Site saturation mutagenesis of 500 human protein domains reveals the contribution of protein destabilization to genetic disease

Antoni Beltran[1] , Xiang'er Jiang[2,3,4], Yue Shen[2,3,4] and Ben Lehner[1,5,6,7*]

1 Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain
2 BGI Research, Changzhou 213299, China
3 BGI Research, Shenzhen 518083, China
4 Guangdong Provincial Key Laboratory of Genome Read and Write, BGI Research, Shenzhen 518083, China
5 University Pompeu Fabra (UPF), Barcelona, Spain
6 Institució Catalana de Recerca i estudis Avançats (ICREA), Barcelona, Spain
7 Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK
*e-mail: bl11@sanger.ac.uk

Abstract

Missense variants that change the amino acid sequences of proteins cause one third of human genetic diseases[1]. Tens of millions of missense variants exist in the current human population, with the vast majority having unknown functional consequences. Here we present the first large-scale experimental analysis of human missense variants across many different proteins. Using DNA synthesis and cellular selection experiments we quantify the impact of >500,000 variants on the abundance of >500 human protein domains. This dataset - Human Domainome 1.0 - reveals that >60% of pathogenic missense variants reduce protein stability. The contribution of stability to protein fitness varies across proteins and diseases, and is particularly important in recessive disorders. We show how stability measurements can be combined with protein language models to annotate functional sites and that measurements made on a small number of proteins can be used to accurately predict stability changes across entire protein families using energy models. Domainome 1.0 demonstrates the feasibility of assaying human protein variants at scale and provides a large consistent reference dataset for clinical variant interpretation and the training and benchmarking of computational methods.

Introduction

The human genome encodes >20,000 proteins. Missense variants in nearly 5,000 of these proteins cause Mendelian diseases[2], however the functional consequences of nearly all missense variants in nearly all proteins are unknown[3–5]. Given the current human population size, most variants compatible with life are present in someone currently alive[6–8], making the large-scale experimental analysis of variant function a central challenge for human genetics[7–9]. However, experiments to date have mostly quantified variant effects in one or a few proteins[7,10]. Despite recent improvements[11–14], computational variant effect predictors are not deemed to provide sufficient evidence to classify clinical variants as pathogenic or benign[15]. They also do not identify the molecular mechanisms by which variants cause disease, information that is important for therapy development and clinical trial design. Whereas many disease variants are likely to destabilize proteins and reduce their abundance[8,16], others may affect specific molecular interactions, or cause gain-of-function phenotypes[17,18].

Most human proteins contain multiple independently folding structural units called domains[19,20]. For example, the human genome encodes more than 200 homeodomains that bind DNA to control gene expression and more than 250 PDZ domains that mediate protein-protein interactions[21,22]. The small size of protein domains (median ~100 amino acids) and their independent folding make them a useful target for large-scale experimental measurement of variant effects[23].

Here, using a highly-validated assay that quantifies the effects of variants on protein abundance in cells[24] we perform the first large-scale mutagenesis of human protein domains. In total we report the impact of >500,000 missense variants on the stability of >500 different human domains. This dataset, which we refer to as Human Domainome 1.0, provides a large reference dataset for the interpretation of clinical genetic variants, and for benchmarking and training computational methods for stability variant effect prediction. We use the dataset to quantify the contribution of stability changes to human genetic disease and how this varies across proteins and diseases. We also show how stability measurements can be combined with protein language models to annotate functional sites across proteins, and that measurements made on a small number of proteins can be used to accurately predict stability changes across entire protein families.

Results

Massively parallel saturation mutagenesis of human protein domains

We used mMPS synthesis technology (manuscript in preparation) to construct a library of 1,230,584 amino acid (aa) variants in 1,248 structurally diverse human protein domains (Supp. Table 1). In this library, which we refer to as Human Domainome 1.0, every amino acid is mutated to all other 19 aa at every position in each domain (Fig. 1a). Sequencing the library revealed it to be high quality, with coverage of 91% of designed aa substitutions (Supp. Table 1, Extended Data Fig. 1a). To quantify the impact of these variants on domain stability, we used an in cell selection system, abundance protein fragment complementation assay (aPCA)[24,25]. In aPCA, the protein domain of interest is expressed fused to a fragment of an essential enzyme and its concentration linearly determines the cellular growth rate over at least three orders of magnitude[24] (Fig. 1b). The effects of variants on protein abundance are quantified using high-throughput sequencing to measure the change in variant frequencies between the input and output cell populations in selection experiments (Fig. 1a). This strategy thus allows pooled cloning, transformation and selection of hundreds of thousands of variants in diverse proteins in a single experiment (Fig. 1a).

In total, we performed 27 transformation, selection and sequencing experiments (three independent replicates of nine sub-libraries). After filtering (see Methods), the final dataset consists of cellular abundance measurements for 563,534 variants in 522 human protein domains (Supp. Table 2). Abundance measurements were highly reproducible (median Pearson's correlation coefficient, r = 0.85 between replicates for all variants, Fig. 1c, Extended Data Fig. 1b). They also correlated well to independent *in vitro* measurements of protein fold stability[26,27] (median Spearman's rho = 0.73 with folding free energy changes ($\Delta\Delta G$), median percent explainable variance in aPCA scores = 77%, n=10 domains, Fig. 1f,g, Extended Data Fig. 1c). Moreover, they correlated well with high-throughput stability measurements from protease sensitivity assays[23] (median rho = 0.65, median percent explainable variance in aPCA scores = 60%, n = 13 domains, Fig. 1g, Extended Data Fig. 1d).

The 522 domains are structurally diverse, constituting 195 all alpha, 127 all beta, 47 alpha+beta, 1 alpha/beta, and 148 metal-binding zinc finger domains (Fig. 1e). In total, they cover 127 different domain families, including 14 families with 10 or more homologous domains (Fig. 2), and 97 families with only one or two domains. 275 of the domains are encoded by human disease genes with 108 domains containing annotated pathogenic variants. Across the dataset, mutations in the buried cores of the domains are more detrimental than in their surfaces (Fig. 1d), with mutations to polar amino acids having stronger destabilizing effects in cores and mutations to hydrophobics having stronger effects in surfaces (Extended Data Fig. 1e). Mutations to proline are the most detrimental overall (Fig. 1d), both in core and surface residues, with highly destabilizing effects in beta strands and helices and milder effects in coils (Extended Data Fig. 1e,f). The full mutagenesis dataset is presented in Extended Data Fig. 2.

Evaluation of variant effect predictors on protein abundance

Our dataset represents a nearly 10-fold increase in stability measurements for mutations in human protein domains (Fig. 3a) and so an opportunity to evaluate how well computational variant effect predictors (VEPs) predict changes in stability. Several general VEPs provide reasonable prediction of abundance changes, including the protein language model ESM1v[12] (median rho = 0.48) and the autoregressive transformer Tranception[13] (median rho= 0.44, Fig. 3b,c). Amongst dedicated stability predictors, the graph neural network thermoMPNN[28] performs best (median rho = 0.50, Fig. 3b,c), and is the best predictor overall. Interestingly, all tested stability predictors perform poorly on small zinc finger domains that require metal binding for stability (Extended Data Fig. 3). After excluding zinc fingers, thermoMPNN is still the best performing method overall (median rho = 0.57, Extended Data Fig. 3).

The contribution of stability to protein fitness

Fold stability is just one of many biophysical properties that contribute to protein function and protein sequence conservation during evolution. In particular, many proteins bind other proteins, nucleic acids and small molecules or catalyze enzymatic reactions, and these functions often trade off with stability[29]. The extent to which selection on protein sequences is driven by changes in stability rather than other biophysical properties is an important open question[30].

To address this, we compared experimentally quantified stability to evolutionary fitness quantified by ESM1v across >500,000 variants in >500 domains (Methods). We find that protein stability accounts for a median of 31% of the variance in protein fitness across all domains (Fig. 3d). However, the contribution of stability to fitness varies across domain families, with stability making a larger contribution to the fitness of all-beta domains (43% of the variance) than to that of all-alpha (27%) and mixed domains (27%) (Fig. 3e). This is consistent with the lower structural tolerance to mutations of beta sheets compared to alpha helices[31], and suggests that stability is a more important determinant of the fitness landscapes of all-beta domains. The contribution of stability to protein fitness also varies in domains with different molecular functions. For example, stability has a larger contribution to fitness in SH3, WW, PHD finger, and Pointed protein-protein interaction domains, and a lower contribution in DNA-binding HMG-box domains, nuclear hormone-type C4 zinc fingers, and homeodomains (Fig. 3f).

The contribution of stability to pathogenicity

Domainome 1.0 contains 3,652 variants with clinical annotations. 621 of these are classified as pathogenic/likely pathogenic (henceforth pathogenic), 322 as benign/likely benign (henceforth benign) and 2,709 as variants of uncertain significance (VUS), with 116 domains containing at least one pathogenic variant. Pathogenic variants are unevenly distributed across domains, with 75% of pathogenic variants contained in 25% of domains, and 41 domains containing only a single pathogenic variant (Extended Data Fig. 4a).

380/621 pathogenic variants (61%) cause domain destabilization (Z-test, FDR < 0.1) and 303/621 (48%) are strongly destabilizing (FDR < 0.1, normalized aPCA fitness < -0.3). This contrasts with 79/322 (25%) and and 50/322 (16%) of benign variants, respectively. However, the association between pathogenicity and destabilization varies across domain families (Fig. 4b, Extended Data Fig. 4b). For example, many pathogenic mutations in beta-gamma crystallins that cause cataract disease are strongly destabilizing (13/18, odds ratio (OR) = 11.98 compared to benign variants, p<2.2e-16, Fisher's exact test (FET), Fig. 4b). In contrast, a smaller proportion of pathogenic variants are strongly destabilizing in homeodomains (OR = 4.11), HMG-box domains (OR = 2.10), and CUT domains (OR = 1.65) (Fig. 4b), all of which bind DNA, suggesting that many pathogenic variants in these domain families affect other biophysical properties such as DNA binding.

We also quantified the relationship between stability and pathogenicity in all individual domains with at least 20 annotated clinical variants (n = 17). In the LIM domain 2 of FHL1, stability is an excellent classifier of pathogenic variants that cause muscular dystrophy-like disorders (Matthew's correlation coefficient, MCC = 1, Fig. 4c). Similarly, most variants in the SAM domain of TP63 that cause ectodermal dysplasia disorders are destabilizing (MCC = 0.83). In contrast, stability changes are a poorer predictor of variants in the methyl-binding domain (MBD) of MECP2 that cause Rett Syndrome (MCC = 0.4, Fig. 4c), consistent with many of these variants interfering with the methylated DNA binding function of MECP2[32–34]. This again suggests that destabilization is the major disease mechanism for some proteins but much less important for others.

Protein stability in recessive and dominant disorders

Comparing across all domains with at least 20 clinical variants, there is a striking correlation between how well stability explains pathogenicity and how well stability explains evolutionary fitness, as quantified by ESM1v (Pearson's r = 0.81, Fig. 4d). We therefore used the correlation between stability and evolutionary fitness to rank all 108 domains with at least one known pathogenic variant (Fig. 4e,f). Domains where stability is highly predictive of fitness include many PHD finger and PDZ domains (Fig. 4f). In these domains we expect stability changes to be the major driver of pathogenicity. In contrast, stability only poorly predicts the fitness of homeodomains, HMG-box, and nuclear hormone receptor-type zinc finger domains (Fig. 4f), suggesting that other molecular mechanisms will more frequently cause pathogenicity.

The contribution of stability to protein fitness also varies among genes with different modes of inheritance (Fig. 4e,f). In proteins mutated in recessive diseases, a median of 49% of variance in ESM1v predicted fitness is explained by changes in stability (Fig. 4e,f). This is consistent with many mutations in these proteins reducing function by destabilization during both evolution and in disease. In contrast, only 27% of variance in protein fitness is accounted for by stability changes in proteins mutated in dominant disorders, suggesting variants more frequently affect biophysical properties beyond stability. Changes in these additional properties may be the cause of dominance in human diseases[35]. In summary, mutagenesis of >500 domains suggests that

stability changes are an important cause of pathogenicity but that this varies across proteins, with changes in stability particularly important in recessive diseases.

Identification of functional sites

Mutations in binding interfaces, active sites and allosteric control sites typically have larger effects on function than can be accounted for by changes in protein stability[23,25,30]. For example, quantifying the effects of mutations on protein binding and abundance allows binding interfaces and allosteric sites to be comprehensively mapped[25,36]. We reasoned that a similar approach could be used to identify functional sites in hundreds of domains by combining our abundance measurements with evolutionary fitness quantified by protein language models.

For individual domains, protein abundance is non-linearly related to evolutionary fitness predicted by ESM1v (Fig. 5a, Extended Data Fig. 5a). We used sigmoidal curves to model this relationship (n = 426 domains, ESM1v fitness range > 10, WT aPCA fitness percentile < 30, Fig. 5, Extended Data Fig. 5). The residuals to these fits identify mutations with larger or smaller effects on evolutionary fitness than can be accounted for by changes in stability (Fig. 5a,b).

Using this approach identifies a total of 102,231 mutations with larger effects on evolutionary fitness than can be accounted for by changes in stability (24% of the total; Z-test FDR < 0.1, normalized aPCA residual > 0.3). These mutations are enriched in known functional sites annotated in the conserved domains database (CDD, OR = 2.72, FET p < 2.2e-16, n = 3,104 functional sites in 2,800 residues, Fig. 5d, Extended Data Fig. 5b). Defining evolutionary functional sites as residues with a weighted mean residual > 0.3 identifies a total of 5,231 sites in 426 domains. These sites are strongly enriched in CDD annotated sites (OR = 4.50, FET p < 2.2e-16, Fig. 5e) and identify many known DNA, RNA, and protein binding interfaces (Fig. 5e,f). However, these evolutionary functional sites also identify 1,942 sites in 180 domains without existing CDD annotations, and 1,873 additional sites in domains with other CDD annotations (Supp. Table 3).

Interestingly, many of the evolutionary functional sites are located not only in known functional sites but also in residues in close proximity to them. These include ‘second shell’ residues contacting DNA, RNA and protein binding interfaces (Fig. 5c,g). We speculate that these functional sites may be identifying novel allosteric sites, where sequence changes indirectly impact binding via energetic interactions with interface residues[25,36].

Conservation of mutational effects in homologous proteins

An important goal of Domainome 1.0 is to quantify the extent to which mutational effects are conserved in structurally homologous proteins. Mutations in homologous sites might be expected to have very similar effects. However, mutations can also interact energetically, resulting in changes in mutational effects depending upon the sequence context, a phenomenon known as epistasis[37]. If epistasis is prevalent then mutational effects will be poorly conserved in homologous proteins, limiting the extent to which experimental data from some proteins can be

used to predict stability changes in homologs[38]. Domainome 1.0 contains saturation mutagenesis data for five or more domains for 26 different domain families, allowing us to quantify the importance of epistasis for protein stability in structurally diverse protein folds.

To quantify the conservation of mutational effects, we fitted a thermodynamic model based on the Boltzmann partition function to all of the data for a domain family (Fig. 6a). The model assumes that mutations cause the same change in folding energy ($\Delta\Delta G$) in all homologous domains, and that the energetic effects of mutations combine additively with no specific epistasis (Fig. 6a,b). We first fitted this model to homeodomains, the most abundant domain family in the dataset, with 37 homologs. The model provides very good prediction of mutational effects across all 37 homeodomains (Pearson's r = 0.78 by ten-fold cross validation, Fig. 6c,d). The inferred free energy changes can be considered homolog-averaged mutational effects, providing an energy model describing the entire family (Fig. 6e,f). A linear model provides similarly good performance but with biased prediction residuals (r = 0.78, Extended Data Fig. 6). We additionally evaluated the performance of the Boltzmann energy model by leaving out single protein domains from the training dataset (n = 37 models) and found similarly good performance (median Pearson's r = 0.74). Predictive performance was, as expected, better for domains with more similar sequence to the training dataset, but reasonable across a wide range of sequence divergence (Pearson's r = 0.73, Fig. 6g).

The genetic architecture of protein stability across protein families

Extending this analysis to all 26 families with at least five homologs in Domainome 1.0, results in similarly accurate Boltzmann energy models for all families (median Pearson's r = 0.80, median percent of explainable variance = 80.7%, Fig. 6i, Extended Data Fig. 7a,b). The performance on individual left out homologs is also very good (median Pearson's r = 0.66, median percent of explainable variance = 73.5%, Fig. 6j). For most domains, predictions were best for the domains most similar to the training dataset but also good for domains with higher sequence divergence (Extended Data Fig. 7c). Predictions using these energy models were better than those made with thermoMPNN, the top performing stability predictor on our dataset (Fig. 6j).

The excellent performance of these additive energy models is both useful and important: it demonstrates that epistasis makes only a small contribution to protein stability across these levels of sequence divergence. Combinatorial mutagenesis of individual proteins suggests a similar conclusion[39]. The decay of predictive performance with sequence divergence does, however, suggest a role for epistasis in the evolution of protein stability. Indeed, we identify 25,410 mutations with evidence of epistasis as variants with large residuals to the Boltzmann model fits (FDR < 0.1 Z-test, | residual to the fit | > 0.05 h^{-1}, Extended Data Fig. 8a,b). These epistatic variants are enriched in the buried cores of protein domains (OR = 2.71, FET p = 1.05e-5) and depleted from protein surfaces (OR = 0.19, FET p = 9.79e-14, Extended Data Fig. 8c), across the full range of measured stabilities (Extended Data Fig. 8d) and protein families (Extended Data Fig. 8e). This data suggests that genetic interactions are more important for the

evolution of protein cores, consistent with these residues having a larger number of structural contacts than solvent-exposed residues (Extended Data 8f).

Energy models identify destabilizing mutations across entire domain families

Finally, the good performance of the Boltzmann energy models across all 26 families suggests that we can use them to provide proteome-wide stability predictions for these domain families. Using the models we made predictions for an additional 4,107,436 variants in 7,271 domains (Extended Data Fig. 9), including an additional 13,878 clinical variants, of which 1,310 are pathogenic, 951 are benign, and 11,617 are variants of uncertain significance (Fig. 7a,b). Consistent with the results for clinical variants with experimentally measured stability changes, 686 (52%) of these pathogenic variants are predicted to reduce protein stability (FDR < 0.1, OR = 2.95 compared to benign), and 452 (34%) are predicted to have strongly destabilizing effects (scaled $\Delta\Delta G > 0.3$, FDR < 0.1, OR = 5.95 compared to benign, Fig. 7b,c). These high-quality stability predictions provide a resource for interpreting clinical variants for entire protein families and suggest a strategy for expanding the Domainome proteome-wide by experimentally mutagenizing representative examples for all families.

Discussion

We have presented here the first large-scale experimental analysis of mutational effects in human proteins. Human Domainome 1.0 demonstrates the feasibility of quantifying human protein variant effects at scale and provides high confidence measurements of changes in protein abundance for 563,534 variants in 522 structurally diverse protein domains. The dataset increases the number of stability measurements for human variants about ten-fold and serves as a large and standardized reference dataset for the interpretation of clinical variants and for benchmarking and training computational methods.

An important goal of this project was to evaluate the contribution of protein stability to genetic disease and sequence evolution. We found that at least 60% of pathogenic missense variants reduce protein stability but that this varies quite extensively across domains and diseases, with stability changes particularly important in recessive disorders. Similarly, we found that the contribution of stability to evolutionary fitness varies across protein families, with a larger contribution in all-beta protein domains compared to other fold types. How this relates to protein evolvability and the importance of trade-offs between stability and other molecular functions will be interesting avenues for future research. A further goal of Domainome 1.0 was to quantify the conservation of mutational effects in structurally homologous proteins. Fitting additive energy models to the data for domain families revealed that mutational effects on stability are largely conserved in homologous domains, with a small contribution from epistasis that increases with sequence divergence. This energetic additivity allows proteome-wide prediction of stability changes for entire families and suggests an efficient strategy to complete a first draft of the Domainome by mutagenizing representative examples for every family.

