Human Missense Variation is Constrained by Domain Structure and Highlights Functional and Pathogenic Residues

MacGowan, Stuart; Madeira, Fabio; Britto Borges, Thiago; Schmittner, Melanie S.; Cole, Christian; Barton, Geoffrey

Published in:
BioRxiv

DOI:
10.1101/127050

Publication date:
2017

Document Version
Early version, also known as pre-print

Human Missense Variation is Constrained by Domain Structure and Highlights Functional and Pathogenic Residues

Stuart A. MacGowan[1,2], Fábio Madeira[1], Thiago Britto-Borges[1], Melanie S. Schmittner[1], Christian Cole[1] and Geoffrey J. Barton[1,2]

Human genome sequencing has generated population variant datasets containing millions of variants from hundreds of thousands of individuals[1-3]. The datasets show the genomic distribution of genetic variation to be influenced on genic and sub-genic scales by gene essentiality,[1,4,5] protein domain architecture[6] and the presence of genomic features such as splice donor/acceptor sites.[2] However, the variant data are still too sparse to provide a comparative picture of genetic variation between individual protein residues in the proteome.[1,6] Here, we overcome this sparsity for ~25,000 human protein domains in 1,291 domain families by aggregating variants over equivalent positions (columns) in multiple sequence alignments of sequence-similar (paralagous) domains[7,8]. We then compare the resulting variation profiles from the human population to residue conservation across all species[9] and find that the same tertiary structural and functional pressures that affect amino acid conservation during domain evolution constrain missense variant distributions. Thus, depletion of missense variants at a position implies that it is structurally or functionally important. We find such positions are enriched in known disease-associated variants (OR = 2.83, $p \approx 0$) while positions that are both missense depleted and evolutionary conserved are further enriched in disease-associated variants (OR = 1.85, $p = 3.3 \times 10^{-17}$) compared to those

that are only evolutionary conserved (OR = 1.29, $p = 4.5 \times 10^{-19}$). Unexpectedly, a subset of evolutionary Unconserved positions are Missense Depleted in human (UMD positions) and these are also enriched in pathogenic variants (OR = 1.74, $p = 0.02$). UMD positions are further differentiated from other unconserved residues in that they are enriched in ligand, DNA and protein binding interactions (OR = 1.59, $p = 0.003$), which suggests this stratification can identify functionally important positions. A different class of positions that are Conserved and Missense Enriched (CME) show an enrichment of ClinVar risk factor variants (OR = 2.27, $p = 0.004$). We illustrate these principles with the G-Protein Coupled Receptor (GPCR) family, Nuclear Receptor Ligand Binding Domain family and In Between Ring-Finger (IBR) domains and list a total of 343 UMD positions in 211 domain families. This study will have broad applications to: (a) providing focus for functional studies of specific proteins by mutagenesis; (b) refining pathogenicity prediction models; (c) highlighting which residue interactions to target when refining the specificity of small-molecule drugs.

Variant densities and the sparsity problem

Human sequencing projects are beginning to shed light on the patterns of genetic variation that are present in human populations.[1,2] One way in which these studies enhance the understanding of inter-individual variation is by characterising different densities of single-nucleotide variants (SNVs) and short insertion and deletions (indels) at different genomic loci. Analysis of large cohort variation datasets has revealed that genes differ in their tolerance of non-synonymous and loss-of-function variation.[1,4] Within protein-coding genes, regions that encode protein domains are less tolerant of non-synonymous variants than inter-domain coding regions and are more prone to

disease variants.[6] The 60,706 sample Exome Aggregation Consortium[1] study yielded ~125 variants per kilobase, rendering a per nucleotide comparison impossible since most single nucleotides have zero variants. Variant sparsity can also be addressed by aggregating over pseudo-paralogous positions. For example, aligning nucleotide sequences on start codons reveals that start codons have fewer variants than adjacent sites, while the 5'-UTR is more variable than the CDS and every third base in a codon variable.[2] These differences are observed because the pressures imposed by those genomic features are common to each individual aligned sequence.

Residue resolution through protein family aggregation

Multiple sequence alignments (MSA) are a well established way to identify position-specific features in a family of homologous sequences. Figure 1A illustrates schematically how an MSA containing multiple human paralogs can be used to aggregate SNVs from multiple loci in a position specific manner. This process condenses the sparse variant counts from single sequences into dense variant counts for the domain family. Similar approaches have been adopted to identify low frequency cancer driver mutations,[10-12] and find sites in domains where pathogenic mutations cluster.[13] To perform a comprehensive analysis of protein domains, germline variation data retrieved from Ensembl[14,15] was aggregated with respect to the domain families in Pfam.[8] Pfam contains 16,035 domain families and of these families 6,088 contain at least one human sequence and 1,376 have at least five after adjusting for duplicate sequences (see Methods). Figures 1B-C show that even though most human sequence residues in Pfam domains have zero variants, after aggregation most Pfam domain family positions have at least two variants.

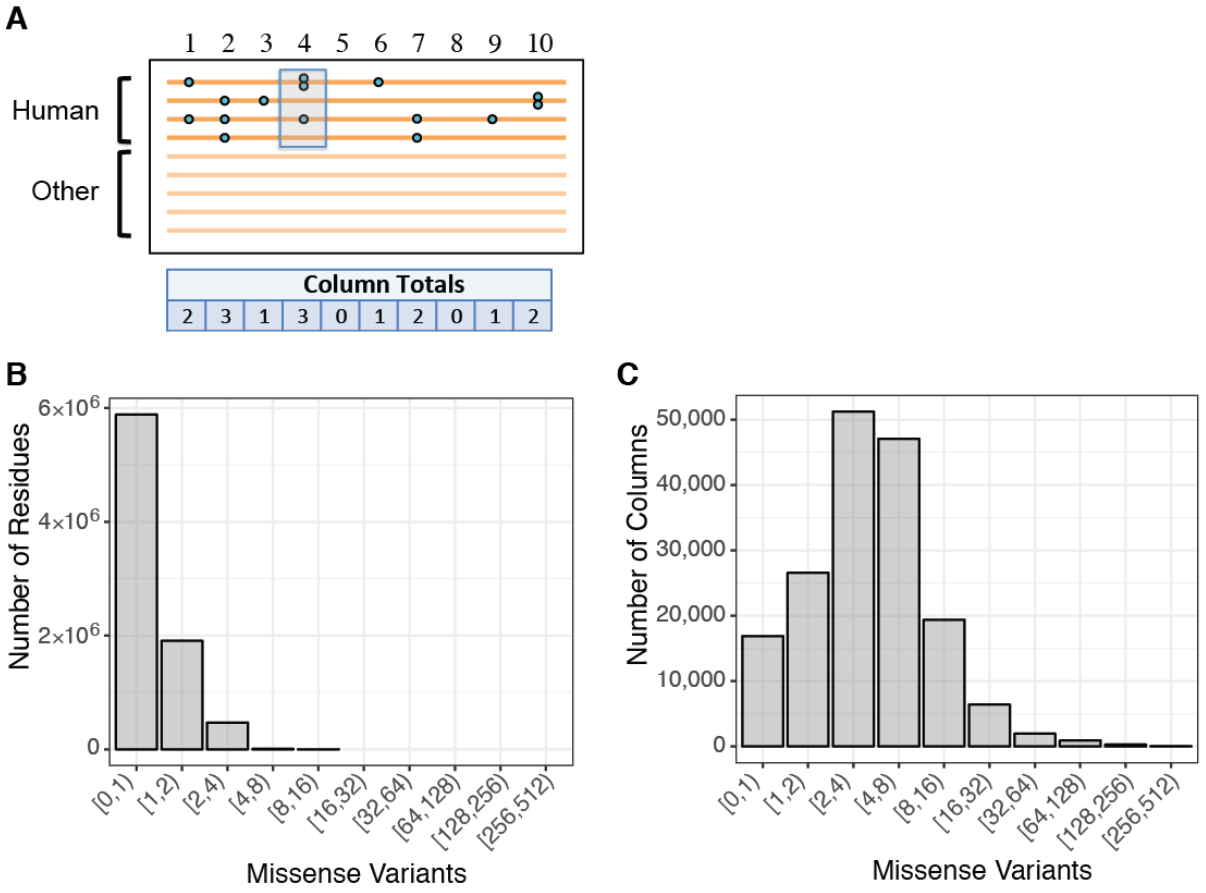


Figure 1: Variant aggregation over protein family alignments. A. Schematic illustration of a protein family alignment. Each line represents a human or non-human sequence and human sequences can have zero or more variants (blue circles). Few variants are observed at each alignment position per sequence but the column totals are larger. B. Distribution of variants per human residue in all Pfam sequences (2,927,499 missense variants, 8,264,091 residues; no filters applied). C. Distribution of variants per alignment column in Pfam alignments (955,636 missense variants, 159,296 columns; includes only columns with at least five human residues).

SNV density is correlated with evolutionary conservation

Accurate predictions of structure and function can be made from MSAs[16-18] because these features impose constraints on accepted mutations in domain families. These constraints can be inferred from patterns in residue conservation scores,[9] which quantify the extent of residue or physicochemical property conservation at each position in the alignment. In protein domain family MSAs, which can contain orthologs

and paralogs in varying proportions, these scores are interpreted as the degree of evolutionary conservation in each site of the domain family and are different to conservation scores for alignments that contain only closely related orthologs because of greater functional divergence. Throughout this text, the term evolutionary conservation refers to the conservation of residues during domain family evolution and accounts for orthologous and paralogous evolutionary process as captured in the Pfam alignments.

Figure 2A shows the correlation between the domain family column variant counts and the Shenkin divergence score ($V_{Shenkin}$)[19] in the SH2 domain family (PF00017). The number of missense variants increases with increasing residue divergence (i.e., decreasing conservation) whilst the frequency of synonymous variation remains constant with respect to column conservation. Extended Data Figs. 1 and 2 illustrate this behaviour on the SH2 alignment and crystal structure and show that in this example, the protein's secondary and tertiary structures and domain-domain interactions are common factors constraining both conservation and population constraint. This demonstrates that the missense variant distribution is subject to the same structural and functional constraints over generational timescales that affect amino acid substitution frequencies over evolutionary timescales. In contrast, the distribution of synonymous variation is not affected because these variants are silent at the protein structure level. Figure 2B shows that this result extends to other protein families by illustrating that the $V_{Shenkin}$ regression coefficients for each family are distributed around zero for synonymous variant totals and are typically positive for missense variants.

