Supporting Information

Bromberg et al. 10.1073/pnas.1216613110

SI Methods

Datasets. All variant counts are in Table 1. Databases are summarized in Table S3.

From the latest (March 2007) version of the Protein Mutant Database (PMD), we extracted, for each single amino acid variant, all experimentally derived annotations of "Disease," "Structure," "Stability," and "Function." We assumed all Disease annotations to imply that the variant is disease-associated. For Structure, Stability, and Function, PMD reports effect gradations as: (*i*) "[=]," wild type; (*ii*) "[-]," negative-effect levels (e.g., "[- - -]" is "severely reduced"); (*iii*) "[+]," positive-effect levels (e.g., "[+ +]" is "moderately increased"); (*iv*) "[0]," inactive. For each variant, we combined (*i*) all entries from the same sequence (100% identical over entire length) and (*ii*) "knockout" and "severe" into one "severe" annotation. For each variant, the most nonneutral Function effect of either (+ or −) direction determined whether the variant was included into the "PMDneutral," "PMDmild," "PMDmoderate," or "PMDsevere" set. All variants experimentally annotated to affect Structure or Stability formed the "PMDstr" set.

To create the "EC" (Enzyme Commission) set, we used the data from screening for nonacceptable polymorphisms (SNAP) training, as described in ref. 1. Briefly, we aligned orthologous enzymes (not in PMD) of the same Enzyme Commission (2) number and selected alignments with sequence identity >40% and HSSP distance >0 (3). We annotated residue differences between orthologs as neutral variants. To make the "PMD/EC-Human" set we installed polymorphism phenotyping (PolyPhen)-2 (version July 2012) and used all default parameters to make predictions (binary predictions gauged at pph_prob = 0.432 threshold, associated with a 10% false positive rate in the PolyPhen-2 v2.1.0r367 documentation) for all variants in human PMD/EC proteins. We also obtained sorting intolerant from tolerant (SIFT) (as in ref. 1, binary predictions gauged at the default 0.05 threshold) and SNAP scores. We compiled the subset of these variants for which all three methods made a prediction (96% of total).

SNAP has never been trained to predict as neutral a nonvariant (synonymous) SNP of the type "amino acid X to X." We can, thus, consider the degree to which SNAP incorrectly predicts synonymous SNPs to affect function as an upper limit of the accuracy in identifying neutral variants. We applied SNAP to the synonymous substitutions for all residues in the PMDneutrals set, e.g., if PMD included a neutral I301T variant, the "Synonymous" set had a neutral I301I variant in the same protein. The 14,651 PMDneutral variants include, in some cases, multiple variants at a single site. The Synonymous set, created from the PMDneutral sites that are not affected by any PMD nonneutral variants, thus encompasses 9,228 individual residue positions (Table 1).

The 4,041 LacI repressor (4) and the 2,015 lysozyme (5) experimentally annotated variants were split into four severity classes: "L&Lneutral," "L&Lmild," "L&Lmoderate," and "L&Lsevere." We combined all variants in both proteins [LacI repressor and lysozyme (L&L)] into one set with three effect classes, with milds and moderates forming a single intermediate class.

All SwissVar (June 2011) (6) "polymorphism" and "disease" variants were accumulated in the "SWISSPROTdisease" and "SWISSPROTpolymorphism" sets. For the FullProt set (100 randomly selected human enzymes; Table S3), we created 19 amino acid substitutions per position. Variants impossible via a SNP (e.g., tryptophan [TGG] to tyrosine [TAT/TAC]) went into the "FullProt-Imp" set. All those that were possible (e.g., glycine [$\underline{G}$GT] to arginine [$\underline{C}$GT]; where the underlined base represents the substitution) went into "FullProt-SNP."

We used the 1000 Genomes ("1KG") SNAP predictions from the SNPdbe (SNP DataBase of Effects) (7) database (May 2011 release of 1000 Genomes data). From HapMap (Haplotype Map) (8), we selected only the nonsynonymous (ns) SNPs found in all available populations with a reference allele frequency >0.1 (2,600 nsSNPs). We mapped each nsSNP to the dbSNP (Single Nucleotide Polymorphism Database) (9) and made SNAP predictions for all possible protein isoforms (4,209 sequence/variant pairs, "HapMap"). Because of sequence length limitations (section below, *SNAP Runs*), we were unable to make predictions for 113 (3%) of these variants.