To maximize diversity and experimental efficiency, Domainome 1.0 focussed on structurally diverse small protein domains. In future work it will be important to extend the approach to longer domains and to full length proteins. Advances in DNA synthesis[40], assembly[41] and mutagenesis[42,43] should facilitate this. In this study we have reported the effects of Domainome 1.0 variants on cellular protein abundance. However the same libraries can be re-used in future work to quantify variant effects on additional molecular traits, for example abundance in different cellular contexts[8,44,45], *in vitro* fold stability[23], protein-protein[46,47] and protein-nucleic acid[48] interactions, protein localisation[49], aggregation[50], and allostery[25,36]. It will also be important to include more complex genetic variants such as insertions and deletions[51] and to extend to proteome-wide coverage, including extracellular and transmembrane proteins[52].

Domainome 1.0 is a first important step in the comprehensive experimental analysis of human protein variants. It forms part of an ongoing global effort to determine the consequences of every mutation in every human protein and to produce reference atlases for the mechanistic interpretation of clinical variants[7]. Beyond human genetics, it is also part of a broader effort to produce large, well-calibrated datasets that quantify how changes in sequence alter the biophysical properties of proteins. We believe that such multimodal biophysical measurements for millions of proteins and variants will enable machine learning approaches to be effectively

brought to bear on the generative functions of molecular biology, allowing accurate prediction and engineering from sequence[53].

Figure legends

Figure 1: Mutating the Human Domainome.
a. Experimental strategy for multiplexed generation of sequence-to-stability maps of human protein domains based on pooled cloning, transformation and selection of pooled saturation mutagenesis libraries. **b.** DHFR complementation assay to measure in vivo stability of variant human protein domains (abundancePCA[24], aPCA). Variants causing unfolding and degradation of the target domain result in a decrease in concentration in the cell, leading to impaired cell growth. **c.** Replicate correlations of aPCA fitness scores. **d.** Comparison of aPCA fitness score distributions of synonymous, missense, and nonsense variants (top panel), core (rSASA < 25%) and surface variants (bottom left panel), and mutations to proline and other missense variants (bottom right panel). **e.** Structural similarity network of domains retained in the final dataset. Nodes represent protein domains, and edges represent foldseek hit probabilities in pairwise structural alignments. Colors correspond to SCOP structural classes. **f.** Correlations between in vitro $\Delta\Delta$Gs and aPCA scores in four representative domains. **g.** Distribution of fraction of explainable variance in aPCA scores explained by in vitro $\Delta\Delta$Gs[26,27] or stabilities derived by a high-throughput proteolysis assay[23].

Figure 2: Deep mutational scans of protein homologs.
Five examples of deep mutational scanning aPCA datasets of the most abundant protein families. Heatmaps depict the effects of mutating every residue in the protein domains (x axis) to all possible 19 amino acids (y axis).

Figure 3: The contribution of protein stability to evolutionary fitness.
a. Total number of stability measurements for human single missense variants in the Domainome 1.0 compared to previous comprehensive datasets of protein stability. **b.** Correlations between aPCA normalized fitness and ESM1v predictions (top panel) or thermoMPNN stability predictions (bottom panel) for the YBX1 cold-shock domain. **c.** Comparison of performance of variant effect and protein stability predictors. The distribution of Spearman's r across all domains is shown for each predictor, with the median Spearman's r labeled in text. **d.** Quantification of the degree of sequence constraint attributable to protein stability across protein families with at least 5 measured homologs. Colors correspond to SCOP structural classes. **e-f.** Quantification of the degree of sequence constraint attributable to protein stability across domains stratified by SCOP structural classes (**e**) and their molecular functions (**f**).

Figure 4: The contribution of destabilization to genetic disease.
a. Distributions of normalized aPCA fitness value of pathogenic, benign, uncertain, and gnomad variants (allele frequency > 10^{-5}). **b.** Proportion of strongly destabilizing variants (FDR < 0.1, normalized fitness < - 0.3) in the full dataset (top bars), and in several protein families. **c.**

Classification of clinical variants using stability effects. Density plots in yellow depict the distribution of aPCA normalized fitness of all variants in each domain. Pathogenic variants are marked as vertical red lines, and benign and gnomad variants (allele frequency > 10^{-5}) as vertical blue lines. **d.** Correlation between classification performance of stability on clinical variants (MCC), and the fraction of explainable variance in aPCA fitness explained by ESM1v for domains with at least 20 clinical and gnomad variants (allele frequency > 10^{-5}). **e.** fraction of explainable variance in aPCA fitness explained by ESM1v in domains grouped by the mode of inheritance of their associated disorders. **f.** Ranking of all 108 domains containing at least 1 pathogenic variant by the fraction of explainable variance in aPCA fitness explained by ESM1v (top domains = highly explained by stability, bottom domains = poorly explained by stability).

Figure 5: Identification of functional sites.
a. Sigmoidal curves to model the relationship between stability and overall fitness of variants in the dsRNA binding domain of ADARB1. Data points are colored by their residuals to the fit. **b.** Heat map depicting the residuals to the fit of all measured variants. The dsRNA binding site residues are marked with blue bold letters. The blue boxes capture all the contiguous stretches of sequence containing the binding site. **c.** AlphaFold2 predicted structure of the dsRNA binding domain of ADARB1, with residues coloured by the weighted mean residuals to the fit. Residues forming the dsRNA binding site are marked with a red silhouette. **d.** Distribution of residuals for mutations in annotated functional sites, and in other protein residues. **e.** Distribution of weighted mean residuals of protein residues in annotated functional sites, and of other residues. **f.** Enrichment of functional sites (residues with weighted mean residuals > 0.3) in several types of CDD annotated functional sites. Error bars depict the 95% confidence interval for the odds ratio of enrichment. **g.** AlphaFold2 predicted structures of representative domains containing functional sites, with residues coloured by the weighted mean residuals. Residues that belong to functional sites are marked with a red silhouette.

Figure 6: The genetic architecture of protein stability.
a. Two-state folding equilibrium and corresponding thermodynamic model. ΔG_f, Gibbs free energy of folding; K_f, folding equilibrium constant; p_f, fraction folded; ff, nonlinear function of ΔG_f; R, gas constant; T, temperature in Kelvin. **b.** Neural network architecture used to fit thermodynamic models to protein families. **c.** Relationship between predicted fitness and additive trait (ΔG) in a Boltzmann model fit to the homeodomain family (PF00046). **d.** Correlation between observed fitness values and MoCHI fitness predictions of the homeodomain family for test set variants (10-fold cross validation; variants held out in any of the 10 training folds are shown). **e.** Structure of a representative homeodomain, with residues colored by the weighted mean homolog-averaged $\Delta\Delta G$ of mutations at each position. **f.** Heatmap depicting inferred homolog-averaged folding $\Delta\Delta G$ defining the stability of the homeodomain fold. **g.** Performance of the energy model on held out homeodomains as a function of the average genetic distance to the training set. **h.** Structures of representative domains of all modeled families with residues colored by the weighted mean homolog-averaged $\Delta\Delta G$ of mutations at each position. **i.** Summary of energy model performance on held out variants across protein families (10-fold cross validation). **j.** Summary of energy model

performance on held out homologs across protein families, compared to the performance of the top stability predictor (thermoMPNN).

Figure 7: Extension of stability predictions to entire domain families.
a. Number of pathogenic clinical variants covered by the predictions of homolog-averaged folding ΔΔGs. Colors represent whether variants are present in the aPCA fitness dataset (red) or not (blue). **b.** Distributions of scaled folding ΔΔG values of pathogenic, benign, uncertain, and gnomad variants (allele frequency > 10^{-5}). **c.** Proportion of strongly destabilizing variants (FDR < 0.1, scaled folding ΔΔG > 0.3) across classes of clinical variants.

Supplementary figure legends

Extended Data Figure 1: Mutating the human Domainome at scale.
a. Histograms depicting sequencing coverage of the 9 site saturation mutagenesis libraries selected in this study. **b.** Correlations (Pearson's r) between replicates for the 9 libraries. **c.** Correlations (Spearman's rho) between aPCA scores and in vitro ΔΔGs for individual protein domains. **d.** Correlations (Spearman's rho) between aPCA scores and ΔΔGs derived through a high-throughput proteolysis assay for individual protein domains. **e.** Distributions of stability effects of mutations in protein cores and surfaces across the entire Domainome 1.0. Colors represent the physicochemical properties of the mutated amino acids (polar/charged = D, E, K, N, Q, R, S, T, hydrophobic = A, F, I, L, M, V, W, Y). **f.** Distributions of stability effects of mutations in different secondary structure types.

Extended Data Figure 2: Deep mutational scans of 522 protein domains.
Domains are ranked according to data quality (see Methods).

Extended Data Figure 3: The contribution of protein stability to evolutionary constraint.
a. Comparison of performance of variant effect and protein stability predictors in non-zinc finger domains (left) and zinc finger domains (right). The distribution of Spearman's rho across all domains is shown for each predictor, with the median Spearman's rho labeled in text. **b.** Correlation between Pearson's r and Spearman's rho-based variance explained estimates of the contribution of stability to protein evolutionary constraint.

Extended Data Figure 4: The contribution of protein destabilization to genetic disease.
a. Gini plot depicting the cumulative number of pathogenic variants present in Domainome 1.0 domains ranked by pathogenic variant counts (from highest to lowest). **b.** Proportions of non-destabilizing, mildly destabilizing (FDR < 0.1, 0 > normalized aPCA fitness > -0.3), and strongly destabilizing (FDR < 0.1, normalized aPCA fitness < -0.3) variants in the pathogenic and benign sets, for several protein families. Families with at least 10 variants in the pathogenic and in the benign classes are shown.

Extended Data Figure 5: Identifying evolutionary functional sites using language models and abundance measurements.

a. Representative examples of linear model and sigmoid model fits to the relationship between ESM1v predictions and abundance. **b.** Residuals to the fit of protein domain variants in several classes of annotated functional sites. Classes of functional sites with at least 50 residues annotated across the full domainome dataset are shown.

Extended Data Figure 6: Comparison of linear and Boltzmann energy model fits to a protein family.
Correlations between observed fitness values and energy model predictions for test set variants held out in any of the 10 training folds, for an energy model with a linear activation function (a) or a Boltzmann activation function (b) at the additive trait layer.

Extended Data Figure 7: Thermodynamic models of protein families.
a. Correlations between observed fitness values and energy model fitness predictions for test set variants held out in any of the 10 training folds, for Boltzmann models across protein families. **b.** Inferred homolog-averaged folding $\Delta\Delta$Gs across protein families, normalized such that the 2.5th percentile of the data corresponds to a scaled $\Delta\Delta$G of -1. **c.** Correlation between model performance on left out protein domains, and the averaged genetic distance of the left-out domains to the rest of the training set, quantified as the hamming distance.

Extended Data Figure 8: Protein cores are more epistatic than surfaces.
a-b. Representative examples of an epistatic site (a) and a non-epistatic site (b) in the PDZ family. Epistatic mutations (FDR < 0.1, | residual | > 0.05 h^{-1}) are shown in blue, and non-epistatic mutations are shown in pink. The epistatic site is enriched in epistatic mutations. **c.** Distributions of enrichment of epistatic mutations in protein core sites (rSASA<25% in >75% of homologs), surface sites (rSASA>25% in >75% of homologs), or changing sites. **d.** Enrichment of epistatic mutations in core (n = 368), surface (n = 1,039), and changing (n = 193) sites as a function of the average fitness of all variants per site. The running median of each class of sites across the fitness range (window size = 0.02 h^{-1}) is shown as a colored line. **e.** Distribution of enrichment of epistatic mutations in core, surface, and changing sites across protein families. **f.** Distribution of total contact numbers per residue in core (n = 87,316) and surface (n = 193,797) residues across the Domainome, estimated from getcontacts.

Extended Data Figure 9: Proteome-wide extension of stability predictions.
a. Number of measured and predicted protein domains, by family. **b.** Number of measured and predicted protein variants, by family.

Data availability

All DNA sequencing data have been deposited in the Gene Expression Omnibus under the accession number GSE265942. All aPCA measurements and their associated errors are available as Supplementary Table 2. Weighted mean residuals of the comparisons between abundance and evolutionary fitness predictions are available as Supplementary Table 3.

Homolog-averaged $\Delta\Delta G$ values and their associated errors mapped to homologous domains proteome-wide are available as Supplementary Table 4.

Code availability

Source code used to perform all analyses and to reproduce all figures in this work is available at: https://github.com/lehner-lab/domainome. Files required to reproduce the analyses can be downloaded at https://zenodo.org/records/11043643.

Author contributions

A.B. designed the libraries and performed all experiments and analyses. X.J. and Y.S. synthesized the variant libraries. A.B. and B.L. conceived the project, designed analyses, and wrote the manuscript, with input from all authors.

Competing interests

B.L. is a founder and shareholder of ALLOX.

Acknowledgements

A.B. and B.L. were funded by a European Research Council (ERC) Advanced (883742) grant, the Spanish Ministry of Science and Innovation (LCF/PR/HR21/52410004, EMBL Partnership, Severo Ochoa Centre of Excellence), the Bettencourt Schueller Foundation, the AXA Research Fund, Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR, 2017 SGR 1322), and the CERCA Program/Generalitat de Catalunya. A.B was funded by an EMBO (ALTF 183-2020) and a Marie Skłodowska-Curie (101030961) fellowship. X.J. and Y.S were funded by the National Natural Science Foundation of China (No.32322047) and Jiangsu Provincial Department of Science and Technology (No.BM2023009). We thank all members of the Lehner Lab for helpful discussions and suggestions.

Materials and methods

Media

- LB: 10 g/L Bacto-tryptone, 5 g/L Yeast extract, 10 g/L NaCl. Autoclaved 20 min at 120ºC.
- YPDA: 20 g/L glucose, 20 g/L Peptone, 10 g/L Yeast extract, 40 mg/L adenine sulphate. Autoclaved 20 min at 120ºC.
- SORB: 1 M sorbitol, 100 mM LiOAc, 10 mM Tris pH 8.0, 1 mM EDTA.
- Filter sterilized (0.2 mm Nylon membrane, ThermoScientific).
- Plate mixture: 40% PEG3350, 100 mM LiOAc, 10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0. Filter sterilized.
- Recovery medium: YPD (20 g/L glucose, 20 g/L Peptone, 10 g/L Yeast extract) +0.5 M sorbitol. Filter sterilized.
- SC -URA: 6.7 g/L Yeast Nitrogen base without amino acid, 20 g/L glucose, 0.77 g/L complete supplement mixture drop-out without uracil. Filter sterilized.
- SC -URA/ADE: 6.7 g/L Yeast Nitrogen base without amino acid, 20 g/L glucose, 0.76 g/L complete supplement mixture drop-out without uracil, adenine and methionine. Filter sterilized.
- MTX competition medium: SD –URA/ADE + 200 µg/mL methotrexate (BioShop Canada Inc., Canada), 2% DMSO.
- DNA extraction buffer: 2% Triton-X, 1% SDS, 100mM NaCl, 10mM Tris-HCl pH8, 1mM EDTA pH8.

Library design and cloning

We sampled PFAM-annotated human protein domains (version 34.0) of intracellular human proteins (not defined as “extracellular”, “transmembrane”, or “secreted” in UniProt), prioritizing proteins that had at least one annotated pathogenic variant ClinVar[3]. We additionally included protein domains with *in vitro* stability measurements to benchmark the assay. We designed the library in two rounds: we first designed and selected two libraries (A1 and B3) containing single mutants of a total of 485 domains that had been previously tested to grow as WTs. Based on the results of the first libraries we designed a second set of libraries (C1 to C7) including only domains with an average AlphaFold2 pLDDT > 50 and at least 10% of residues buried in the core (rSASA < 0.25). This second set contained 631 domains that had been previously tested to grow as wildtype, and 132 additional domains not previously tested.

The domain sequences were codon optimized with emboss backtranseq[54] using a *S. cerevisiae* codon usage table, and excluding the GCT alanine codon to prevent the appearance of HindIII restriction sites. The domain sequences were fused to 5´ gggctgctctagaatggctagc and 3´ aagcttggcggtggcgggtctg constant regions containing NheI and HindIII restriction sites for cloning. Libraries were synthesized by BGI using the mMPS technology (manuscript in preparation). Libraries were cloned at a variant coverage of ~100x or greater by restriction digestion and ligation into pGJJ162, an aPCA assay plasmid where the DHFR3 fragment is fused at the C-terminus of the target protein domains and the fusion is driven by the CYC

promoter, and the DHFR1,2 fragment is expressed at high levels as is driven by the GPD promoter.

Large-scale transformation and competition assay

Variant libraries were transformed in triplicate with a coverage of 20x or greater (with a mean coverage across libraries ~170x). For each transformation, we grew a 1 L YPDA culture of late log phase *S. cerevisiae* BY4741 cells (OD_{600}~0.8-1), harvested the cells by centrifugation (5 minutes, 4000 g), resuspended in 43 mL SORB medium and incubated for 30 min on a shaker at room temperature. 875 µL 10 mg/mL previously boiled (5 min, 100 ºC) ssDNA was added to the cells, followed by 17.5 µg library plasmid DNA and 175 mL Plate Mixture. The mix was incubated for 30 min on a shaker at room temperature. 17.5 mL DMSO were then added, and the cells were split in 50 mL Falcon tubes for 20 min heat shock at 42ºC. Following incubation, cells were pooled and harvested by centrifugation, the supernatant was discarded with a pump, and cells were resuspended in 250 mL Recovery Media and incubated at 30ºC for 1h. Cells were then centrifuged for 3 min at 3,000 g and transferred into 1 L SC –URA. 10 µL of this culture were immediately plated onto SC -URA selective plates to monitor transformation efficiency. The rest of the culture was incubated for one or two overnights at 30ºC.

SC -URA cultures were used to inoculate a 1 L culture of SC -URA-ADE at an OD_{600}=0.2-0.4, which was grown overnight (input culture). Cells from this culture were inoculated in 1 L SC -URA/ADE +200 µg/mL MTX to select stably expressed protein domain variants. The remaining input cells grown SC -URA/ADE were harvested and frozen for DNA extraction. MTX cultures were left to grow overnight to an OD_{600}=1.6-2.5, corresponding approximately to 5 generations, harvested, and frozen for DNA extraction (outputs).

DNA extraction, plasmid quantification and sequencing library preparation

Total DNA was extracted from yeast pellets equivalent to 50 mL of cells at OD_{600}=1.6 as described in our previous work[25,36]. Plasmid concentrations in the resulting samples were quantified by against a standard curve of known concentrations by qPCR, using oGJJ152 and oGJJ153 as qPCR primers that amplify in the origin of replication of the aPCAa assay plasmid.