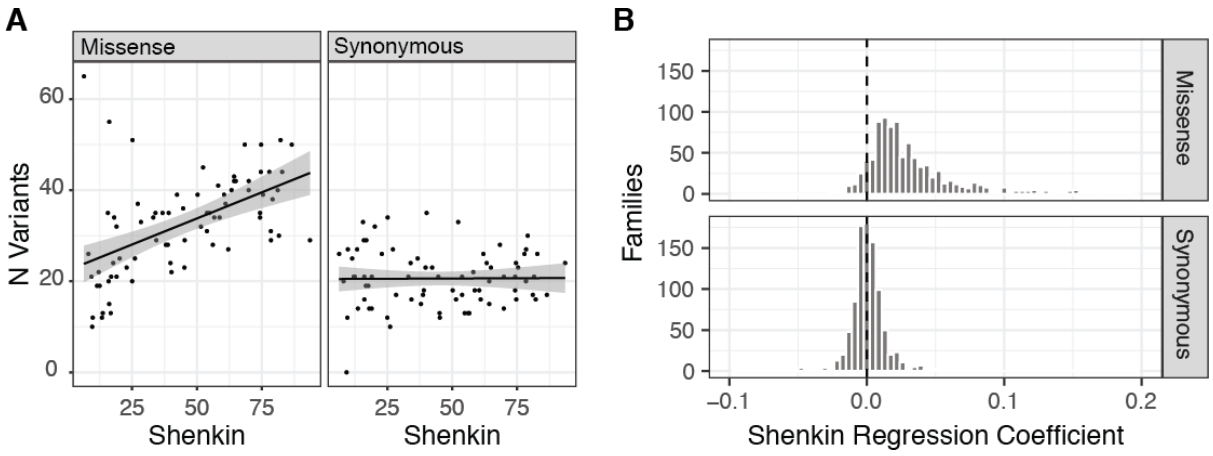


Figure 2: Relationship between column variant totals and $V_{Shenkin}$. A. Variant counts vs. $V_{Shenkin}$ for missense (left panel) and synonymous variants (right panel) for the SH2 domain (PF00017). The regression lines show least-squares fits and the shaded regions indicate standard errors of prediction. B. Histograms showing the distributions of $V_{Shenkin}$ regression coefficients for linear models fitting column variant totals to $V_{Shenkin}$ and column human residue occupancies for protein families with > 50 included alignment columns (n = 934).

Properties of sites relatively depleted or enriched for missense variation

Domain family alignment columns were classified as missense depleted or missense enriched by testing whether a column possessed significantly more or less missense variation than observed elsewhere in the alignment (see Methods). Figure 3A shows that with respect to ClinVar[20] variant annotations missense depleted columns have higher rates of 'pathogenic' (Fisher OR = 2.83, $p \approx 0$) and 'likely pathogenic' variants (OR = 2.17, $p = 1.9 \times 10^{-12}$) compared to other sites, indicating that diversity is suppressed in positions that are critical for function. Variant enriched columns possess proportionally more 'risk factor' variants (Fisher OR = 1.66, $p = 0.017$). This may suggest that there is generally an increased chance of co-segregating phenotypic differences at sites with relatively high population diversity.

For comparison, Figure 3B shows the equivalent ClinVar association tests for columns classified by their evolutionary conservation as measured by Valdar's score (C_{Valdar}).[9] For pathogenic variants, conserved vs. unconserved columns display the same behaviour as missense depleted vs. enriched columns, which is concordant with previous work and expected since most missense depleted columns are also conserved. However, the column classification schemes yield almost opposite trends with respect to the distribution of ClinVar risk factor variants. There is a slight tendency for risk factor variants to occur more frequently in evolutionary conserved columns (OR = 1.47, p = 0.194), which contrasts with their higher frequencies in columns that are relatively enriched for missense variation.

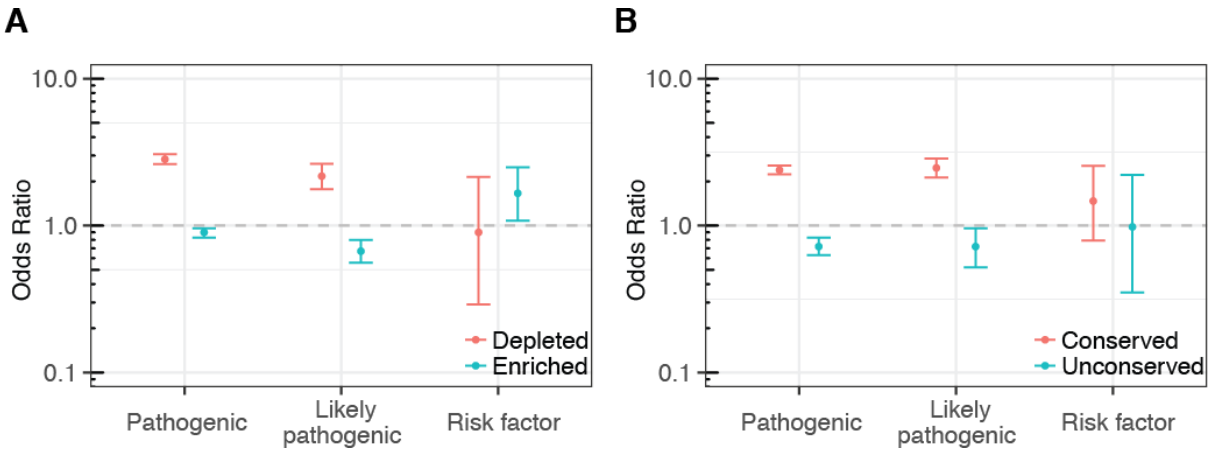


Figure 3: Properties of missense depleted and enriched domain family alignment columns. Odds ratios and 95% C.I. for enrichment of variants with specific ClinVar terms that affect residues found in A. missense depleted (p < 0.1; see methods) or enriched (p < 0.1) domain family alignment columns and B. conserved (C_{Valdar} in 1st decile) or unconserved columns (C_{Valdar} in 10th decile).

The Conservation Plane: Combining column variant class and conservation

Although the distribution of missense variants within domains is typically concordant with the evolutionary conservation profile (Figure 2B), the two metrics are not

redundant and cross-classification of alignment columns by both yields residue categories with interesting properties. Figure 4A shows the distribution of ClinVar annotated pathogenic variants between columns classified as unconserved-missense depleted (UMD), unconserved-missense enriched (UME), conserved-missense depleted (CMD) and conserved-missense enriched (CME). Conserved and unconserved columns that are neither missense depleted or enriched, i.e. have an average number of missense variants for the family, are also shown. It shows that: 1) all conserved sites are enriched for pathogenic variants but CMD sites are more so (CME: OR = 1.24, $p = 1.6 \times 10^{-5}$; CMD: OR = 1.85, $p = 3.3 \times 10^{-17}$) and 2) the UMD subset of unconserved residues are enriched for pathogenic variants to an extent comparable to conserved residues (OR = 1.74, p = 0.02). The UMD classification identifies sites where residues have varied throughout the evolution of the domain family but the specific residue adopted by each domain is now under negative selection in human. This implies that residues in this column class could be enriched for specificity determinants. A structural analysis of 270 UMD sites found in 160 families provides some support for this hypothesis. We compared these sites to UME columns from the same families and found that UMD columns were enriched for ligand, domain-domain and nucleotide interactions (OR = 1.59, p = 0.003) and tended to be less accessible to solvent (OR = 1.73, $p = 2.0 \times 10^{-04}$; Extended Data Table 1). Figure 4C illustrates an example of a protein family where UMD residues indicate known ligand-binding sites. The Rhodopsin-like receptor family (PF00001) contains 11 UMD sites, five of which occur in sequence in the centre of Helix 3 and form interactions with ligands in many structures (e.g. residues in column 780 interact with ligands in 23 distinct proteins; Extended Data Table 2) and includes a Na^{2+} binding residue. Extended Data Fig. 3 shows another example of ligand binding site identification in the nuclear receptor ligand binding domain family (NR-LBD; PF00104). Additionally, Extended Data

Fig. 4 shows UMD sites in the NR-LBD family that are not directly involved in ligand binding but instead mediate strong intra-domain cross-helical interactions that vary dramatically between domains. Structures of intact DNA-bound nuclear receptors suggest that in some proteins these residues interact with the LBD-DNA binding domain linker and thus may mediate the ligand induced DNA binding response (Not shown. For an example see Glu 295 and Ser 332 in PDB ID: 3e00 chain D.).[21] These important interactions may not be detected by residue co-variation analysis[18] because the UMD site interacts with residues aligned in different columns in each domain. One UMD site is seen in the IBR domain (PF01485). In the E3 ubiquitin-protein ligase parkin, this is Glu370 that recent structural studies suggest is at the interface with Ubiquitin[22] and so likely to be important in mediating this interaction. All other UMD classified sites can be found in Supplementary Data Table 1. Together, these findings show that human missense variation can stratify unconserved alignment columns to identify a small number of residues likely to be important for function and specificity.

