutral variants may be more fluid than initially believed; an example being the fact that a number of missense variants thought to lead to severe Mendelian childhood disease were identified in nominally healthy individuals in the ExAC database [6]. Variant impact predictors play an important role in the identification/prioritisation of potential disease-associated variants. However these have, in the past, been trained predominantly to detect the difference between disease-associated and common variants, neglecting the

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

difference between disease-associated and rare variants; thus it has been suggested that these do not perform so well when distinguishing rare neutral variants from those which are pathogenic [7]. Moreover, whether common variants have more functional impact than rare variants is hotly debated [8, 9]. A handful of recent studies have attempted to verify whether predictions are functionally accurate, by performing *in vivo* saturation mutagenesis. Here it has been shown that current predictive methods are limited in accuracy [10, 11]. As the majority of such methods rely heavily on evolutionary, sequence-based information, cases where pathogenic mutations localise to non-conserved positions are often incorrectly predicted. This has been proven to be problematic, particularly in the case of compensated pathogenic mutations; such mutations are present as the wild-type in other species, where their pathogenic effects are negated by another variant [12].

Due to these observations, it becomes increasingly pressing to understand the molecular characteristics of variants in health and disease; including differences in the characteristics of driver and passenger somatic cancer mutations, and in the impact of rare and common population variants. Analyses of the localisation of variants to protein structure, taking into account their proximity to functional sites (e.g. post-translational modifications, or PTMs) [13, 14, 15, 16, 17, 18], have shown to be effective in uncovering the impact of variants at the molecular level [19]. In the field of cancer research, protein structure-based methods have been used to successfully predict cancer driver genes [20, 21], as validated by a recent large-scale study by Bailey et al. [22]. Despite such success, this does not appear to have been applied to other classes of variants (i.e. population and Mendelian disease-associated variants).

Only a few studies [15, 23] have taken advantage of the recently available large-scale data, and compare, using structural bioinformatics methods, disease-associated variants with somatic cancer variants, and variants found in the general population. Furthermore, recent papers propose that previously observed trends, which suggest that somatic cancer single amino acid variants (SAVs) are enriched in protein-protein interaction sites, could be due to biases caused by the tendency of disease-associated variants to localise to those proteins which are most experimentally studied [15]. Therefore robust statistical methods to treat these comparisons are urgently needed. Additionally, large-scale proteomics and transcriptomics datasets have been generated in recent years, but, to the best of our knowledge, studies which incorporate this information in the analysis of the impact of genetic variants are yet to appear.

These factors have motivated us to undertake an integrative analysis which compares disease-associated variants, including somatic cancer variants and germline disease-associated variants, with variants found at different frequencies in the general population. A unique feature of our analysis is in addressing the interplay between macroscopic features, such as proteomics data and functional pathways, with microscopic features, such as protein structural localisation. In particular, we have made use of recently published protein half-life data [24], along with protein abundance [25], thermal stability [26] and transcriptomics data [27], to uncover underexplored biophysical and biochemical principles governing the impact of variants.

The ZoomVar database, which underlies our analysis, enables users to structurally annotate SAVs and to calculate the enrichment of SAVs in protein structural regions. It can be queried directly via the web interface (`http://fraternalilab.kcl.ac.uk/ZoomVar`) or programmatically using the REST script downloadable from the site.

3 Methods

3.1 Data sources

3.1.1 Variant data

ClinVar (dbSNP BUILD ID 149) variant data [28], COSMIC coding mutations (v80) [29] and gnomAD exome data [30], all mapped to the GRCh37 genome build, were obtained in variant call format (VCF). The ClinVar dataset contains variants submitted through clinical channels. Only variants with CLINSIG codes 4 and 5 (probably pathogenic and pathogenic) were selected for further analysis. To ensure the quality of our dataset, we selected only variants with "variant suspect reason code" of 0 (unspecified). Additionally, all variants labelled as being somatic were filtered from this dataset. All variant datasets were mapped to Ensembl protein sequences [31] using the Variant Effect Predictor (VEP) [32], and further mapped to canonical UniProt sequences and the respective structures/homologs.

3.1.2 Protein-protein interaction networks

A large non-redundant protein-protein interaction network (UniPPIN) [33] was used. This incorporates non-redundant data amalgamated from IntACT [34], BioGRID [35], STRING [36], DIP [37] and HPRD [38], as well as recent large-scale experimental studies [39, 40, 41].

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

3.1.3 Protein sequences and structures

The biounit database of the Protein Data Bank (PDB) was downloaded on 28/04/2017. For mapping purposes, in this study, both the canonical UniProt human protein sequences [42] (for mapping to structures and protein-protein interaction networks) and Ensembl protein sequences [31] (for mapping variant datasets) were used.

3.1.4 Gene and protein annotations

Gene sets for KEGG pathways were obtained from the MSigDB database [43]. Oncogene and tumour suppressor gene annotations were taken from Supplementary Table S2A from Vogelstein et al. [44]. Cancer drivers were taken from the Cancer Gene Census (CGC) (COSMIC v84). Genes from both tiers 1 and 2 were included. Conversions between gene symbols, Entrez gene identifiers and UniProt accession numbers were performed using the biomartR package [45]. A list of DNA-binding domains was obtained from the review by Vaquerizas et al. [46]. These domains were mapped from InterPro [47] IDs to PFAM IDs using conversion tables in PFAM (v31).

3.1.5 Protein-drug interaction mapping

A mapping of protein-drug interactions was obtained from DrugBank (v5.0.11) [48] (under "Target Drug-UniProt Links") and filtered for human proteins. Drugs were mapped to a PFAM domain-type if at least one domain of that type occurs in a protein a drug is known to interact with. It is, of course, possible that a drug may only interact directly with another domain-type within the protein. However, this approach was chosen due to the fact that if only domain-drug interactions with supporting structural information are accepted, the data becomes both sparse and biased towards structurally resolved domains.

3.1.6 Protemics and transcriptomics data

Protein thermal stability and half-life data were obtained from separate large-scale studies by the Savitski lab [26, 24]. Gene expression quantification (Reads Per Kilobase of transcript per Million mapped reads [RPKM]) counts per sample (v6p) was downloaded from the GTEx portal [27] and grouped by tissue, according to the sample metadata provided. For each tissue type, we quantified the gene-wise proportion of samples with an RPKM equal to zero. Only those genes with zero counts in < 10 % of samples were retained for our analysis. Protein abundance data (protein per million [ppm]), integrated for each tissue/sample type were obtained from PaxDb [25].

3.2 ZoomVar Database

3.2.1 Identification of resolved structures/homologs

Canonical UniProt human protein sequences were assigned resolved structures/homologs from the PDB biounit database [49] using BLAST [50]. BLAST searches were carried out using both the entire protein sequences and domain sequences, which were defined by scanning UniProt sequences against the PFAM seed library [51] using HMMER [52]. Hits were only accepted with sequence identity > 30 % and E-value < 0.001. T-COFFEE [53] was used to obtain a residue level mapping of queries to structure hits. The quotient solvent accessible surface area [Q(SASA)] of each structure residue was computed using POPS [54].

3.2.2 Mapping of Ensembl proteins

Ensembl protein sequences were mapped to UniProt protein sequences [42], using the UniProt ID mapping. Additionally, if UniProt and Ensembl sequences were not of the same length, the sequences were aligned using T-COFFEE [53] to obtain a per residue mapping. Stretcher [55] was used to align those sequences which were too long to align using T-COFFEE.

3.2.3 Identification of interaction complexes

For each interaction in our protein-protein interaction network, resolved binary interaction complexes and homologues were identified using the BLAST search results. As an example, if protein A and B are annotated as interacting in UniPPIN, and their structure homologues A' and B' are located in a resolved structural complex (and at least one residue from each protein is involved in a shared interface), residues from A and B are mapped onto A' and B' to infer their interaction interface.

The partner-specific regression formula from HomPPI [56] was used to assign a score and zone to each interaction interface inferred in this way. Residues involved in interfaces were assigned using POPSCOMP [57]. Only those residues with a change in SASA > 15 Å^2 were annotated as interface residues.

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

3.2.4 Determination of per-residue binding partners

A protein may interact with multiple other proteins. For each of these interactions, a maximum of 10 corresponding best hits (ordered by HomPPI defined score [56]), located in the best populated zone, were considered. If a residue was located at the interaction interface, in at least half of these structures, it was annotated as interacting with that specific protein, otherwise it was annotated as non-interacting.

3.2.5 Mapping of variant data

Variants in each dataset were annotated according to protein region localisation using the ZoomVar database. For certain analyses the COSMIC data was divided into "driver" and "non-driver" subsets, taking drivers as variants which map to all proteins from both tier 1 and tier 2 of the Cancer Gene Census (CGC) (COSMIC v84). The non-driver subset contains all other variants.

3.2.6 Definition of regions

We defined several types of protein and domain regions as described below.

Interface regions were considered to be composed of residues which bind to at least one protein interaction partner. Core residues were defined as those with a Q(SASA) < 0.15. Surface residues were defined as those with a Q(SASA) ≥ 0.15 which do not take part in protein-protein interaction interfaces.

Disordered protein regions were predicted using DISOPRED3 [58]. Intra-domain ordered regions were defined as those regions predicted to be ordered which lie within PFAM defined domains. Intra-domain disordered regions were defined as regions predicted to be disordered which lie within PFAM domains. Inter-domain disordered regions were defined as those regions not located within PFAM defined domains which are predicted to be disordered. Any residues with structural coverage were filtered from the inter-domain disordered regions as it was thought that these could potentially belong to domains which have not been defined by PFAM.

Ubiquitination and phosphorylation sites were obtained from PhosphoSitePlus [59]. Each site was mapped to the structural template with the highest identity. Regions close to phosphorylation and ubiquitination sites were defined as those within 8 Å in 3D space.

3.2.7 Creation of ZoomVar database

All data, including per-residue mappings, were stored in the ZoomVar MySQL database [60]. A web interface and REST architecture was implemented, using the Django framework [61] to allow users to query this. It is available at `http://fraternalilab.kcl.ac.uk/ZoomVar`.

3.3 Calculation of SAV enrichment

The binomial cumulative distributive function (see Equation 1) was used to assess the SAV enrichments of individual proteins, domains or domain-types, and the 2-tailed binomial test was used to assess the significance of enrichment/depletion. In this formula k is the number of observed SAVs which localise to a region, n is the total number of SAVs which localise to all regions of interest (all_regions) and p is ratio of the size of the region (number of amino acids) to the size of all_regions:

$$P(N(SAVs_{region}) \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \tag{1}$$

Hereafter, we refer to the binomial CDF as the variant enrichment score (VES).

The calculations were performed for the regions defined in the table below:

level/region	all_regions
protein	all proteins
domain	all domains
PFAM domain-type	all PFAM domain types
protein region	union of all regions of a particular protein
domain region	union of all regions of a particular domain
PFAM domain-type region	union of all regions of a particular PFAM domain-type

Table 1: The anatomy of the protein levels considered in our analysis. N.B. at the protein region, domain region and domain-type region levels, if the region of interest is the "core", the union of all regions will be the "core", "surf" and "interact". The same logic applies to other regions of interest (see Fig. 1).

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

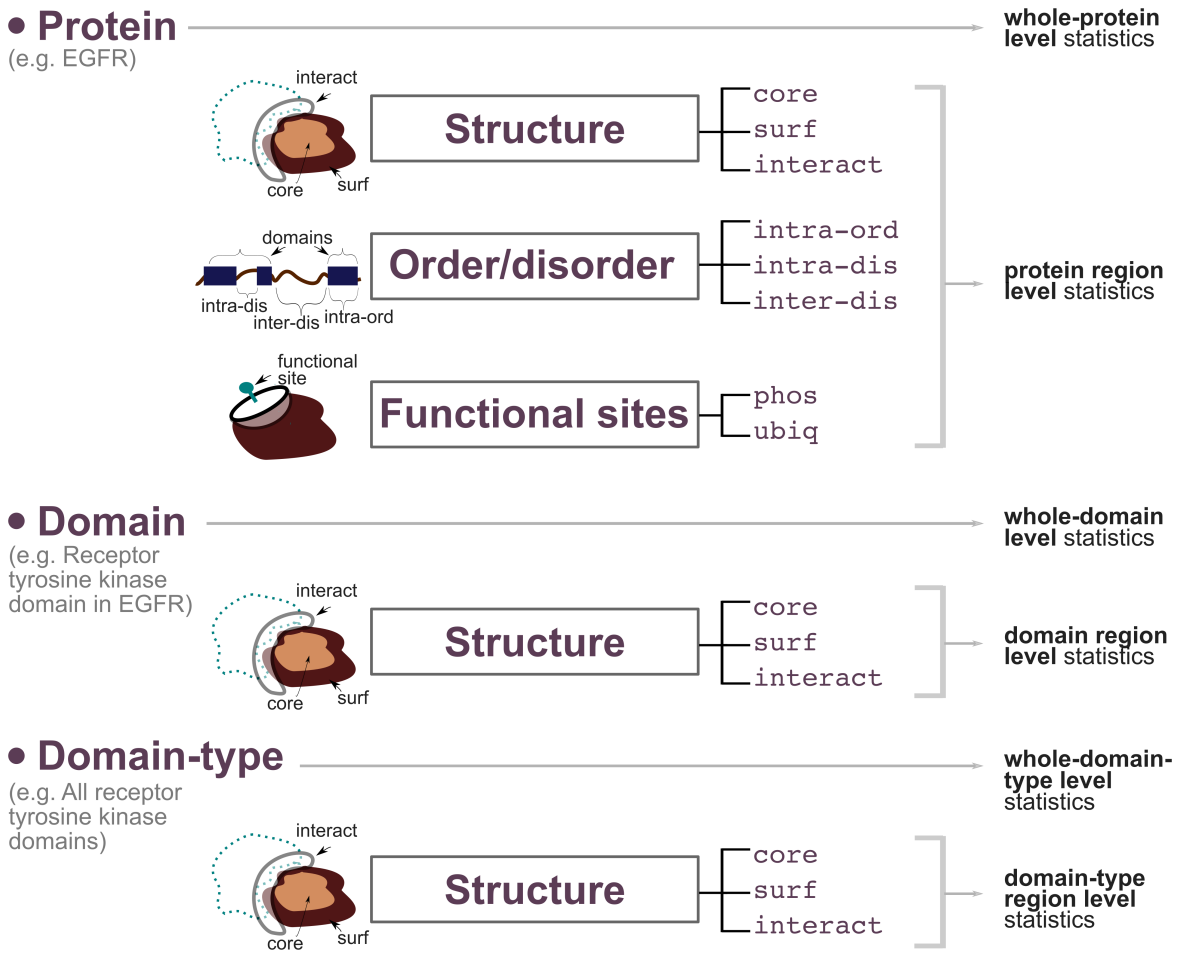

Figure 1: The regions used for the analysis of variant enrichment at the protein region, domain region and domain-type region levels.