To generate the sequencing libraries, we performed two rounds of PCR amplification. In the first round, we used primer pools (oTB595+ and oTB748+) flanking the inserts that introduce frame-shifting nucleotides between the Illumina adapters and the sequencing region of interest. To maintain variant representation, we carried out eight 100 µL PCR1 reactions per sample, each of which starting with 125 million plasmid molecules that we amplified for 8 cycles. The reactions were column-purified (QIAquick PCR purification kit, QIAGEN), and the purified product was amplified further using the standard i5 and i7 primers to add the remainder of the Illumina adapter sequences and the demultiplexing indices (dual indexing) unique to each sample. We carried out a total of 8 100 µL PCR2 reactions per sample, each starting with 20-40 ng of purified product, that was amplified for 8 more cycles. The resulting amplicons were run on a 2% agarose gel to quantify and pool the samples for joint purification, and to ensure the

specificity of the amplification and check for any potential excess amplification problems. The final libraries were size selected by electrophoresis on a 2% agarose gel, and gel-purified (QIAEX II Gel Extraction Kit ,QIAGEN). The amplicons were subjected to Illumina paired end 2x150 sequencing on a NextSeq2000 instrument at the CRG Genomics facility.

Sequencing data processing and normalization

FastQ files from paired-end sequencing of all aPCA selections were processed with DiMSum[55] v1.2.11 (https://github.com/lehner-lab/DiMSum) using default settings with minor adjustments. The option "barcodeIdentityPath" was used to specify a variants file in order to recover designed variants only (NNK mutations in the protein domains present in each sublibrary), and starting and final culture optical densities and selection times were specified to infer absolute growth rates across libraries.

Data filtering and normalization

We computed several quality control metrics on a per-domain basis. First, the wild-type position in the growth rate distribution, as the difference between the wild-type and the 95th percentile of the distribution, divided by the difference between the 95th and 5th percentile in the distribution (growth rate 90% range). Second, the Pearson's correlation coefficient between replicates for missense variants. Third, the correlation between growth rate and solvent accessibility for all variants in each domain. And fourth, the correlation between growth rate and hydrophobicity of all variants in each domain (measured as the first principal component of a comprehensive table of amino acid properties[56]). These four quality metrics were highly correlated with each other, and we integrated them using principal component analysis. The first PC explained 64% of the variance, and was used as a single quality metric to rank domains.

We retained all domains ranked 600 or lower, with a missense Pearson's $r > 0.485$, and with more than 50% of variants measured (with at least 10 counts in at least 1 replicate), resulting in 538 domains. We removed from this set 16 additional domains that were not compatible with 2-state folding, either showing many large effect abundance-increasing mutations (O75364_PF00046_64, O75956_PF09806_73, P10242_PF00249_89, P52952_PF00046_140, Q13263_PF00643_205, Q86TZ1_PF13181_60, Q8IX03_PF00397_1, Q8NDW8_PF13181_799, Q9Y2H9_PF17820_968, Q9Y6M9_PF05347_15), narrow fitness ranges (P35637_PF00641_421, EHEE-rd2-0005_1, HHH-rd2-0133_1), or clear correlations to hydrophobicity (E9PAV3_PF19026_2040, HEEH-rd3-0223_1, Q5VTD9_PF00096_193), resulting in a final set of 522 domains.

We normalized the growth rates and growth rate errors within each protein domain by linearly scaling the data such that the WT normalized fitness equals zero, and the 2.5th percentile of the distribution of growth rates of all missense variants plus the WT is equivalent to a normalized fitness of -1.

In all boxplots shown in the manuscript, the central line represents the median, the upper and lower hinges correspond to the first and third quartiles, with the upper whisker extending from the hinge to the largest value no further than 1.5 * IQR (inter-quartile range) from the hinge, and the lower whisker extending from the hinge to the smallest value no further than 1.5 * IQR from the hinge.

Structural similarity network representation

We generated a structural distance matrix based on foldseek[57] hit probabilities. We computed all pairwise alignments between the domains in the final retained set by first creating a foldseek database 'foldseek_domainome_db' containing all domains, and then searching all domains against the database using foldseek easy-search *.pdb foldseek_domainome_db foldseek_easy_allvsall tmp --format-output "query,target,alntmscore,qtmscore,ttmscore,prob" --exhaustive-search TRUE -e inf.

To visualize the similarity network, we imported the domains as nodes and the foldseek probabilities as edges into Gephi[58], and applied two layout algorithms: first the Fruchterman Reingold algorithm to equilibrium, which resulted in a clear separation of the different SCOP classes, followed by a Force Atlas 2 layout algorithm (preventing node overlap), which accentuated the separation between the different domain family clusters within each SCOP class.

Comparison to reference ΔΔG datasets

To compare aPCA measurements with *in vitro* ΔΔGs, we used domains containing at least 10 variants in more than a single residue measured in vitro and in aPCA, and with a range of at least 2 kcal/mol for the in vitro measurements, and of 0.075 h^{-1} log growth rate units for the aPCA measurements (n = 10 domains). To compare aPCA measurements with ΔΔGs derived from high-throughput proteolysis deep mutational scans[23], we used protein domains with an overlap (overlapping length / total alignment length) of at least 80%, and an aPCA measurement range of 0.075 h^{-1} log growth rate units (n = 12). The total number of human missense variants measured in the mega-scale dataset was retrieved from the high-confidence dataset excluding double mutants and domain duplicates (from different PDB entries and genetic backgrounds of the same PDB entry).

Comparison to variant effect predictors

We generated ESM1v[12] predictions (https://github.com/facebookresearch/esm) using the domain sequences alone as input, and using sliding windows across the human proteome (size=1000 aa, step=250 aa). For each domain, we used the predictions corresponding to the window in which the domain is most centered. We obtained EVE[11], popEVE[59] and Tranception[13] scores from https://pop.evemodel.org/. We used precomputed RaSP scores[60], and generated DDMut[61] and thermoMPNN[28] predictions with the aid of AlphaFold2 structure predictions. We

used Spearman's rho as an agreement metric as the relationship between the predictions and aPCA fitness.

To estimate the fraction of explainable variance in aPCA fitness explained by ESM1v predictions, we computed R^2 between aPCA replicates, and R^2 between ESM1v predictions and each of the aPCA replicates. We estimated the fraction of variance explained as the mean ESM1v-aPCA R^2 divided by the mean inter-replicate R^2. We also calculated equivalent estimates using Spearman's rho, which estimate the fraction of the variance explained in the ranks. The two were highly correlated (r = 0.94) and we report the Pearson's r-based version for ease of interpretation.

Analysis of the stability effects of pathogenic variants

We identified destabilizing variants using a one-tailed z-test. Z-scores were calculated as the normalized fitness divided by the normalized fitness error, a p-value was derived on the basis of a normal distribution, and FDR multiple test correction was applied. Destabilizing variants were defined as variants with FDR < 0.1, and strongly destabilizing variants as FDR < 0.1 and a normalized aPCA fitness < -0.3. We obtained clinical variant annotations from ClinVar (January 2024 version) and from the UniProt API. Variant annotations were highly consistent between the two sources, and were merged. We tested for enrichments of clinical variant classes in stability classes using a two-tailed Fisher's exact test.

To estimate to what extent destabilization explains pathogenic mutations in individual domains, we used AUROC and Mathew's correlation coefficient (MCC), which were highly correlated. To increase statistical power, we included gnomad variants with an allele frequency of at least 10^{-5} as benign. To extend the analysis to a larger number of domains with small numbers of pathogenic variants, we used the fraction of explainable variance in aPCA fitness explained by ESM1v as described above. MCCs based on clinical and gnomad variants correlated well with the fraction of explainable variance in domains with at least 20 clinical and gnomad variants, validating the approach.

Analysis of functional sites

We carried out this analysis for domains where (1) the WT is above the 30th percentile in the fitness distribution, and (2) the range between the 5th and 95th percentile of the distribution of ESM1v predicted fitness is > 10 (n = 426 domains). We used sigmoid curves to model the relationship between normalized aPCA fitness (stability) and predicted function (ESM1v) of all variants in each individual domain, with an upper bound of 0 (WT aPCA normalized fitness) and a lower bound of -1:

$$f_{aPCA} = -1 + \frac{1}{1+e^{-(f_{ESM1v} - xmid)/scal}}$$

Where f_{aPCA} is the aPCA normalized fitness, f_{ESM1v} is the ESM1v predicted fitness, *xmid* is the midpoint of the sigmoid and *scal* is the steepness parameter. To prioritize fitting the low stability variants, we weighted the fit by the aPCA normalized fitness:

$$w_i = max(f_{aPCA}) - min(f_{aPCA}) - (f_{aPCA,i} + 1)$$

We used a two-tailed z-test to identify mutations whose effects on fitness cannot be accounted for by stability effects. We calculated z-scores as the aPCA residuals to the fit divided by the aPCA error, derived p-values based on the normal distribution, and performed multiple testing correction using Benjamini-Hochberg's false discovery rate (FDR). We additionally calculated per-residue mean residuals weighted by the aPCA normalized fitness error.

We obtained residue-level functional site annotations corresponding to the Conserved Domains Database (CDD)[62] using the InterPro API.

Thermodynamic modeling of protein domain families

We used MoCHI[63] to fit two-state thermodynamic models on a per-family basis (for families with at least 5 homologs and variants with a mean count >29). We specified a neural network architecture consisting of a single additive trait layer for shared folding energies across the family and a shared linear transformation layer. We used a "TwoStateFractionFolded" transformation derived from the Boltzmann distribution function that relates energies to proportions of folded protein. To map homologous positions across families, we used PFAM alignments. To input into MoCHI, we recoded each domain WT sequence as an indel sequence as long as the PFAM alignment, plus additional positions in the alignment with variants that encode the WT identity of each homolog. Mutations in each domain were encoded in the corresponding alignment position. This design allows MoCHI's hot-encoding of both the mutations and the WT identities simultaneously, and the joint fitting of mutation $\Delta\Delta G$ and starting ΔG of each homolog. We trained the model using a 10-fold cross-validation approach and evaluated the performance on held-out data.

We additionally fitted a linear model, and a "TwoStateFractionFolded" model with shared energies across all homologs of a family but with domain-specific linear transformations. These models resulted in highly correlated inferred energies to the original Boltzmann model (Pearson's r = 0.972 for the linear model, and r = 0.938 for the Boltzmann model with domain-specific linear transformations) and similar performances on held out data across families. We discarded the linear model due to highly biased residuals, and chose the "TwoStateFractionFolded" model without homolog-specific linear scaling as the simplest of the remaining models.

We further evaluated the performance of the "TwoStateFractionFolded" model in families with at least 10 homologs by training the model leaving out a single domain at a time. The genetic distance (hamming distance and BLOSUM62 distance) of each left-out homolog to each of the

homologs that went into training was calculated and averaged, to compare to model performance. The fraction of explainable variance in aPCA scores accounted for by the models was estimated as the R^2 between observed and predicted fitness divided by the R^2 between replicates.

Epistasis analysis of protein domain families

We identified epistatic variants with large residuals to the MoCHI fits using a two-tailed z-test. We calculated z-scores as the residuals to the fit divided by the aPCA fitness errors, derived p-values on the basis of a normal distribution, and applied FDR multiple testing correction. Epistatic mutations were defined as those with |residual| > 0.05 and FDR < 0.1. Enrichments of alignment sites in epistatic variants were calculated, and significantly enriched sites were identified using a two-tailed Fisher's exact test. Epistatic sites were defined as those with a $\log_2(OR)>1.5$ and a Fisher's Exact test FDR < 0.05. We classified alignment sites as core residues (rSASA < 25% in at least 75% of homologs), surface residues (rSASA > 25% in at least 75% of homologs), or changing residues (the rest). We defined contacts using getcontacts[64].

Supplementary tables

Supplementary table 1: Library design: sequences, library statistics.
Supplementary table 2: Fitness scores and errors.
Supplementary table 3: Weighted mean residuals of abundance to evolutionary fitness predictions.
Supplementary table 4: Homolog-averaged $\Delta\Delta G$ predictions across families mapped to homologous domains proteome-wide.

Supplementary table 5: Plasmids

pGJJ162	aPCA empty	Available upon request (Material Transfer Agreement required).

Supplementary table 6: Oligonucleotides

qPCR quantification oligos	
oGJJ152	GCCTACATACCTCGCTCTGC
oGJJ153	CAACCCGGTAAGACACGACT
Frameshifting PCR1 oligos	
oGJJ595	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTGCTCTAGAATGGCTAGC
oGJJ595_+1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNGCTGCTCTAGAATGGCTAGC
oGJJ595_+2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNGCTGCTCTAGAATGGCTAGC
oGJJ595_+3	ACACTCTTTCCCTACACGACGCTCTTCCGATCTHWAGCTGCTCTAGAATGGCTAGC
oGJJ595_+4	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNHTAGCTGCTCTAGAATGGCTAGC
oGJJ595_+5	ACACTCTTTCCCTACACGACGCTCTTCCGATCTSSSAAGCTGCTCTAGAATGGCTAGC
oGJJ748	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCCGCCACCGCCAAG
oGJJ748_+1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNCCCGCCACCGCCAAG
oGJJ748_+2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTNCCCGCCACCGCCAAG
oGJJ748_+3	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNTGCCCGCCACCGCCAAG
oGJJ748_+4	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGWWWCCCGCCACCGCCAAG
oGJJ748_+5	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTACTWWCCCGCCACCGCCAAG
oGJJ748_+6	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCATWWWCCCGCCACCGCCAAG
oGJJ748_+7	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTATADDCCCGCCACCGCCAAG
oGJJ748_+8	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTGGGDWWWCCCGCCACCGCCAAG

References

1. Karbassi, I. *et al.* A Standardized DNA Variant Scoring System for Pathogenicity Assessments in Mendelian Disorders. *Hum. Mutat.* **37**, 127–134 (2016).
2. Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Res.* **47**, D1038–D1043 (2018).
3. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–5 (2014).
4. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
5. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
6. Shirts, B. H., Pritchard, C. C. & Walsh, T. Family-Specific Variants and the Limits of Human Genetics. *Trends Mol. Med.* **22**, 925–934 (2016).
7. Fowler, D. M. *et al.* An Atlas of Variant Effects to understand the genome at nucleotide resolution. *Genome Biol.* **24**, 147 (2023).
8. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).
9. Findlay, G. M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, (2018).
10. Weile, J. *et al.* A framework for exhaustively mapping functional missense variants. *Mol. Syst. Biol.* (2017) doi:10.15252/msb.20177908.
11. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).
12. Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv* 2021.07.09.450648 (2021) doi:10.1101/2021.07.09.450648.
13. Notin, P. *et al.* Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. (2022).
14. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* (2023) doi:10.1126/science.adg7492.
15. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
16. Redler, R. L., Das, J., Diaz, J. R. & Dokholyan, N. V. Protein Destabilization as a Common Factor in Diverse Inherited Disorders. *J. Mol. Evol.* **82**, 11–16 (2016).
17. Backwell, L. & Marsh, J. A. Diverse Molecular Mechanisms Underlying Pathogenic Protein Mutations: Beyond the Loss-of-Function Paradigm. *Annu. Rev. Genomics Hum. Genet.* **23**, 475–498 (2022).
18. Badonyi, M. & Marsh, J. A. Proteome-scale prediction of molecular mechanisms underlying dominant genetic diseases. *bioRxiv* 2023.09.08.556798 (2024)

doi:10.1101/2023.09.08.556798.
19. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
20. Andreeva, A., Kulesha, E., Gough, J. & Murzin, A. G. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res.* **48**, D376–D382 (2020).
21. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).
22. Amacher, J. F., Brooks, L., Hampton, T. H. & Madden, D. R. Specificity in PDZ-peptide interaction networks: Computational analysis and review. *J Struct Biol X* **4**, 100022 (2020).
23. Tsuboyama, K. *et al.* Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**, 434–444 (2023).
24. Levy, E. D., Kowarzyk, J. & Michnick, S. W. High-resolution mapping of protein concentration reveals principles of proteome architecture and adaptation. *Cell Rep.* **7**, 1333–1340 (2014).
25. Faure, A. J. *et al.* Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175–183 (2022).
26. Nikam, R., Kulandaisamy, A., Harini, K., Sharma, D. & Gromiha, M. M. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Res.* **49**, D420–D424 (2021).
27. Xavier, J. S. *et al.* ThermoMutDB: a thermodynamic database for missense mutations. *Nucleic Acids Res.* **49**, (2021).
28. Dieckhaus, H., Brocidiacono, M., Randolph, N. Z. & Kuhlman, B. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proc. Natl. Acad. Sci. U. S. A.* **121**, e2314853121 (2024).
29. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **19**, 596–604 (2009).
30. Høie, M. H., Cagiada, M., Beck Frederiksen, A. H., Stein, A. & Lindorff-Larsen, K. Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. *Cell Rep.* **38**, 110207 (2022).
31. Abrusán, G. & Marsh, J. A. Alpha Helices Are More Robust to Mutations than Beta Strands. *PLoS Comput. Biol.* **12**, e1005242 (2016).
32. Goffin, D. *et al.* Rett syndrome mutation MeCP2 T158A disrupts DNA binding, protein stability and ERP responses. *Nat. Neurosci.* **15**, 274–283 (2011).
33. Heckman, L. D., Chahrour, M. H. & Zoghbi, H. Y. Rett-causing mutations reveal two domains critical for MeCP2 function and for toxicity in MECP2 duplication syndrome mice. *Elife* **3**, (2014).
34. Brown, K. *et al.* The molecular basis of variable phenotypic severity among common missense mutations causing Rett syndrome. *Hum. Mol. Genet.* **25**, 558–570 (2016).
35. Veitia, R. A., Caburet, S. & Birchler, J. A. Mechanisms of Mendelian dominance. *Clin. Genet.* **93**, 419–428 (2018).
36. Weng, C., Faure, A. J., Escobedo, A. & Lehner, B. The energetic and allosteric landscape for KRAS inhibition. *Nature* **626**, 643–652 (2024).
37. Domingo, J., Baeza-Centurion, P. & Lehner, B. The Causes and Consequences of Genetic Interactions (Epistasis). *Annu. Rev. Genomics Hum. Genet.* **20**, 433–460 (2019).