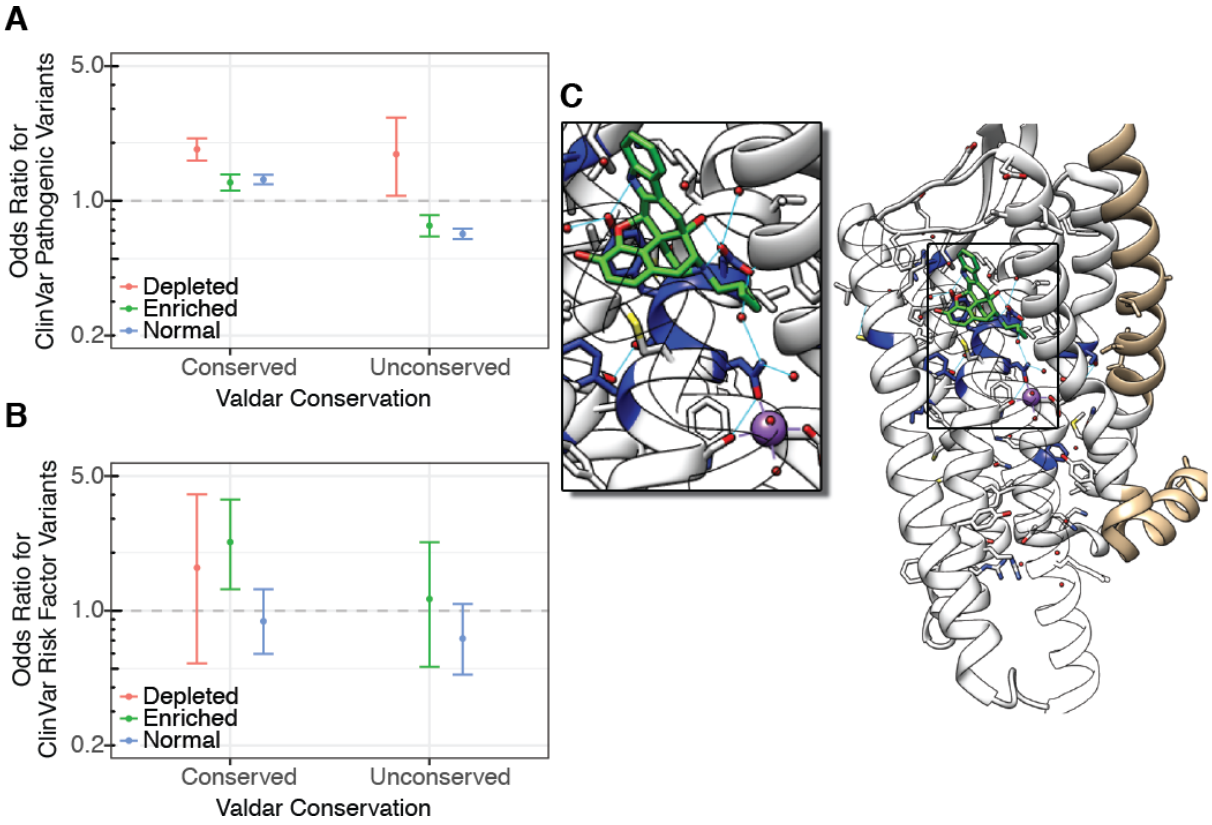


Figure 4: Classification of domain residues by evolutionary conservation and relative population variation. A. Odds ratios for ClinVar pathogenic variants in missense depleted ($p < 0.1$; see methods), enriched ($p < 0.1$) or normal ($p \geq 0.1$) alignment columns that were either conserved (C_{Valdar} < median) or unconserved (C_{Valdar} > median). B. Odds ratios for ClinVar risk factor variants in different column classes. UMD columns are not shown as there are zero risk factor variants in this column class; the ClinVar risk factor OR and 95 % C.I. for UMD columns is 0 [0, 14]. C. Illustration of UMD residues (blue) in the Rhodopsin-like receptors (PF00001) mapped to a structure of the Delta-type opioid receptor (PDB ID: 4n6h).[23] Amongst the 11 UMD residues are several involved in ligand binding and one that coordinates the bound sodium ion; residues 249-287 are hidden for clarity.

Another striking feature of residues in columns with discordant levels of evolutionary conservation and population diversity was found. Figure 4B shows the odds ratios of observing ClinVar risk factor variants in columns classed according to evolutionary conservation and whether they are relatively enriched in missense variants or not and highlights that CME sites are significantly enriched in risk factor variants (OR = 2.27, p =

209 0.004). This is consistent with the previous observation that missense enriched
210 columns were enriched for risk factor variants and that conserved columns showed a
211 tendency toward risk factor enrichment (Figure 3) but the combined effect is much
212 stronger. To our knowledge this is the first time that a feature marking residues pre-
213 disposed to carrying risk factor variants has been identified.
214
215 With further development, the conservation plane may yield insight into the
216 evolutionary forces acting on individual sites in protein domain families. Although this
217 will require consideration of each family's phylogeny coupled with more detailed
218 variation metrics (e.g., considering allele frequencies, heterozygosity,
219 missense/synonymous ratios (d_N/d_S), McDonald–Kreitman test[24] and derivatives) our
220 results offer clues as to which evolutionary signatures are being detected. Given the
221 recognised effects of different types of selection upon intra- and interspecific
222 variability,[25] we can loosely associate: CMD sites with negative selection and sites
223 affected by selective sweeps; UMD sites with positive selection (here, domain
224 specialisation) and CME sites with balancing selection. Whilst these associations are
225 speculative, the structural features and disease associations of those classes are
226 congruent with these evolutionary processes.[25-27] A few immediate practical
227 applications follow from the missense-depletion and conservation plane class
228 associations. For variant pathogenicity prediction, the results extend the work of
229 Gussow and coworkers[4,6] and open the door to hierarchical classification where the
230 impact of a variant can be can be judged in genic, sub-genic architecture, and now,
231 residue level contexts on the basis of population variation. In protein feature prediction,
232 the ability to identify functionally important residues that are classically unconserved
233 could help to identify allosteric and surface interaction sites, whilst a metric that is

sensitive to specificity determining residues should prove useful in understanding enzyme active sites and other functional sites in more detail.

Methods

Datasets, Mapping and Filtering

Protein family alignments were downloaded from Pfam (v29)[7,8] and parsed using Biopython (v1.66, with patches #768 #769)[28] and conservation scores were calculated by AACons via JABAWS (v2.1).[29] The human sequences in the alignment were mapped to the corresponding full UniProt sequences to create keys between UniProt sequence residue numbers and Pfam alignment column numbers. For each human sequence, germline population variants were retrieved from Ensembl 84[14,15] via the Ensembl API using ProteoFAV.[30] Ensembl variants are provided with indexes to UniProt sequence residue numbers and were thus mapped to Pfam alignment columns.

Ensembl variation agglomerates variants and annotation data from a variety of sources including dbSNP (v146), 1KG, ESP and ExAC[1]. A full description of the variant sources present in Ensembl 84 is available at http://mar2016.archive.Ensembl.org/info/genome/variation/sources_documentation.html. Ensembl provides numerous annotations including the predicted protein consequences (i.e. missense, synonymous, stop gained, etc.), minor allele frequency (MAF) and ClinVar[20] disease status. These annotations were used to filter the Pfam-mapped variants for the collection of variant sub-class alignment column statistics. For example, this is how the number of ClinVar 'pathogenic' missense variants in each alignment column was calculated.

Pfam (v29) contains 16,035 domain family alignments. Variants were gathered and mapped to the alignments for the 6,088 families that contain at least one human sequence. For inclusion in this analysis, a minimum threshold of five human sequences was adopted corresponding to 2,939 protein families. However, some of these families do not meet this criterion after sequence duplication correction (see below) leaving 1,376 families. Finally, alignment column conservation scores could not be obtained for 85 of the families, resulting in a final dataset of 1,291 protein families. These families contain an estimated 25,158 human protein domains. Only columns with ≥ 5 human residues (i.e., non-gap) were considered, corresponding to 159,296 alignment columns. This filter was applied in all analyses reported in this work.

Variant Duplication

Some alignments contained variants that mapped to multiple sequences due to sequence duplication. For example, in PF00001 all variants that mapped to the human sequence P2Y11/45-321 (P2Y purinoceptor 11) from the P2RY11 gene are duplicated in A0A0B4J1V8/465-741 because this sequence contains the same 7 transmembrane receptor domain as P2Y11 as a result of A0A0B4J1V8 being the product of a read-through transcript that includes the P2RY11 gene. This means there are two copies of the P2RY11 7 transmembrane receptor domain in the alignment and its variant profile is doubly weighted. Another example in this family comes from human sequences MSHR/55-298 (Melanocortin receptor 1), G3V4F0/55-298 and A0A0B4J269, which all are mapped to the same genomic loci. Accordingly, sequence duplication was accounted for by de-duplicating variants and sequences before summing over columns.

Statistical Analyses

The statistical analyses were all performed using R version 3.2.2. Regressions were calculated by the *lm* function from the *stats* library. Odds ratios and Fishers exact *p* values were calculated with the *fisher.test* function from the *stats* library. Plots were produced with *ggplot2*.

Alignment Column Classification

Columns were classified as depleted, enriched or neutral with respect to the column variant totals relative to the average for the other columns in the alignment. For each alignment column *x*, a 2 × 2 table was constructed of the form *a*, *b*, *c*, *d* with elements: *a*. the number of variants mapped to residues in column *x*, *b*. the total number of variants mapped to all other alignment columns, *c*. the number of human residues in column *x* and *d*. the total number of human residues in the rest of the alignment. Application of the R *stats* function *fisher.test* to each table yielded an odds ratio > 1 if the column contained more than the alignment average number of variants per human residue or OR < 1 if there were fewer than the average number of variants per human residue. The function also provided the *p* value afforded by Fisher's exact test. This meant that for a given $p_{threshold}$ columns with $p \geq p_{threshold}$ were considered normal and columns with $p < p_{threshold}$ were considered depleted if OR < 1 or enriched if OR > 1. Notably, in addition to the effect size, *p* is sensitive to data availability (i.e., variant counts) and alignment column occupancy. In this work, $p_{threshold} = 0.1$ unless otherwise specified.

Structural Analysis of Evolutionary Unconserved and Missense Depleted Residues

Columns were classified as unconserved-missense depleted (UMD) or unconserved-missense enriched (UME) if they displayed significant residue diversity ($V_{Shenkin}$ in 4th

quartile) and were missense depleted or enriched, respectively. The 343 columns in 211 families that met these criteria were subjected to an automated analysis where the flagged residues were mapped to PDB structures via SIFTS;[31] 270 columns from 160 families were mapped to at least one PDB structure. Biological units were obtained from the PDBe in mmCIF format. When multiple biological units were available for a particular asymmetric unit, the preferred biological unit ID was obtained by querying the PDBe API.[32] Atoms were considered to interact if they were within 5 Å. A residue was considered to participate in a domain interaction if it interacted with a Pfam domain on a different PDB chain. Residue relative solvent accessibilities (RSAs) were calculated from the DSSP accessible surface[33] as described in Tien *et al*.[34] and were classified as surface (RSA > 25%), partially exposed (5% < RSA ≤ 25%) or core (RSA ≤ 5%).

The results of the automated analysis were supplemented by a manual structural analysis using a workflow enabled by the Jalview multiple sequence alignment workbench[35] and the UCSF Chimera molecular graphics program.[36] Jalview feature files identifying the UMD columns were generated. When the feature files were loaded onto the appropriate alignment in Jalview, the residues in the UMD columns were highlighted for the user. Jalview was then used to find PDB structures for the sequences in the alignment that were then visualised in UCSF Chimera. Jalview automatically mapped the UMD residue annotations to the PDB structure so that the residues could be assessed in their structural context. UCSF Chimera was used to identify other residues in the structure that were hydrogen bonded to, or had a Van der Waals distance < 1 Å with, a side-chain atom of any UMD residues present. The residues were then classified according to any contacts made as either: ligand binding, ion binding, inter-domain

interaction, intra-domain interaction or surface residue. This analysis found that of those families with UMD residues, 19% had at least one UMD site involved in ligand-binding whilst 42% had a site directly involved in domain-domain interactions.