We downloaded from PolyPhen the set of predictions for all possible nsSNPs in all human genes (hg19; ∼150M variants). The variants unique at the protein sequence level from this set made up the "PolyPhen-SNP" set (∼134 million variants). From the latter, all SNPs found in dbSNP release 135 and mapping unambiguously to a single genome location, made up the "PolyPhen-natural" set (70,338 variants). For both *PolyPhen* sets we recorded the PolyPhen-2 HumDiv model (10) pph2_prob predictions, ranged 0–1, and gauged binary predictions at the pph2_prob = 0.432 cutoff (as above).

SNAP Runs. We made SNAP predictions for all datasets (except PolyPhen-SNP and PolyPhen-natural) using default parameters. SNAP fails for proteins with over 6,000 residues. These were, therefore, excluded from all sets (56 proteins total). SNAP scores range from −100 (strong prediction for neutrality) to +100 (strong prediction of effect), with scores >0 indicating functional effects, whereas scores ≤0 indicate neutral predictions. As a consequence of the neural-network training, scores of exactly 0 come from both (neutral/non) prediction sides and encompass twice as many variants as expected. Here, for all figures, we divide the 0 scores into equal numbers of +0.001s and −0.001s. Note that all of these are still neutral for all other references. To maintain the same axis heights in the panels of Figs. 1 and 3, the distributions are normalized to the total number of mutations in each particular set (Table 1).

Statistical Significance Measurements. For all statistical comparisons, independence of datasets is assumed. To maintain this condition, overlapping variants were removed from the largest set of each pair of the compared datasets. Thus, dataset sizes are different for all compared pairs (Tables S1 and S2). For each pair of datasets, we used (*i*) the Welch's *t* test (11) to find differences between distribution means (assuming normal distribution), (*ii*) Mann–Whitney *U* test (12) to find median differences, and (*iii*) the two-sample Kolmogorov–Smirnov (K-S) test (13, 14) to measure the overall similarity between distributions. We assumed two dataset prediction distributions to be similar/identical (drawn from the same distribution) if the *P* value was >0.05 [>0.0025 with Benjamini–Hochberg correction for multiple comparisons (15)]. The K-S test is a nonparametric test that reports the maximum difference between the cumulative distributions of two datasets. Unlike the *t* test or the Mann–Whitney test, K-S tests for a number of deviations including, different medians, different variances, and different distribution shapes (16). Because K-S takes distribution shape into account, we have used it even though it is intended for continuous distributions, although our distributions are discrete. The *P* values for the K-S distance metric (KSD), therefore, mean little (17, 18), whereas the approximate set-pair KSDs we do report can be compared between pairs.

1. Bromberg Y, Rost B (2007) SNAP: Predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res* 35(11):3823–3835.
2. Bairoch A (2000) The ENZYME database in 2000. *Nucleic Acids Res* 28(1):304–305.
3. Rost B (2002) Enzyme function less conserved than anticipated. *J Mol Biol* 318(2):595–608.
4. Markiewicz P, Kleina LG, Cruz C, Ehret S, Miller JH (1994) Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *J Mol Biol* 240(5):421–433.
5. Rennell D, Bouvier SE, Hardy LW, Poteete AR (1991) Systematic mutation of bacteriophage T4 lysozyme. *J Mol Biol* 222(1):67–88.
6. Mottaz A, David FP, Veuthey AL, Yip YL (2010) Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics* 26(6):851–852.
7. Schaefer C, Meier A, Rost B, Bromberg Y (2012) SNPdbe: Constructing an nsSNP functional impacts database. *Bioinformatics* 28(4):601–602.
8. International HapMap Consortium (2003) The International HapMap Project. *Nature* 426(6968):789–796.
9. Sherry ST, et al. (2001) dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res* 29(1):308–311.
10. Adzhubei IA, et al. (2010) A method and server for predicting damaging missense mutations. *Nat Methods* 7(4):248–249.
11. Welch BL (1947) The generalization of Student's problem when several different population variances are involved. *Biometrika* 34(1-2):28–35.
12. Mann HB, Whitney DR (1947) On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann Math Stat* 18(1):50–60.
13. Kolmogorov A (1933) Sulla determinazione empirica di una legge di distribuzione. *G Ist Ital Attuari* 4:83–91. Italian.
13. Smirnov N (1948) Tables for estimating the goodness of fit of empirical distributions. *Ann Math Stat* 19(2):279–281.
14. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Series B* 57(1):289–300.
15. Lehmann EL, D'Abrera HJM (2007) *Nonparametrics: Statistical Methods Based on Ranks* (Springer, New York).
16. Jeng M (2006) Error in statistical tests of error in statistical tests. *BMC Med Res Methodol* 6:45.
17. Marsaglia G, Tsang WW, Wang J (2003) Evaluating Kolmogorov's distribution. *J Stat Softw* 8(18):1–4.