38. Park, Y., Metzger, B. P. H. & Thornton, J. W. Epistatic drift causes gradual decay of predictability in protein evolution. *Science* (2022) doi:10.1126/science.abn6895.
39. Faure, A. J., Martí-Aranda, A., Hidalgo-Carcedo, C., Schmiedel, J. M. & Lehner, B. The genetic architecture of protein stability. *bioRxiv* 2023.10.27.564339 (2023) doi:10.1101/2023.10.27.564339.
40. Hoose, A., Vellacott, R., Storch, M., Freemont, P. S. & Ryadnov, M. G. DNA synthesis technologies to close the gene writing gap. *Nature Reviews Chemistry* **7**, 144–161 (2023).
41. Holston, A. S., Hinton, S. R., Lindley, K. A., Kearns, N. C. & Plesa, C. Degenerate DropSynth for Simultaneous Assembly of Diverse Gene Libraries and Local Designed Mutants. *bioRxiv* 2023.12.11.569291 (2023) doi:10.1101/2023.12.11.569291.
42. Mighell, T. L., Toledano, I. & Lehner, B. SUNi mutagenesis: Scalable and uniform nicking for efficient generation of variant libraries. *PLoS One* **18**, e0288158 (2023).
43. Macdonald, C. B. *et al.* DIMPLE: deep insertion, deletion, and missense mutation libraries for exploring protein variation in evolution, disease, and biology. *Genome Biol.* **24**, 36 (2023).
44. Malik, A., Mueller-Schickert, A. & Bardwell, J. C. A. Cytosolic Selection Systems To Study Protein Stability. *J. Bacteriol.* **196**, 4333 (2014).
45. Foit, L. *et al.* Optimizing protein stability in vivo. *Mol. Cell* **36**, (2009).
46. Diss, G. & Lehner, B. The genetic landscape of a physical interaction. *Elife* **7**, (2018).
47. Cheng, F. *et al.* Comprehensive characterization of protein-protein interactions perturbed by disease mutations. *Nat. Genet.* **53**, (2021).
48. Ichikawa, D. M. *et al.* A universal deep-learning model for zinc finger design enables transcription factor reprogramming. *Nat. Biotechnol.* **41**, 1117–1129 (2023).
49. Lacoste, J. *et al.* Pervasive mislocalization of pathogenic coding variants underlying human disorders. *bioRxiv* doi:10.1101/2023.09.05.556368.
50. Seuma, M., Lehner, B. & Bolognesi, B. An atlas of amyloid aggregation: the impact of substitutions, insertions, deletions and truncations on amyloid beta fibril nucleation. *Nat. Commun.* **13**, 7084 (2022).
51. Topolska, M., Beltran, A. & Lehner, B. Deep indel mutagenesis reveals the impact of amino acid insertions and deletions on protein stability and function. *bioRxiv* 2023.10.06.561180 (2024) doi:10.1101/2023.10.06.561180.
52. Coyote-Maestas, W., Nedrud, D., He, Y. & Schmidt, D. Determinants of trafficking, conduction, and disease within a K channel revealed through multiparametric deep mutational scanning. *Elife* **11**, (2022).
53. Bronstein, M. The Road to Biology 2.0 Will Pass Through Black-Box Data. *Towards Data Science* https://towardsdatascience.com/the-road-to-biology-2-0-will-pass-through-black-box-data-bbd00fabf959 (2024).
54. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
55. Faure, A. J., Schmiedel, J. M., Baeza-Centurion, P. & Lehner, B. DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. *Genome Biol.* **21**, 207 (2020).
56. Bolognesi, B. *et al.* The mutational landscape of a prion-like domain. *Nat. Commun.* **10**,

4162 (2019).
57. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).
58. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *ICWSM* **3**, 361–362 (2009).
59. Orenbuch, R. *et al.* Deep generative modeling of the human proteome reveals over a hundred novel genes involved in rare genetic disorders. *medRxiv* 2023.11.27.23299062 (2023) doi:10.1101/2023.11.27.23299062.
60. Blaabjerg, L. M. *et al.* Rapid protein stability prediction using deep learning representations. (2023) doi:10.7554/eLife.82593.
61. Zhou, Y., Pan, Q., Pires, D. E. V., Rodrigues, C. H. M. & Ascher, D. B. DDMut: predicting effects of mutations on protein stability using deep learning. *Nucleic Acids Res.* **51**, W122–W128 (2023).
62. Wang, J. *et al.* The conserved domain database in 2023. *Nucleic Acids Res.* **51**, (2023).
63. Faure, A. J. & Lehner, B. MoCHI: neural networks to fit interpretable models and quantify energies, energetic couplings, epistasis and allostery from deep mutational scanning data. *bioRxiv* 2024.01.21.575681 (2024) doi:10.1101/2024.01.21.575681.
64. Venkatakrishnan, A. J. *et al.* Uncovering patterns of atomic interactions in static and dynamic structures of proteins. *bioRxiv* 840694 (2019) doi:10.1101/840694.

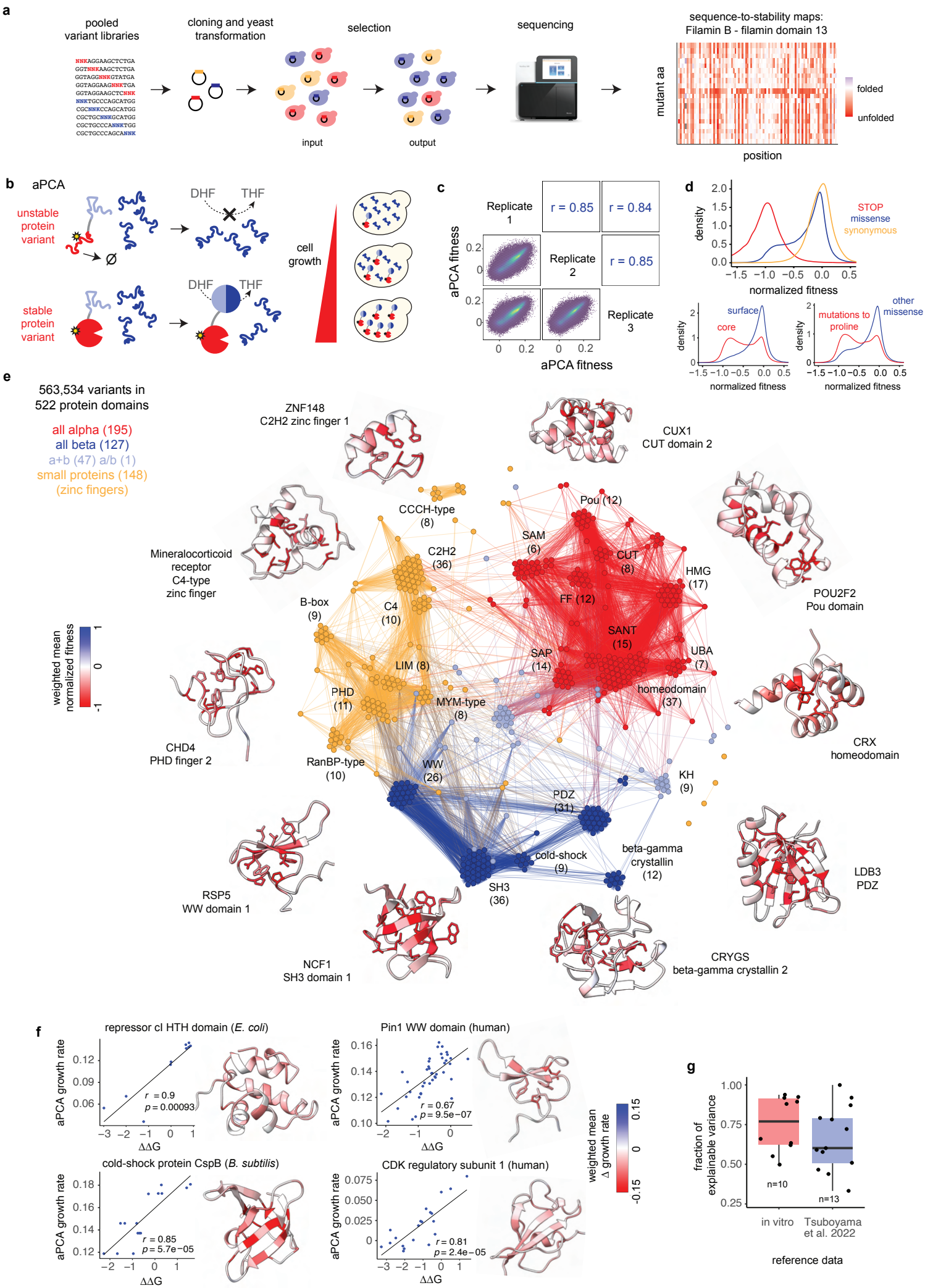

a
pooled variant libraries
cloning and yeast transformation
selection
sequencing
sequence-to-stability maps: Filamin B - filamin domain 13
input
output
mutant aa
position
folded
unfolded
b
aPCA
unstable protein variant
stable protein variant
DHF
THF
cell growth
c
Replicate 1
Replicate 2
Replicate 3
r = 0.85
r = 0.84
r = 0.85
aPCA fitness
d
density
STOP
missense
synonymous
normalized fitness
surface
core
mutations to proline
other missense
e
563,534 variants in 522 protein domains
all alpha (195)
all beta (127)
a+b (47) a/b (1)
small proteins (148)
(zinc fingers)
weighted mean normalized fitness
ZNF148 C2H2 zinc finger 1
CUX1 CUT domain 2
Mineralocorticoid receptor C4-type zinc finger
POU2F2 Pou domain
CHD4 PHD finger 2
CRX homeodomain
RSP5 WW domain 1
LDB3 PDZ
NCF1 SH3 domain 1
CRYGS beta-gamma crystallin 2
CCCH-type (8)
C2H2 (36)
B-box (9)
C4 (10)
LIM (8)
PHD (11)
RanBP-type (10)
MYM-type (8)
WW (26)
SH3 (36)
cold-shock (9)
PDZ (31)
beta-gamma crystallin (12)
KH (9)
homeodomain (37)
UBA (7)
SANT (15)
SAP (14)
FF (12)
HMG (17)
CUT (8)
SAM (6)
Pou (12)
f
repressor cI HTH domain (E. coli)
Pin1 WW domain (human)
cold-shock protein CspB (B. subtilis)
CDK regulatory subunit 1 (human)
aPCA growth rate
ΔΔG
r = 0.9
p = 0.00093
r = 0.67
p = 9.5e−07
r = 0.85
p = 5.7e−05
r = 0.81
p = 2.4e−05
weighted mean Δ growth rate
g
fraction of explainable variance
n=10
n=13
in vitro
Tsuboyama et al. 2022
reference data

Figure 1

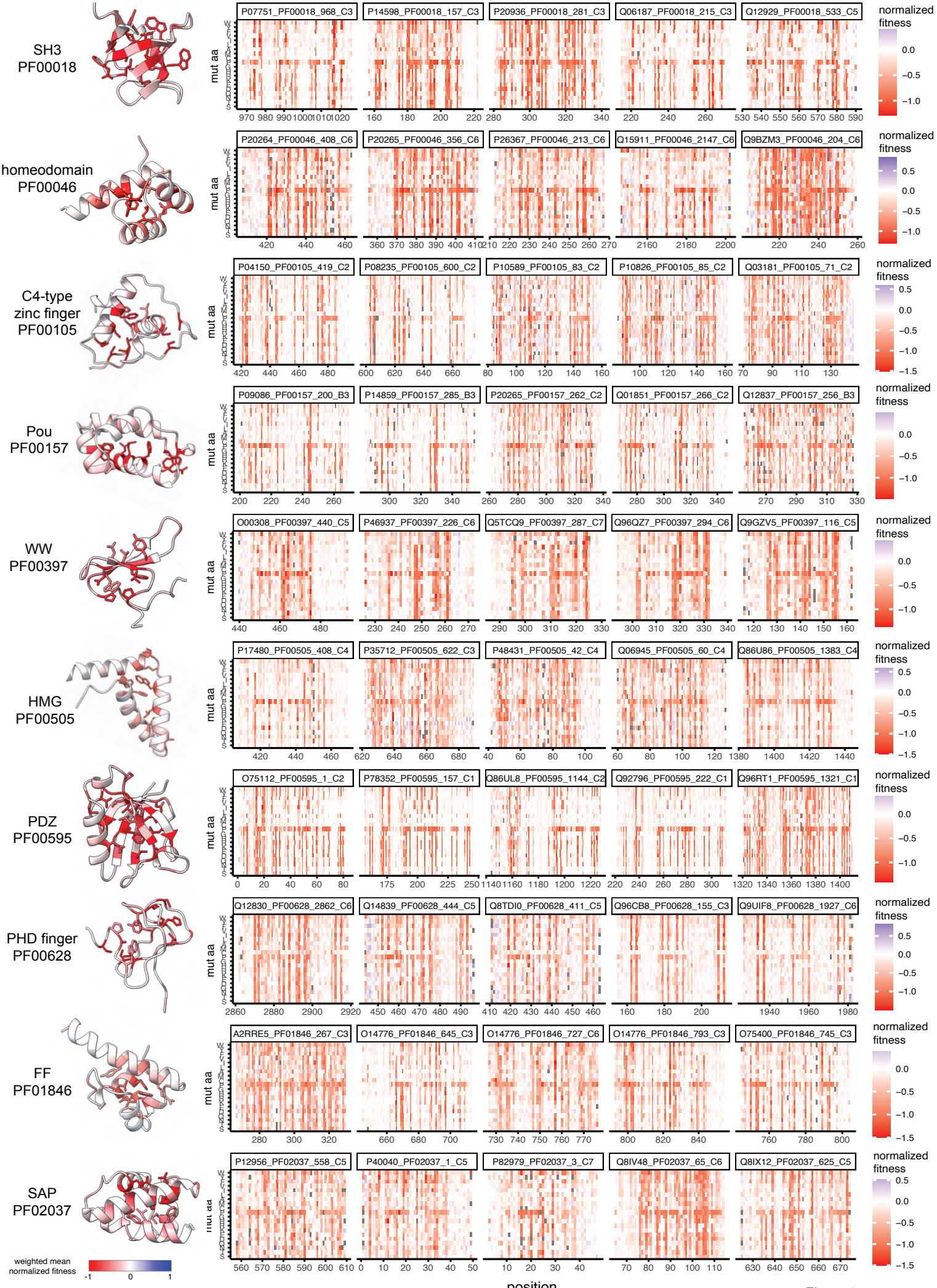

SH3
PF00018
P07751_PF00018_968_C3
P14598_PF00018_157_C3
P20936_PF00018_281_C3
Q06187_PF00018_215_C3
Q12929_PF00018_533_C5
homeodomain
PF00046
P20264_PF00046_408_C6
P20265_PF00046_356_C6
P26367_PF00046_213_C6
Q15911_PF00046_2147_C6
Q9BZM3_PF00046_204_C6
C4-type
zinc finger
PF00105
P04150_PF00105_419_C2
P08235_PF00105_600_C2
P10589_PF00105_83_C2
P10826_PF00105_85_C2
Q03181_PF00105_71_C2
Pou
PF00157
P09086_PF00157_200_B3
P14859_PF00157_285_B3
P20265_PF00157_262_C2
Q01851_PF00157_266_C2
Q12837_PF00157_256_B3
WW
PF00397
O00308_PF00397_440_C5
P46937_PF00397_226_C6
Q5TCQ9_PF00397_287_C7
Q96QZ7_PF00397_294_C6
Q9GZV5_PF00397_116_C5
HMG
PF00505
P17480_PF00505_408_C4
P35712_PF00505_622_C3
P48431_PF00505_42_C4
Q06945_PF00505_60_C4
Q86U86_PF00505_1383_C4
PDZ
PF00595
O75112_PF00595_1_C2
P78352_PF00595_157_C1
Q86UL8_PF00595_1144_C2
Q92796_PF00595_222_C1
Q96RT1_PF00595_1321_C1
PHD finger
PF00628
Q12830_PF00628_2862_C6
Q14839_PF00628_444_C5
Q8TDI0_PF00628_411_C5
Q96CB8_PF00628_155_C3
Q9UIF8_PF00628_1927_C6
FF
PF01846
A2RRE5_PF01846_267_C3
O14776_PF01846_645_C3
O14776_PF01846_727_C6
O14776_PF01846_793_C3
O75400_PF01846_745_C3
SAP
PF02037
P12956_PF02037_558_C5
P40040_PF02037_1_C5
P82979_PF02037_3_C7
Q8IV48_PF02037_65_C6
Q8IX12_PF02037_625_C5
mut aa
normalized fitness
position
weighted mean normalized fitness

Figure 2

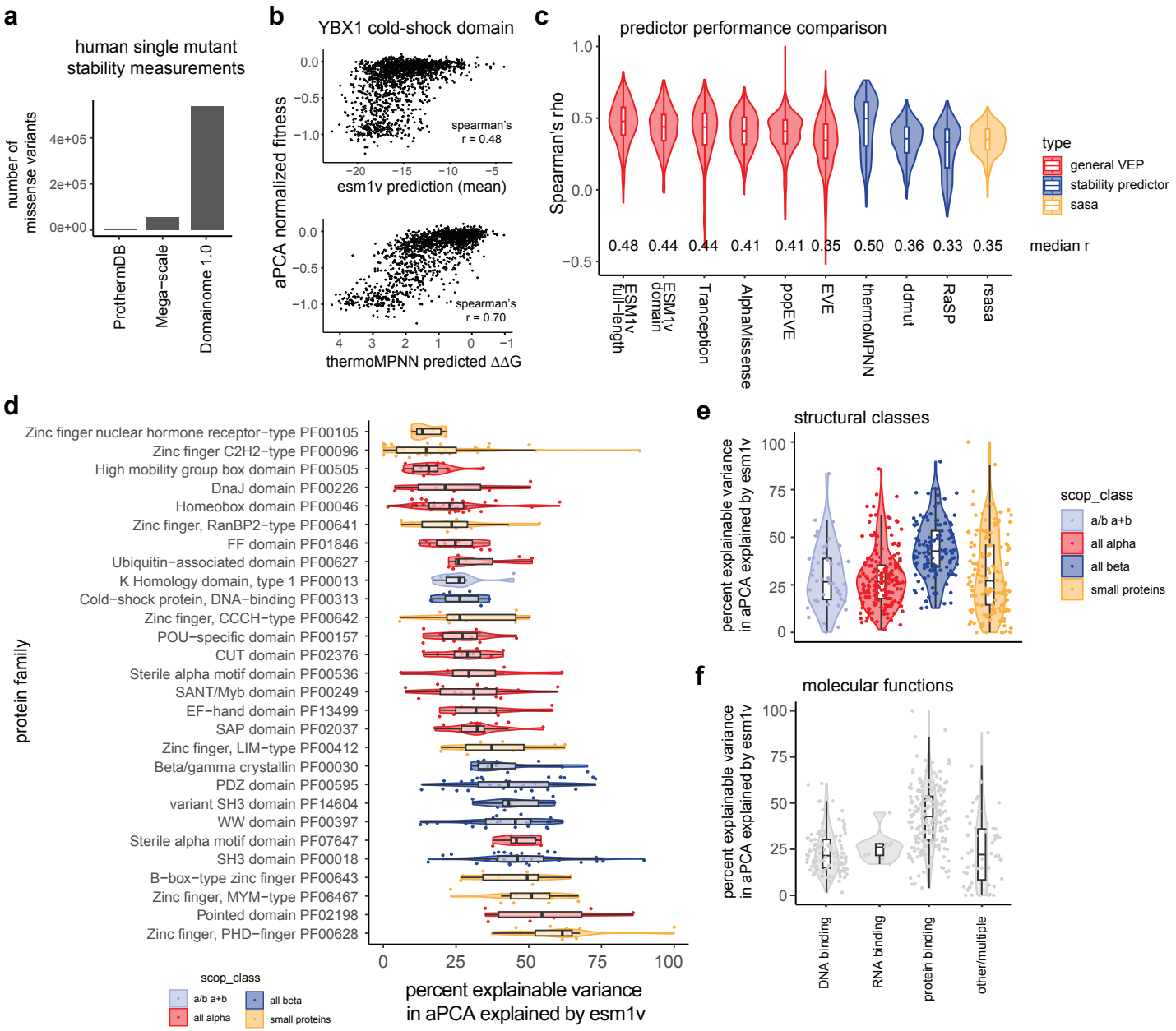

a
human single mutant stability measurements
number of missense variants
4e+05
2e+05
0e+00
ProthermDB
Mega-scale
Domainome 1.0
b
YBX1 cold-shock domain
aPCA normalized fitness
spearman's r = 0.48
esm1v prediction (mean)
spearman's r = 0.70
thermoMPNN predicted ΔΔG
c
predictor performance comparison
Spearman's rho
type
general VEP
stability predictor
sasa
0.48 0.44 0.44 0.41 0.41 0.35 0.50 0.36 0.33 0.35
median r
ESM1v full-length
ESM1v domain
Tranception
AlphaMissense
popEVE
EVE
thermoMPNN
ddmut
RaSP
rsasa
d
protein family
Zinc finger nuclear hormone receptor-type PF00105
Zinc finger C2H2-type PF00096
High mobility group box domain PF00505
DnaJ domain PF00226
Homeobox domain PF00046
Zinc finger, RanBP2-type PF00641
FF domain PF01846
Ubiquitin-associated domain PF00627
K Homology domain, type 1 PF00013
Cold-shock protein, DNA-binding PF00313
Zinc finger, CCCH-type PF00642
POU-specific domain PF00157
CUT domain PF02376
Sterile alpha motif domain PF00536
SANT/Myb domain PF00249
EF-hand domain PF13499
SAP domain PF02037
Zinc finger, LIM-type PF00412
Beta/gamma crystallin PF00030
PDZ domain PF00595
variant SH3 domain PF14604
WW domain PF00397
Sterile alpha motif domain PF07647
SH3 domain PF00018
B-box-type zinc finger PF00643
Zinc finger, MYM-type PF06467
Pointed domain PF02198
Zinc finger, PHD-finger PF00628
scop_class
a/b a+b
all alpha
all beta
small proteins
percent explainable variance in aPCA explained by esm1v
e
structural classes
percent explainable variance in aPCA explained by esm1v
scop_class
a/b a+b
all alpha
all beta
small proteins
f
molecular functions
percent explainable variance in aPCA explained by esm1v
DNA binding
RNA binding
protein binding
other/multiple