Code availability

The code used in this study will be available from the Barton Group GitHub repository at https://github.com/bartongroup/ on journal publication. The software was not designed for portability and may not function as intended in all environments, but the source code illustrates our methodology. We are currently developing a production version that will enable users to apply our methods to their own alignments to be released in the same repository.

Data availability

The multiple sequence alignments and human variation data that underlie and support the findings of this study are available from Pfam, http://pfam.xfam.org/ and Ensembl 84, http://www.Ensembl.org/, respectively. The calculated data, including alignment column variation statistics and residue conservation scores are presently available from the corresponding author upon request whilst a web resource is under development. The UMD columns are also identified in the supplementary material.

References

1 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).
2 Telenti, A. *et al.* Deep sequencing of 10,000 human genomes. *Proc Natl Acad Sci U S A* **113**, 11901-11906, doi:10.1073/pnas.1613365113 (2016).
3 Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

4 Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* **9**, e1003709, doi:10.1371/journal.pgen.1003709 (2013).

5 MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828, doi:10.1126/science.1215040 (2012).

6 Gussow, A. B., Petrovski, S., Wang, Q., Allen, A. S. & Goldstein, D. B. The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biol* **17**, 9, doi:10.1186/s13059-016-0869-4 (2016).

7 Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res* **42**, D222-230, doi:10.1093/nar/gkt1223 (2014).

8 Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44**, D279-285, doi:10.1093/nar/gkv1344 (2016).

9 Valdar, W. S. Scoring residue conservation. *Proteins* **48**, 227-241, doi:10.1002/prot.10146 (2002).

10 Melloni, G. E. *et al.* LowMACA: exploiting protein family analysis for the identification of rare driver mutations in cancer. *BMC Bioinformatics* **17**, 80, doi:10.1186/s12859-016-0935-7 (2016).

11 Miller, M. L. *et al.* Pan-Cancer Analysis of Mutation Hotspots in Protein Domains. *Cell Syst* **1**, 197-209, doi:10.1016/j.cels.2015.08.014 (2015).

12 Yang, F. *et al.* Protein domain-level landscape of cancer-type-specific somatic mutations. *PLoS Comput Biol* **11**, e1004147, doi:10.1371/journal.pcbi.1004147 (2015).

13 Peterson, T. A., Park, D. & Kann, M. G. A protein domain-centric approach for the comparative analysis of human and yeast phenotypically relevant mutations. *BMC Genomics* **14 Suppl 3**, S5, doi:10.1186/1471-2164-14-S3-S5 (2013).

14 Yates, A. *et al.* Ensembl 2016. *Nucleic Acids Res* **44**, D710-716, doi:10.1093/nar/gkv1157 (2016).

15 Chen, Y. *et al.* Ensembl variation resources. *BMC Genomics* **11**, 293, doi:10.1186/1471-2164-11-293 (2010).

16 Cuff, J. A. & Barton, G. J. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* **40**, 502-511 (2000).

17 Mistry, J., Bateman, A. & Finn, R. D. Predicting active site residue annotations in the Pfam database. *BMC Bioinformatics* **8**, 298, doi:10.1186/1471-2105-8-298 (2007).

18 Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766, doi:10.1371/journal.pone.0028766 (2011).

19 Shenkin, P. S., Erman, B. & Mastrandrea, L. D. Information-theoretical entropy as a measure of sequence variability. *Proteins* **11**, 297-313, doi:10.1002/prot.340110408 (1991).

20 Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* **44**, D862-868, doi:10.1093/nar/gkv1222 (2016).

21 Chandra, V. *et al.* Structure of the intact PPAR-gamma-RXR- nuclear receptor complex on DNA. *Nature* **456**, 350-356, doi:10.1038/nature07413 (2008).

22 Kumar, A. *et al.* Parkin–phosphoubiquitin complex reveals cryptic ubiquitin-binding site required for RBR ligase activity. *Nature Structural & Molecular Biology (in press)*, doi:10.1038/nsmb.3400 (2017).

23 Fenalti, G. *et al.* Molecular control of delta-opioid receptor signalling. *Nature* **506**, 191-196, doi:10.1038/nature12944 (2014).

24 McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**, 652-654, doi:10.1038/351652a0 (1991).

25 Nielsen, R. Molecular signatures of natural selection. *Annu Rev Genet* **39**, 197-218, doi:10.1146/annurev.genet.39.073003.112420 (2005).

26 Fay, J. C. Disease consequences of human adaptation. *Appl Transl Genom* **2**, 42-47, doi:10.1016/j.atg.2013.08.001 (2013).

27 Worth, C. L., Gong, S. & Blundell, T. L. Structural and functional constraints in the evolution of protein families. *Nat Rev Mol Cell Biol* **10**, 709-720, doi:10.1038/nrm2762 (2009).

28 Cock, P. J. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423, doi:10.1093/bioinformatics/btp163 (2009).

29 Troshin, P. V., Procter, J. B. & Barton, G. J. Java bioinformatics analysis web services for multiple sequence alignment--JABAWS:MSA. *Bioinformatics* **27**, 2001-2002, doi:10.1093/bioinformatics/btr304 (2011).

30 Britto-Borges, T., Madeira, F., MacGowan, S. A. & Barton, G. J. ProteoFAV: fast structural data integration. *Manuscript in preparation* (2017).

31 Velankar, S. *et al.* SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res* **41**, D483-489, doi:10.1093/nar/gks1258 (2013).

32 Velankar, S. *et al.* PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Res* **44**, D385-395, doi:10.1093/nar/gkv1047 (2016).

33 Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637, doi:10.1002/bip.360221211 (1983).

34 Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J. & Wilke, C. O. Maximum allowed solvent accessibilites of residues in proteins. *PLoS One* **8**, e80635, doi:10.1371/journal.pone.0080635 (2013).

35 Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191, doi:10.1093/bioinformatics/btp033 (2009).

36 Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612, doi:10.1002/jcc.20084 (2004).

37 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948, doi:10.1093/bioinformatics/btm404 (2007).

38 Xu, W., Doshi, A., Lei, M., Eck, M. J. & Harrison, S. C. Crystal structures of c-Src reveal features of its autoinhibitory mechanism. *Mol Cell* **3**, 629-638 (1999).

39 Huet, T., Fraga, R., Mourino, A., Moras, D. & Rochel, N. Design, Chemical synthesis, Functional characterization and Crystal structure of the sidechain analogue of 1,25-dihydroxyvitamin D3. *To be published*, doi:10.2210/pdb3ogt/pdb (2011).

40 Souza, P. C. *et al.* Identification of a new hormone-binding site on the surface of thyroid hormone receptor. *Mol Endocrinol* **28**, 534-545, doi:10.1210/me.2013-1359 (2014).

41 Capelli, D. *et al.* Structural basis for PPAR partial or full activation revealed by a novel ligand binding mode. *Sci Rep* **6**, 34792, doi:10.1038/srep34792 (2016).

42 Blind, R. D. *et al.* The signaling phospholipid PIP3 creates a new interaction surface on the nuclear receptor SF-1. *Proc Natl Acad Sci U S A* **111**, 15054-15059, doi:10.1073/pnas.1416740111 (2014).

43 Raaijmakers, H. C., Versteegh, J. E. & Uitdehaag, J. C. The X-ray structure of RU486 bound to the progesterone receptor in a destabilized agonistic conformation. *J Biol Chem* **284**, 19572-19579, doi:10.1074/jbc.M109.007872 (2009).

44 Kallen, J. *et al.* Evidence for ligand-independent transcriptional activation of the human estrogen-related receptor alpha (ERRalpha): crystal structure of ERRalpha ligand binding domain in complex with peroxisome proliferator-activated receptor coactivator-1alpha. *J Biol Chem* **279**, 49330-49337, doi:10.1074/jbc.M407999200 (2004).

45 Wisely, G. B. *et al.* Hepatocyte nuclear factor 4 is a transcription factor that constitutively binds fatty acids. *Structure* **10**, 1225-1234 (2002).

46 le Maire, A. *et al.* A unique secondary-structure switch controls constitutive gene repression by retinoic acid receptor. *Nat Struct Mol Biol* **17**, 801-807, doi:10.1038/nsmb.1855 (2010).

Acknowledgements

We thank the Jalview development team for their help with streamlining the visualisation of alignment and structural data and Jim Procter for additional assistance in using AACons and useful discussions regarding evolutionary theory in relation to multiple sequence alignments. We also thank Helen Walden, Maurice van Steensel, Ulrich Zachariae, Owen Vickery, David Gray and Alessio Ciulli for discussions about specific protein families. This work was supported by Wellcome Trust Strategic Awards [098439/Z/12/Z and WT097945], Wellcome Trust Doctoral Training Account [100150/Z/12/Z], Wellcome Trust Biomedical Resources Grant [101651/Z/13/Z], Coordenação de Aperfeiçoamento de Pessoal de Nível Superior studentship [CAPES process 1529/12-9] and Biotechnology and Biological Sciences Research Council Grants [BB/J019364/1, BB/L020742/1].