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

For each analysis at the whole protein or whole domain level, all UniProt proteins/domains (except for immunoglobulin and T cell receptors) containing SAVs in any of the datasets analysed, were considered to be the background proteome (all_regions). Proteins belonging to immunoglobulin and T cell receptor gene family products were filtered from all analyses (HGNC definition [62]), to avoid the inclusion of variants which could have arisen from the process of affinity maturation.

For all calculations of enrichment and simulations involving protein or domain regions (e.g. core, surface and interface), cases where the region is of size 0, or that the protein/domain contains no SAVs, were omitted in this analysis.

The overall SAV enrichment of protein regions, for each data set, was also calculated using a density-based metric (see Equation 2).

$$P(SAV_{region}) = \frac{(N(SAVs)_{region}/size_{region})}{(N(SAVs)_{all_regions}/size_{all_regions})} \tag{2}$$

Here 95 % confidence intervals were estimated via bootstrapping (10,000 iterations). The 2-tailed significance of enrichment/depletion was estimated by simulation. 10,000 simulations were carried out for each dataset, in which the number of variants which localise to a given protein was kept constant, but their location within the protein randomised. The regional density of variants was calculated for each simulation and compared to the actual value in order to derive a p-value. Simulations were performed in this way, keeping the number of SAVs which localise to each protein fixed, in order to overcome bias which stems from the assumption that variants are uniformly distributed throughout the proteome.

3.4 Enrichment analysis of gene sets

3.4.1 Gene set enrichment

Enrichment analyses were performed using Gene Set Enrichment Analysis, using the implementation provided by the R FGSEA package [63]. Given an enrichment statistic for each query gene, the GSEA algorithm outputs a score per gene set, which quantifies the enrichment of query genes in the sets examined. This is then normalised by the size of the gene set, to give a normalised enrichment score (NES).

We utilise the centred VES, as the enrichment statistic which is input into the GSEA algorithm. Here, the centred VES is simply obtained by subtracting 0.5, therefore proteins with the expected number of SAVs have a centred VES of 0. At the whole protein level only sets with n $\geq$ 25 were considered. At the protein subregion level, variant enrichment data exists for a smaller number of proteins, due to incomplete structural coverage of the proteome. In order to perform a complete comparison between pathway enrichment at different levels, all pathways analysed at the whole protein level were also analysed at the protein subregion level.

3.4.2 Definition of pathway clusters

The pathway normalised enrichment scores (NESs), calculated at the whole protein level for each dataset, were used to perform K-means clustering of KEGG pathways [64]. The R package NbClust [65] was used to determine the optimum number of clusters.

3.5 Analysis of expression, abundance, and stability data

Spearman correlations of protein-wise and region SAV enrichments with expression levels (RPKM), abundance (ppm), half-life (hours), thermal stability (Tm), and density (mean contacts of core α carbons) were calculated. Additionally, gene set enrichment analysis was performed as in Section 3.4.1, but using the metrics in the table below as enrichment statistics.

thermal stability	Tm - mean(Tm)
abundance	log(ppm + 1) - mean(ppm + 1)
expression	log(median(RPKM) + 1) - mean(log(median(RPKM) + 1))
half life	log(hours) - mean(log(hours))

Table 2: Proteomics and transcriptomics-based metrics used as enrichment statistics for GSEA analysis.

Here it can be seen that the mean value for each quantity of interest was subtracted to obtain values centred around 0, allowing both pathway enrichment and depletion to be assessed.

3.6 Statistics and data visualisation

The majority of data analyses were performed in the R statistical programming environment. All corrections for multiple testing have been done using the Benjamini-Hochberg method in R (p.adjust function). Bootstrapping

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

was performed using the boot package (function boot) [68]. Spearman correlations were performed using the SpearmanRho function of the DescTools package [69]. Heatmaps were produced with either the heatmap.2 function in the gplots package [70] or the ComplexHeatmap package [71], in which clustering, wherever shown, was performed with hierarchical clustering (hclust function) using default parameters unless otherwise stated. Circos plots were generated with the Circos package [72]. Additionally, binomial CDFs were calculated and two-tailed binomial tests performed using the NumPy package in Python [73].

4 Results

We present a multidimensional analysis of single amino acid variants (SAVs) observed in the general population (gnomAD database) [30], in comparison to somatic cancer-associated SAVs from the COSMIC database [29] and disease-associated SAVs from the ClinVar database [28]. Throughout this analysis we further divide the gnomAD data into its constituent common and rare variants, to investigate whether there are differences between these two subsets.

We ask whether the enrichment of variants is associated with specific structural features and functional pathways, and whether results differ for population and disease-associated variants. In particular we investigate the interplay between variant enrichment and proteomics features; for example, we explore whether disease-associated variants preferentially target the core of less thermally stable proteins, as these might be more prone to destabilisation, leading to complete/partial unfolding. Finally, we use these features to understand whether rare population variants demonstrate characteristics which are more similar to common population variants or disease-associated variants. Such exploration of the interplay of the microscopic versus macroscopic features of proteins is novel in the field.

Our analysis explores the enrichment of SAVs at different levels, constituting what we define as a protein-centric anatomy of variants in health and disease, as illustrated in Fig. 2. We employ a similar approach to that used in the prediction of cancer driver genes [74]: the SAV enrichment of individual proteins/regions has been modelled using a binomial distribution (Methods Equation 1), whereas global trends in the distribution of SAVs have been investigated by calculating SAV density (Methods Equation 2). The binomial cumulative distribution function quantifies the enrichment of variants (Fig. 2g) and is referred to as the Variant Enrichment Score (VES). This is assessed statistically using a two-tailed test (see Section 3.3). Additionally, the significance of the enrichment/depletion of SAVs, in terms of their density, is assessed by comparison to simulated SAV distributions, in which the number of SAVs is kept identical to that observed in the data, but their positions within the protein are randomised. This goes beyond similar studies (e.g. [16, 17, 18]) and addresses biases which could result from the increased study and structural coverage of disease-related proteins.

A summary of the numbers of SAVs investigated in each dataset is given in Table 3, and a more detailed breakdown is given in the Supplementary Materials Section S4.1.

Region	Common	Rare	COSMIC	ClinVar
protein	54,571	3,806,698	1,731,030	21,272
surf	12,151	966,409	491,179	8,558
interact	403	38,108	22,205	768
core	2,789	296,291	152,356	5,194
intra-ord	20,650	1,575,286	755,683	14,620
intra-dis	2,984	197,482	96,914	1,211
inter-dis	17,352	1,045,997	439,437	1,128
phos	1,661	158,192	82,364	2,362
ubiq	440	52,250	25,778	607

Table 3: Numbers of SAVs which localise to different protein regions in the studied datasets.

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

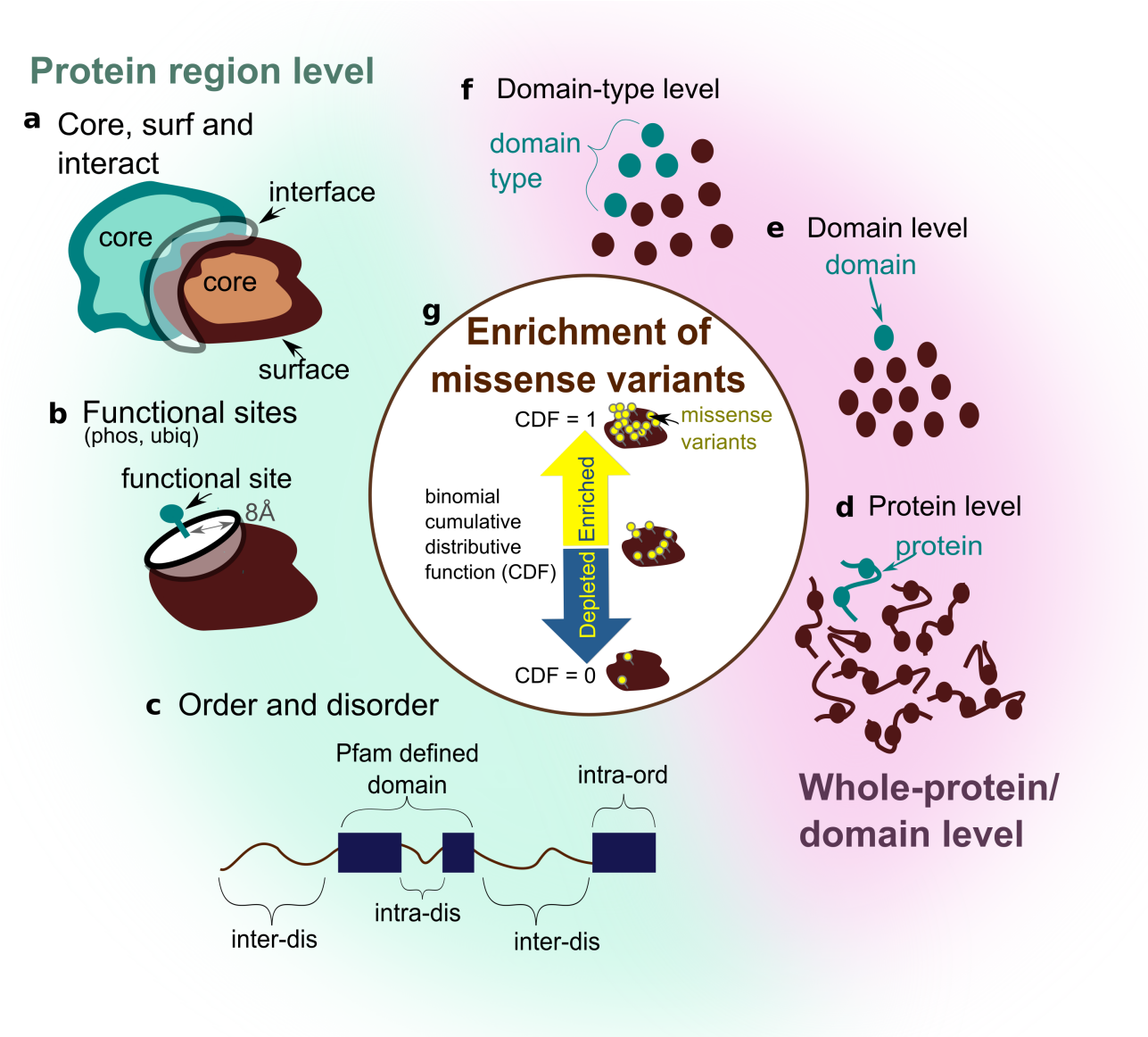

Figure 2: Enrichment statistics are calculated at different levels. At the protein region level, the number of SAVs in a region is compared to the number of SAVs in the rest of the protein. Regions are defined as: a) core, surface (surf) and interface (interact) regions of a protein; b) regions close to functional sites, and; c) regions predicted to be ordered or disordered which lie either within or outside of PFAM defined domains. d,e) At the protein/domain level the number of SAVs in a protein or domain is compared to the number of SAVs in the whole dataset which localise to defined proteins/domains. f) At the domain-type level the number of SAVs in a particular PFAM defined domain type, are compared to the number of SAVs which localise to all domains. g) The calculation of enrichment at the different levels is statistically assessed using the binomial distribution. The binomial cumulative distributive function constitutes a Variant Enrichment Statistic (VES) with value range 0 to 1, which quantifies enrichment.

4.1 Disease-associated and population variants target different functional pathways

We first investigated whether variants from each dataset target proteins which are involved in distinct functional pathways. To do this we performed KEGG [43] functional pathway analysis, by ranking proteins using their whole-protein VESs (see Fig. 2d) calculated for each dataset, and using the Gene Set Enrichment Analysis (GSEA) algorithm [43] (see Section 3.4.1).