Figure 3

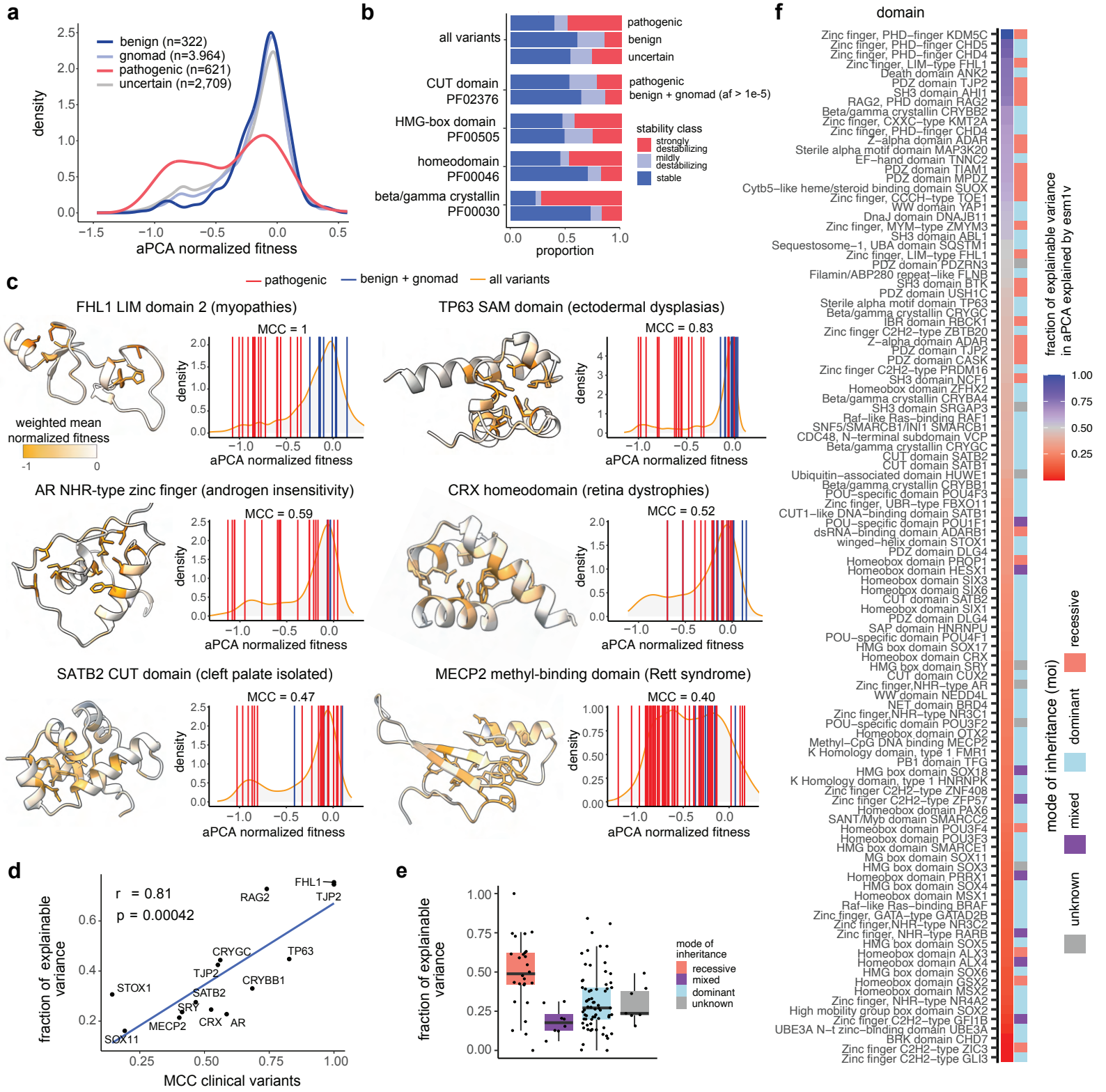

a
benign (n=322)
gnomad (n=3.964)
pathogenic (n=621)
uncertain (n=2,709)
density
aPCA normalized fitness
b
all variants
CUT domain PF02376
HMG-box domain PF00505
homeodomain PF00046
beta/gamma crystallin PF00030
pathogenic
benign
uncertain
pathogenic
benign + gnomad (af > 1e-5)
stability class
strongly destabilizing
mildly destabilizing
stable
proportion
c
pathogenic
benign + gnomad
all variants
FHL1 LIM domain 2 (myopathies)
MCC = 1
weighted mean normalized fitness
TP63 SAM domain (ectodermal dysplasias)
MCC = 0.83
AR NHR-type zinc finger (androgen insensitivity)
MCC = 0.59
CRX homeodomain (retina dystrophies)
MCC = 0.52
SATB2 CUT domain (cleft palate isolated)
MCC = 0.47
MECP2 methyl-binding domain (Rett syndrome)
MCC = 0.40
aPCA normalized fitness
d
r = 0.81
p = 0.00042
FHL1
RAG2
TJP2
CRYGC
TP63
TJP2
CRYBB1
STOX1
SATB2
SRY
MECP2
CRX
AR
SOX11
fraction of explainable variance
MCC clinical variants
e
fraction of explainable variance
mode of inheritance
recessive
mixed
dominant
unknown
f
domain
fraction of explainable variance in aPCA explained by esm1v
mode of inheritance (moi)
recessive
dominant
mixed
unknown

Figure 4

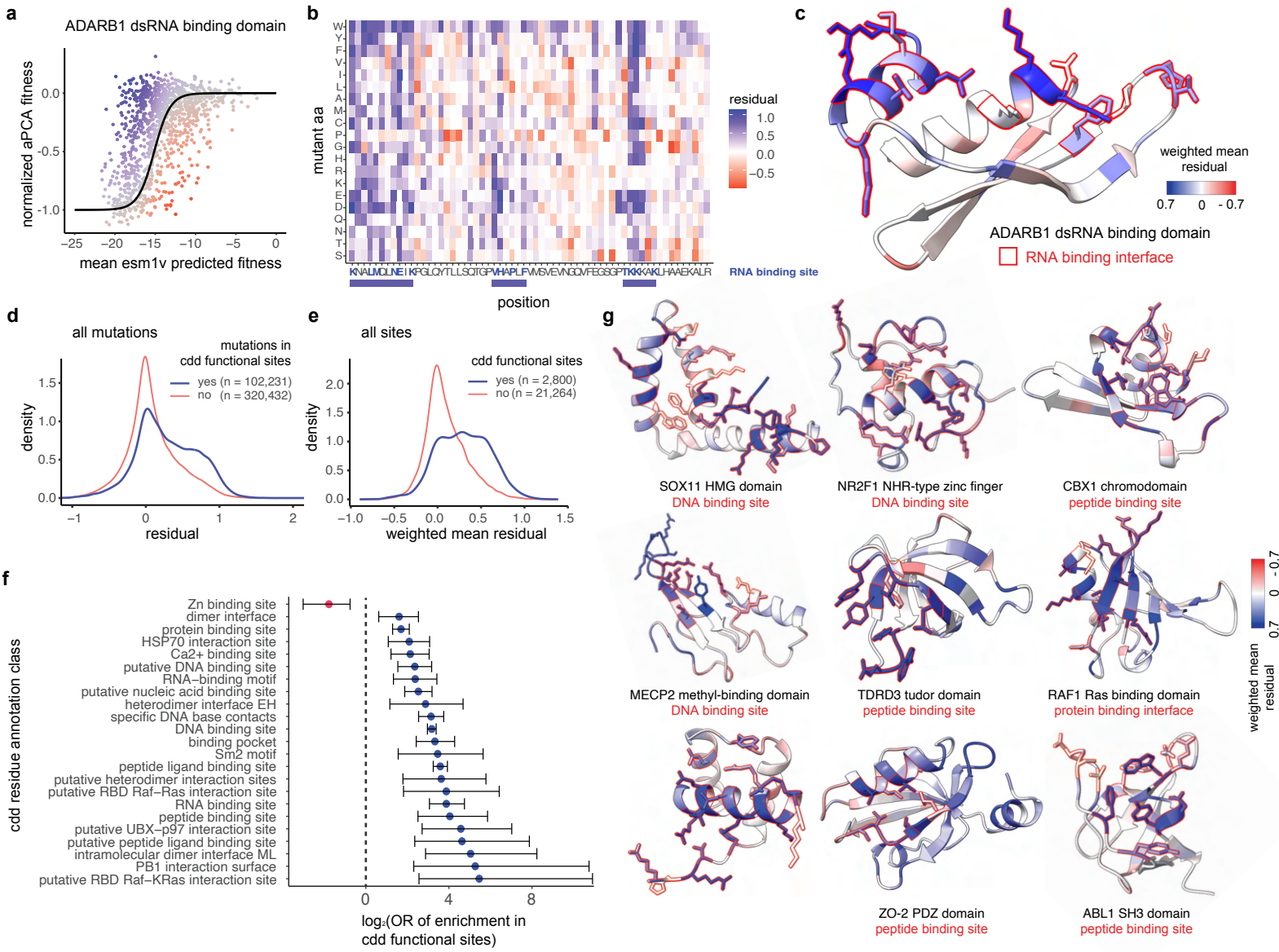

a
ADARB1 dsRNA binding domain
normalized aPCA fitness
mean esm1v predicted fitness
b
mutant aa
position
residual
RNA binding site
c
weighted mean residual
ADARB1 dsRNA binding domain
RNA binding interface
d
all mutations
mutations in cdd functional sites
yes (n = 102,231)
no (n = 320,432)
density
residual
e
all sites
cdd functional sites
yes (n = 2,800)
no (n = 21,264)
weighted mean residual
f
cdd residue annotation class
Zn binding site
dimer interface
protein binding site
HSP70 interaction site
Ca2+ binding site
putative DNA binding site
RNA−binding motif
putative nucleic acid binding site
heterodimer interface EH
specific DNA base contacts
DNA binding site
binding pocket
Sm2 motif
peptide ligand binding site
putative heterodimer interaction sites
putative RBD Raf−Ras interaction site
RNA binding site
peptide binding site
putative UBX−p97 interaction site
putative peptide ligand binding site
intramolecular dimer interface ML
PB1 interaction surface
putative RBD Raf−KRas interaction site
log2(OR of enrichment in cdd functional sites)
g
SOX11 HMG domain
DNA binding site
NR2F1 NHR-type zinc finger
DNA binding site
CBX1 chromodomain
peptide binding site
MECP2 methyl-binding domain
DNA binding site
TDRD3 tudor domain
peptide binding site
RAF1 Ras binding domain
protein binding interface
ZO-2 PDZ domain
peptide binding site
ABL1 SH3 domain
peptide binding site
weighted mean residual

Figure 5

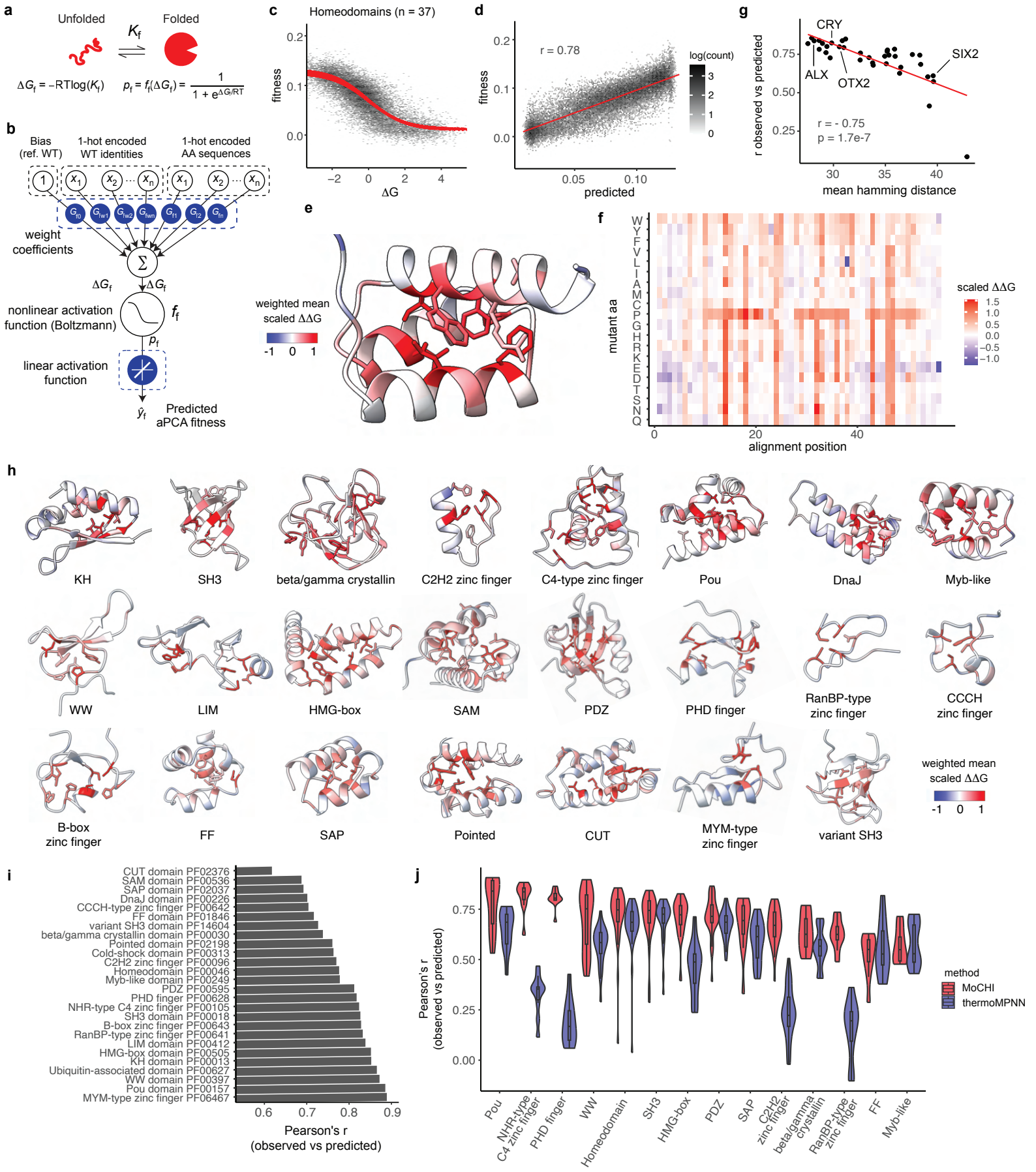

a
Unfolded
Folded
K_f
$\Delta G_f = -RT\log(K_f)$
$p_f = f_f(\Delta G_f) = \frac{1}{1 + e^{\Delta G_f/RT}}$
b
Bias (ref. WT)
1-hot encoded WT identities
1-hot encoded AA sequences
weight coefficients
nonlinear activation function (Boltzmann)
linear activation function
Predicted aPCA fitness
c
Homeodomains (n = 37)
fitness
ΔG
d
r = 0.78
fitness
predicted
log(count)
e
weighted mean scaled ΔΔG
f
mutant aa
alignment position
scaled ΔΔG
g
CRY
SIX2
ALX
OTX2
r = - 0.75
p = 1.7e-7
r observed vs predicted
mean hamming distance
h
KH
SH3
beta/gamma crystallin
C2H2 zinc finger
C4-type zinc finger
Pou
DnaJ
Myb-like
WW
LIM
HMG-box
SAM
PDZ
PHD finger
RanBP-type zinc finger
CCCH zinc finger
B-box zinc finger
FF
SAP
Pointed
CUT
MYM-type zinc finger
variant SH3
weighted mean scaled ΔΔG
i
CUT domain PF02376
SAM domain PF00536
SAP domain PF02037
DnaJ domain PF00226
CCCH-type zinc finger PF00642
FF domain PF01846
variant SH3 domain PF14604
beta/gamma crystallin domain PF00030
Pointed domain PF02198
Cold-shock domain PF00313
C2H2 zinc finger PF00096
Homeodomain PF00046
Myb-like domain PF00249
PDZ PF00595
PHD finger PF00628
NHR-type C4 zinc finger PF00105
SH3 domain PF00018
B-box zinc finger PF00643
RanBP-type zinc finger PF00641
LIM domain PF00412
HMG-box domain PF00505
KH domain PF00013
Ubiquitin-associated domain PF00627
WW domain PF00397
Pou domain PF00157
MYM-type zinc finger PF06467
Pearson's r (observed vs predicted)
j
Pearson's r (observed vs predicted)
method
MoCHI
thermoMPNN
Pou
NHR-type C4 zinc finger
PHD finger
WW
Homeodomain
SH3
HMG-box
PDZ
SAP
C2H2 zinc finger
beta/gamma crystallin
RanBP-type zinc finger
FF
Myb-like