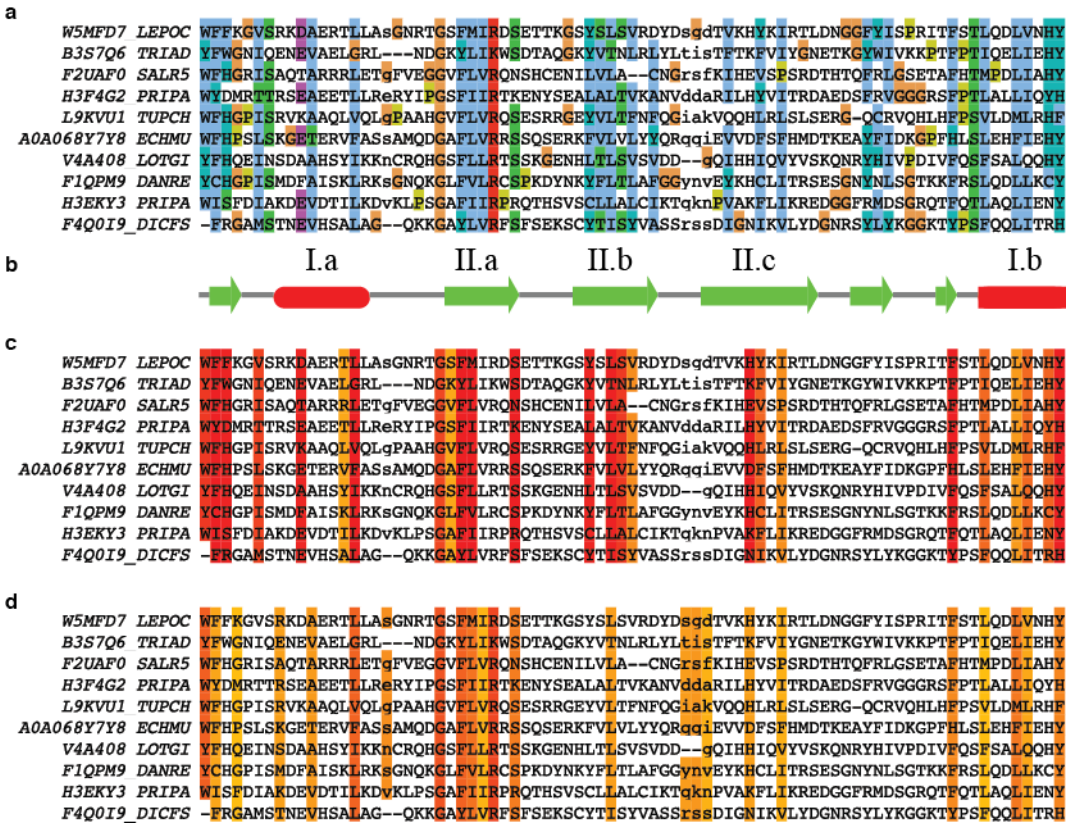
Author contributions

S.A.M. designed and performed the study, analysed data and wrote the manuscript. F.M. contributed software to collect variation data and collected the interaction and RSA data for UMD and UME residues. T.B. contributed software to collect variation data. M.S. performed the manual structural analysis of UMD residues. C.C. analysed data. G.J.B. designed the study, analysed data and wrote the manuscript.

Author information

1. Division of Computational Biology, School of Life Sciences, University of Dundee, Dundee, UK. 2. Centre for Dermatology and Genetic Medicine, School of Life Sciences, University of Dundee, Dundee, U.K. Correspondence should be addressed to G.J.B. (g.j.barton@dundee.ac.uk). The authors declare no competing financial interests.

Extended Data Captions

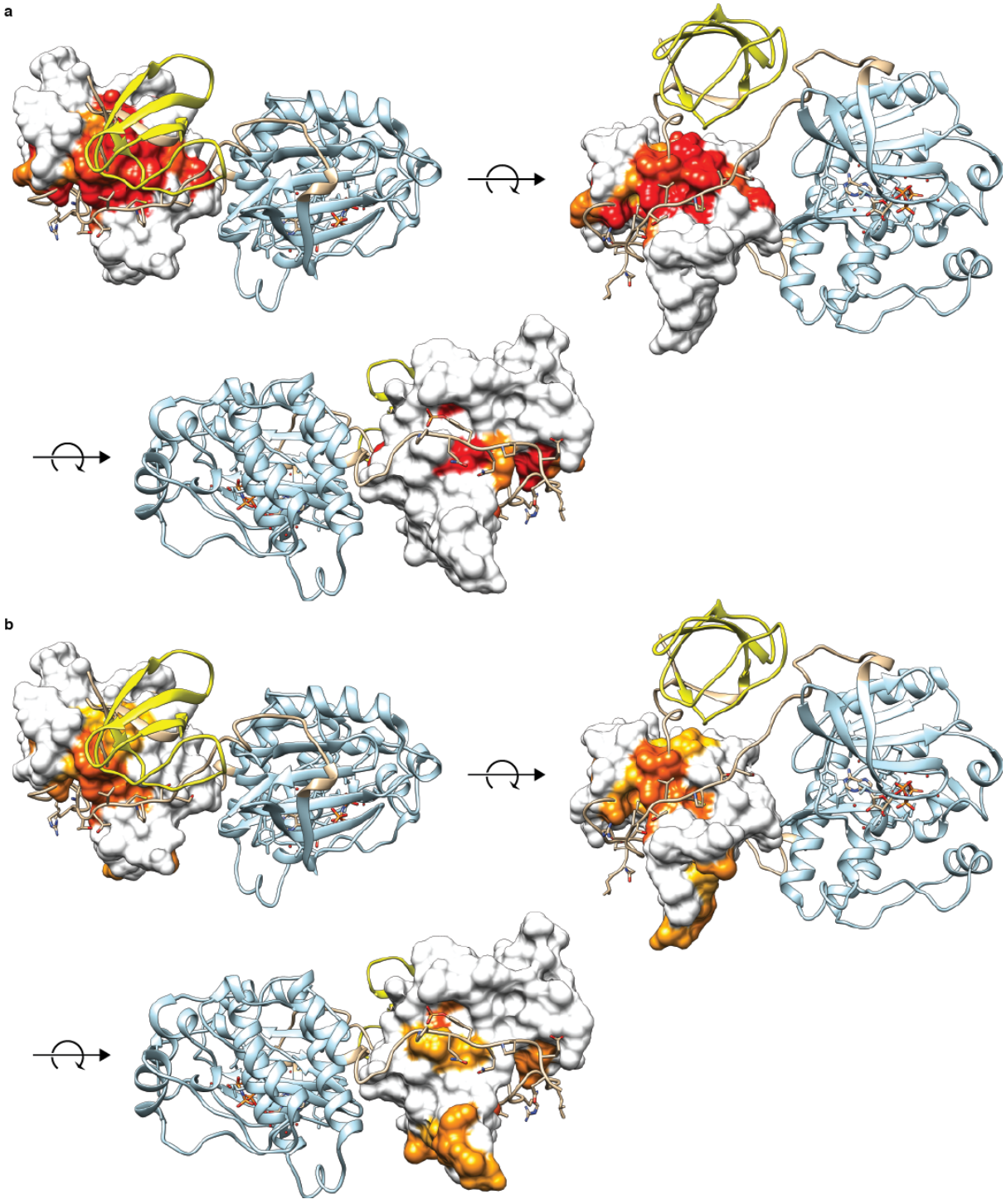


Extended Data Figure 1: An extract of the SH2 alignment (PF00017.21) showing the influence of secondary structure constraints upon evolutionary conservation and missense depletion. a. Alignment with Clustal X[37] colouring where blue indicates hydrophobic residue conservation. b. Consensus secondary structure from Pfam (v31);[7,8] labelled elements indicate the archetypal SH2 partially buried helices (I.a and I.b) and β-strands (II.a-c). c. Missense depleted columns with $P \leq 0.2$. d. Columns with $V_{Shenkin} \leq 20$. The pattern of conserved hydrophobic residues in a are indicative of the structural constraints imposed by the secondary structure elements in b. These structural constraints are known to produce patterns in conservation metrics like

516 $V_{Shenkin}$ in d. These constraints also influence the distribution of missense depleted

517 columns in c. Figure created with Jalview.[35]