The pathway enrichment data, for each mutation dataset, were subjected to clustering and Principal Component Analysis (PCA) (see Section 3.4.2). In Fig. 3a it can be seen that variant enrichment segregates pathways into three clusters. Strikingly each pathway cluster appears to have distinct characteristics. The cluster visualised in orange is primarily composed of terms associated with cancer, growth and proliferation, whereas that coloured in pink contains pathways associated with splicing, transcription, translation and metabolic terms.

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Pathways associated with sensory perception and the immune response are found in the final "green" cluster. A handful of metabolic pathways also localise to this cluster, however, these appear to be more associated with environmental response and adaptation than those pathways found in the "pink" cluster; for example, pathways associated with the metabolism of drugs and xenobiotics are found here. For brevity, the "orange", "pink" and "green" clusters will be termed the "proliferative", "nucleotide processing" and "response" clusters respectively, for the remainder of this text. A list of pathways assigned to each cluster is given in the Supplementary Materials Section S4.2.

This visualisation (Fig. 3a) also reveals that both the common and rare subsets of the gnomAD database associate mostly with the "response" cluster, whereas COSMIC data localises between the clusters associated with response and proliferation. ClinVar data associates (as revealed by the localisation of factor loadings) with the "nucleotide processing" cluster, between both the "response" and "proliferation" clusters. Strikingly, the population variant datasets (gnomAD rare and common) are clearly separated from the disease-associated variant datasets by the first principal component (PC1), whereas COSMIC variants are separated from ClinVar variants along the third principal component (PC3) (see Fig. 4a).

These trends of functional distinction are further visualised in the Circos plot (Fig. 3b) [72]. Here it can be clearly seen that the gnomAD data only shows significant enrichment for pathways belonging to the "response" cluster, whereas the COSMIC data shows enrichment for pathways belonging to this cluster and those belonging to the "proliferative" cluster. The ClinVar dataset displays enrichment for pathways belonging to all three clusters; uniquely showing enrichment for pathways within the "nucleotide processing" cluster.

We went on to extend this analysis to the protein region level (Figures 4 and S4). Here we find that proteins enriched in gnomAD variants at the surface (Fig. 4b) are significantly enriched in pathways belonging to the "proliferative" cluster. Moreover, this enrichment is shared between common and rare variants (albeit not significant for common variants in individual pathways after FDR correction). Proteins with surfaces enriched in disease-associated variants (from COSMIC and ClinVar) are, contrastingly, not enriched in "proliferative" cluster pathways. However, no such pattern emerges for the protein core and interface (Fig. 4c and S4b), suggesting that population variants avoid disrupting the function of proliferation-related proteins by preferentially localising to the surface. Interestingly, the "nucleotide processing" cluster does not show such a marked enrichment of variants which localise to the surface in the gnomAD database, a possible indication that these pathways are more robust to disruption than those in the proliferative cluster. These data show that there is clearly an interplay between variant localisation at the macroscopic level (functional pathways) and the microscopic level (structural regions).

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

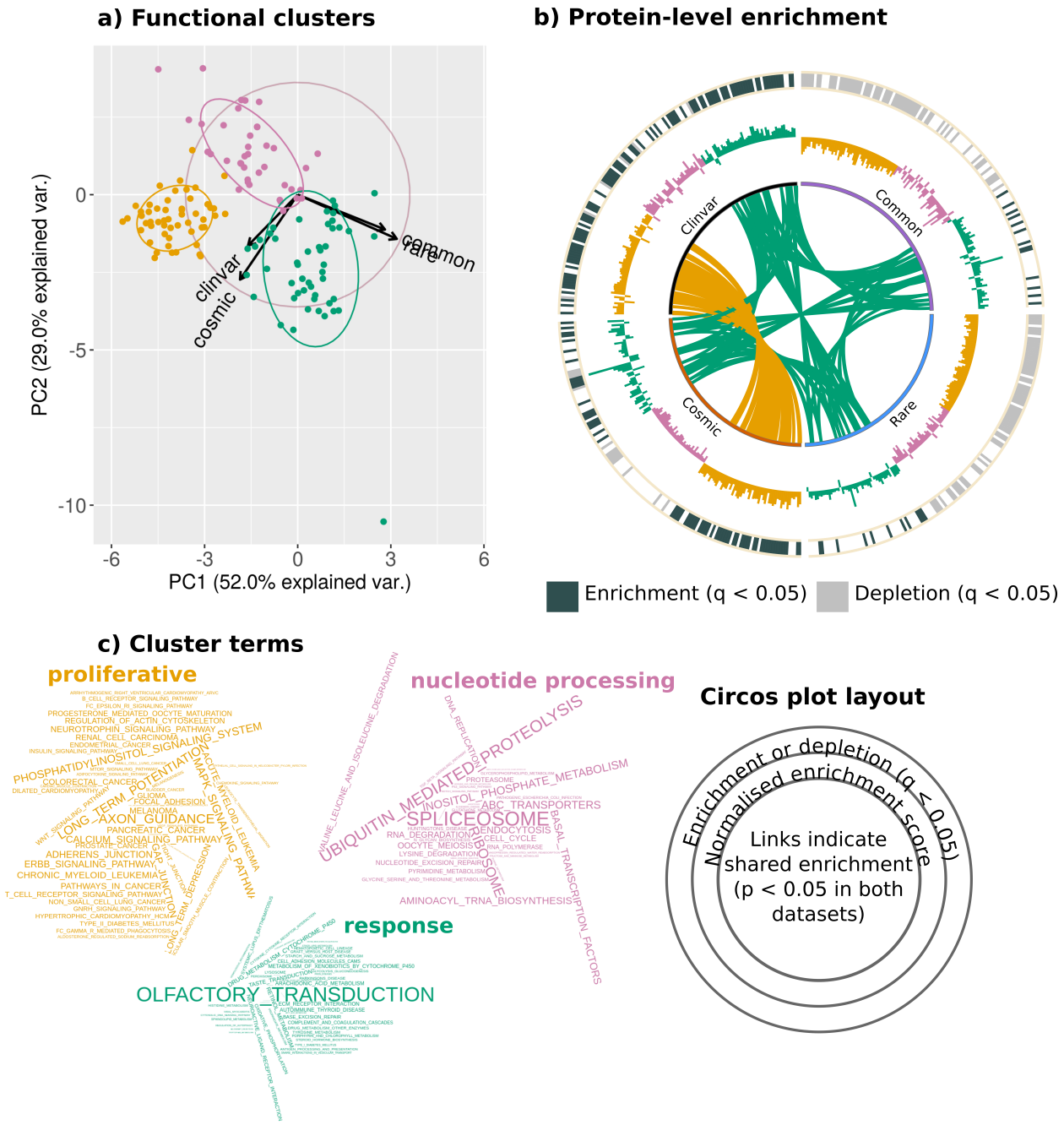

Figure 3: Functional analysis of proteins according to variant enrichment. a) At the whole protein level, KEGG pathways form 3 distinct clusters (K-means), as projected the first two principal components of the PCA. COSMIC, ClinVar and gnomAD (rare/common) data can be clearly separated by pathway enrichment, as evidenced by the visualisation of factor loadings (arrows). b) Enrichment for each dataset, at the whole protein level, visualised on a Circos plot (see Circos plot layout at the bottom-right). The Normalised Enrichment Score for each pathway is plotted as a bar graph (the further from the centre, the more positive) in the middle layer of the plot. In the outermost layer of the plot, significant enrichment (dark grey) or depletion (light grey) of a pathway (q-value < 0.05) is depicted. In the centre of the plot, links indicate enrichment (p-value < 0.05) shared between datasets. c) Pathway terms visualised by cluster, sized by their cluster uniqueness score. This is defined as the average of the Euclidian distances (calculated in 4D) to the two other cluster centres.

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

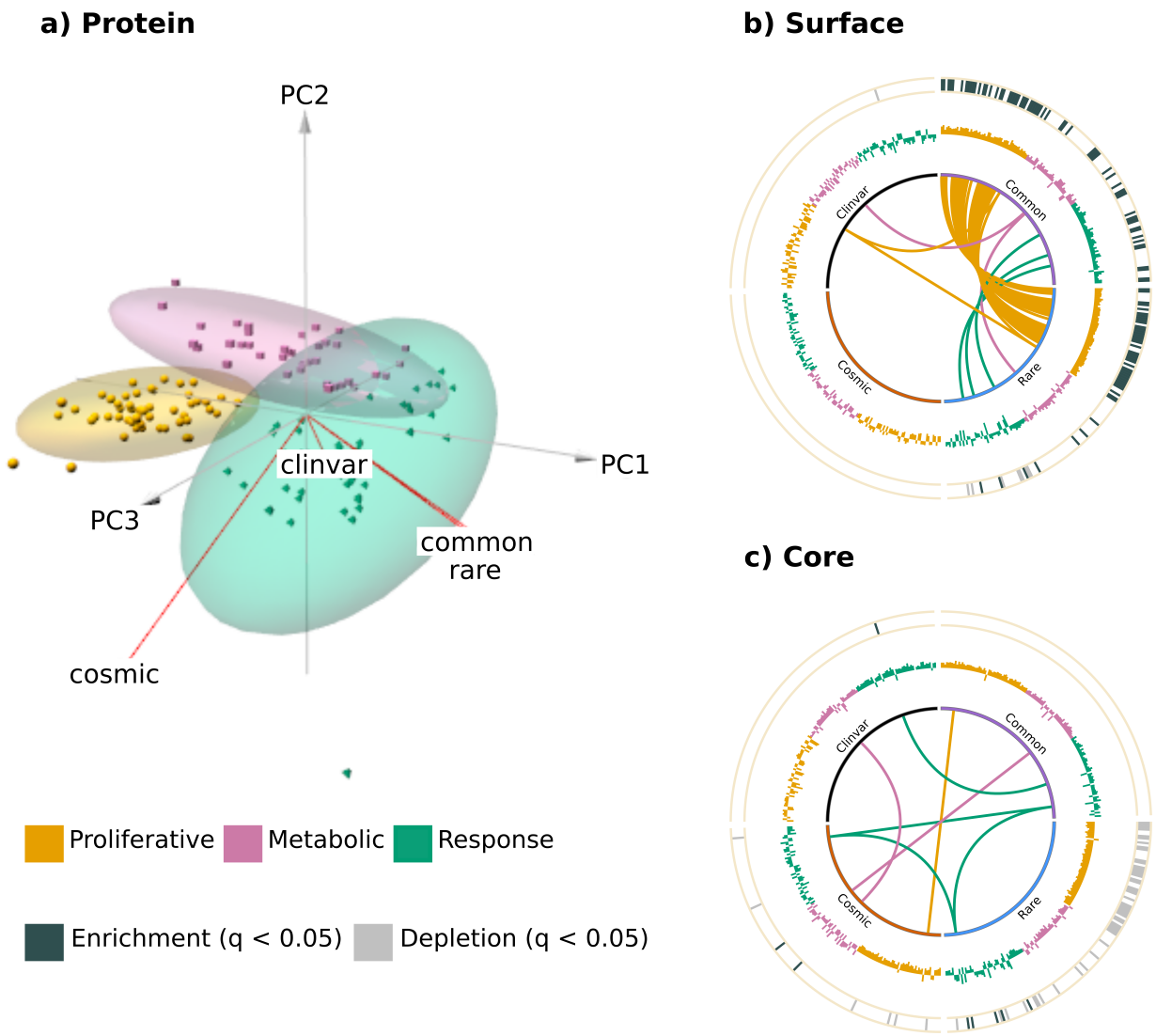

Figure 4: Functional analysis of proteins according to variant enrichment at the protein, surface and core region levels. A 3D visualisation of protein level data presented in Fig. 3a highlights the fact that COSMIC and ClinVar data are separated by the third principal component. b-c) Functional analysis of proteins according to variant enrichment at the protein surface and core, visualised on a Circos plot. The normalised enrichment score for each pathway is plotted as a bar graph (the further from the centre, the more positive) in the middle layer of the plot. In the outermost layer of the plot, significant enrichment (dark grey) or depletion (light grey) of a pathway (q-value < 0.05) is depicted. In the centre of the plot, links indicate enrichment (p-value < 0.05) shared between datasets.

4.2 Population and disease-associated variants localise to different protein regions

We then zoomed in to view trends in the enrichment of variants at the microscopic level. Specifically, we catalogued the enrichment of variants in core, surface and interface regions; intra-domain ordered regions (intra-ord), intra-domain disordered regions (intra-dis), and inter-domain disordered regions (inter-dis); and regions close to (≤ 8 Å) of phosphorylation sites and ubiquitination sites.