Figure 6

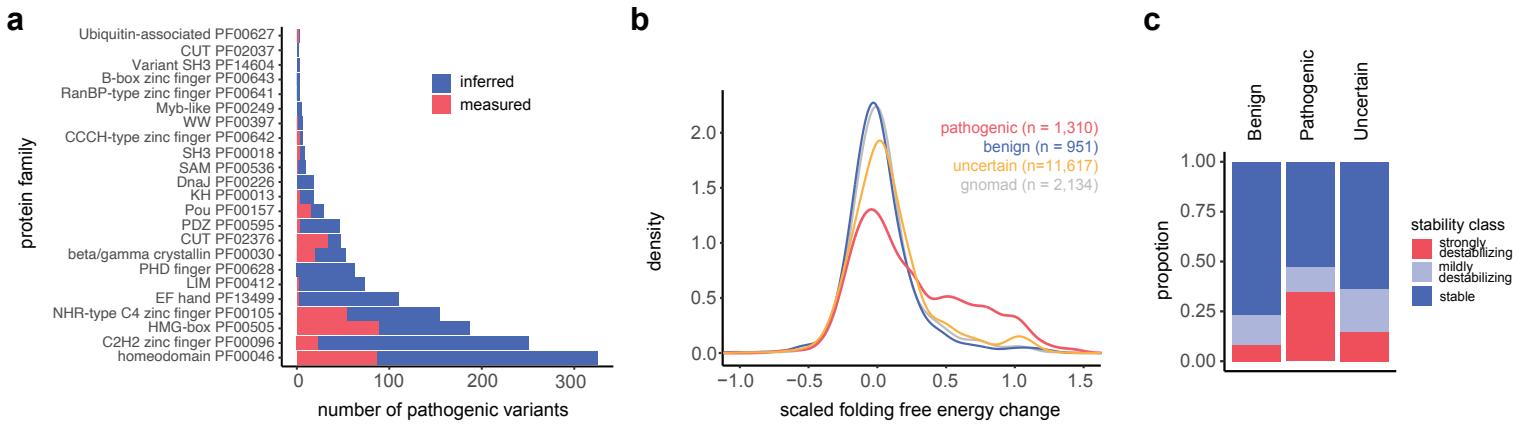
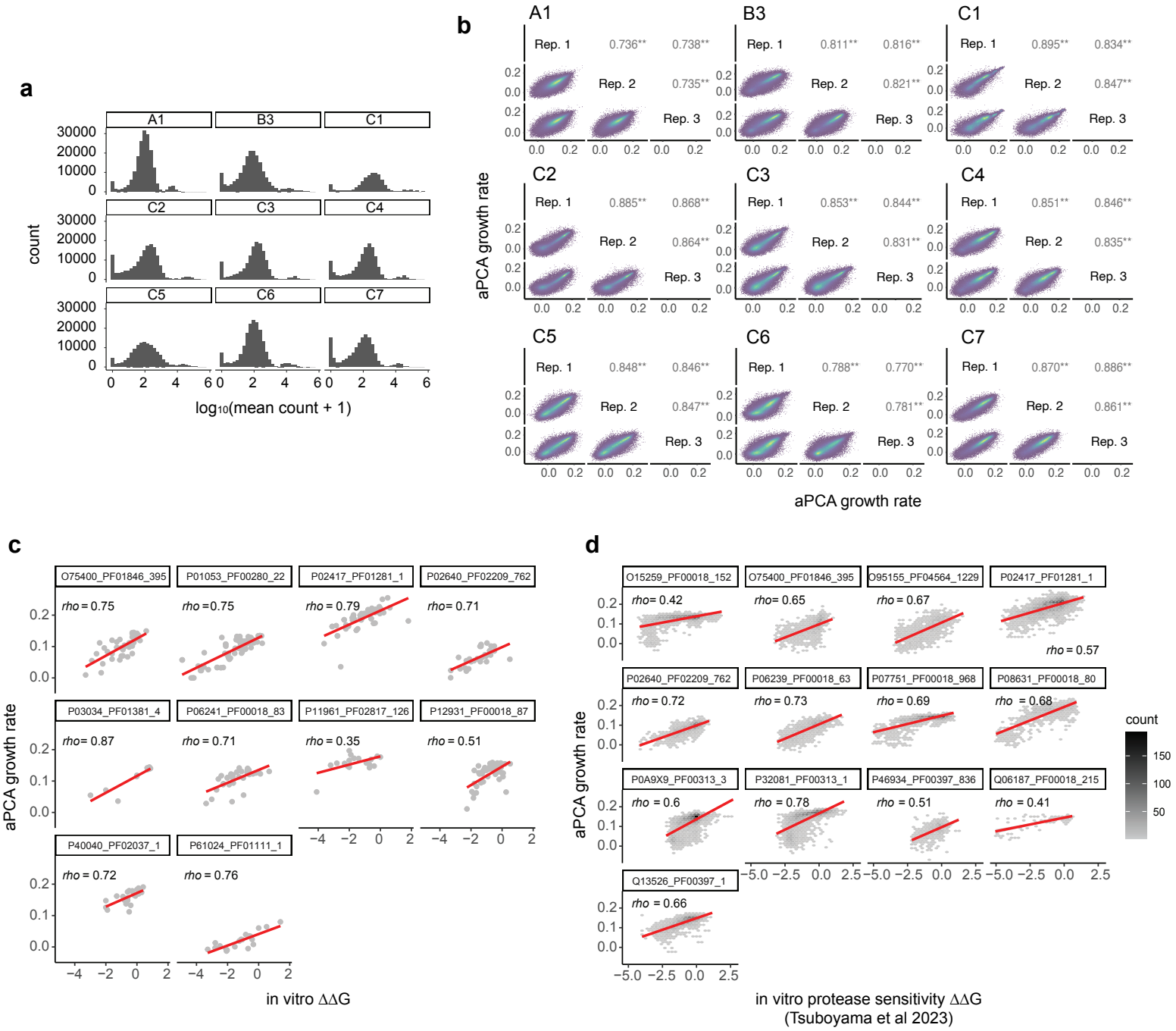

a
Ubiquitin-associated PF00627
CUT PF02037
Variant SH3 PF14604
B-box zinc finger PF00643
RanBP-type zinc finger PF00641
Myb-like PF00249
WW PF00397
CCCH-type zinc finger PF00642
SH3 PF00018
SAM PF00536
DnaJ PF00226
KH PF00013
Pou PF00157
PDZ PF00595
CUT PF02376
beta/gamma crystallin PF00030
PHD finger PF00628
LIM PF00412
EF hand PF13499
NHR-type C4 zinc finger PF00105
HMG-box PF00505
C2H2 zinc finger PF00096
homeodomain PF00046
protein family
inferred
measured
0
100
200
300
number of pathogenic variants
b
pathogenic (n = 1,310)
benign (n = 951)
uncertain (n=11,617)
gnomad (n = 2,134)
density
2.0
1.5
1.0
0.5
0.0
−1.0
−0.5
0.0
0.5
1.0
1.5
scaled folding free energy change
c
Benign
Pathogenic
Uncertain
propotion
1.00
0.75
0.50
0.25
0.00
stability class
strongly destabilizing
mildly destabilizing
stable

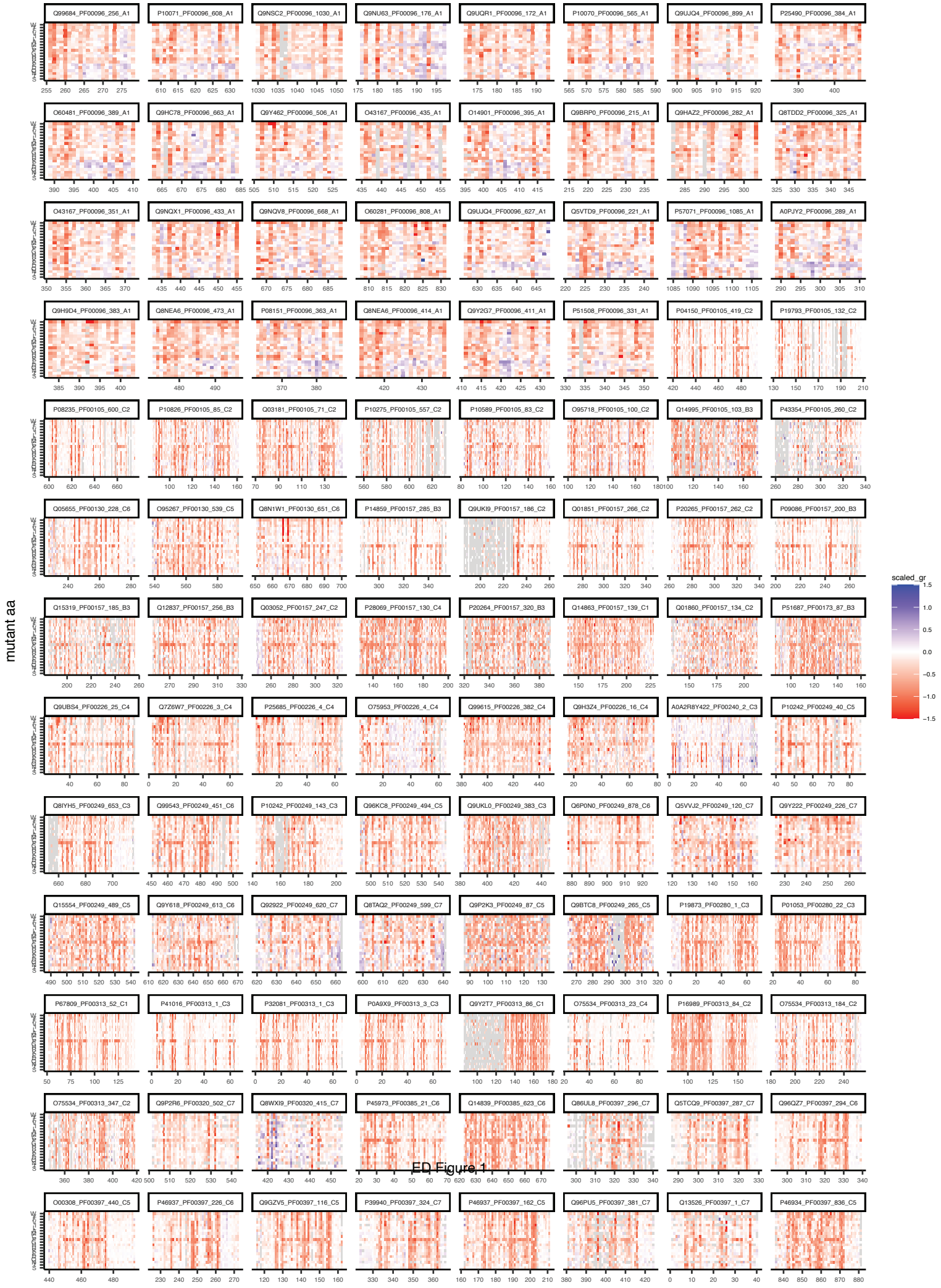
Figure 7


a
A1
B3
C1
C2
C3
C4
C5
C6
C7
count
log10(mean count + 1)
b
Rep. 1
Rep. 2
Rep. 3
aPCA growth rate
c
O75400_PF01846_395
P01053_PF00280_22
P02417_PF01281_1
P02640_PF02209_762
P03034_PF01381_4
P06241_PF00018_83
P11961_PF02817_126
P12931_PF00018_87
P40040_PF02037_1
P61024_PF01111_1
in vitro ΔΔG
d
O15259_PF00018_152
O75400_PF01846_395
O95155_PF04564_1229
P02417_PF01281_1
P02640_PF02209_762
P06239_PF00018_63
P07751_PF00018_968
P08631_PF00018_80
P0A9X9_PF00313_3
P32081_PF00313_1
P46934_PF00397_836
Q06187_PF00018_215
Q13526_PF00397_1
count
in vitro protease sensitivity ΔΔG
(Tsuboyama et al 2023)
e
domain cores
domain surfaces
mutant amino acid
polar
hydrophobic
proline
density
normalized aPCA fitness
f
secondary structure
Bridge
Coil
Helix
Strand
Turn
density
normalized aPCA fitness

ED Figure 1

Q05195_PF00010_58_C6
P50539_PF00010_70_C6
Q96I24_PF00013_75_C2
Q96AE4_PF00013_98_C2
Q96AE4_PF00013_184_C2
Q92945_PF00013_142_C2
Q06787_PF00013_217_C4
Q15365_PF00013_278_C2
Q5U5Q3_PF00013_325_C3
P61978_PF00013_385_C2
Q86XN8_PF00013_273_C3
P14598_PF00018_157_C3
Q06187_PF00018_215_C3
P20936_PF00018_281_C3
Q12929_PF00018_533_C5
P07751_PF00018_968_C3
P62993_PF00018_159_C5
A1X283_PF00018_222_C3
O43586_PF00018_361_C5
Q08881_PF00018_172_C3
Q12965_PF00018_1051_C5
P15498_PF00018_592_C3
O43295_PF00018_745_C5
Q7Z6B7_PF00018_744_C5
Q6XZF7_PF00018_1516_C3
P15498_PF00018_782_C3
Q8N157_PF00018_1053_C5
P19878_PF00018_241_C3
P16885_PF00018_772_C5
P06241_PF00018_83_C3
Q5HYK7_PF00018_664_C5
Q9Y5K6_PF00018_271_C3
A1X283_PF00018_155_C6
P06239_PF00018_63_C6
O15259_PF00018_152_C3
Q9UNF0_PF00018_428_C4
P51451_PF00018_59_C3
Q15811_PF00018_1003_C5
P08631_PF00018_80_C5
Q15811_PF00018_909_C3
Q15811_PF00018_1156_C6
Q6XZF7_PF00018_242_C3
P00519_PF00018_64_C5
P12931_PF00018_87_C5
Q6XZF7_PF00018_146_C3
P02549_PF00018_979_C5
P19878_PF00018_456_C3
Q8N1P7_PF00030_1358_B3
P05813_PF00030_31_C1
P07315_PF00030_91_B3
P53673_PF00030_14_B3
P07315_PF00030_4_B3
P22914_PF00030_96_B3
P07316_PF00030_1_C2
P11844_PF00030_85_C1
P07316_PF00030_87_C1
P43320_PF00030_106_C2
P07320_PF00030_90_B3
P53674_PF00030_57_C1
O95793_PF00035_183_C3
P78563_PF00035_80_C4
Q9BZM3_PF00046_204_C6
P20265_PF00046_356_C6
P20264_PF00046_408_C6
P26367_PF00046_213_C6
Q15911_PF00046_2147_C6
Q9H161_PF00046_216_C6
P78424_PF00046_609_C6
O43186_PF00046_41_C6
P14653_PF00046_208_C6
Q5SQQ9_PF00046_102_C6
Q15911_PF00046_2244_C6
O95343_PF00046_212_C5
Q15699_PF00046_134_C6
Q9NPC8_PF00046_131_C5
O95076_PF00046_155_C6
O75360_PF00046_71_C6
Q96QS3_PF00046_330_C6
Q15911_PF00046_2643_C6
Q86UP3_PF00046_2562_C5
Q86UP3_PF00046_2086_C6
O14813_PF00046_92_C6
P35548_PF00046_144_C6
Q15475_PF00046_131_C5
O95475_PF00046_134_C5
Q86UP3_PF00046_2183_C6
P32243_PF00046_40_C6
P56915_PF00046_162_C6
P23759_PF00046_219_C6
Q9C0A1_PF00046_1859_C6
Q8NFW5_PF00046_73_C6
P54821_PF00046_96_C6
Q9UBX0_PF00046_110_C5
O15266_PF00046_119_C6
P28360_PF00046_174_C6
P49335_PF00046_280_C6
P39880_PF00046_1246_C5
Q9H2P0_PF00046_770_C7
Q9HC78_PF00096_716_A1
P37275_PF00096_933_A1
Q9UJQ4_PF00096_383_A1
Q9HC78_PF00096_635_A1
Q99684_PF00096_313_A1
P51508_PF00096_555_A1
mutant aa
position
scaled_gr
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