Table S1. Statistical significance (two-tailed *t* test and Mann–Whitney *U* test *P* value) of SNAP score distributions between variant datasets

		PMD/EC						PMD/EC-human						L&L			SP				FullProt	
		EC	Syn	Neut	Mild	Mod	Sev	EC	Neut	Mild	Mod	Sev	PMDstr	Neut	Inter	Sev	Dis	Poly	1KG	HapMap	SNP	Imp
PMD/EC	EC		–17	0	0	0	0		0	0	0	0	0	0	–320	0	0	0	0	0	0	0
	Syn	–33		0	0	0	0	0.6	0	0	0	0	0	0	–307	0	0	0	0	0	0	0
	Neut	0	0		–273	0	0	0		–131	–85	0	–169	–112	–20	–168	0	–95	0	–11	**0.44**	–154
	Mild	0	0	–281		–92	0	0	–141		–13	–214	0.09	0	–05	–47	0	–160	0	0	0	–95
	Mod	0	0	0	–91		–60	0	–270	–68		–7	–78	0	–44	0.007	–27	–276	0	0	0	–237
	Sev	0	0	0	0	–72		0	0	–297	–27		0	0	–93	–11	–30	0	0	0	0	0
PMD/EC-human	EC		0.4	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Neut	0	0		–147	–269	0	0		–94	–80	0	–102	–67	–19	–153	0	–44	0	–4	**0.3**	–70
	Mild	0	0	–129		–75	0	0	–96		–13	–134	0.04	0	–4	–45	–228	–40	0	–176	–189	–27
	Mod	0	0	–92	–12		–33	0	–84	–14		–09	–15	–169	–17	–05	–16	–45	0	–102	–88	–36
	Sev	0	0	0	–217	–9		0	0	–147	–10		–158	0	–57	0.6	–6	0	0	0	0	0
PMDstr		0	0	–188	0.78	–66	0	0	–106	0.7	–11	–145		0	0.001	–43	–306	–53	0	–198	–229	–30
L&L	Neut	0	0	–111	0	0	0	0	–63	–295	–166	0	0		–83	–283	0	–276	0	–58	–135	0
	Inter	0	0	–23	–05	–48	–118	0	–23	0.002	–16	–65	–4	–86		–39	–74	0.005	0	–31	–21	**0.33**
	Sev	0	0	–197	–54	–4	–12	0	–166	–55	–7	0.93	–40	–272	–42		0.003	–107	0	–190	–172	–92
SP	Dis	0	0	0	0	–19	–65	0	0	–247	–15	0.37	–290	0	–88	0.37		0	0	0	0	0
	Poly	0	0	–79	–239	0	0	0	–44	–61	–61	0	–96	0	–06	–159	0		0	–146	0	–37
1KG		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0
HapMap		0	0	–14	0	0	0	0	–05	–183	–120	0	–201	–57	–40	–228	0	–146	0		–15	–201
FullProt	SNP	0	0	**0.42**	0	0	0	0	***0.01***	–160	–83	0	–241	–127	–20	–187	0	–216	0	–20		0
	Imp	0	0	–162	–125	–276	0	0	–76	–27	–39	0	–48	0	**0.15**	–118	0	–85	0	–208	0	

For all detailed dataset descriptions, see *SI Methods*. Dataset names have been abbreviated to keep table within limits. For reference, in order from left to right and top to bottom of table rows and columns, the datasets are as follows: EC; Synonymous; PMDneutrals, milds, moderates, severes; PMD/EC-Human ECs, neutrals, milds, moderates, severes; PMDstr; L&L neutrals, intermediates, severes; SWISSPROT (SP) disease, polymorphisms; 1KG; HapMap; FullProt-SNP, FullProt-Imp. All negative values represent exponents (i.e., –X indicates 10^{-x}). Dark gray cells indicate cell identity, gray cells represent the results of the Mann–Whitney *U* test, and white cells represent the results of a two-tailed *t* test. All results of a single test are symmetric across the diagonal and, therefore, are reported only once. Values highlighted in bolded roman text are >0.05; these are standard cutoff values used to signify whether two sets of values come from the same distribution: higher than cutoff is likely the same distribution. To account for multiple comparisons (Benjamini–Hochberg correction), we also highlight (in bolded italic text) significant values of >0.0025 and >0.0026 (0.05/20 or 0.05/19, as appropriate).