In agreement with previous research, we find disease-associated (ClinVar) variants to be enriched in both protein cores and interfaces, but depleted on protein surfaces (see Fig. 5a and Supplementary Fig. S4 [15, 16, 17, 18]. This reflects the potential disruption, caused by such mutations, of structurally and functionally important protein regions. The enrichment of Clinvar variants in structurally important sites is further demonstrated by their preferential targeting of residues which are highly connected when considering network representations of protein structure, as shown in Section S2.1 of the Supplementary Materials. GnomAD variants (both common and rare) and somatic non-driver variants display the opposite trend, most likely as variants which localise to protein surfaces are less likely to impact on protein structure and function than either core or interface mutations. Somatic driver variants follow trends closer to ClinVar variants, with slight, but significant, depletion on the

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

surface, but enrichment in the core. Protein interfaces are enriched in disease-associated variants but depleted of gnomAD rare variants. GnomAD common variants appear neither significantly enriched nor depleted, however this may result from the relative sparsity of the data; fewer variants are shared between many individuals (this is clearly evidenced by the numbers in Table 3). Interestingly, COSMIC non-driver variants appear depleted in interacting interfaces. However, it becomes clear that they are actually significantly enriched when compared to simulated null distributions (Fig. 5a inset), and that this enrichment is due to a small subset of proteins which harbour a large number of variants at interface regions. Genes to which these variants reside may be putative driver genes (see Supplementary Materials Section S4.3), as a number of known driver genes are enriched in variants in protein interface regions [16, 22, 74], and this phenomenon has been exploited by by Porta-Pardo et al. [74] to identify cancer driver genes.

A more detailed per-protein analysis can bring finer granularity into the comparison of variant enrichment. Therefore we look at a curated list of oncogenes and tumour suppressor genes (TSGs) (see Section 3.1.4) [44]. Several studies have suggested that proteins encoded by oncogenes (which are activated upon mutation) and tumour-suppressor genes (TSGs, which are inactivated upon mutation) tend to be enriched in mutations in different protein regions [15, 16, 75]. We found that clustering based on VESs broadly classifies these proteins into two groups, one comprising of proteins enriched in mutations mainly at protein-protein interaction interfaces and protein surfaces, and another group of proteins generally enriched in mutations in the core (some of these proteins also show enrichment in mutations in interacting interfaces, but a clear depletion at the surface is evident) (Fig. 5b). Interestingly, we observe a statistically significant (Fisher-exact test p-value = 0.004199) segregation of these two groups in terms of cancer driver status: the first group of proteins are mainly (17 out of 24) products of oncogenes, and the other mainly those of TSGs (17 out of 25). These results are consistent with the hypotheses that activating mutations in oncogenes are likely to affect particular functions by targeting specific interactions, whilst inactivating mutations in TSGs abrogate protein function [16, 75]. Taking the oncogenes and TSGs as two separate groups, the GSEA result confirms a similar trend; moreover, it can also be seen that the disease-associated datasets (ClinVar and COSMIC) show opposite patterns of enrichment in comparison to the gnomAD data (Supplementary Fig. S6) [15, 16, 75]. These results confirm that our approach reproduces previous results and highlights clear, robust trends.

On analysis of variant enrichment in ordered and disordered regions, we again observe clear segregation between disease and population variants (see Fig. 5a). ClinVar and COSMIC variants are depleted in inter-domain disordered regions and enriched in intra-domain ordered regions. In contrast, gnomAD variants (both rare and common) appear enriched in inter-domain disordered regions and depleted in intra-domain ordered regions. GnomAD common and rare variants show similar trends to one another, which are distinct to those of disease-associated variants. Using quantitative statistical measures, these results suggest, as intuitively one would expect, that variants are more likely to be pathogenic if they fall within ordered domain regions.

The density of variants close to PTMs is also shown in Fig. 5a. Here, ClinVar variants appear enriched when considering the density of SAVs close to phosphorylation sites, but not significantly so in comparison to simulations. The large bootstrap confidence interval suggests this may be due to the sparsity of the data available. A similar observation is seen for COSMIC driver variants; however, COSMIC non-driver variants, which appear depleted according to variant density, are significantly enriched close to phosphorylation sites in comparison to simulated null distributions (Supplementary Fig. S5). This indicates that, in agreement with a number of other studies [76, 77], the disruption of phosphorylation sites may play a particularly important role in cancer. In contrast to phosphorylation sites, all data sets appear depleted of variants close to ubiquitination sites (Supplementary Fig. S5).

These analyses conclude that the enrichment of missense variants at various structural features consistently segregate population variants from disease-associated ones. For the majority of structural regions defined here, the greatest, most consistent distinction is always seen between common and ClinVar variants, in conditions where the data are not too sparse.

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

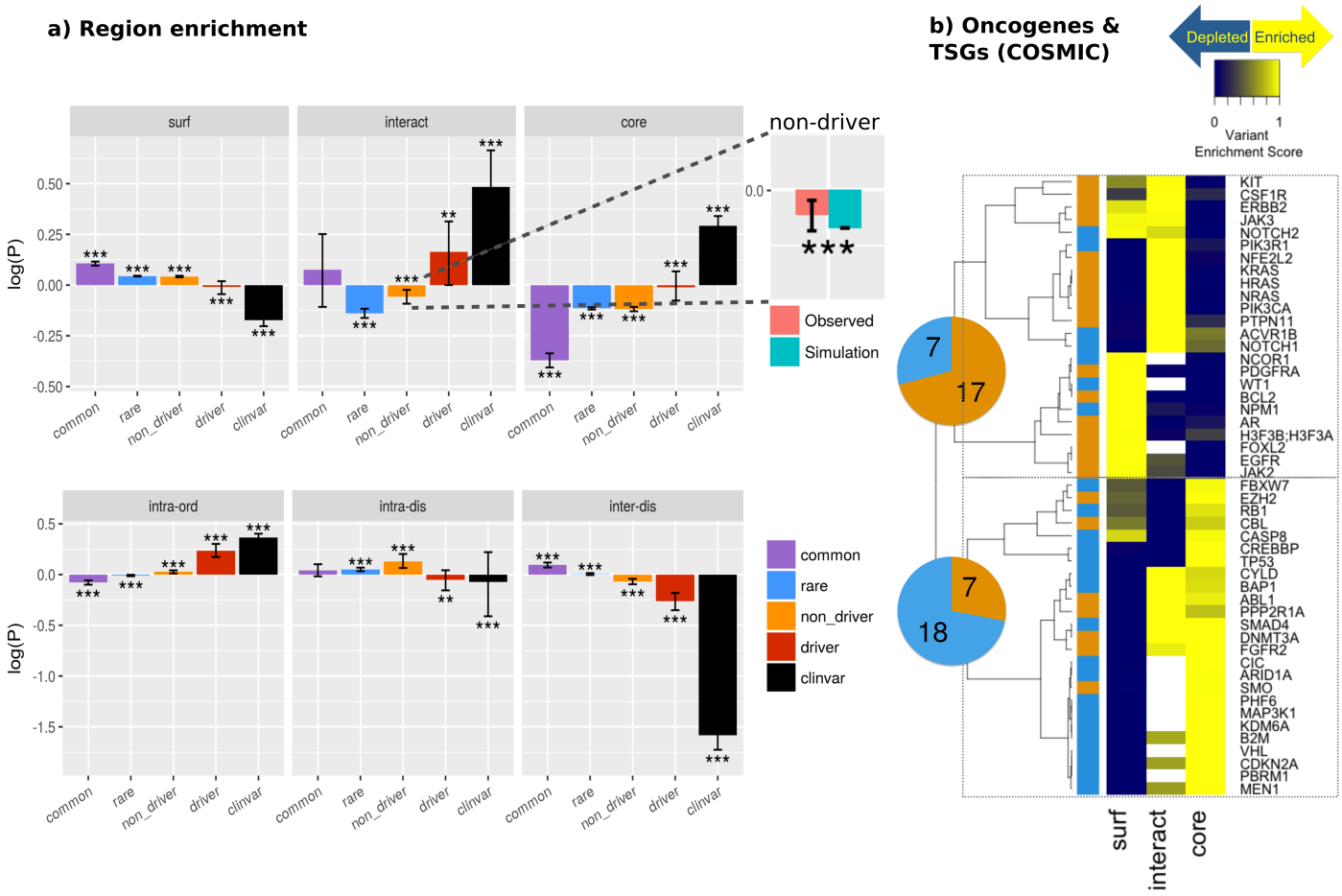

Figure 5: The localisation of variants to protein regions and CATH domain architectures. (a) The density of mutations in different protein regions, calculated using Equation 1. Error bars depict 95% confidence intervals obtained from bootstrapping. Significance was calculated by comparison to simulated SAV distributions (significance level indicated by: * q-value < 0.05, ** q-value < 0.001, *** q-value < 0.0001). The inset shows the observed density of non-driver variants at interacting interfaces in comparison to the simulated density. (b) Enrichment of COSMIC missense variants in protein core, surface and interface regions, across a list of annotated oncogene and tumour suppressor gene (TSG) products. The genes were grouped into two clusters using hierarchical clustering (see dendrogram by rows), with the pie charts enumerating the number of oncogenes (orange) and TSGs (blue) in each cluster. (c) The enrichment of CATH architectures in PFAM domains according to SAV enrichment. Results are depicted at the domain level and the domain region level. * indicates q-value < 0.05.

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

4.3 Towards a domain-centric landscape of variant enrichment

We then proceeded from the protein level to examine variant enrichment at the domain level. First, we compare variant localisation across protein domain types, using PFAM domain definitions. As depicted in Fig. 2, we calculated the variant enrichment at the amalgamated whole-domain and structural region (core/surface/interaction sites) levels.

Strikingly, characteristic patterns of variant enrichment appear. Fig. 6 depicts the union of the top 20 most variant-enriched domains for each data set. Here it can be seen that a small number of domains appear enriched in variants primarily only in the COSMIC and ClinVar data sets. These include known drug targets such as kinase and ion channel domains. A handful of domains, which are only enriched in COSMIC variants, include the Cadherin_tail and Laminin_G_2 domain, both of which play an important role in cancer [78, 79]. A larger number of domains are variant enriched in both the COSMIC and gnomAD dataset (rare and common variants). Some domains (e.g. Serpin, UDPGT, Collagen and EGF_CA) contain variants from all datasets or all datasets with the exception of COSMIC. In such domains, it is likely that the precise structural localisation of a variant determines whether it plays a pathogenic role. Intriguingly a few domain types, such as NPIP and NUT appear only enriched in common variants. This could suggest that these domains take part in functions for which it is desirable to maintain diversity within a population; however, little is known about either domain type [80, 81]. Thereby this further highlights the bias in study towards those domains associated with disease, rather than those enriched in population variants.

It also becomes apparent that the global trends in variant localisation to the core, surface and interface regions, observed in Section 4.2 are recapitulated here. Again the majority domains are enriched in gnomAD (rare and common) variants at the surface but ClinVar variants at the core. Although COSMIC variants show a trend broadly similar to gnomAD variants, it is clear that a larger proportion of domain-types are enriched at the core or interface. These include domain-types with known cancer driver associations, such as the P53 and VHL domains [82]. The observed patterns of variant enrichment are further highlighted by considering variant localisation to CATH architectures and by a case study on DNA-binding domains, both presented in the Supplementary Materials (see Sections S2.3 and S2.2).

We wished to understand how the targeting of domains by drugs and small molecules mapped to the landscape of variant enrichment we previously observed. To investigate this we used the protein-drug mapping provided in the DrugBank database, as detailed in Section 3.1.5. As already extensively pointed out [83], the targeting of domain-types by existing drugs is highly biased towards a small number of domain types, such as GPCRs and kinase domains. Indeed, we observe a large number of drugs targeting proteins containing 7tm (GPCR) domains. These domains are enriched in variants from the gnomAD and COSMIC database but are devoid of disease-associated ClinVar variants. Interestingly it has recently been shown that genetic variants in such domains (GPCRs), identified in the general population, may be associated with differential drug response between individuals [84]. Therefore we show that our domain-centric landscape of variant localisation highlights, for each domain type, implications useful for both understanding variant impact and motivating therapeutic design (see discussion).