ED Figure 2-1



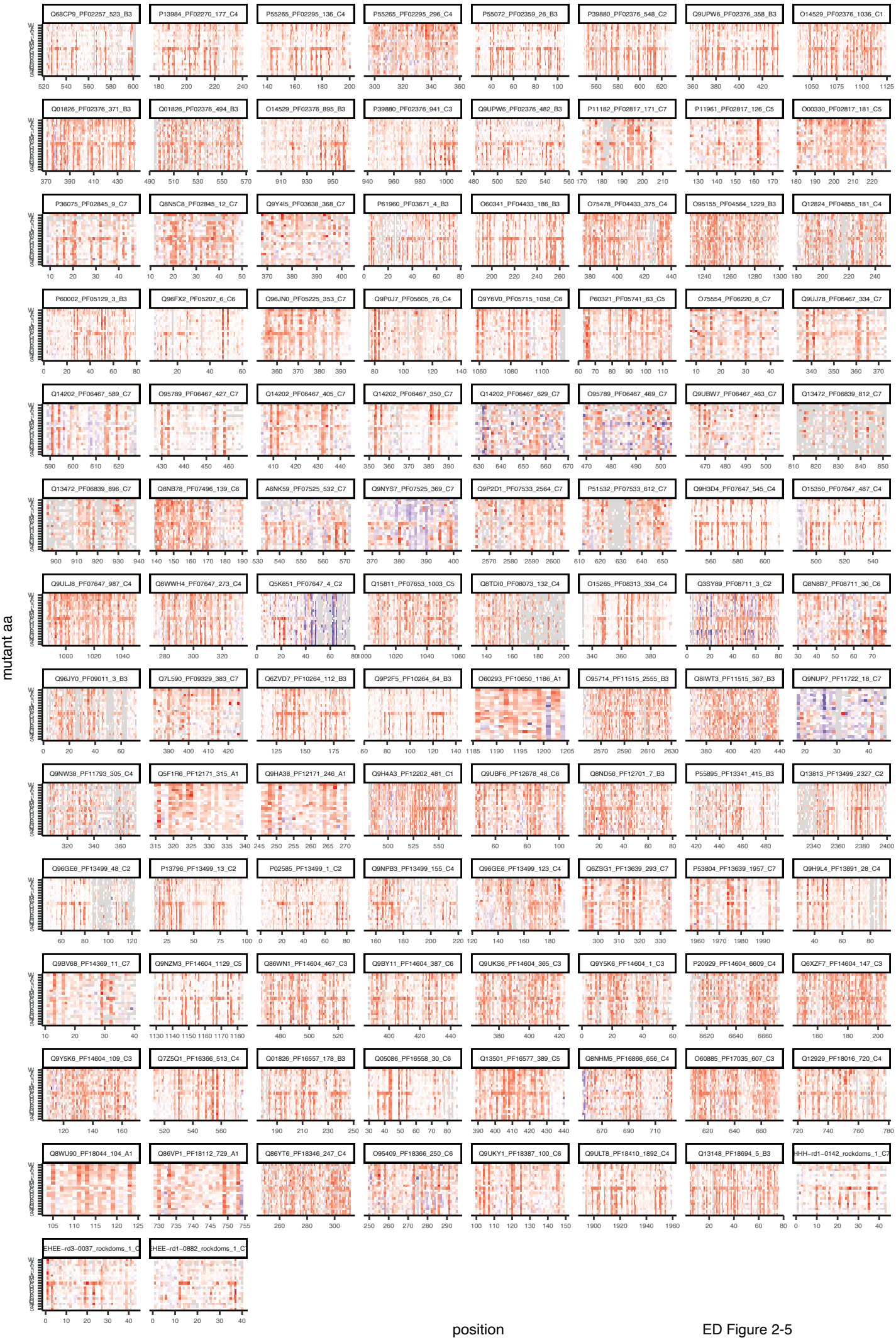
ED Figure 1

ED Figure 2-2

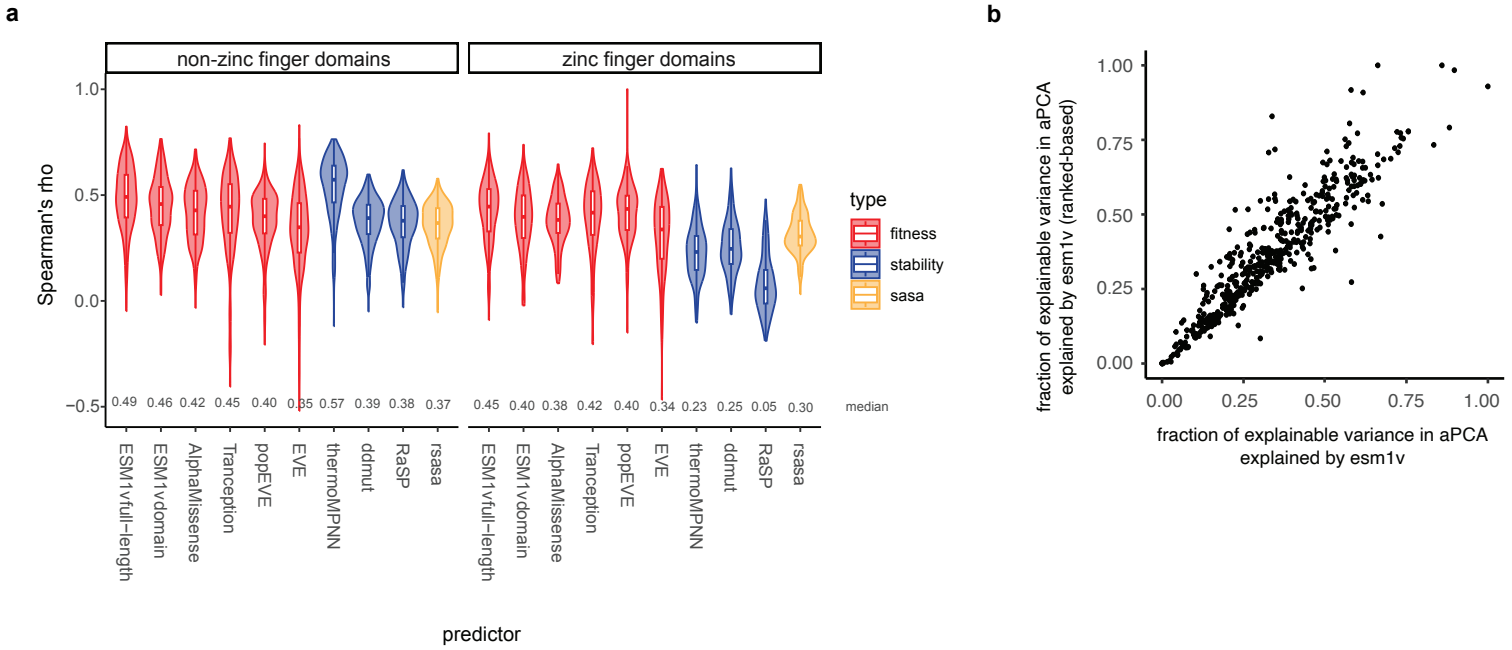
P46937_PF00397_165_C5
Q8N3X1_PF00397_593_C7
Q96J02_PF00397_474_C5
Q9P2P5_PF00397_800_C7
Q9BYW2_PF00397_2382_C5
Q9H0M0_PF00397_343_C6
Q8N3X1_PF00397_200_C3
Q96QZ7_PF00397_352_C5
Q9HAU4_PF00397_150_C5
Q96PU5_PF00397_539_C5
Q96PU5_PF00397_184_C5
Q96J02_PF00397_428_C5
Q6ZWJ1_PF00397_493_C6
O95817_PF00397_11_C5
O75400_PF00397_176_C5
O94929_PF00412_149_C5
Q15942_PF00412_504_C4
Q13642_PF00412_160_C5
Q14192_PF00412_38_C4
Q7Z4I7_PF00412_13_C3
Q96JY6_PF00412_284_C5
Q13642_PF00412_99_C4
O94929_PF00412_210_C5
P07814_PF00458_827_C6
P12081_PF00458_8_C7
P17480_PF00505_408_C4
Q86U86_PF00505_1383_C4
P48431_PF00505_42_C4
Q06945_PF00505_60_C4
P35712_PF00505_622_C3
P35716_PF00505_50_C3
O94993_PF00505_338_C4
P35711_PF00505_557_C4
Q9H6I2_PF00505_69_C4
P41225_PF00505_140_C4
P35713_PF00505_87_C4
Q969G3_PF00505_67_C4
O15405_PF00505_256_C4
P54277_PF00505_572_C4
P57073_PF00505_103_C4
Q05066_PF00505_61_C4
Q9NP66_PF00505_104_C4
Q12955_PF00531_4094_B3
P16157_PF00531_1405_B3
Q01484_PF00531_3571_B3
P35251_PF00533_405_B3
Q8WXI2_PF00536_9_C4
Q7Z6G8_PF00536_812_C4
O94885_PF00536_630_C2
Q68DC2_PF00536_774_C4
Q9NYL2_PF00536_330_C2
O94830_PF00536_389_C3
P22492_PF00538_44_B3
Q92734_PF00564_6_C1
P19878_PF00564_353_B3
Q13163_PF00564_15_C1
Q9H7E2_PF00567_554_C6
Q13501_PF00569_120_C6
Q9Y4J8_PF00569_235_C4
O75112_PF00595_1_C2
P78352_PF00595_157_C1
Q92796_PF00595_222_C1
Q13424_PF00595_79_C1
Q96RT1_PF00595_1321_C1
Q86UL8_PF00595_1144_C2
Q9UPQ7_PF00595_244_C1
Q14CM0_PF00595_75_C2
Q86UL8_PF00595_774_C1
O14936_PF00595_488_C2
Q14160_PF00595_857_C1
P78352_PF00595_309_C1
O75970_PF00595_693_C1
O75970_PF00595_1003_C1
Q8TEW0_PF00595_457_C1
Q9Y6N9_PF00595_207_C1
Q9UPQ7_PF00595_414_C1
O75970_PF00595_372_C1
Q92796_PF00595_383_C2
Q13368_PF00595_132_C1
Q61234_PF00595_80_C2
Q9UDY2_PF00595_26_C1
O75970_PF00595_1860_C2
Q9ULL8_PF00595_7_C2
Q13009_PF00595_843_C2
Q8TEW8_PF00595_380_C1
P78352_PF00595_60_C1
Q86UL8_PF00595_601_C2
Q9UDY2_PF00595_509_C1
O14640_PF00595_246_C1
Q9Y3R0_PF00595_666_C1
Q9H0A6_PF00612_189_A1
Q8IZT6_PF00612_1877_A1
Q9ULZ3_PF00619_114_B3
O43353_PF00619_438_B3
P54727_PF00627_186_C7
Q9H347_PF00627_613_C7
P54727_PF00627_361_C7
O95835_PF00627_98_C7
P61086_PF00627_158_C7
Q7Z6Z7_PF00627_1316_C7
P22681_PF00627_854_C6
Q12830_PF00628_2862_C6
Q14839_PF00628_444_C5
Q96CB8_PF00628_155_C3
mutant aa
position
scaled_gr
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

ED Figure 2-3

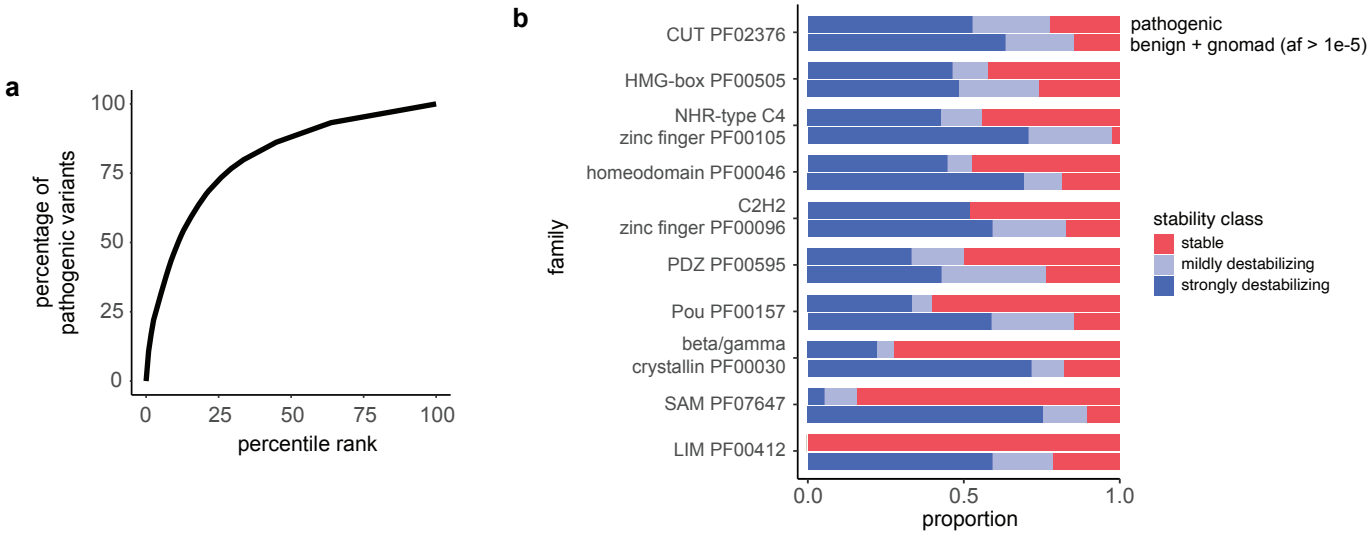
ED Figure 2-4


Q68CP9_PF02257_523_B3
P13984_PF02270_177_C4
P55265_PF02295_136_C4
P55265_PF02295_296_C4
P55072_PF02359_26_B3
P39880_PF02376_548_C2
Q9UPW6_PF02376_358_B3
O14529_PF02376_1036_C1
Q01826_PF02376_371_B3
Q01826_PF02376_494_B3
O14529_PF02376_895_B3
P39880_PF02376_941_C3
Q9UPW6_PF02376_482_B3
P11182_PF02817_171_C7
P11961_PF02817_126_C5
O00330_PF02817_181_C5
P36075_PF02845_9_C7
Q8N5C8_PF02845_12_C7
Q9Y4I5_PF03638_368_C7
P61960_PF03671_4_B3
O60341_PF04433_186_B3
O75478_PF04433_375_C4
O95155_PF04564_1229_B3
Q12824_PF04855_181_C4
P60002_PF05129_3_B3
Q96FX2_PF05207_6_C6
Q96JN0_PF05225_353_C7
Q9P0J7_PF05605_76_C4
Q9Y6V0_PF05715_1058_C6
P60321_PF05741_63_C5
O75554_PF06220_8_C7
Q9UJ78_PF06467_334_C7
Q14202_PF06467_589_C7
O95789_PF06467_427_C7
Q14202_PF06467_405_C7
Q14202_PF06467_350_C7
Q14202_PF06467_629_C7
O95789_PF06467_469_C7
Q9UBW7_PF06467_463_C7
Q13472_PF06839_812_C7
Q13472_PF06839_896_C7
Q8NB78_PF07496_139_C6
A6NK59_PF07525_532_C7
Q9NYS7_PF07525_369_C7
Q9P2D1_PF07533_2564_C7
P51532_PF07533_612_C7
Q9H3D4_PF07647_545_C4
O15350_PF07647_487_C4
Q9ULJ8_PF07647_987_C4
Q8WWH4_PF07647_273_C4
Q5K651_PF07647_4_C2
Q15811_PF07653_1003_C5
Q8TDI0_PF08073_132_C4
O15265_PF08313_334_C4
Q3SY89_PF08711_3_C2
Q8N8B7_PF08711_30_C6
Q96JY0_PF09011_3_B3
Q7L590_PF09329_383_C7
Q6ZVD7_PF10264_112_B3
Q9P2F5_PF10264_64_B3
O60293_PF10650_1186_A1
O95714_PF11515_2555_B3
Q8IWT3_PF11515_367_B3
Q9NUP7_PF11722_18_C7
Q9NW38_PF11793_305_C4
Q5F1R6_PF12171_315_A1
Q9HA38_PF12171_246_A1
Q9H4A3_PF12202_481_C1
Q9UBF6_PF12678_48_C6
Q8ND56_PF12701_7_B3
P55895_PF13341_415_B3
Q13813_PF13499_2327_C2
Q96GE6_PF13499_48_C2
P13796_PF13499_13_C2
P02585_PF13499_1_C2
Q9NPB3_PF13499_155_C4
Q96GE6_PF13499_123_C4
Q6ZSG1_PF13639_293_C7
P53804_PF13639_1957_C7
Q9H9L4_PF13891_28_C4
Q9BV68_PF14369_11_C7
Q9NZM3_PF14604_1129_C5
Q86WN1_PF14604_467_C3
Q9BY11_PF14604_387_C6
Q9UKS6_PF14604_365_C3
Q9Y5K6_PF14604_1_C3
P20929_PF14604_6609_C4
Q6XZF7_PF14604_147_C3
Q9Y5K6_PF14604_109_C3
Q7Z5Q1_PF16366_513_C4
Q01826_PF16557_178_B3
Q05086_PF16558_30_C6
Q13501_PF16577_389_C5
Q8NHM5_PF16866_656_C4
O60885_PF17035_607_C3
Q12929_PF18016_720_C4
Q8WU90_PF18044_104_A1
Q86VP1_PF18112_729_A1
Q86YT6_PF18346_247_C4
O95409_PF18366_250_C6
Q9UKY1_PF18387_100_C6
Q9ULT8_PF18410_1892_C4
Q13148_PF18694_5_B3
HHH-rd1-0142_rockdoms_1_C7
EHEE-rd3-0037_rockdoms_1_C
EHEE-rd1-0882_rockdoms_1_C
mutant aa
position
scaled_gr
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

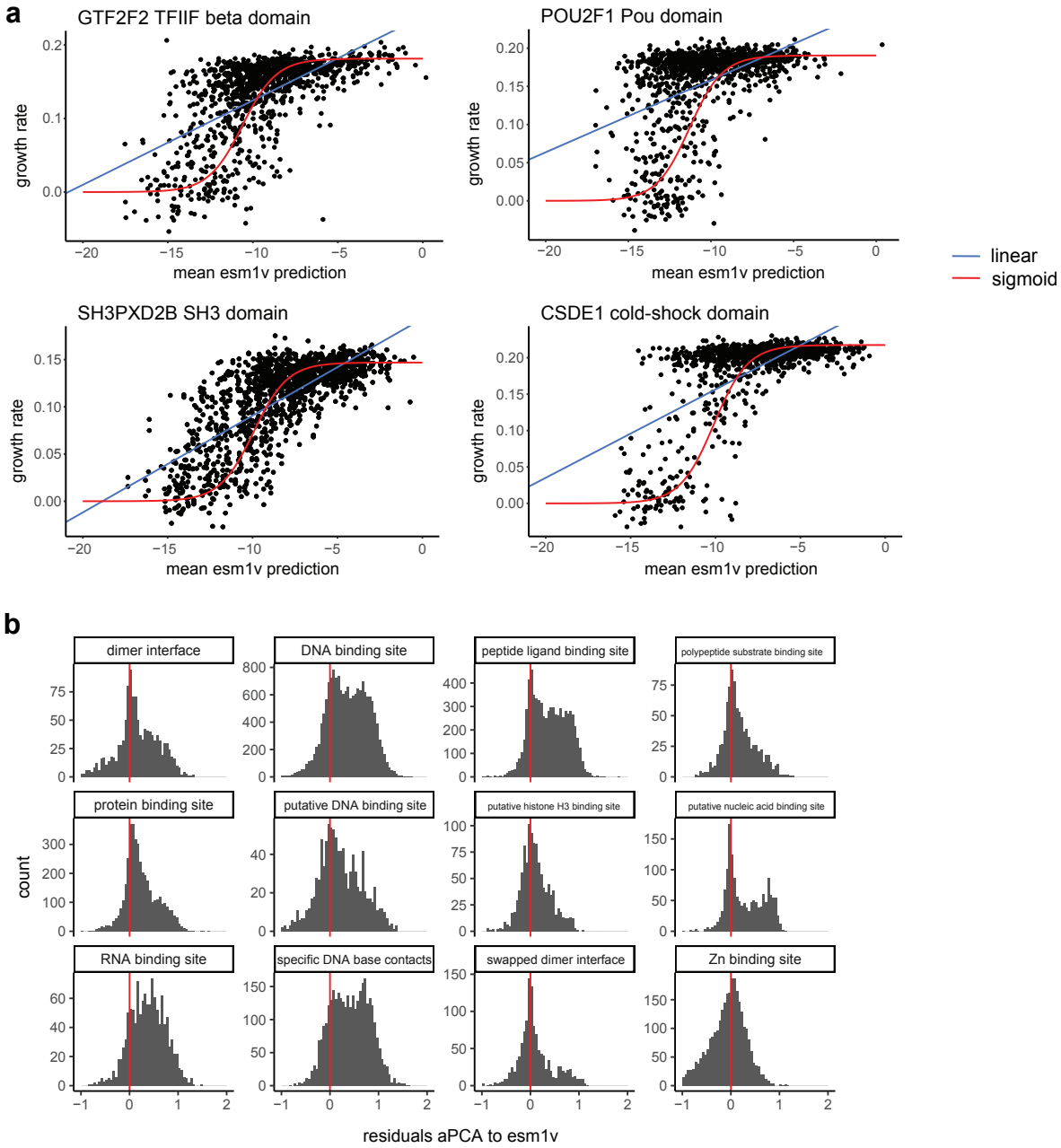
ED Figure 2-5


a
non-zinc finger domains
zinc finger domains
Spearman's rho
1.0
0.5
0.0
−0.5
0.49 0.46 0.42 0.45 0.40 0.35 0.57 0.39 0.38 0.37
0.45 0.40 0.38 0.42 0.40 0.34 0.23 0.25 0.05 0.30
median
ESM1vfull−length
ESM1vdomain
AlphaMissense
Tranception
popEVE
EVE
thermoMPNN
ddmut
RaSP
rsasa
predictor
type
fitness
stability
sasa
b
fraction of explainable variance in aPCA explained by esm1v (ranked-based)
1.00
0.75
0.50
0.25
0.00
fraction of explainable variance in aPCA explained by esm1v

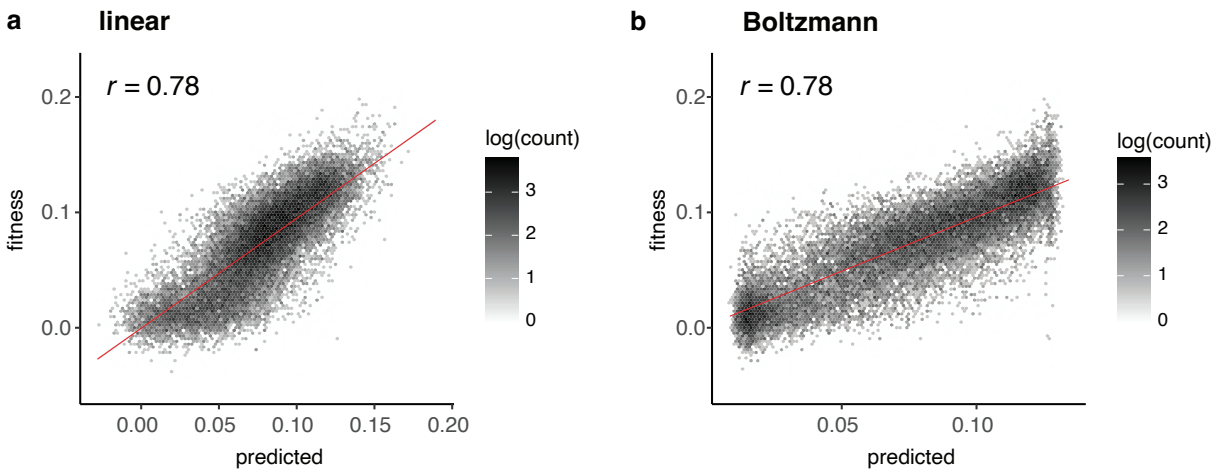
ED Figure 3


a
percentage of pathogenic variants
100
75
50
25
0
0
25
50
75
100
percentile rank
b
CUT PF02376
HMG-box PF00505
NHR-type C4 zinc finger PF00105
homeodomain PF00046
C2H2 zinc finger PF00096
PDZ PF00595
Pou PF00157
beta/gamma crystallin PF00030
SAM PF07647
LIM PF00412
family
0.0
0.5
1.0
proportion
pathogenic
benign + gnomad (af > 1e-5)
stability class
stable
mildly destabilizing
strongly destabilizing