Table S2. KSD between score distributions of variant datasets

		PMD/EC						PMD/EC-human						L&L			SP				FullProt	
		EC	Syn	Neut	Mild	Mod	Sev	EC	Neut	Mild	Mod	Sev	PMDstr	Neut	Inter	Sev	Dis	Poly	1KG	HapMap	SNP	Imp
PMD	EC		0.08	0.60	0.70	0.78	0.83		0.61	0.71	0.76	0.79	0.67	0.47	0.67	0.79	0.83	0.71	0.95	0.63	0.57	0.68
PMD/EC	Syn			0.59	0.69	0.78	0.83	0.03	0.60	0.70	0.76	0.79	0.67	0.46	0.66	0.79	0.83	0.69	0.95	0.61	0.56	0.67
	Neut				0.20	0.34	0.45	0.56		0.18	0.30	0.37	0.20	0.19	0.14	0.39	0.39	0.11	0.88	0.09	0.04	0.13
	Mild					0.15	0.27	0.68	0.22		0.11	0.18	0.03	0.36	0.07	0.20	0.20	0.16	0.86	0.29	0.17	0.11
	Mod						0.13	0.76	0.39	0.18		0.06	0.15	0.49	0.23	0.06	0.07	0.31	0.86	0.45	0.31	0.25
	Sev							0.81	0.47	0.30	0.18		0.25	0.58	0.32	0.08	0.09	0.42	0.86	0.53	0.41	0.37
PMD/EC-	EC								0.59	0.69	0.74	0.78	0.66	0.44	0.65	0.77	0.82	0.68	0.95	0.59	0.55	0.66
human	Neut									0.20	0.32	0.39	0.22	0.17	0.16	0.41	0.43	0.11	0.88	0.07	0.06	0.16
	Mild										0.13	0.21	0.05	0.36	0.06	0.23	0.22	0.13	0.87	0.26	0.16	0.08
	Mod											0.10	0.12	0.44	0.16	0.11	0.12	0.27	0.86	0.38	0.26	0.21
	Sev												0.19	0.52	0.24	0.03	0.04	0.34	0.86	0.46	0.34	0.28
PMDstr														0.36	0.07	0.19	0.21	0.17	0.86	0.29	0.16	0.12
L&L	Neut														0.31	0.52	0.56	0.28	0.89	0.16	0.21	0.32
	Inter															0.26	0.27	0.11	0.87	0.22	0.12	0.06
	Sev																0.05	0.37	0.86	0.48	0.36	0.30
SP	Dis																	0.36	0.86	0.48	0.36	0.29
	Poly																		0.88	0.16	0.14	0.07
1KG																				0.89	0.87	0.87
HapMap																					0.13	0.21
FullProt	SNP																					0.11
	Imp																					

For all detailed dataset descriptions, see *SI Methods*. Dataset names have been abbreviated to keep table within limits. For reference, in order from left to right and top to bottom of table rows and columns, the datasets are as follows: EC; Synonymous; PMDneutrals, milds, moderates, severes; PMD/EC-Human ECs, neutrals, milds, moderates, severes; PMDstr; L&L neutrals, intermediates, severes; SWISSPROT (SP) disease, polymorphisms; 1KG; HapMap; FullProt-SNP, Full-Prot-Imp.

Table S3. Summary of database contents

Database	Description
Swiss-Prot	Protein-centered, manually annotated, and reviewed subset of UniProt knowledgebase. Includes annotations of naturally occurring variants evaluated for disease predisposition (UniVar).
PMD	Scientific literature (article)-centered collection of variants annotated for disease association and effects on structure, function, and stability of proteins
dbSNP	The Single Nucleotide Polymorphism database is an archive of many simple genetic polymorphisms with some population frequency annotation
SNPdbe	SNPdbe is a database that joins related bits of knowledge about nsSNPs, currently distributed throughout various databases, into a single resource
1000 Genomes	A database cataloging genetic variation in >1,000 healthy individuals from different ethnic subgroups (includes frequency annotation)
HapMap	A set of variants describing the haplotype map of the human genome: the common variants in human DNA

Other Supporting Information Files

Dataset S1 (XLSX)