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

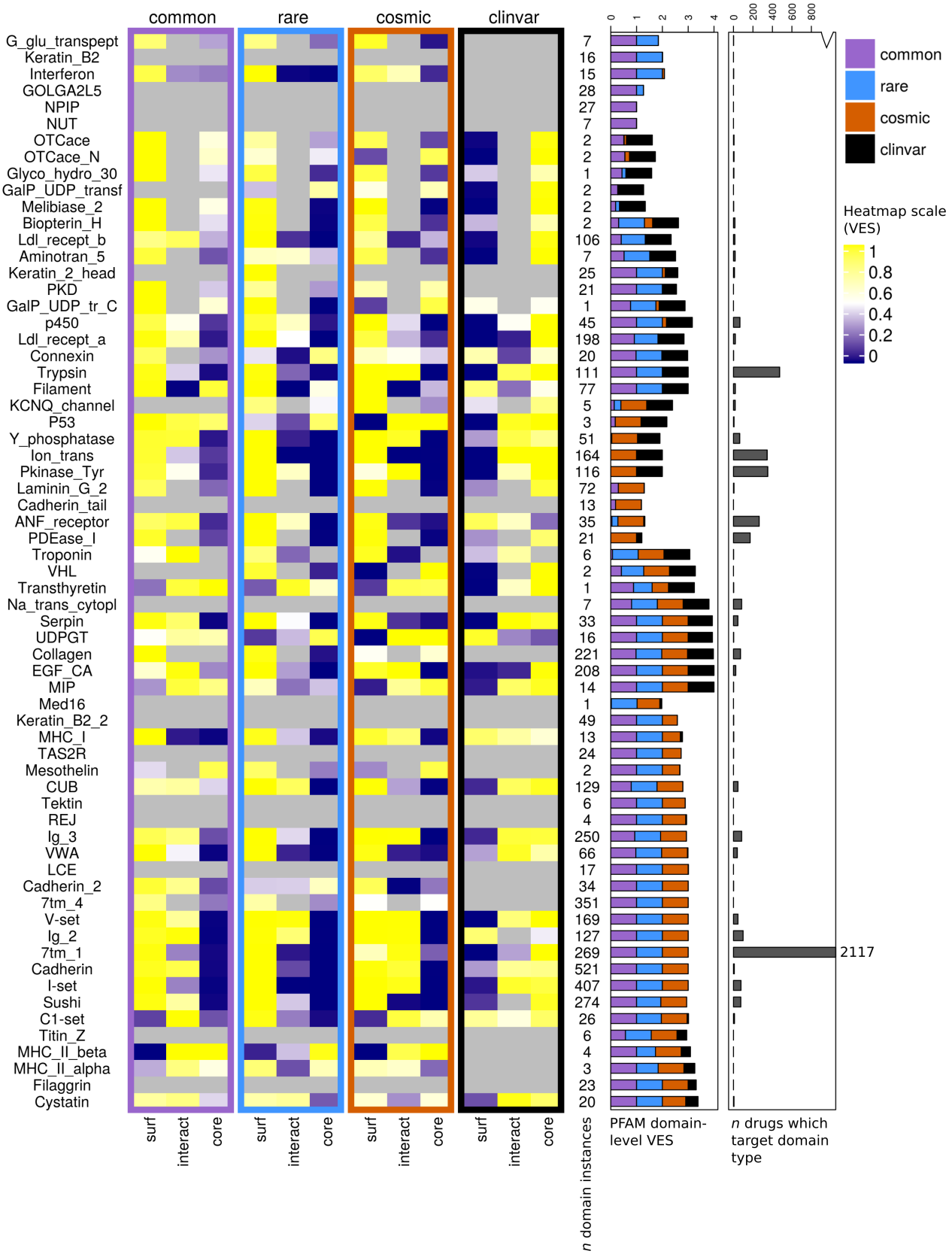

Figure 6: A domain-centric landscape of variant enrichment. Here the union of the top 20 most enriched domains for each data set is depicted: each row corresponds to a PFAM domain type. The plots are arranged in the same way as in Fig. 5, showing, from the left to the right, heatmaps of regional (surface, interface and core) enrichments for variants from the gnomAD rare, gnomAD common, COSMIC and ClinVar datasets. Domain instances and enrichment at the whole protein level (stacked bars) for each data set are shown. The number of drugs known to target proteins containing each domain type is depicted in the rightmost bar graph. Note the cut numeric axis; the number of drugs which target the only outlier, the 7tm_1 domain, is noted on the plot.

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

4.4 Proteomics and transcriptomics features associate with variant localisation

Proteins, of course, do not function in isolation but in the crowded environment of the cell. In our analysis so far we have viewed proteins through their three dimensional and functional properties; however, we have to consider that proteins may be present in the cell in different quantities, display different turnover rates and possess different melting temperatures. All of these factors can crucially affect the stability and the fitness of a protein to perform its function. Here we have made use of large-scale proteomics data, including protein abundance data from PaxDb [25] and data describing both protein half-lives and thermal stability from the Savitski lab [26, 24], together with transcriptomics data (GTEx database [27]), to explore relationships between these features and variant localisation. Please note that the numbers of proteins and SAVs which underlie each comparison are described in the Supplementary Materials Section S4.5.

Our results show that the protein-wise enrichment of disease-associated variants displays positive correlations with protein abundance, expression, half-life and thermal stability, whereas population variants exhibit the opposite trend (see Fig. 7 and Supplementary Figures S7-S8). It is important to recall here that the protein-wise enrichment of variants is calculated in comparison to the entire proteome (all UniProt proteins which contain SAVs in any of the datasets; see Fig. 2d).

However, zooming into the enrichment of variants in the core of protein structures, we found that in comparison to all regions of proteins with resolved structure, rare population variants demonstrate a positive correlation with abundance and thermal stability, whereas disease-associated variants negatively correlate with this (see Fig. 7). These results prove robust across multiple tissue types. Analogous correlations for variant enrichments at protein surfaces display opposite trends to those observed at the protein cores. Due to the relative sparsity of variants which map to protein interfaces, we believe it is difficult to draw robust conclusions from any trends observed for correlations of proteomics data with variant enrichment at protein-protein interaction sites.

Our results, at the "core" region level, for gnomAD rare and ClinVar variants suggest that disease-associated variants might preferentially localise to the core of unstable proteins, as these might be more easily destabilised to a degree at which function is deleteriously impacted. This possibility is further explored in the discussion. Similarly to the ClinVar data, the gnomAD common data also show negative correlations for variants occurring at the protein core; this could potentially give weight to the argument presented by Mahlich et al. [8] that common variants could affect molecular function more than rare variants. However, we believe this is more likely to be due to the fact that very few common variants localise to protein cores, as shown by Fig. 5, resulting in sparse statistics (i.e. the correlation is calculated over Variant Enrichment Scores which are already very low).

One might expect that mutations would be less easily accommodated in cores of densely packed proteins, which would have higher thermal stability. To assess this we calculate the mean number of Cα contacts within 8 Å of core residues, as a proxy for protein density. We find a significant correlation between this metric and protein thermal stability (vehicle_1: $\rho = 0.168$, q-value = 1.464e-12; vehicle_2: $\rho = 0.185$, q-value = 1.529e-13).

If we correlate this metric of core density (see Supplementary Materials S1.2 for details) with the core Variant Enrichment Score, we find a significant negative correlation for the gnomAD common dataset. No other datasets show significant correlations with core density, however a clear trend emerges in which correlations become progressively more positive in the order of gnomAD common, gnomAD rare, somatic driver, somatic non-driver and ClinVar (see Supplementary Fig. S9). This suggests variants may be more deleterious if they localise to a packed core. Again the complexity of the interplay between features is highlighted, as the higher stability of proteins with more packed cores suggests that destabilisation, to a degree which is physiologically relevant, may be more difficult to achieve. Although core packing and thermal stability are correlated, the correlation value (ρ) is low. Therefore, this feature is clearly not the only determinant of protein stability.

The results we see at the whole protein level, where the disease-associated ClinVar data clearly show a more positive correlation with Tm, are, at a first sight, more difficult to explain. However, work by the Picotti lab [85] has demonstrated that more stable proteins are generally more abundant. In agreement with this, we find significant correlations between the protein abundance and thermal stability data (see Supplementary Materials Section S4.6). Moreover, we do see significant positive correlations of protein-wise variant enrichment with protein abundance, in our analysis (see Fig. 7b). This suggests that the preferential localisation of ClinVar variants to more stable proteins could be attributed to the higher abundance of such proteins.

Interestingly, it can be seen that the trends observed at both the protein level and core region level, are less pronounced for cell line data and break down for extracellular fluids (saliva and urine). Moreover, the trend is most evident for tissues containing long-lived cell-types, such as the brain, ovary and testis. Transcriptomics data (see Fig. S7) again reinforces this picture, albeit with less contrast between data sets (particularly at the protein core).

Finally, we wanted to understand whether correlations with these proteomic and transcriptomic features could be associated with the specific functional roles of the involved proteins. This was achieved by investigating the association of these proteomic and transcriptomic features with biological pathways, using the GSEA algorithm. For the majority of proteomic and transcriptomic features, no clear associations with the functional clusters identified in Fig. 3 can be detected (see Supplementary Figures S11-S13). An exception to this is

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

protein thermal stability: pathways which belong to the "proliferative" cluster are clearly enriched in proteins of lower stability than the other two clusters (see Fig. 7c). This suggests that proliferation-related proteins may be vulnerable to disruption by mutations which target their already unstable cores. Moreover, this agrees with the idea proposed in Section 4.1, that "proliferative" cluster proteins may be less robust to disruption.

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

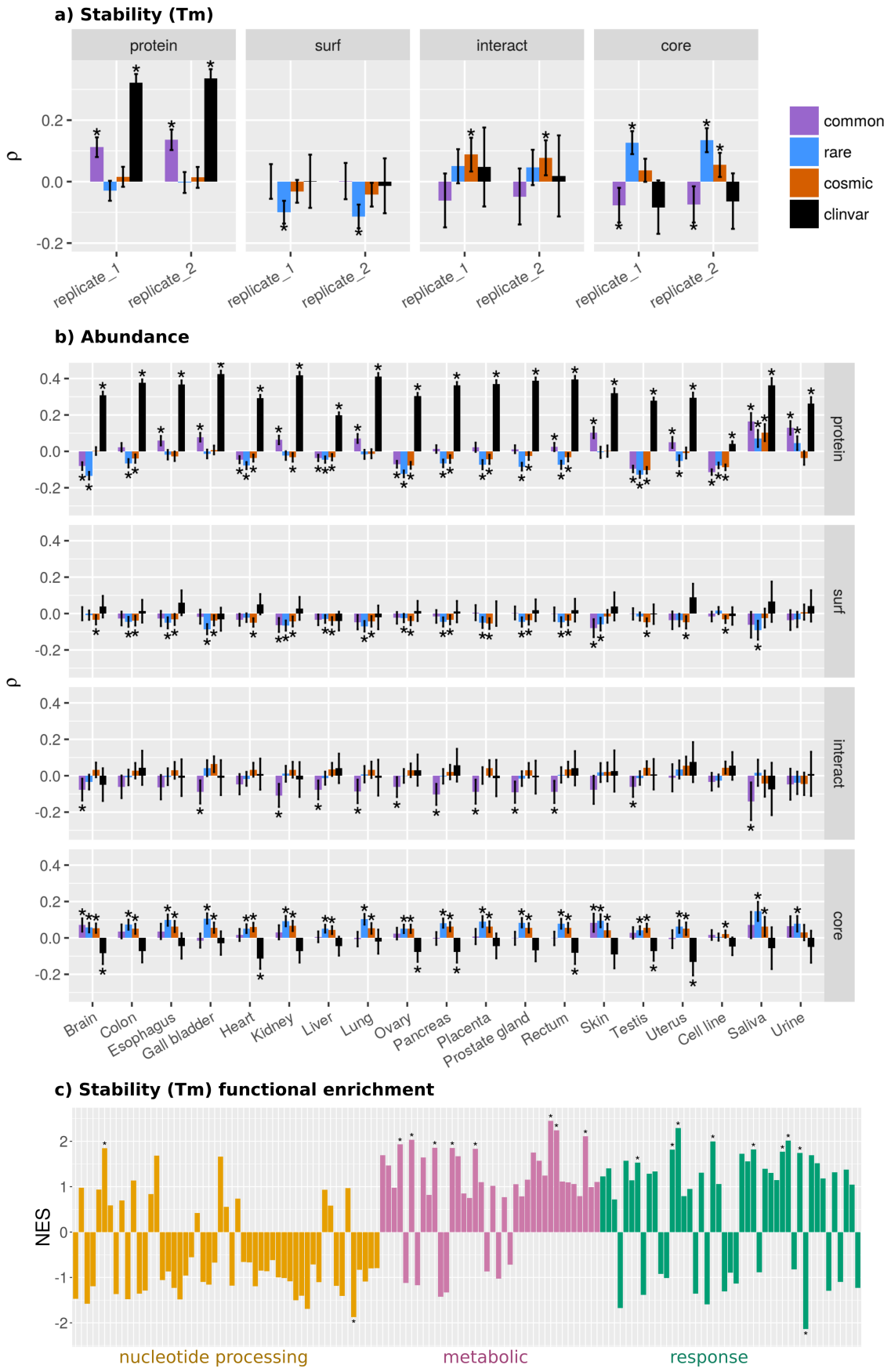

Figure 7: The protein-wise enrichment of SAVs in comparison to protein abundance, expression and stability. Spearman correlations for SAV enrichment (quantified as VESs) at different levels with a) protein melting temperature (Tm) and b) protein abundance (ppm). Error bars indicate 95 % confidence intervals. * indicates q-value < 0.05. c) Functional enrichment of proteins in KEGG pathways according to Tm. The Normalised Enrichment Score (NES) is shown on the y-axis. Pathways have been mapped to the 3 clusters defined in Section 4.1.

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

4.5 Rare variants are similar to common variants

Throughout the majority of analyses, performed both at the macroscopic and microscopic levels, the greatest segregation of data can be seen between common and disease-associated variants (see Figures 3-5). Rare variants show characteristics more similar to common variants, both in terms of the functional pathways they target, and in terms of the protein regions they localise to (core, surface and interface, order and disorder). If more stringent minor allele frequency (MAF) thresholds are used to define rare variants, their properties move towards those of disease-associated variants, but still remain closest to those of common variants (see Fig. 8 and Supplementary Fig. S14). A visible separation between common and rare variants, especially in the pathway analysis, can only be seen if an extreme MAF cutoff (<0.00001) is used.

