

Supplementary Table 1. Results of mapping SNPs from dbSNP onto protein sequences from SwissProt/TrEMBL. Due to gene duplication and alternatively-spliced isoforms, some proteins are aligned to the genome in more than one location. The 98.8% identity between amino-acid residues and translated codons in the protein-to-genome alignments indicates that these mappings are of high quality. The genomic locations of all SNPs were taken from dbSNP build 120.

	all			dbSNP-validated		
	total	SwissProt	TrEMBL	total	SwissProt	TrEMBL
dbSNP SNPs resulting in non-synonymous substitutions	70,153	20,832	49,321	28,043	7,730	20,313
Proteins with nsSNPs aligned to genomic DNA	24,944	6,476	18,468	14,551	3,561	10,990
Alignments of nsSNP-containing proteins to genomic DNA	25,313	6,543	18,770	14,707	3,589	11,118
Percent identity of amino acids and translated codons in the alignments	0.988	0.974	0.994	0.988	0.973	0.994

Supplementary Table 2. Results of automated comparative modeling of nsSNP-containing protein sequences with MODPIPE. A protein may have several models, each covering a distinct region of its sequence. A covered position is one in which we can identify an equivalent (aligned) residue in the structure of the template protein used to build the model.

	Total	SwissProt	TrEMBL
protein sequences modeled	13,393	4,628	8,765
models produced	40,777	14,552	26,225
models that cover nsSNPs	17,874	6,944	10,930
protein sequences with models that cover nsSNPs	8,931	3,372	5,559
nsSNPs covered in the models	13,062	8,494	4,568
models that cover validated nsSNPs	8,725	3,272	5,453
proteins sequences with models that cover validated nsSNPs	4,593	1,686	2,907
validated nsSNPs covered in the models	4,907	3,013	1,894

Supplementary Table 3. (a) Sample structural annotations for non-synonymous SNPs. Features used in predicting destabilizing SNPs include the secondary structure and relative solvent accessibility of the SNP position in protein structure models and the proximity of the SNP's equivalent residue in the PDB template structure to ligands and domain-domain interfaces.

dbSNP ID	Variant residues	SNP position	Swiss-Prot ID	DSSP code	Relative solvent exposure [%]	Template PDB code	Equiv. residue number	Target-Template sequence identity [%]	Model assessment score	Ligand	Domain-domain interface [PIBASE]
rs5948	IN	219	P12429	H	35	1axn	220	100	1.00	-	-
rs17585	PL	156	P13489	T	52	1a4yA	169	100	1.00	-	c.10.1.1- c.10.1.1
rs698	IVF	349	P00326	E	57	1dehA	349	95	1.00	-	-
rs8192462	EQ	123	P00568	H		3adk	123	95	1.00	-	-
rs1131215	YD	98	P30464	H		1efxA	116	92	1.00	-	-
rs5030621		161	P00338	L	2	9ldtA	163	92	1.00	NAD	c.2.1.5- d.162.1.1
rs1131218	SN	92	Q04826	H	35	1alnA	77	91	1.00	-	d.19.1.1- b.1.1.4
rs11575344	PL	210	P20711	L	65	ljs3A	210	88	1.00	-	-
rs1126478	KR	48	P02788	H	76	ljw1A	28	71	1.00	-	-
rs1042140	KE	97	P0440	H	3	2iadB	71	64	1.00	-	-
rs4740046	NS	68	Q8N1Q1	E	11	1hcb	67	61	1.00	-	-
rs3752566	RH	2039	P08922	H	43	1p4oA	1062	47	1.00	-	-
rs753856	DE	562	P17066	T	69	1dkyB	55	46	1.00	-	-
rs5030752	IV	971	P07814	S	46	1hc7A	33	44	1.00	-	d.104.1.1- d.104.1.1
rs2748210	QK	126	P09466	E	19	1bebA	108	44	1.00	-	-
rs1804495	LF	303	P05543	E	1	1qlpA	286	43	1.00	-	-
rs11970	FY	418	Q9BRV0	E	16	1igtB	435	39	1.00	-	-
rs1801394	IM	39	Q7Z4M8	B	0	1ja1A	98	30	1.00	-	-