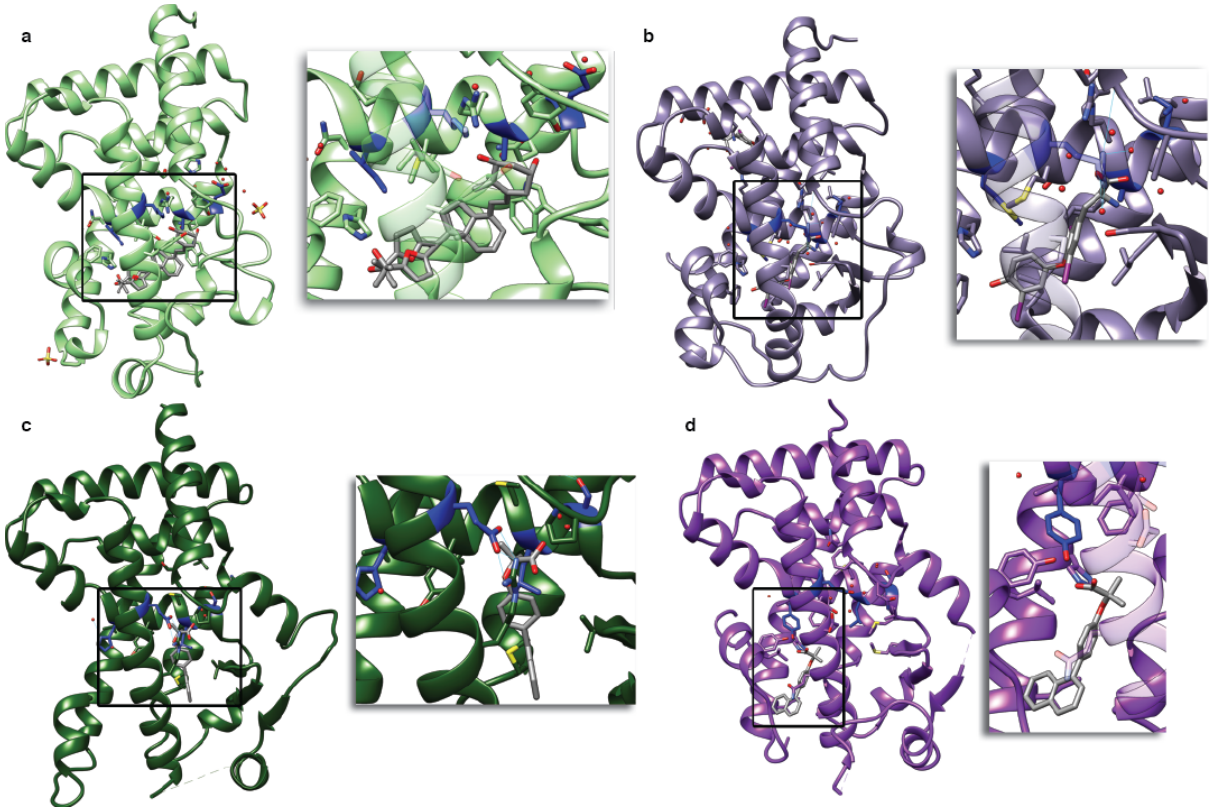
518



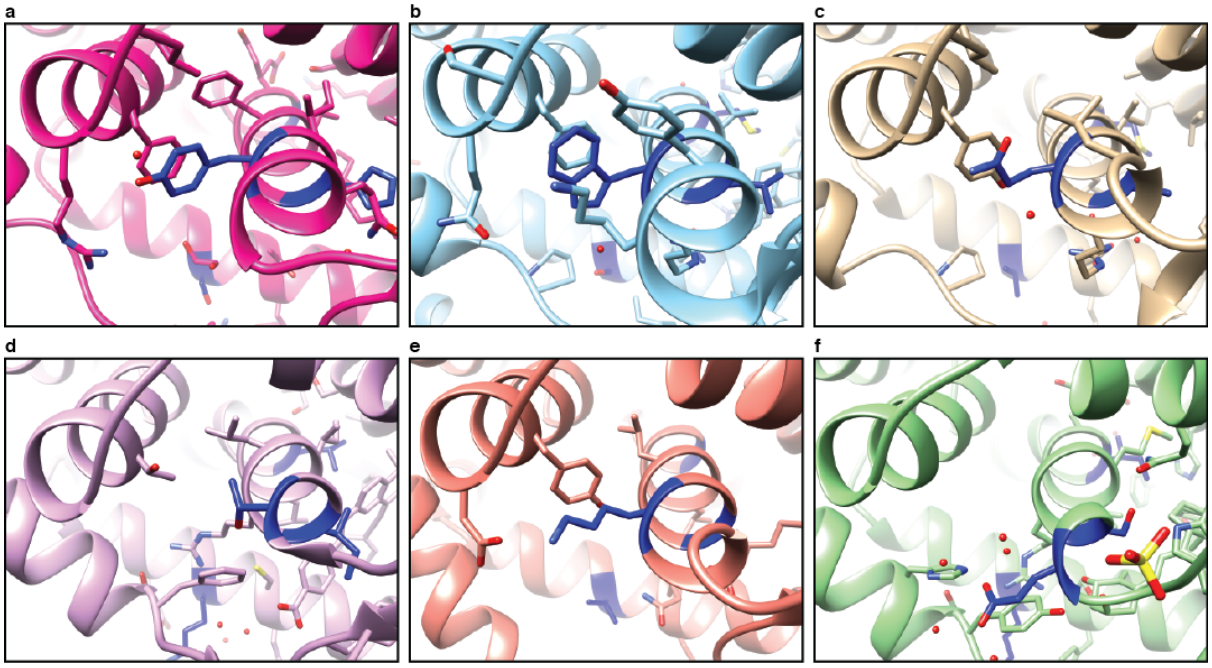
519

520 Extended Data Figure 2: Inter-domain interactions of the SH2 domain in inactivated Src
521 (PDB ID: 2src).[38] The surface of the SH2 domain (PF00017) is coloured red to yellow
522 corresponding to a. missense depletion *P* over range [0, 0.2) and b. $V_{Shenkin}$ over range
523 [0, 20); white surface regions are outside these ranges. The sub-panels show
524 interactions with the Src SH3 domain (yellow), kinase-SH2 linker (tan) and the tail

525 region including phosphorylated-Tyr (tan). Residues that interact with the SH2 domain
526 are displayed as sticks. Figure created with Jalview[35] and UCSF Chimera.[36]
527



Extended Data Figure 3: Examples of UMD residues (blue) involved in ligand-binding in the nuclear receptor ligand binding domains protein family (PF00104). a. VDR in complex with a calcitriol analog (3ogt).[39] b. THα in complex with triiodothyronine (4lnx).[40] c. PPARγ (5hzc) and d. PPARα (5hyk) in complex with the PPAR pan-agonist AL29-26.[41] The ligand is in VdW contact with the unconserved-depleted L330 in PPARγ and with Y314 in PPARα. Note that the substitution at the unconserved-depleted site H323 in PPARγ to Y314 in PPARα is related to the activity specificity of these two receptors with respect to AL29-26.[41] Figure created with UCSF Chimera[36] and Jalview.[35]



Extended Data Figure 4: Local environments of the UMD residue of H5 distal to the ligand binding pocket (blue). π-π interactions between residues a. Tyr A312 and Phe A368 in SF-1 (4qk4),[42] b. Trp A765 and Phe A818 of PR (2w8y)[43] and c. Gln A371 and Tyr A422 of ERRα (1xb7).[44] Equivalent residues also form salt-bridge interactions with H8 illustrated by e) Lys A185 and Asp A233 of HNF-4γ (1lv2).[45] In other proteins these strong, specific interactions are replaced with general hydrophobic contacts such as in d. Thr B275, which is in contact with both Phe B199 and Thr B326 of RARα (3kmz)[46] and the same interactions are observed in RARγ (e.g. see 1fcx, not shown). f. Lastly, the negatively charged Glu A277 found in this position of VDR (3ogt)[39] forms a potential salt-bridge with His A139 and pi-pi interactions with Tyr A143. This results in a radically different interaction topology where the site binds to a different helix. Figure created with UCSF Chimera[36] and Jalview.[35]

554 Extended Data Table 1: Differences in the structural properties of unconserved residues

555 differentiated by their human missense variation classification.

556

	Observed in one or more mapped PDB		Not observed in any mapped PDB		OR[c]	p[c]
Residue Counts[a, e]	UMD[d]	UME[d]	UMD	UME		
Ligand	765	1,448	5,579	14,454	1.37	6.4×10^{-11}
Domain	649	1,312	5,695	14,590	1.27	3.5×10^{-06}
Ligand, domain or nucleotides	1,338	2,549	5,006	13,353	1.40	0
Core	1,635	1,995	4,709	13,907	2.42	0
Part-exposed	2,584	4,526	3,760	11,376	1.73	0
Surface	3,213	11,742	3,131	4,160	0.36	0
Column Counts[b, e]						
Ligand	156	407	114	357	1.20	0.23
Domain (inter-chain)	131	328	139	436	1.25	0.18
Ligand, domain or nucleotides	201	494	69	279	1.59	0.0033
Core	179	406	91	358	1.73	2.0×10^{-04}
Part-exposed	231	607	39	157	1.53	0.03
Surface	253	735	17	29	0.59	0.12

557 a. Protein residues are counted in possession of the row feature if it is observed in *any* mapped PDB

558 residue and are counted as lacking the feature if it is not observed in any of its mapped PDB residues.

559 Residues that did not map to at least one PDB structure are not counted. For example, 765 UMD residues

560 map to at least one PDB structure and bind a ligand in at least one of these structures whilst 5,579 UMD

561 residues also map to at least one PDB structure but do not bind a ligand in any of them. b. Pfam columns

562 are counted in possession of the row feature if it is observed in *any* mapped PDB residue that is aligned in

563 the column and are counted as lacking the feature if it is not observed in any of its mapped PDB residues

564 present in the column. Columns that did not contain at least one residue that mapped to a PDB structure

565 were not counted. For example, 156 UMD columns contain at least one residue that maps to a PDB

566 structure that shows the residue is in contact with a ligand whilst 114 UMD columns contain at least one

567 residue that maps to a PDB structure but a ligand interaction is not observed in any mapped structure.

568 Note that the column statistics are not sensitive to family size variability. c. Fisher's test of association

569 between column classification (UMD or UME) and structural property; OR > 1 indicates enrichment of the

570 row feature in the UMD class. For example, the enrichment of ligand binding residues in UMD columns

571 compared to UME columns (OR = 1.20; p = 0.23) is calculated from the contingency table [(156, 407),

572 (114, 357)]. d. Unconserved-missense depleted (UMD) residues were defined as mapping to Pfam

573 columns with $V_{Shenkin}$ in 4th quartile for the protein family that are also missense depleted (see Methods)

574 whilst unconserved-missense enriched (UME) residues are equally divergent but missense enriched. e.

575 See Methods for feature definitions.

576

Extended Data Table 2: Example proteins with protein, ligand or nucleotide binding interactions involving residues in unconserved-missense depleted (UMD) columns from selected families (see Supplementary Data Table 1 for all families with discovered UMD columns).

Family	Col.[a]	Res.[b]	Protein[c]	Ligand[c]	Nucleotide[c]
PF00001	525	89		AA2AR_HUMAN (5iu8) [6]	
	575	98		ACM3_RAT (4u14) [6]	
	584	99			
	780	129		5HT1B_HUMAN* (4iar) [23]	
	792	130		5HT1B_HUMAN* (4iaq) [23]	
	808	131		5HT2B_HUMAN (4nc3) [4]	
	818	132		5HT2B_HUMAN (4nc3) [7]	
	832	134		ACM2_HUMAN (4mqs) [7]	
	1075	176		5HT2B_HUMAN (4ib4) [8]	
	1141	187		AA2AR_HUMAN (2ydo) [4]	
	1328	211	ACM3_RAT (4u15) [2]	AA2AR_HUMAN (4eiy) [6]	
PF00076	291	148	B3GWA1_CAEEL* (5ca5) [13]	B3GWA1_CAEEL* (5ca5) [10]	CELF1_HUMAN (3nmr) [29]
PF00104	190	715	ESR1_HUMAN (2jf9) [4]	A0A0B4J1T2_HUMAN* (2amb) [17]	
	312	743	NR4A1_HUMAN (3v3e) [3]	ANDR_HUMAN (1e3g) [35]	
	330	750	NR4A1_HUMAN (3v3e) [2]	ANDR_HUMAN (1t5z) [39]	
	332	752	NR1I3_MOUSE (1xnx) [3]	A0A0B4J1T2_HUMAN* (1t5z) [12]	

a. Pfam alignment column number, b. UniProt residue number for aligned residue of asterisked sequence in columns 4-6. For example, in the PF00001 (Rhodopsin-like receptor family) the numbering corresponds to 5HT1B_HUMAN and in PF00076 it corresponds to B3GWA1_CAEEL. This additional numbering allows the distance between UMD residues to be assessed in sequence space, which is obscured by gaps in Pfam alignment column indexes. c. Example protein and PDB structure where this interaction is observed. Number in parenthesis indicates how many domains in total have at least one PDB structure that provides evidence for the interaction. For example, the first row indicates that the AA2AR_HUMAN residue aligned in column 525 of PF00001 is in contact with a ligand in PDB 5iu8 and there are a total of 6 domains that display this interaction type in at least one PDB structure. Additionally, residue 89 of 5HT1B_HUMAN maps to column 525 of PF00001.

Supplementary Data

Supplementary Data Table 1: Example proteins with protein, ligand or nucleotide binding interactions involving residues in unconserved-missense depleted (UMD) columns. See table end for footnotes.