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

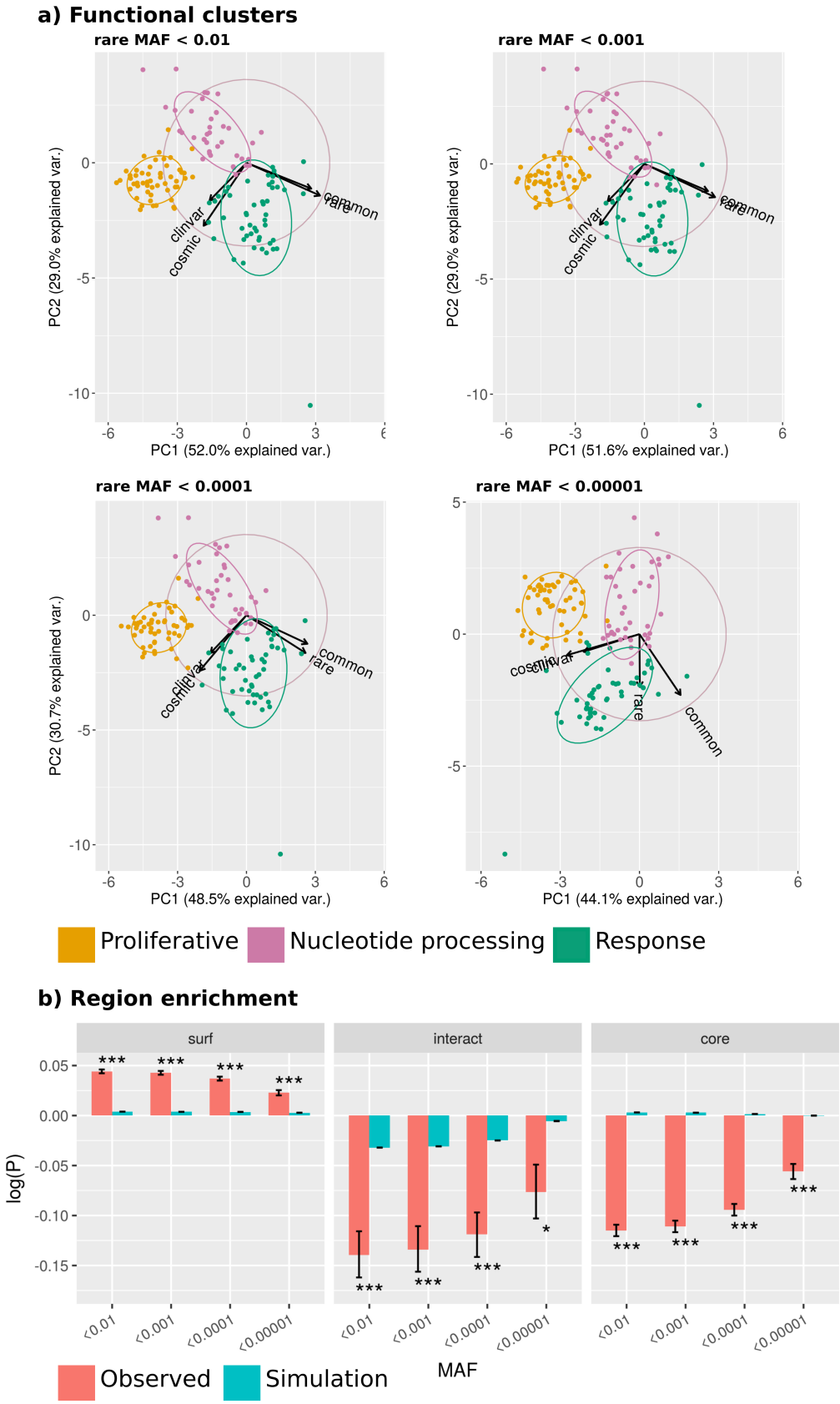

Figure 8: Rare variants are similar to common variants. (a) As for Fig. 3a, but with increasingly stringent minor allele frequencies (MAFs) used to define rare variants. (b) The localisation of rare variants to protein core, surface and interface regions. Rare variants have been defined using different MAF cut-offs as shown on the x-axes. Results for simulated null distributions are also shown.

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

5 Discussion and conclusions

Throughout this work, we show that SAVs in the general population, considered 'nominally healthy', show properties distinct from those in disease cohorts, both at the macroscopic (omics features and functional pathways) and microscopic levels (protein structural localisation). Additionally, although we uncover a spectrum in these properties of variants, which ranges from common population variants to disease-associated ClinVar variants, we find that the properties of rare variants remain close to those of common variants. These findings contrast with other observations [8], which suggest that common variants have more impact on molecular function than rare variants. Common variants appear closer in character to disease-associated variants than to rare variants, only for certain proteomics properties, such as the thermal stability and abundance of the targeted proteins, as discussed in Section 4.4. However, we consider these results inconclusive, due to the sparsity of the data. Alhuzimi et al. [9] suggest that the properties of genes enriched in rare population variants are similar to those enriched in disease-associated variants, and are thus good candidates for harbouring unknown disease associations. Instead, we show that such proteins are, from the annotated functional pathways, most similar to those enriched in common variants (Fig. 8). Moreover our results, which show that variants maintained within a population target functions which are mainly associated with response to the environment (Figs 3a and 4), agree with results from evolutionary studies reviewed in [86].

We have dissected the levels of variant enrichment in diverse datasets and across different protein levels (Fig. 2). Such a detailed anatomy of variant enrichment in health and disease provides a unique link between the cataloguing of mutations, and understanding both their mechanistic and functional effects. This supplies invaluable information to researchers studying specific proteins or domains, or focusing on proteins involved in a particular function (e.g. DNA binding; Fig. S3). By analysing the enrichment of variants in protein regions (core, surface, interface, disorder and disorder, PTM vicinity), we recapitulate trends observed by previous studies (e.g. in the comparison of oncogenes and TSGs; Fig. 5b) [16, 18, 17, 15, 75], but also shed light on the debate as to whether somatic cancer variants are enriched in interface regions, by simulating null-distributions of variants. The simulations we have performed show that it is essential to consider that variants from different datasets are not uniformly randomly distributed throughout the proteome. Through density-based metrics we find somatic cancer variants are not enriched in protein interfaces, however using a simulation-based approach we do find an enrichment (Fig. 5a). A similar simulation-based approach was taken by Gress et al., [15], but they found no significant enrichment for COSMIC variants in interface regions. Whilst they analysed a filtered set of mutations likely to play a driver role, we investigated all somatic variants and addressed separately mutations that localise to defined driver and non-driver genes. Our enrichment calculations were rigorous, and directly compared against null (n = 10,000) simulations to assess statistical significance. Throughout this analysis, we have, of course, been limited by the number of proteins with available structural data, although this has been enriched by considering homologous structures. We are also still limited by the structural coverage of protein interactions; although enough data exists to uncover broad trends, our analyses at a finer granularity, which probed protein-protein interaction sites, generally lacked statistical power. Moreover, it is likely that a more detailed picture will emerge if variant localisations to proteins involved in different classes of interactions are probed (e.g. transient vs permanent interactions). We envisage that the recent advances in cryo-EM [87], and the integration of structural data derived by a variety of techniques [88], will further increase the structural coverage of the protein-protein interaction network, enabling such finer-grained analyses in the future.

Our analysis at the macromolecular level, which probes associations between the enrichment of variants and proteomic features, is, to the best of our knowledge, unprecedented, and has only been made possible due to the recent release of large-scale proteomics data [25, 24, 26, 85]. We observe correlations which suggest an interplay between variant enrichment, protein abundance and thermal stability. First, disease-associated variants localise preferentially to proteins which are highly expressed and abundant (Fig. 7). These results complement a body of research which concludes that the rate of protein evolution correlates negatively with protein expression and abundance [89]. The extent of this anti-correlation has been found to be tissue-specific; those tissues with a high neuron density demonstrating the highest anti-correlation [90]. Consistent with this, we found the largest negative correlation for the protein-wise enrichment of rare variants, from the gnomAD dataset, with protein abundance in the brain, and, interestingly also in the ovary and testis, which both harbour long-lived germline progenitor cells (Fig. 7b; Fig. S7); purportedly the lifespan long-lived cells render them more sensitive to the toxicity of misfolded proteins. Second, we see a trend which suggests disease-associated variants preferentially localise to the core in less thermally stable proteins, most probably as these are more easily destabilised to an extent at which function is lost or impaired (Fig. 7a). Hence two competing trends emerge; variants which localise to less abundant proteins have greater disruptive potential, conversely, those which localise to thermally unstable proteins (which are normally less abundant [85]) may be able to deleteriously destabilise such proteins more easily. It is conceivable that the chemical nature of the particular missense variant plays an important role here: e.g. if a variant at the protein surface alters the "stickiness" of the protein and promotes non-specific interactions, this is likely to be most detrimental if the affected protein is present in great abundance. This highlights the importance of evaluating the interplay of macroscopic and microscopic features when estimating

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the potential impact of variants on protein function and stability.

The relationship between variant localisation and protein stability is of importance, as a number of algorithms have used the change in protein stability upon mutation ($\Delta\Delta$G) as a proxy for variant impact. Our results indicate that the baseline stability of the wild-type protein may also be important when considering the phenotypic relevance of a change in stability upon mutation. From their analysis of the ProTherm database, Serahijos et al. [91] found that mutations in more stable proteins generally led to greater destabilisation ($\Delta\Delta$G variation). They interpret this as suggesting that proteins which have evolved to become more stable are in a state closer to their peak stability, where any changes will result in drastic destabilisation. Similarly, Pucci and Rooman [92] used temperature dependent statistical potentials to investigate the thermal stability of the structurome (all proteins with resolved structure), and concluded that mutations in proteins which are highly thermally stable lead to a larger decrease in thermal stability, compared with those in less thermally stable proteins. We believe that our results point to the fact that, even under a scenario in which mutations in proteins with higher stability result in a greater change in stability, a mutation in an already unstable protein is more likely to result in complete/partial unfolding under physiological conditions. These factors should be brought into consideration when interpreting the impact of missense variants.

We show that greater insight into the properties of variants in health and disease can be obtained by combining protein structural and functional pathway information. For example, as discussed in Section 4.1, it can be clearly seen that population variants are most enriched on the surface of proteins which take part in pathways we have defined as belonging to the "proliferative" cluster (Fig. 3d). Moreover, pathways belonging to this cluster also appear to be enriched in proteins with less thermal stability (Fig. 7c), suggesting a possible mechanistic basis underlying the localisation of variants (variants tend to localise to the surface and avoid disrupting the core of these already unstable proteins). This indicates that the combinatorial use of such features may aid in both improving the prediction of a variant's impact on phenotype, and in assessing the molecular mechanisms underlying this.

Ultimately, the goal should reach beyond the identification of variants which underlie a disease phenotype, to the use of this information in the development of therapeutic strategies. Here we envisage that our domain-centric landscape of variant enrichment (Fig. 6), which includes the mapping of targeted drugs, besides providing another feature for the characterisation of variants, will allow for more informed decisions in selecting new therapeutic targets. We show that many domains are enriched in either COSMIC and/or ClinVar variants, but few or no drugs exist to target these proteins. This could offer a starting point to prioritise drug discovery efforts for these domain-types. For domain-types already targetable by drugs, our analysis highlight domains to which multiple disease-associated variants localise, which could give scope for drug repurposing or redesign. Additionally, targets with few population variants could be selected, to minimise differential drug response due to genetic differences between individuals.

In conclusion, our work highlights the complex interplay between different factors which may determine variant pathogenicity, at both the macroscopic and microscopic levels. We believe that these insights will prove important in the prediction of which variants drive disease phenotypes. Moreover, the ZoomVar database, which we have made available at `http://fraternalilab.kcl.ac.uk/ZoomVar`, will facilitate users in the structural analysis of variants, and provides precomputed data underlying all analyses presented here. Further advancement in the structural coverage of the proteome, and the exploitation of high throughput proteomics technologies, such as those pioneered by the Savitski and Picotti labs [26, 85], will ultimately offer a finer-grained picture of features which segregate variants in "health" and "disease".

6 Acknowledgements

This research was supported by the British Heart Foundation (RE/13/2/30182 to FF and AL), Croucher Foundation Hong Kong (to JCN) and the Medical Research Council (MR/L01257X/1 to FF).

References

[1] Adrián Blanco-Gómez, Sonia Castillo-Lluva, María Del Mar Sáez-Freire, Lourdes Hontecillas-Prieto, Jian Hua Mao, Andrés Castellanos-Martín, and Jesus Pérez-Losada. Missing heritability of complex diseases: Enlightenment by genetic variants from intermediate phenotypes. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 38(7):664–73, 07 2016.

[2] Santhosh Girirajan. Missing heritability and where to find it. *Genome biology*, 18(1):89, 05 2017.

[3] Wei-Feng Guo, Shao-Wu Zhang, Li-Li Liu, Fei Liu, Qian-Qian Shi, Lei Zhang, Ying Tang, Tao Zeng, and Luonan Chen. Discovering personalized driver mutation profiles of single samples in cancer by network control strategy. *Bioinformatics (Oxford, England)*, 34(11):1893–1903, Jun 2018.

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[4] Lorenzo Bomba, Klaudia Walter, and Nicole Soranzo. The impact of rare and low-frequency genetic variants in common disease. *Genome biology*, 18(1):77, 04 2017.

[5] Arun Prasad Pandurangan, David B Ascher, Sherine E Thomas, and Tom L Blundell. Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. *Biochemical Society transactions*, 45(2):303–311, 04 2017.