Supplementary Table 3. (b) Twenty-five non-synonymous SNPs that putatively destabilize protein structure. The sequence identity of the SwissProt/TrEMBL protein (target) and PDB structure used to build the model (template) and a model assessment score are provided as reliability measures for the annotation. UAPC unfavorable accessible surface potential change.

dbSNP ID	Variant amino acids	SNP position in protein	Protein SwissProt/TrEMBL ID	Target-template sequence identity	PDB code	Model assessment score	Reason
rs5643	WR	276	P11806	100	1hnnA	1.00	buried charge
rs9696578	RL	668	P06396	95	1d0nA	1.00	domain interface, buried charge
rs14777	LP	330	P40925	95	5mdhA	1.00	domain interface, helix breaker
rs5030849	RQ	261	P00439	93	1phzA	1.00	domain interface, buried charge
rs1801266	RW	235	Q12882	93	1gte	1.00	near ligand FAD, buried charge
rs5030621	SR	161	P00338	92	9ldtA	1.00	near ligand NAD, buried charge
rs11546624	DGNS	69	Q8WU19	93	1ffxA	1.00	near ligand GTP, buried charge
rs11558370	QP	122	Q8WUW7	91	1a49A	1.00	helix breaker
rs11549172	PLTI	102	P20571	90	1yagA	1.00	domain interface, UAPC at buried position
rs2295474	CY	1317	P47989	89	1fo4A	1.00	domain interface, UAPC at buried position
rs167447	HL	216	P04746	87	1hx0A	1.00	near ligand Acarbose molecule, buried charge
rs3179181	NEDK	87	P18462	86	1efxA	1.00	buried charge
rs1131215	YD	73	Q29763	84	1qqdA	1.00	buried charge
rs513694	DG	15	AAP35520	83	1fbl	1.00	buried charge
rs1043657	AT	142	O43488	79	1gveA	1.00	near ligand citric acid, UAPC at buried position
rs1111335	RW	75	Q7Z373	79	1bg3A	1.00	buried charge
rs1059517	QK	168	Q9GJ45	75	1kjmA	1.00	buried charge
rs1049110	QR	161	Q8MGQ8	67	2iadB	1.00	UAPC at buried position
rs11548056	IT	68	AAP35812	67	1mfiA	1.00	near interface, UAPC at buried position
rs2240572	HR	107	Q96PY2	60	1d0nA	1.00	UAPC at buried position
rs2020950	DG	162	P29122	59	1p8jA	1.00	near ligand calcium ion, buried charge
rs4986891	RQ	128	P11509	51	1nr6A	1.00	near ligand heme, buried charge
rs1555696	LR	106	Q8TE74	51	1kcgC	1.00	buried charge
rs2234953	EK	172	P30711	51	1ljr	1.00	domain interface, buried charge
rs1801272	LH	160	P11509	51	1nr6	1.00	buried charge

Supplementary Table 3. (c) Twenty-five non-synonymous SNPs (not found in our training set) predicted as disease-associated with high confidence by the support vector machine (SVM). The SVM classifies each example with a discriminant score. In our implementation, negative scores predict disease association while positive scores predict a neutral or positive nsSNP. The absolute value of the score provides a confidence measure for the prediction. In three-fold cross-validated testing on our benchmark set, the SVM had 80.5% prediction accuracy. Protein names and functions are taken from the SwissProt database (Boeckmann, et al., 2003).