Family	Col.[a]	Res.[b]	Protein[c]	Ligand[c]	Nucleotide[c]
PF00001	525	89		AA2AR_HUMAN (5iu8) [6]	
	575	98		ACM3_RAT (4u14) [6]	
	584	99			
	780	129		5HT1B_HUMAN* (4iar) [23]	
	792	130		5HT1B_HUMAN* (4iaq) [23]	
	808	131		5HT2B_HUMAN (4nc3) [4]	
	818	132		5HT2B_HUMAN (4nc3) [7]	
	832	134		ACM2_HUMAN (4mqs) [7]	
	1075	176		5HT2B_HUMAN (4ib4) [8]	
	1141	187		AA2AR_HUMAN (2ydo) [4]	
	1328	211	ACM3_RAT (4u15) [2]	AA2AR_HUMAN (4eiy) [6]	
PF00004	557	172	A4YHC5_METS5* (4d80) [12]		DPA44_BPT4 (3u60) [1]
	690	197	CLPC_BACSU (3pxg) [4]	FTSH_THET8 (1ixz) [1]	
PF00011	104	58	HS16B_WHEAT* (1gme) [7]		
PF00018	142	94	DLG4_RAT (2xkx) [6]	ABL1_HUMAN* (1bbz) [9]	
	176	104	DLG4_RAT (2xkx) [6]	ABL1_HUMAN* (4j9d) [8]	
PF00022	1610	190	ACTS_RABIT* (1o18) [4]	ACTS_RABIT* (2a3z) [2]	
	2486	281	ACTS_RABIT* (1o18) [1]	ACTS_RABIT* (2asm) [1]	
	2690	316	ACTS_RABIT* (1o18) [2]	ACTS_RABIT* (1s22) [2]	
PF00023	62	537	TRPA1_HUMAN (3j9p) [3]	ANK1_HUMAN* (1n11) [7]	
PF00024	56	304	FA11_HUMAN* (2j8j) [1]		
PF00029	179	61	CXB2_HUMAN* (2zw3) [1]		
PF00031	266	102	CYTC_HUMAN* (1tij) [1]	CYTC_HUMAN* (3qrd) [1]	
PF00042	196	90	CYGB_HUMAN* (2dc3) [6]	GLOB6_CAEEL (3mvc) [7]	
PF00043	397	183		GSTM1_RAT (3fyg) [3]	
	426	192	MCA3_HUMAN (5bmu) [1]	D2WL63_POPTR* (5f05) [3]	
PF00045	102	229	MMP9_HUMAN (1itv) [1]	HEMO_RABIT* (1qhu) [4]	
PF00047	148	54	CD4_HUMAN* (3j70) [3]	CD4_HUMAN* (2nxy) [2]	
PF00049	105	73	INS_BOVIN (2a3g) [4]	IGF1_HUMAN* (1imx) [3]	
PF00059	75	222	C209B_MOUSE* (3zhg) [7]	CLC1B_HUMAN (3wsr) [10]	
	171	245	CD209_HUMAN (1k9i) [8]	C209B_MOUSE* (4c9f) [4]	
PF00063	1233	226		F1RQI7_PIG* (4pjm) [5]	
PF00074	116	83		ECP_HUMAN* (4a2o) [2]	
PF00076	291	148	B3GWA1_CAEEL* (5ca5) [13]	B3GWA1_CAEEL* (5ca5) [10]	CELF1_HUMAN (3nmr) [29]
PF00079	263	101		A1AT_HUMAN* (1hp7) [2]	
	490	134	ILEU_HORSE (1hle) [5]	ANT3_HUMAN (1sr5) [5]	

Family	Col.[a]	Res.[b]	Protein[c]	Ligand[c]	Nucleotide[c]
PF00100	764	591	TGBR3_RAT* (3qw9) [2]	TGBR3_RAT* (3qw9) [1]	
PF00102	1290	255	PTN1_HUMAN* (2cm3) [2]	PTN11_HUMAN (4gwf) [2]	
PF00104	190	715	ESR1_HUMAN (2jf9) [4]	A0A0B4J1T2_HUMAN* (2amb) [17]	
	312	743	NR4A1_HUMAN (3v3e) [3]	ANDR_HUMAN (1e3g) [35]	
	330	750	NR4A1_HUMAN (3v3e) [2]	ANDR_HUMAN (1t5z) [39]	
	332	752	NR1I3_MOUSE (1xnx) [3]	A0A0B4J1T2_HUMAN* (1t5z) [12]	
PF00118	198	76	A8JE91_CHLRE* (5cdi) [6]		
	435	138	A8JE91_CHLRE* (5cdi) [10]	CH60_ECOLI (1xck) [1]	
	781	207	A8JE91_CHLRE* (5cdi) [6]	CH602_MYCTU (3rtk) [1]	
PF00125	468	102	CENPA_HUMAN* (3an2) [17]		
PF00134	226	229	CCND3_HUMAN (3g33) [2]	CCNC_HUMAN (3rgf) [2]	
	578	296	CCNA2_HUMAN* (1jsu) [6]		
PF00135	496	103	EST1_HUMAN (1mx1) [1]	ACES_HUMAN* (4ey7) [3]	
	522	110		ACES_MOUSE (4b84) [4]	
	1571	265		ACES_MOUSE (2ha0) [2]	
PF00149	925	125	J3K8M7_COCIM (5b8i) [9]	ASM3A_MOUSE (5fc1) [10]	MRE11_METJA (4tug) [1]
	926	126	G0RYR3_CHATD (4yke) [6]	ASM_MOUSE (5hqn) [4]	MRE11_METJA (4tug) [1]
	932	128	MRE11_METJA (4tug) [3]	A6THC4_KLEP7* (3jyf) [11]	
PF00151	289	74			
	345	96		LIPP_HUMAN* (1lpb) [2]	
	795	202			
PF00155	290	87	1A12_SOLLC* (1iax) [51]	AAT_ECOLI (3qn6) [8]	
	341	95		AADAT_HUMAN (2r2n) [7]	
	698	151		AAT_ECOLI (3zzk) [4]	
	804	162	1A12_SOLLC* (1iax) [28]	AAT_ECOLI (3qpg) [4]	
PF00157	118	184			PO5F1_MOUSE* (3l1p) [1]
PF00160	122	72		Q7RRM6_PLAYO (2b71) [1]	
	978	217	C6XII3_HIRBI* (5ex1) [1]	PPIA_HUMAN (4ipz) [2]	
PF00168	247	26		CAR1_ARATH* (5a52) [27]	
	312	39	UN13A_RAT (2cjt) [1]	DYSF_HUMAN (4ihb) [9]	
	579	77	SYT1_HUMAN (2k8m) [1]	CAR1_ARATH* (5a52) [12]	
PF00170	234	148	HY5_ARATH* (2oqq) [1]		
PF00171	886	139	A1U6U7_MARHV* (3rh9) [43]		
	1260	209		B1XMM6_SYNP2 (4it9) [1]	
PF00173	198	64	CYB5B_HUMAN* (3ner) [1]		
PF00194	389	138	CAH12_HUMAN* (1jcz) [2]	CAH2_HUMAN (2q38) [3]	
PF00209	1454	271	O67854_AQUAE* (3tt1) [2]		
	1877	333		Q9KDT3_BACHD (4us3) [1]	
PF00211	228	902	ADCY2_RAT* (1u0h) [4]	ADCYA_HUMAN (4clu) [1]	
PF00258	155	563	NOS2_HUMAN* (3hr4) [1]		
PF00270	1238	199	DBP5_YEAST* (3rrm) [4]	DDX3X_HUMAN (4pxa) [1]	DD19B_HUMAN (3fht) [2]
	1405	220	DBP5_YEAST* (3rrm) [2]	DBP5_YEAST* (3pew) [3]	DD19B_HUMAN (3fht) [2]
	1591	253	DBP5_YEAST* (3rrm) [4]	DBP5_YEAST* (3pew) [5]	DBP5_YEAST* (3pew) [5]
PF00293	556	143	AP4A_HUMAN (4ijx) [5]	8ODP_HUMAN* (3q93) [8]	

Family	Col.[a]	Res.[b]	Protein[c]	Ligand[c]	Nucleotide[c]
PF00300	485	321	F262_HUMAN* (5htk) [4]	F262_HUMAN* (5htk) [2]	
	913	367	PGAM1_HUMAN (4gpi) [4]	F262_HUMAN* (5htk) [2]	
PF00307	437	229	ACTN3_HUMAN (3lue) [4]	ACTN2_HUMAN* (4d1e) [4]	
PF00350	868	152	DRP1A_ARATH* (3t34) [2]		
PF00365	119	630	PFKA1_YEAST* (3o8o) [3]		
PF00378	543	71	B1MEE0_MYCA9 (3qxz) [12]	A0R747_MYCS2 (3moy) [3]	
	861	92	B1MIA8_MYCA9 (3rsi) [4]	A0QS88_MYCS2* (4qfe) [6]	
PF00386	272	193	ADIPO_MOUSE (1c28) [10]	ADIPO_HUMAN* (4dou) [6]	
	370	228	ADIPO_MOUSE (1c28) [3]	C1QT5_HUMAN (4nn0) [3]	
PF00406	502	74	KAD_FRATT* (4pzl) [2]	KAD1_HUMAN (1z83) [2]	
	853	118			
PF00412	129	49		LHX4_MOUSE* (3mmk) [2]	
PF00413	898	210	MMP13_HUMAN (2ozr) [2]	MMP1_HUMAN* (1hfc) [7]	
PF00431	304	221	C1S_HUMAN (1nzi) [3]		
	444	243	A2VCV7_RAT* (5ckn) [1]	A2VCV7_RAT* (5ckn) [8]	
PF00454	1001	359	PK3CA_HUMAN (4jps) [1]	P4K2A_HUMAN* (4pla) [1]	
PF00481	335	265			
	341	266			
	511	323	P2C16_ARATH* (3rt0) [3]		
	823	376		Q7PP01_ANOGA (2i0o) [1]	
PF00501	1787	113		Q6ND88_RHOPA (4fut) [1]	
	1924	137		C6W5A4_DYAFD* (4gs5) [1]	
PF00566	375	331		RBG1L_HUMAN (3hzj) [1]	
	843	431			
	934	448	GYP1_YEAST* (2g77) [1]		
PF00629	497	758		NRP1_HUMAN* (5l73) [1]	
PF00630	716	2308	FLNA_HUMAN* (2brq) [2]	FLNA_HUMAN* (2w0p) [1]	
PF00641	27	77			ZRAB2_HUMAN* (3g9y) [1]
PF00643	80	150	PML_HUMAN* (2mvw) [3]	PML_HUMAN* (2mvw) [11]	
PF00644	100	1624		PAR14_HUMAN* (3se2) [1]	
	224	1657	TNKS2_HUMAN (4hkk) [1]	PAR14_HUMAN* (3se2) [4]	
PF00685	218	62		ST4A1_HUMAN* (1zd1) [1]	
	745	125			
PF00688	580	140			
	784	195		GDF2_MOUSE* (4ycg) [1]	
PF00690	244	58		AT2A1_RABIT* (1su4) [1]	
PF00704	939	203	CHID1_HUMAN (3bxw) [1]	A8GFD6_SERP5* (4ptm) [4]	
PF00754	384	377		NRP2_HUMAN* (5dq0) [2]	
PF00777	338	225	SIA8C_HUMAN* (5bo6) [1]		
PF00786	43	95	PAK1_HUMAN* (1f3m) [1]		
PF00787	104	39		NCF4_HUMAN* (1h6h) [3]	
PF00822	353	65	CLD4_HUMAN* (5b2g) [1]		
PF00850	480	71		B2JF16_BURP8* (5ji5) [2]	
	493	79		B2JF16_BURP8* (5ji5) [4]	
PF00855	102	1104	HDGF_HUMAN (2nlu) [1]	BRPF1_HUMAN* (5c6s) [4]	