[6] Rong Chen, Lisong Shi, Jörg Hakenberg, Brian Naughton, Pamela Sklar, Jianguo Zhang, Hanlin Zhou, Lifeng Tian, Om Prakash, Mathieu Lemire, Patrick Sleiman, Wei-Yi Cheng, Wanting Chen, Hardik Shah, Yulan Shen, Menachem Fromer, Larsson Omberg, Matthew A Deardorff, Elaine Zackai, Jason R Bobe, Elissa Levin, Thomas J Hudson, Leif Groop, Jun Wang, Hakon Hakonarson, Anne Wojcicki, George A Diaz, Lisa Edelmann, Eric E Schadt, and Stephen H Friend. Analysis of 589,306 genomes identifies individuals resilient to severe mendelian childhood diseases. *Nature biotechnology*, 34(5):531–8, 05 2016.

[7] Nilah M Ioannidis, Joseph H Rothstein, Vikas Pejaver, Sumit Middha, Shannon K McDonnell, Saurabh Baheti, Anthony Musolf, Qing Li, Emily Holzinger, Danielle Karyadi, Lisa A Cannon-Albright, Craig C Teerlink, Janet L Stanford, William B Isaacs, Jianfeng Xu, Kathleen A Cooney, Ethan M Lange, Johanna Schleutker, John D Carpten, Isaac J Powell, Olivier Cussenot, Geraldine Cancel-Tassin, Graham G Giles, Robert J MacInnis, Christiane Maier, Chih-Lin Hsieh, Fredrik Wiklund, William J Catalona, William D Foulkes, Diptasri Mandal, Rosalind A Eeles, Zsofia Kote-Jarai, Carlos D Bustamante, Daniel J Schaid, Trevor Hastie, Elaine A Ostrander, Joan E Bailey-Wilson, Predrag Radivojac, Stephen N Thibodeau, Alice S Whittemore, and Weiva Sieh. Revel: An ensemble method for predicting the pathogenicity of rare missense variants. *American journal of human genetics*, 99(4):877–885, Oct 2016.

[8] Yannick Mahlich, Jonas Reeb, Maximilian Hecht, Maria Schelling, Tjaart Andries Petrus De Beer, Yana Bromberg, and Burkhard Rost. Common sequence variants affect molecular function more than rare variants? *Scientific reports*, 7(1):1608, May 2017.

[9] Eman Alhuzimi, Luis G Leal, Michael J E Sternberg, and Alessia David. Properties of human genes guided by their enrichment in rare and common variants. *Human mutation*, 39(3):365–370, Mar 2018.

[10] Line Lykke Andersen, Ewa Terczyńska-Dyla, Nanna Mørk, Carsten Scavenius, Jan J Enghild, Klara Höning, Veit Hornung, Mette Christiansen, Trine H Mogensen, and Rune Hartmann. Frequently used bioinformatics tools overestimate the damaging effect of allelic variants. *Genes and immunity*, Dec 2017.

[11] M Miller, Y Bromberg, and L Swint-Kruse. Computational predictors fail to identify amino acid substitution effects at rheostat positions. *Scientific reports*, 7:41329, Jan 2017.

[12] Luisa Azevedo, Matthew Mort, Antonio C Costa, Raquel M Silva, Dulce Quelhas, Antonio Amorim, and David N Cooper. Improving the in silico assessment of pathogenicity for compensated variants. *European journal of human genetics : EJHG*, 25(1):2–7, 01 2016.

[13] Anna Laddach, Joseph Chi-Fung Ng, Sun Sook Chung, and Franca Fraternali. Genetic variants and protein-protein interactions: a multidimensional network-centric view. *Current opinion in structural biology*, 50:82–90, Jan 2018.

[14] Hui-Chun Lu, Julián Herrera Braga, and Franca Fraternali. Pinsnps: structural and functional analysis of snps in the context of protein interaction networks. *Bioinformatics (Oxford, England)*, 32(16):2534–6, 08 2016.

[15] A Gress, V Ramensky, and O V Kalinina. Spatial distribution of disease-associated variants in three-dimensional structures of protein complexes. *Oncogenesis*, 6(9):e380, Sep 2017.

[16] H Billur Engin, Jason F Kreisberg, and Hannah Carter. Structure-based analysis reveals cancer missense mutations target protein interaction interfaces. *PloS one*, 11(4):e0152929, 2016.

[17] Mu Gao, Hongyi Zhou, and Jeffrey Skolnick. Insights into disease-associated mutations in the human proteome through protein structural analysis. *Structure (London, England : 1993)*, 23(7):1362–9, Jul 2015.

[18] Alessia David, Rozami Razali, Mark N Wass, and Michael J E Sternberg. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous snps. *Human mutation*, 33(2):359–63, Feb 2012.

[19] Douglas E V Pires, David B Ascher, and Tom L Blundell. mcsm: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics (Oxford, England)*, 30(3):335–42, Feb 2014.

[20] Eduard Porta-Pardo and Adam Godzik. e-driver: a novel method to identify protein regions driving cancer. *Bioinformatics (Oxford, England)*, 30(21):3109–14, Nov 2014.

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[21] Eduard Porta-Pardo, Thomas Hrabe, and Adam Godzik. Cancer3d: understanding cancer mutations through protein structures. *Nucleic acids research*, 43(Database issue):D968–73, Jan 2015.

[22] Matthew H Bailey, Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C Wendl, Jaegil Kim, Brendan Reardon, Patrick Kwok-Shing Ng, Kang Jin Jeong, Song Cao, Zixing Wang, Jianjiong Gao, Qingsong Gao, Fang Wang, Eric Minwei Liu, Loris Mularoni, Carlota Rubio-Perez, Niranjan Nagarajan, Isidro Cortés-Ciriano, Daniel Cui Zhou, Wen-Wei Liang, Julian M Hess, Venkata D Yellapantula, David Tamborero, Abel Gonzalez-Perez, Chayaporn Suphavilai, Jia Yu Ko, Ekta Khurana, Peter J Park, Eliezer M Van Allen, Han Liang, Michael S Lawrence, Adam Godzik, Nuria Lopez-Bigas, Josh Stuart, David Wheeler, Gad Getz, Ken Chen, Alexander J Lazar, Gordon B Mills, Rachel Karchin, and Li Ding. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385.e18, Apr 2018.

[23] R Michael Sivley, Xiaoyi Dou, Jens Meiler, William S Bush, and John A Capra. Comprehensive analysis of constraint on the spatial distribution of missense variants in human protein structures. *American journal of human genetics*, 102(3):415–426, Mar 2018.

[24] Holger Franken, Toby Mathieson, Dorothee Childs, Gavain M A Sweetman, Thilo Werner, Ina Tögel, Carola Doce, Stephan Gade, Marcus Bantscheff, Gerard Drewes, Friedrich B M Reinhard, Wolfgang Huber, and Mikhail M Savitski. Thermal proteome profiling for unbiased identification of direct and indirect drug targets using multiplexed quantitative mass spectrometry. *Nature protocols*, 10(10):1567–93, Oct 2015.

[25] Mingcong Wang, Christina J Herrmann, Milan Simonovic, Damian Szklarczyk, and Christian von Mering. Version 4.0 of paxdb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*, 15(18):3163–8, Sep 2015.

[26] Toby Mathieson, Holger Franken, Jan Kosinski, Nils Kurzawa, Nico Zinn, Gavain Sweetman, Daniel Poeckel, Vikram S Ratnu, Maike Schramm, Isabelle Becher, Michael Steidel, Kyung-Min Noh, Giovanna Bergamini, Martin Beck, Marcus Bantscheff, and Mikhail M Savitski. Systematic analysis of protein turnover in primary cells. *Nature communications*, 9(1):689, 02 2018.

[27] GTEx Consortium. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–5, Jun 2013.

[28] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, Wonhee Jang, Kenneth Katz, Michael Ovetsky, George Riley, Amanjeev Sethi, Ray Tully, Ricardo Villamarin-Salomon, Wendy Rubinstein, and Donna R Maglott. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, 44(D1):D862–8, Jan 2016.

[29] Simon A Forbes, David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, Sally Bamford, Charlotte Cole, Sari Ward, Chai Yin Kok, Mingming Jia, Tisham De, Jon W Teague, Michael R Stratton, Ultan McDermott, and Peter J Campbell. Cosmic: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research*, 43(Database issue):D805–11, Jan 2015.

[30] Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, Taru Tukiainen, Daniel P Birnbaum, Jack A Kosmicki, Laramie E Duncan, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, Joanne Berghout, David N Cooper, Nicole Deflaux, Mark DePristo, Ron Do, Jason Flannick, Menachem Fromer, Laura Gauthier, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, Mitja I Kurki, Ami Levy Moonshine, Pradeep Natarajan, Lorena Orozco, Gina M Peloso, Ryan Poplin, Manuel A Rivas, Valentin Ruano-Rubio, Samuel A Rose, Douglas M Ruderfer, Khalid Shakir, Peter D Stenson, Christine Stevens, Brett P Thomas, Grace Tiao, Maria T Tusie-Luna, Ben Weisburd, Hong-Hee Won, Dongmei Yu, David M Altshuler, Diego Ardissino, Michael Boehnke, John Danesh, Stacey Donnelly, Roberto Elosua, Jose C Florez, Stacey B Gabriel, Gad Getz, Stephen J Glatt, Christina M Hultman, Sekar Kathiresan, Markku Laakso, Steven McCarroll, Mark I McCarthy, Dermot McGovern, Ruth McPherson, Benjamin M Neale, Aarno Palotie, Shaun M Purcell, Danish Saleheen, Jeremiah M Scharf, Pamela Sklar, Patrick F Sullivan, Jaakko Tuomilehto, Ming T Tsuang, Hugh C Watkins, James G Wilson, Mark J Daly, and Daniel G MacArthur. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–91, 08 2016.

[31] Bronwen L Aken, Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, Konstantinos Billis, Carlos García Girón, Thibaut Hourlier, Kevin Howe, Andreas Kähäri, Felix Kokocinski, Fergal J Martin, Daniel N Murphy, Rishi Nag, Magali Ruffier, Michael Schuster, Y Amy Tang, Jan-Hinnerk Vogel, Simon White, Amonida Zadissa, Paul Flicek, and Stephen M J Searle. The ensembl gene annotation system. *Database : the journal of biological databases and curation*, 2016, 2016.

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[32] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham R S Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome biology*, 17(1):122, 06 2016.

[33] Sun Sook Chung, Anna Laddach, N. Shaun Bevan Thomas, and Franca Fraternali. Short loop motif profiling of protein interaction networks in acute myeloid leukaemia. *bioRxiv*, 2018.

[34] Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H Campbell, Gayatri Chavali, Carol Chen, Noemi del Toro, Margaret Duesbury, Marine Dumousseau, Eugenia Galeota, Ursula Hinz, Marta Iannuccelli, Sruthi Jagannathan, Rafael Jimenez, Jyoti Khadake, Astrid Lagreid, Luana Licata, Ruth C Lovering, Birgit Meldal, Anna N Melidoni, Mila Milagros, Daniele Peluso, Livia Perfetto, Pablo Porras, Arathi Raghunath, Sylvie Ricard-Blum, Bernd Roechert, Andre Stutz, Michael Tognolli, Kim van Roey, Gianni Cesareni, and Henning Hermjakob. The mintact project–intact as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, 42(Database issue):D358–63, Jan 2014.

[35] Andrew Chatr-Aryamontri, Rose Oughtred, Lorrie Boucher, Jennifer Rust, Christie Chang, Nadine K Kolas, Lara O'Donnell, Sara Oster, Chandra Theesfeld, Adnane Sellam, Chris Stark, Bobby-Joe Breitkreutz, Kara Dolinski, and Mike Tyers. The biogrid interaction database: 2017 update. *Nucleic acids research*, 45(D1):D369–D379, Jan 2017.

[36] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, Michael Kuhn, Peer Bork, Lars J Jensen, and Christian von Mering. String v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(Database issue):D447–52, Jan 2015.

[37] Ioannis Xenarios, Lukasz Salwínski, Xiaoqun Joyce Duan, Patrick Higney, Sul-Min Kim, and David Eisenberg. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic acids research*, 30(1):303–5, Jan 2002.

[38] Suraj Peri, J Daniel Navarro, Troels Z Kristiansen, Ramars Amanchy, Vineeth Surendranath, Babylakshmi Muthusamy, T K B Gandhi, K N Chandrika, Nandan Deshpande, Shubha Suresh, B P Rashmi, K Shanker, N Padma, Vidya Niranjan, H C Harsha, Naveen Talreja, B M Vrushabendra, M A Ramya, A J Yatish, Mary Joy, H N Shivashankar, M P Kavitha, Minal Menezes, Dipanwita Roy Choudhury, Neelanjana Ghosh, R Saravana, Sreenath Chandran, Sujatha Mohan, Chandra Kiran Jonnalagadda, C K Prasad, Chandan Kumar-Sinha, Krishna S Deshpande, and Akhilesh Pandey. Human protein reference database as a discovery resource for proteomics. *Nucleic acids research*, 32(Database issue):D497–501, Jan 2004.

[39] Pierre C Havugimana, G Traver Hart, Tamás Nepusz, Haixuan Yang, Andrei L Turinsky, Zhihua Li, Peggy I Wang, Daniel R Boutz, Vincent Fong, Sadhna Phanse, Mohan Babu, Stephanie A Craig, Pingzhao Hu, Cuihong Wan, James Vlasblom, Vaqaar un Nisa Dar, Alexandr Bezginov, Gregory W Clark, Gabriel C Wu, Shoshana J Wodak, Elisabeth R M Tillier, Alberto Paccanaro, Edward M Marcotte, and Andrew Emili. A census of human soluble protein complexes. *Cell*, 150(5):1068–81, Aug 2012.