dbSNP ID	Variant residues	SNP position	Protein SwissProt/ TrEMBL ID	Protein name	Protein function	Discriminant score
rs7507257	LP	174	Q9NSA1	Fibroblast growth factor-21 [Precursor]	cell-cell signaling	-1.67
rs5325	LP	303	P09172	Dopamine beta-monoxygenase [Precursor]	synaptic transmission	-1.67
rs9659608	AS	375	P17066	Heat shock protein 6	chaperone	-1.50
rs6943147	ST	592	P19801	Kidney amine oxidase	amine oxidase	-1.25
rs5641	LQ	217	P11086	Phenylethanolamine N-methyltransferase	catecholamine biosynthesis	-1.25
rs3970559	RC	369	O43272	Proline oxidase, mitochondrial [Precursor]	proline metabolism	-1.25
rs8050904	TM	1406	Q7Z442	Polycystic kidney disease 1-like 2	cation channel activity	-1.24
rs1805378	IT	176	P78549	Endonuclease III-like protein 1	nucleotide-excision repair	-1.19
rs3807068	KN	147	Q03924	Zinc finger protein 117 [Fragment]	transcription regulation; zinc ion binding	-1.14
rs6140	IT	331	P24557	Thromboxane-A synthase	prostaglandin to thromboxane conversion	-1.14
rs1801280	IT	114	P11245	Arylamine N-acetyltransferase 2	conjugation reaction critical to rate of drug metabolism	-1.12
rs5030621	SR	160	P00338	L-lactate dehydrogenase	oxioreductase. catalyzes conversion of lactate to pyruvate	-1.10
rs10113875	GR	33	P01566	Interferon alpha-10 [Precursor]	response to virus	-1.10
rs4976	IT	1018	P12821	Angiotensin-converting enzyme, somatic isoform [Precursor]	blood pressure regulation	-1.08
rs6943147	SR	592	P19801	Amiloride-sensitive amine oxidase	amine oxidase activity	-1.05
rs11571098	RW	33	P00797	Renin [Precursor]	endopeptidase, initiates reactions that elevate blood pressure and increase kidney sodium retention	-1.04
rs5030382	EK	469	P05362	Intercellular adhesion molecule-1 [Precursor]	rhinovirus receptor	-0.95
rs2229624	AS	387	P52789	Hexokinase, type II	cell cycle regulation	-0.95
rs3745765	IS	494	P10072	Krueppel-related zinc finger protein 1	putative transcriptional regulation	-0.93
rs1804298	EG	110	P39026	Angiotensin-converting enzyme	increases vasoconstrictor activity of angiotensin; inactivates the vasodilator bradykinin	-0.93
rs963075	RC	246	P48595	Bomapin	putative protease regulation during hematopoiesis	-0.93
rs8176739	RC	199	P16442	Histo-blood group ABO system transferase	glycosyl transferase/basis of the ABO blood group system	-0.92
rs8179183	KN	656	P48357	Leptin receptor [Precursor]	receptor for obesity factor	-0.91
rs2236225	RQ	652	P11586	C-1-tetrahydrofolate synthase	required for biosynthesis of purines, thymidylate, methionine, histidine, pantothenate, and formyl tRNA-Met	-0.90
rs5324	DN	276	P09172	Dopamine beta-monoxygenase [Precursor]	Catecholamine biosynthesis	-0.89

Supplementary Figure 1. (a) The model of human dihydropyrimidine dehydrogenase provides a structural explanation for the association of a SNP (dbSNP ID rs1801266, ARG235TRP) with DPD deficiency (Vreken, et al., 1997) (buried charge change near a ligand). The template for the MODPIPE comparative model is 1gte, a pig DPD. Overall target-template sequence identity is 93%, with 97.26% sequence identity in the region around the ligand. The SNP is shown in spacefill, near the heterocyclic rings of FAD. (b) A SNP in human glutathione S-transferase (GST) theta 1 (dbSNP ID rs2234953, Glu172Lys) lies at a domain interface, produces a buried charge change and an unfavorable change in accessible surface potential at a buried position. The model is based on human GST theta 2 (PDB code 1ljr). Overall target-template sequence identity is 51%, but for the domain-interface residues sequence identity is 67%. In GST theta 2, Glu172 and Arg107 are in close proximity and Arg107 interacts with the thiol sulfur of the GSH ligand (Rossjohn, et al., 1998). These sidechains are conserved in GST theta 1, supporting the idea that the Arg107Glu substitution may be deleterious. The two domains of GST are colored yellow and green, Arg107 is shown in cyan. Interface residues (within 6 Å of the adjacent domain) are shown in spacefill with the SNP in blue. (c) A SNP (dbSNP ID rs1801272, Leu160His) in cytochrome P450 2A6 produces a buried charge change (shown as blue sphere). The model is based on rabbit cytochrome P450 2C5 (PDB code 1nr6). Overall target-template sequence identity is 51%. The accuracy of the local alignment in the vicinity of the SNP is supported by a Leu found at equivalent positions in both the target and template proteins. The clinical literature confirms that this SNP produces an unstable and catalytically inactive enzyme (Yamano, et al., 1990). Graphics produced with Chimera (Huang, et al., 1996).