Family	Col.[a]	Res.[b]	Protein[c]	Ligand[c]	Nucleotide[c]
PF00856	834	1153		EHMT2_HUMAN (3rjw) [3]	
	926	1158		EZH2_HUMAN (4mi0) [4]	
	935	1161	O41094_PBCV1 (1n3j) [2]	O41094_PBCV1 (3kma) [4]	
	2057	1212		EHMT2_HUMAN (3rjw) [3]	
	2140	1219	EZH2_HUMAN (5hyn) [1]	EHMT1_HUMAN* (4i51) [3]	
PF00858	1887	262	ASIC1_CHICK* (2qts) [1]		
PF00878	380	1555		MPRI_HUMAN* (1gqb) [1]	
PF00884	1080	164	BETC_RHIME* (4ug4) [1]		
	1595	300			
PF00899	787	524			
	956	551	UBA1_YEAST (4nnj) [1]	UBA1_SCHPO* (4ii2) [2]	
PF00928	234	205		AP2M1_RAT* (3h85) [1]	
	835	316			
	837	318		AP4M1_HUMAN (3l81) [1]	
	1354	407			
PF00969	24	47	HB2A_MOUSE (3c6l) [1]	2B11_HUMAN* (3pgc) [3]	
PF01055	373	189		GANAB_MOUSE (5f0e) [2]	
	1074	303			
	1092	309			
	1979	492	AGLU_SULSO* (2g3m) [1]		
	2016	497			
PF01094	318	92	ANPRC_HUMAN (1jdp) [3]	ANPRC_HUMAN (1jdn) [1]	
	1009	197	CASR_HUMAN (5fbh) [11]	GRID1_MOUSE (5kc9) [3]	
	1609	298	CASR_HUMAN (5fbh) [2]	GRM7_HUMAN (5c5c) [2]	
	2223	391	CASR_HUMAN (5fbh) [2]	ANPRA_RAT* (1t34) [8]	
PF01150	834	229	ENTP1_RAT* (3zx0) [1]		
PF01237	727	185			
	972	202		KES1_YEAST* (1zhy) [1]	
PF01344	30	130	ESP_ARATH* (5gq0) [6]	KEAP1_HUMAN (3vnh) [17]	
	109	157	ESP_ARATH* (5gq0) [8]	KEAP1_HUMAN (3zgd) [18]	
PF01365	139	2159	RYR1_RABIT* (5t15) [2]	RYR2_MOUSE (4l4i) [1]	
PF01399	312	385	RPN3_YEAST* (3jck) [2]		
PF01433	384	71		AMPN_ECOLI* (3b2p) [1]	
	1059	172		AMPN_ECOLI* (3puu) [2]	
PF01436	43	800		BRAT_DROME* (4zlr) [2]	BRAT_DROME* (4zlr) [4]
PF01485	332	377		ARI1_HUMAN* (2m9y) [8]	
PF01590	389	322		PDE6C_CHICK (3dba) [2]	
	581	360		PDE10_HUMAN* (2zmf) [6]	
PF01602	104	25	AP1B1_HUMAN* (4hmy) [1]		
	343	80			
PF01663	322	218			
	424	237		ENPP1_MOUSE* (4b56) [7]	ENPP2_MOUSE (5hrt) [1]
PF01740	595	524		Q9KN88_VIBCH* (3mgl) [3]	
PF01759	244	1624	CO5_HUMAN* (5hcc) [1]		

Family	Col.[a]	Res.[b]	Protein[c]	Ligand[c]	Nucleotide[c]
PF01833	385	328		Q8A1I2_BACTN (3hrp) [2]	
	386	329		COE1_HUMAN* (3mqi) [3]	
PF01979	2169	325	ADEC2_AGRFC* (3nqb) [2]	Q9X247_THEMA (3ooq) [1]	
PF02023	101	94	MZF1_HUMAN* (2fi2) [2]	PEG3_HUMAN (4bhx) [2]	
PF02210	479	428		NRX1A_BOVIN* (2h0b) [3]	
PF02263	510	189	ATLA1_HUMAN* (4idn) [2]		
	826	252	GBP1_HUMAN (2b92) [1]		
PF02412	109	330		COMP_HUMAN* (3fby) [7]	
PF02770	387	259	ACOX1_ARATH (1w07) [7]	ACDSB_HUMAN* (2jif) [1]	
PF02798	208	38		GSTA4_HUMAN* (3ik7) [3]	
PF02815	316	268	RYR1_RABIT* (5t15) [1]	RYR1_RABIT* (4i0y) [1]	
PF02932	262	262	GBRB3_HUMAN (4cof) [4]		
	360	279	5HT3A_MOUSE* (4pir) [8]		
	421	293	5HT3A_MOUSE* (4pir) [7]		
	485	305	GLRA3_HUMAN (5tio) [2]		
	617	330	5HT3A_MOUSE* (4pir) [8]		
PF03098	1517	495		PERL_BOVIN* (2pt3) [4]	
PF03114	398	84	BIN2_HUMAN (4i1q) [4]		
	541	107		AMPH_HUMAN* (3sog) [1]	
PF03281	606	333		MID51_HUMAN* (4nxt) [1]	
PF03372	477	97	TYDP2_DANRE (4f1h) [1]	APEX1_HUMAN (5dff) [3]	TYDP2_MOUSE (4gz2) [1]
	667	113	TYDP2_DANRE (4f1h) [1]	APEX1_HUMAN (5dff) [5]	
	867	132	APEX1_DANRE* (2o3c) [2]	APEX1_HUMAN (4qh9) [4]	
	868	133	APEX1_DANRE* (2o3c) [2]	APEX1_HUMAN (4qh9) [4]	
	889	140	TYDP2_DANRE (4f1h) [1]	O26314_METTH (3g0a) [4]	
	1086	155	TYDP2_DANRE (4f1h) [1]	APEX1_HUMAN (4qh9) [5]	
	1281	192	TYDP2_DANRE (4f1h) [1]	EXOA_BACSU (5cfe) [3]	
	1409	206	TYDP2_DANRE (4f1h) [1]	APEX1_HUMAN (4qhe) [6]	CNO6L_HUMAN (3ngo) [1]
	1654	247	TYDP2_DANRE (4f1h) [1]	C5C3L1_BEUC1 (4ruw) [4]	
PF03727	719	444		HXK_KLULA* (3o08) [1]	
PF03810	73	49		XPO1_YEAST* (5dhf) [1]	
PF04408	73	502		G0RY84_CHATD* (5d0u) [1]	
PF04547	1815	546			
	2144	600	C7Z7K1_NECH7* (4wis) [1]		
PF04969	111	32	Q8SSJ3_ENCCU* (2o30) [1]		
PF05485	203	48			THAP1_HUMAN* (2ko0) [1]
	208	50			THAP1_HUMAN* (2ko0) [1]
PF07707	213	251		KLH11_HUMAN* (3i3n) [1]	
PF08240	270	85	YHFP_BACSU (1tt7) [1]	ADH1E_HORSE* (7adh) [2]	
PF08441	1216	970	ITAX_HUMAN* (3k6s) [1]		
PF13424	87	124	GPSM2_MOUSE* (4jhr) [5]	GPSM2_MOUSE* (4g2v) [3]	
PF13499	331	59	CALM_HUMAN (2be6) [10]	C4M0U8_ENTHI* (2lc5) [25]	

Family	Col.[a]	Res.[b]	Protein[c]	Ligand[c]	Nucleotide[c]
PF13561	1250	130	A9NFJ2_ACHLI (4nbt) [7]	FABG_VIBCH (4i08) [4]	
	1429	147	A0QQJ6_MYCS2* (3pk0) [77]	A0QQJ6_MYCS2* (3pk0) [21]	
	2287	217	A9CL57_AGRFC (4imr) [27]	A9CJ43_AGRFC (4ibo) [6]	
PF13640	201	338	EGLN1_HUMAN* (5las) [1]	Q81LZ8_BACAN (5hv0) [1]	
PF13848	587	277	PDIA1_HUMAN* (4ju5) [1]		
PF14497	297	173		C5ATQ9_METEA* (4pxo) [3]	
PF14670	26	288		LRP6_HUMAN* (3sov) [1]	
PF16746	179	56	ACAP1_HUMAN* (4ckg) [3]		

a. Pfam alignment column number, b. UniProt residue number for aligned residue of asterisked sequence in columns 4-6. For example, in PF00001 (Rhodopsin-like receptor family) the numbering corresponds to 5HT1B_HUMAN and in PF00004 it corresponds to A4YHC5_METS5. This additional numbering allows the distance between UMD residues to be assessed in sequence space, which is obscured by gaps in Pfam alignment column indexes. c. Example protein and PDB structure where this interaction is observed. Number in parenthesis indicates how many domains in total have at least one PDB structure that provides evidence for the interaction. For example, the first row indicates that the AA2AR_HUMAN residue aligned in column 525 of PF00001 is in contact with a ligand in PDB 5iu8 and there are a total of 6 domains that display this interaction type in at least one PDB structure. Additionally, residue 89 of 5HT1B_HUMAN maps to column 525 of PF00001.