[40] Thomas Rolland, Murat Taşan, Benoit Charloteaux, Samuel J Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, Irma Lemmens, Celia Fontanillo, Roberto Mosca, Atanas Kamburov, Susan D Ghiassian, Xinping Yang, Lila Ghamsari, Dawit Balcha, Bridget E Begg, Pascal Braun, Marc Brehme, Martin P Broly, Anne-Ruxandra Carvunis, Dan Convery-Zupan, Roser Corominas, Jasmin Coulombe-Huntington, Elizabeth Dann, Matija Dreze, Amélie Dricot, Changyu Fan, Eric Franzosa, Fana Gebreab, Bryan J Gutierrez, Madeleine F Hardy, Mike Jin, Shuli Kang, Ruth Kiros, Guan Ning Lin, Katja Luck, Andrew MacWilliams, Jörg Menche, Ryan R Murray, Alexandre Palagi, Matthew M Poulin, Xavier Rambout, John Rasla, Patrick Reichert, Viviana Romero, Elien Ruyssinck, Julie M Sahalie, Annemarie Scholz, Akash A Shah, Amitabh Sharma, Yun Shen, Kerstin Spirohn, Stanley Tam, Alexander O Tejeda, Shelly A Trigg, Jean-Claude Twizere, Kerwin Vega, Jennifer Walsh, Michael E Cusick, Yu Xia, Albert-László Barabási, Lilia M Iakoucheva, Patrick Aloy, Javier De Las Rivas, Jan Tavernier, Michael A Calderwood, David E Hill, Tong Hao, Frederick P Roth, and Marc Vidal. A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226, Nov 2014.

[41] Edward L Huttlin, Lily Ting, Raphael J Bruckner, Fana Gebreab, Melanie P Gygi, John Szpyt, Stanley Tam, Gabriela Zarraga, Greg Colby, Kurt Baltier, Rui Dong, Virginia Guarani, Laura Pontano Vaites, Alban Ordureau, Ramin Rad, Brian K Erickson, Martin Wühr, Joel Chick, Bo Zhai, Deepak Kolippakkam, Julian Mintseris, Robert A Obar, Tim Harris, Spyros Artavanis-Tsakonas, Mathew E Sowa, Pietro De Camilli, Joao A Paulo, J Wade Harper, and Steven P Gygi. The bioplex network: A systematic exploration of the human interactome. *Cell*, 162(2):425–440, Jul 2015.

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[42] Sylvain Poux, Cecilia N Arighi, Michele Magrane, Alex Bateman, Chih-Hsuan Wei, Zhiyong Lu, Emmanuel Boutet, Hema Bye-A-Jee, Maria Livia Famiglietti, Bernd Roechert, and The UniProt Consortium. On expert curation and scalability: Uniprotkb/swiss-prot as a case study. *Bioinformatics (Oxford, England)*, 33(21):3454–3460, Nov 2017.

[43] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–50, Oct 2005.

[44] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *Science (New York, N.Y.)*, 339(6127):1546–58, Mar 2013.

[45] Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21:3439–3440, 2005.

[46] Juan M Vaquerizas, Sarah K Kummerfeld, Sarah A Teichmann, and Nicholas M Luscombe. A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics*, 10(4):252–63, 04 2009.

[47] Robert D Finn, Teresa K Attwood, Patricia C Babbitt, Alex Bateman, Peer Bork, Alan J Bridge, Hsin-Yu Chang, Zsuzsanna Dosztányi, Sara El-Gebali, Matthew Fraser, Julian Gough, David Haft, Gemma L Holliday, Hongzhan Huang, Xiaosong Huang, Ivica Letunic, Rodrigo Lopez, Shennan Lu, Aron Marchler-Bauer, Huaiyu Mi, Jaina Mistry, Darren A Natale, Marco Necci, Gift Nuka, Christine A Orengo, Youngmi Park, Sebastien Pesseat, Damiano Piovesan, Simon C Potter, Neil D Rawlings, Nicole Redaschi, Lorna Richardson, Catherine Rivoire, Amaia Sangrador-Vegas, Christian Sigrist, Ian Sillitoe, Ben Smithers, Silvano Squizzato, Granger Sutton, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, Ioannis Xenarios, Lai-Su Yeh, Siew-Yit Young, and Alex L Mitchell. Interpro in 2017-beyond protein family and domain annotations. *Nucleic acids research*, 45(D1):D190–D199, Jan 2017.

[48] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, Jan 2018.

[49] Helen Berman, Kim Henrick, and Haruki Nakamura. Announcing the worldwide protein data bank. *Nature structural biology*, 10(12):980, Dec 2003.

[50] S F Altschul, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–402, Sep 1997.

[51] Robert D Finn, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Jaina Mistry, Alex L Mitchell, Simon C Potter, Marco Punta, Matloob Qureshi, Amaia Sangrador-Vegas, Gustavo A Salazar, John Tate, and Alex Bateman. The pfam protein families database: towards a more sustainable future. *Nucleic acids research*, 44(D1):D279–85, Jan 2016.

[52] Robert D Finn, Jody Clements, and Sean R Eddy. Hmmer web server: interactive sequence similarity searching. *Nucleic acids research*, 39(Web Server issue):W29–37, Jul 2011.

[53] C Notredame, D G Higgins, and J Heringa. T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–17, Sep 2000.

[54] Luigi Cavallo, Jens Kleinjung, and Franca Fraternali. Pops: A fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic acids research*, 31(13):3364–6, Jul 2003.

[55] E W Myers and W Miller. Optimal alignments in linear space. *Computer applications in the biosciences : CABIOS*, 4(1):11–7, Mar 1988.

[56] Li C Xue, Drena Dobbs, and Vasant Honavar. Homppi: a class of sequence homology based protein-protein interface prediction methods. *BMC bioinformatics*, 12:244, Jun 2011.

[57] Jens Kleinjung and Franca Fraternali. Popscomp: an automated interaction analysis of biomolecular complexes. *Nucleic acids research*, 33(Web Server issue):W342–6, Jul 2005.

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[58] David T Jones and Domenico Cozzetto. Disopred3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics (Oxford, England)*, 31(6):857–63, Mar 2015.

[59] Peter V Hornbeck, Indy Chabra, Jon M Kornhauser, Elzbieta Skrzypek, and Bin Zhang. Phosphosite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, 4(6):1551–61, Jun 2004.

[60] AB MySQL. *Mysql 5.1 reference manual*, 2008.

[61] Django. `http://djangoproject.com`.

[62] Bethan Yates, Bryony Braschi, Kristian A Gray, Ruth L Seal, Susan Tweedie, and Elspeth A Bruford. Genenames.org: the hgnc and vgnc resources in 2017. *Nucleic acids research*, 45(D1):D619–D625, Jan 2017.

[63] Alexey Sergushichev. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv*, 2016.

[64] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1):D353–D361, 01 2017.

[65] Malika Charrad, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(6):1–36, 2014.

[66] Ian Sillitoe, Tony E Lewis, Alison Cuff, Sayoni Das, Paul Ashford, Natalie L Dawson, Nicholas Furnham, Roman A Laskowski, David Lee, Jonathan G Lees, Sonja Lehtinen, Romain A Studer, Janet Thornton, and Christine A Orengo. Cath: comprehensive structural and functional annotations for genome sequences. *Nucleic acids research*, 43(Database issue):D376–81, Jan 2015.

[67] Sameer Velankar, José M Dana, Julius Jacobsen, Glen van Ginkel, Paul J Gane, Jie Luo, Thomas J Oldfield, Claire O'Donovan, Maria-Jesus Martin, and Gerard J Kleywegt. Sifts: Structure integration with function, taxonomy and sequences resource. *Nucleic acids research*, 41(Database issue):D483–9, Jan 2013.

[68] Angelo Canty and B. D. Ripley. *boot: Bootstrap R (S-Plus) Functions*, 2017. R package version 1.3-20.

[69] Andri Signorell et mult. al. *DescTools: Tools for Descriptive Statistics*, 2017. R package version 0.99.19.

[70] Gregory R. Warnes, Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz, and Bill Venables. *gplots: Various R Programming Tools for Plotting Data*, 2016. R package version 3.0.1.

[71] Zuguang Gu, Roland Eils, and Matthias Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 2016.

[72] Martin Krzywinski, Jacqueline Schein, Inanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–45, Sep 2009.

[73] Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

[74] Eduard Porta-Pardo, Luz Garcia-Alonso, Thomas Hrabe, Joaquin Dopazo, and Adam Godzik. A pan-cancer catalogue of cancer driver protein interaction interfaces. *PLoS computational biology*, 11(10):e1004518, Oct 2015.

[75] Henning Stehr, Seon-Hi J Jang, José M Duarte, Christoph Wierling, Hans Lehrach, Michael Lappe, and Bodo M H Lange. The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. *Molecular cancer*, 10:54, May 2011.

[76] Jüri Reimand, Omar Wagih, and Gary D Bader. The mutational landscape of phosphorylation signaling in cancer. *Scientific reports*, 3:2651, Oct 2013.

[77] Aleksandra Olow, Zhongzhong Chen, R Hannes Niedner, Denise M Wolf, Christina Yau, Aleksandr Pankov, Evelyn Pei Rong Lee, Lamorna Brown-Swigart, Laura J van 't Veer, and Jean-Philippe Coppé. An atlas of the human kinome reveals the mutational landscape underlying dysregulated phosphorylation cascades in cancer. *Cancer research*, 76(7):1733–45, 04 2016.

[78] D Menzies and H Ellis. The role of plasminogen activator in adhesion prevention. *Surgery, gynecology and obstetrics*, 172(5):362–6, May 1991.

bioRxiv preprint doi: https://doi.org/10.1101/512764; this version posted January 9, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[79] Manoj Garg, Glenn Braunstein, and Harold Phillip Koeffler. Lamc2 as a therapeutic target for cancers. *Expert opinion on therapeutic targets*, 18(9):979–82, Sep 2014.

[80] Family: Npip (pf06409). `https://pfam.xfam.org/family/PF06409`. [Online; accessed 13-Mar-2018].

[81] Family: Nut (pf12881). `https://pfam.xfam.org/family/PF12881`. [Online; accessed 13-Mar-2018].

[82] Gregg L Semenza. Vhl and p53: tumor suppressors team up to prevent cancer. *Molecular cell*, 22(4):437–9, May 2006.

[83] Rita Santos, Oleg Ursu, Anna Gaulton, A Patrícia Bento, Ramesh S Donadi, Cristian G Bologa, Anneli Karlsson, Bissan Al-Lazikani, Anne Hersey, Tudor I Oprea, and John P Overington. A comprehensive map of molecular drug targets. *Nature reviews. Drug discovery*, 16(1):19–34, 01 2017.

[84] Alexander S Hauser, Sreenivas Chavali, Ikuo Masuho, Leonie J Jahn, Kirill A Martemyanov, David E Gloriam, and M Madan Babu. Pharmacogenomics of gpcr drug targets. *Cell*, 172(1-2):41–54.e19, Jan 2018.

[85] Pascal Leuenberger, Stefan Ganscha, Abdullah Kahraman, Valentina Cappelletti, Paul J Boersema, Christian von Mering, Manfred Claassen, and Paola Picotti. Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. *Science (New York, N.Y.)*, 355(6327), 02 2017.

[86] Lluis Quintana-Murci. Understanding rare and common diseases in the context of human evolution. *Genome biology*, 17(1):225, 11 2016.

[87] Igor Orlov, Alexander G Myasnikov, Leonid Andronov, S Kundhavai Natchiar, Heena Khatter, Brice Beinsteiner, Jean-François Ménétret, Isabelle Hazemann, Kareem Mohideen, Karima Tazibt, Rachel Tabaroni, Hanna Kratzat, Nadia Djabeur, Tatiana Bruxelles, Finaritra Raivoniaina, Lorenza di Pompeo, Morgan Torchy, Isabelle Billas, Alexandre Urzhumtsev, and Bruno P Klaholz. The integrative role of cryo electron microscopy in molecular and cellular structural biology. *Biology of the cell*, 109(2):81–93, Feb 2017.

[88] Stephen K Burley, Genji Kurisu, John L Markley, Haruki Nakamura, Sameer Velankar, Helen M Berman, Andrej Sali, Torsten Schwede, and Jill Trewhella. Pdb-dev: a prototype system for depositing integrative/hybrid structural models. *Structure (London, England : 1993)*, 25(9):1317–1318, 09 2017.

[89] Jianzhi Zhang and Jian-Rong Yang. Determinants of the rate of protein sequence evolution. *Nature reviews. Genetics*, 16(7):409–20, Jul 2015.

[90] D Allan Drummond and Claus O Wilke. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2):341–52, Jul 2008.

[91] Adrian W R Serohijos, Zilvinas Rimas, and Eugene I Shakhnovich. Protein biophysics explains why highly abundant proteins evolve slowly. *Cell reports*, 2(2):249–56, Aug 2012.

[92] Fabrizio Pucci and Marianne Rooman. Improved insights into protein thermal stability: from the molecular to the structurome scale. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2080), Nov 2016.