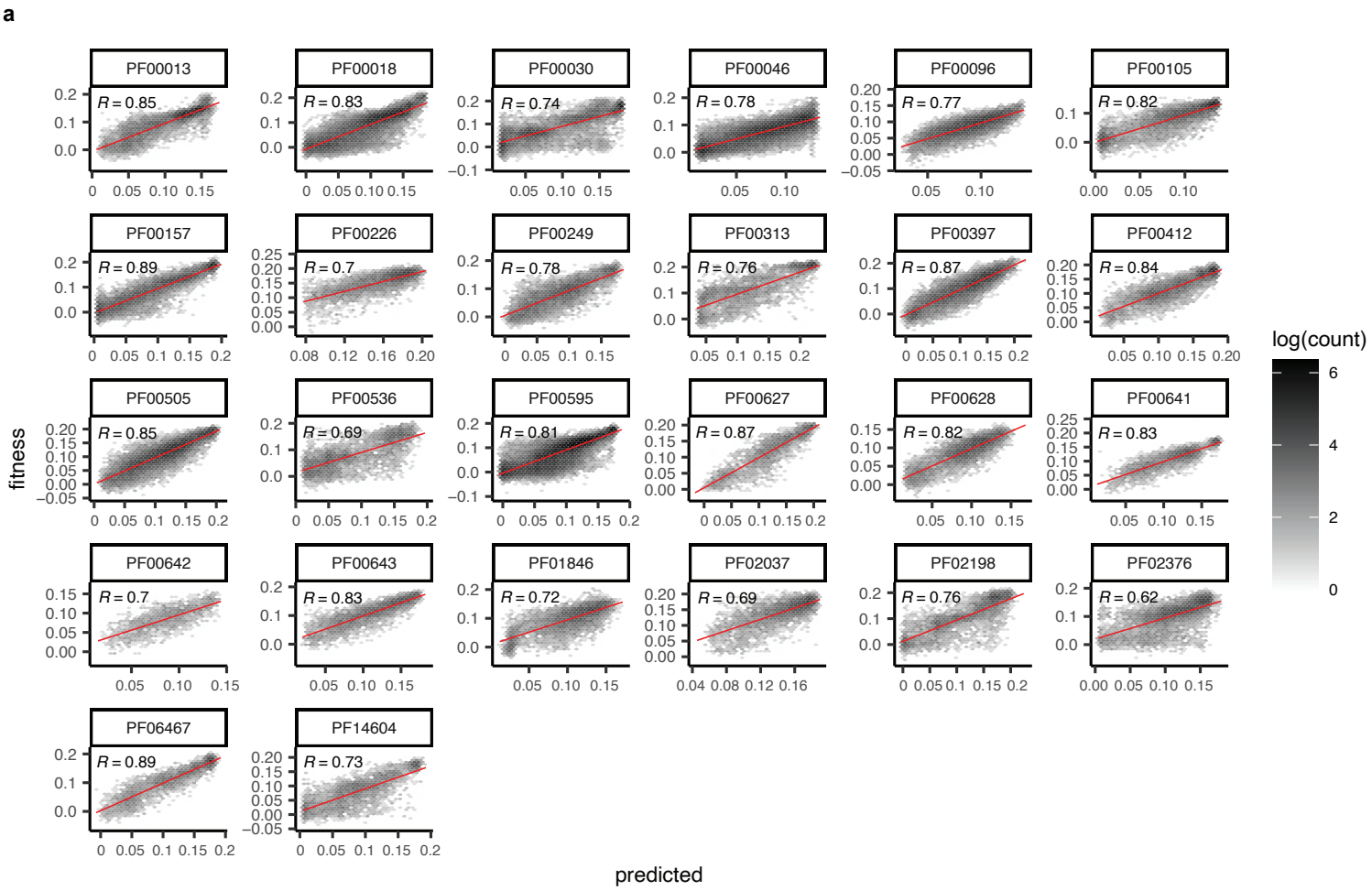
ED Figure 4

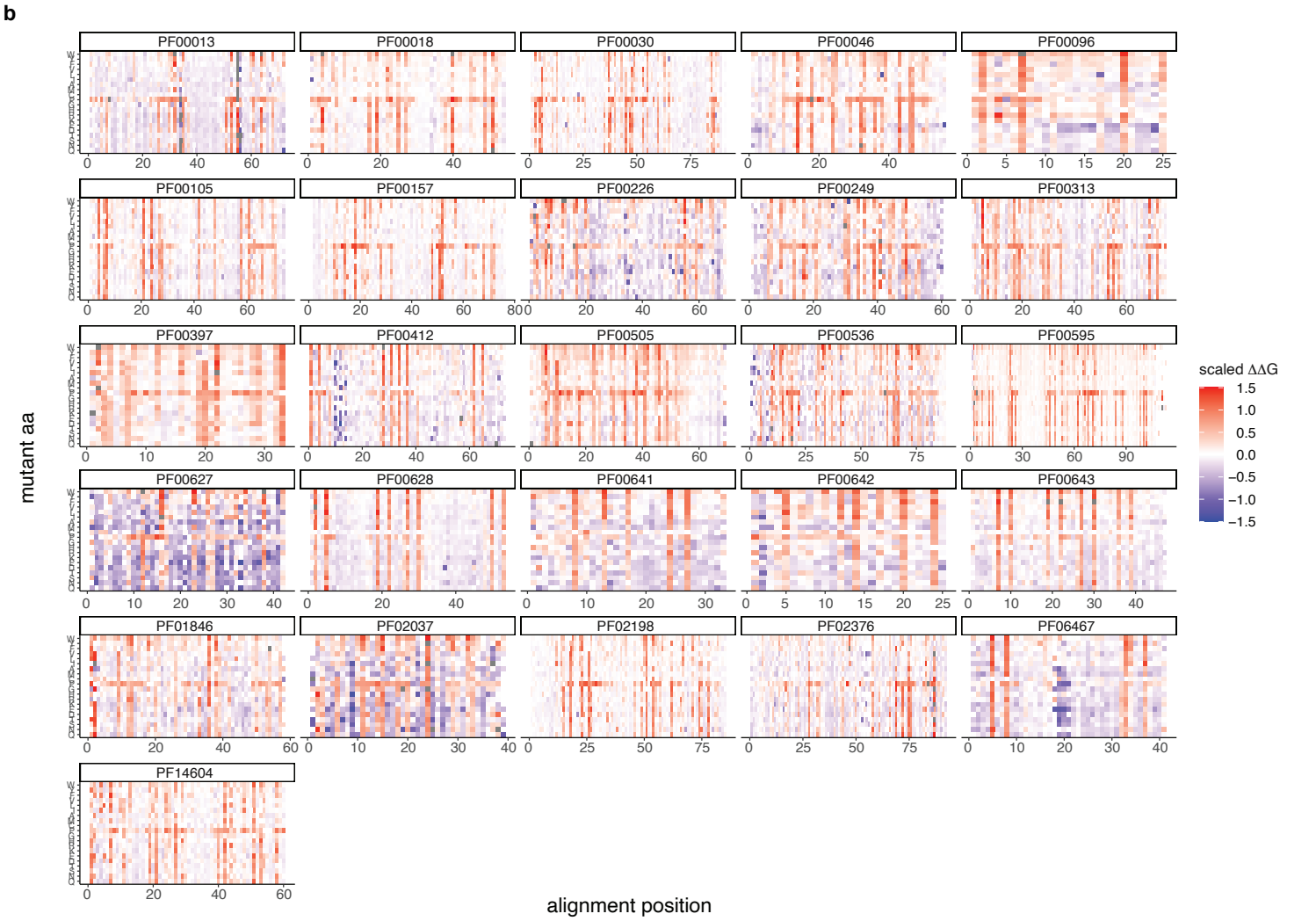

a
GTF2F2 TFIIF beta domain
POU2F1 Pou domain
SH3PXD2B SH3 domain
CSDE1 cold-shock domain
growth rate
mean esm1v prediction
linear
sigmoid
b
dimer interface
DNA binding site
peptide ligand binding site
polypeptide substrate binding site
protein binding site
putative DNA binding site
putative histone H3 binding site
putative nucleic acid binding site
RNA binding site
specific DNA base contacts
swapped dimer interface
Zn binding site
count
residuals aPCA to esm1v

ED Figure 5

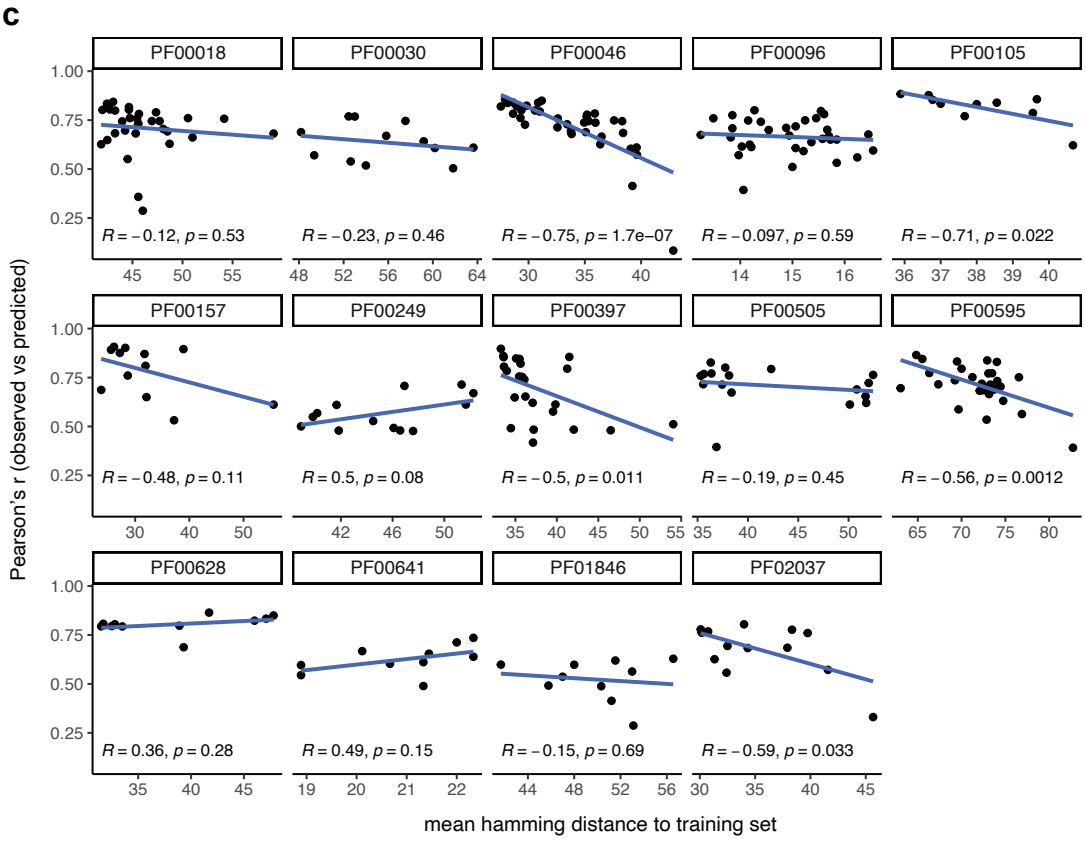

a
linear
r = 0.78
fitness
predicted
log(count)
b
Boltzmann
r = 0.78
fitness
predicted
log(count)

ED Figure 6

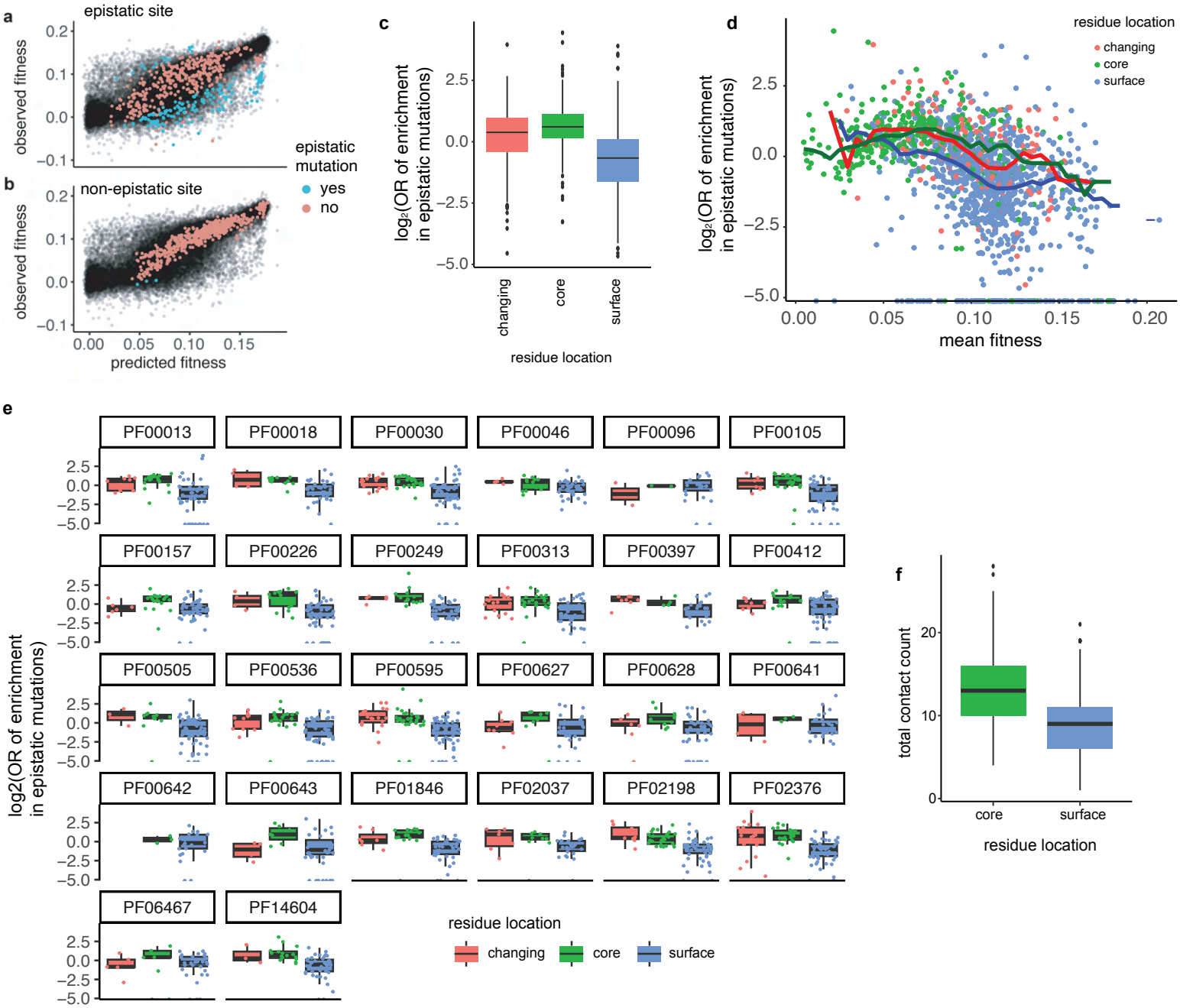

a
PF00013 R = 0.85
PF00018 R = 0.83
PF00030 R = 0.74
PF00046 R = 0.78
PF00096 R = 0.77
PF00105 R = 0.82
PF00157 R = 0.89
PF00226 R = 0.7
PF00249 R = 0.78
PF00313 R = 0.76
PF00397 R = 0.87
PF00412 R = 0.84
PF00505 R = 0.85
PF00536 R = 0.69
PF00595 R = 0.81
PF00627 R = 0.87
PF00628 R = 0.82
PF00641 R = 0.83
PF00642 R = 0.7
PF00643 R = 0.83
PF01846 R = 0.72
PF02037 R = 0.69
PF02198 R = 0.76
PF02376 R = 0.62
PF06467 R = 0.89
PF14604 R = 0.73
fitness
predicted
log(count)


b
PF00013
PF00018
PF00030
PF00046
PF00096
PF00105
PF00157
PF00226
PF00249
PF00313
PF00397
PF00412
PF00505
PF00536
PF00595
PF00627
PF00628
PF00641
PF00642
PF00643
PF01846
PF02037
PF02198
PF02376
PF06467
PF14604
mutant aa
alignment position
scaled ΔΔG

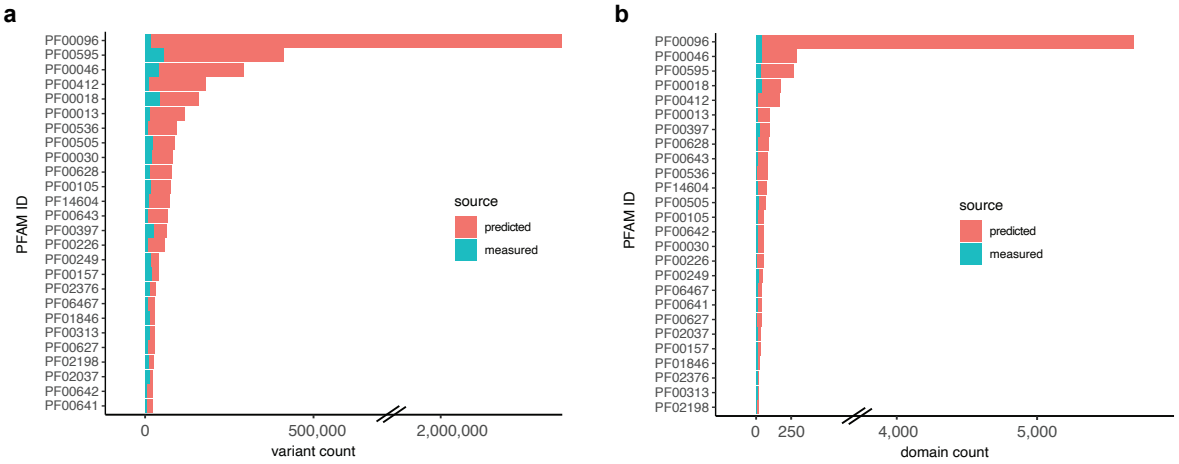
ED Figure 7-1


c
PF00018
PF00030
PF00046
PF00096
PF00105
R = −0.12, p = 0.53
R = −0.23, p = 0.46
R = −0.75, p = 1.7e−07
R = −0.097, p = 0.59
R = −0.71, p = 0.022
PF00157
PF00249
PF00397
PF00505
PF00595
R = −0.48, p = 0.11
R = 0.5, p = 0.08
R = −0.5, p = 0.011
R = −0.19, p = 0.45
R = −0.56, p = 0.0012
PF00628
PF00641
PF01846
PF02037
R = 0.36, p = 0.28
R = 0.49, p = 0.15
R = −0.15, p = 0.69
R = −0.59, p = 0.033
Pearson's r (observed vs predicted)
mean hamming distance to training set

ED Figure 7-2


a
epistatic site
observed fitness
b
non-epistatic site
observed fitness
predicted fitness
epistatic mutation
yes
no
c
log2(OR of enrichment in epistatic mutations)
changing
core
surface
residue location
d
residue location
changing
core
surface
log2(OR of enrichment in epistatic mutations)
mean fitness
e
PF00013
PF00018
PF00030
PF00046
PF00096
PF00105
PF00157
PF00226
PF00249
PF00313
PF00397
PF00412
PF00505
PF00536
PF00595
PF00627
PF00628
PF00641
PF00642
PF00643
PF01846
PF02037
PF02198
PF02376
PF06467
PF14604
log2(OR of enrichment in epistatic mutations)
residue location
changing
core
surface
f
total contact count
core
surface
residue location

ED Figure 8


a
PFAM ID
PF00096
PF00595
PF00046
PF00412
PF00018
PF00013
PF00536
PF00505
PF00030
PF00628
PF00105
PF14604
PF00643
PF00397
PF00226
PF00249
PF00157
PF02376
PF06467
PF01846
PF00313
PF00627
PF02198
PF02037
PF00642
PF00641
0
500,000
2,000,000
variant count
source
predicted
measured
b
PFAM ID
PF00096
PF00046
PF00595
PF00018
PF00412
PF00013
PF00397
PF00628
PF00643
PF00536
PF14604
PF00505
PF00105
PF00642
PF00030
PF00226
PF00249
PF06467
PF00641
PF00627
PF02037
PF00157
PF01846
PF02376
PF00313
PF02198
0
250
4,000
5,000
domain count
source
predicted
measured

ED Figure 9