Supporting Information

Supp. Methods

Mutation prediction analysis
The effect of missense variations was evaluated using different tools and prediction software, including BLOSUM 62, SIFT, POLYPHEN, PANTHER, PMUT and SNAP.

BLOSUM 62 is a scoring matrix for amino acid substitutions; it is based on local alignment of sequences with no less than 62% divergence. The more negative the value, the more deleterious the substitution [Henikoff and Henikoff, 1992].

SIFT (http://blocks.fhcrc.org/sift/SIFT.html; Sorting Intolerant From Tolerant) predicts the functional importance of amino acid substitutions and their potential pathogenicity, based on the alignment of orthologous and/or paralogous protein sequences. SIFT scores were classified as intolerant (0.00–0.05), potentially intolerant (0.051–0.10), borderline tolerant (0.101–0.20), or tolerant (0.201–1.00) according to a classification previously proposed [Ng and Henikoff, 2003; Xi et al., 2004]. The higher the tolerance index, the less functional impact a particular amino acid substitution is likely to have, and vice versa.

POLYPHEN (http://genetics.bwh.harvard.edu/pph/; POLYmorphism PHENotyping) predicts the effect of an amino acid substitution on the structure and function of a protein. POLYPHEN predictions are based on empirical rules that are applied to the sequence, as well as phylogenetic and known structural information that characterize the substitution. It calculates the Position-Specific Independent Counts (PSIC) for the two different alleles and the score for wild type and variant mapping to the known 3D structure [Ramensky et al., 2002].

PANTHER (http://www.pantherdb.org/; Protein ANalysis THrough Evolutionary Relationships) estimates the likelihood of a non-synonymous variant to cause loss of function of the protein. The output, the subPSEC (substitution position-specific evolutionary conservation), is the negative logarithm of the probability ratio of the wild-type and mutant amino acids at a particular position based on a library. This library contains a set of over 5,000 protein families and 30,000 subfamilies, each represented by a multiple sequence alignment and Hidden Markov Model (HMM). PANTHER subPSEC scores are continuous from 0 to -10. A value of 0 is interpreted as a functionally neutral variant; the more negative the subPSEC value, the more deleterious the substitution. The cutoff value suggested is -3 [Thomas et al., 2003; Thomas and Kejariwal, 2004; Thomas et al., 2006].

PMUT (http://mmb2.pcb.ub.es:8080/PMut/) use neural networks that have been trained with a large database of disease-associated and neutral variants to predict the impact of a given amino acid substitution. The output gives a neural network (NN) value between 0 and 1 (the higher this value, the more deleterious the variant) and a confidence value

between 0 and 9 (the higher this value, the more reliable the NN) [Ferrer-Costa et al., 2005].

SNAP (http://cubic.bioc.columbia.edu/services/SNAP/; Single Nucleotide polymorphism Annotation Platform) is a neural-network based method that uses protein information (e.g. conservation, secondary structure, chemical property, etc.) extracted from several public databases (Ensembl, SWISS-PROT, Pfam, etc.) in order to predict the effects of amino acid substitutions [Bromberg and Rost, 2007; Bromberg et al., 2008]. The output gives the user an assessment of the possible effect on the protein, i.e., as "non neutral" or "neutral". It also provides reliability indices.

The effect of splice site variations was also evaluated, using different analysis programs, including the splice site prediction tool from the Berkeley Drosophila Genome Project (**BDGP**) web site (http://www.fruitfly.org/seq_tools/splice.html). This is based on a generalized Hidden Markov Model to predict the strength of the possible splice site, using a neural network that has been trained by a set of 793 unrelated human genes [Reese et al., 1997].
The other splice site prediction tool we employed was **NetGene2**, available from the Center of Biological Sequence analysis (CBS), http://www.cbs.dtu.dk/services/NetGene2/, which is also based on a neural network that was trained by a data set containing 146 genes extracted from GenBank [Hebsgaard et al., 1996].

Supp. Table S1: Prediction analysis of the effects of *HGD* missense variants on protein function

	Exon	Variants	Blossum 62	POLYPHEN		SIFT	PMUT		PANTHER	SNAP			Enzyme Activity [b] (%wt)	Publication
				PSIC score	Exclusion [a]	Score	NN score	Reliability	subPSEC	Prediction	Reliability	Accuracy		
Variants found in this study (Novel and previously published)	1	p.E3A	-1	1.083	Align	0.03	0.4791	0	-1.69	Neutral	4	85%		
		p.L4S	-2	1.694	Align	0.03	0.4798	0	-2.95	Neutral	0	53%		[Phornphutkul et al., 2002]
	3	p.E42A	-1	2.706	Align	0.00	0.6223	2	-3.83	Non-neutral	5	87%	29	[Beltran-Valero de Bernabe et al., 1998]
	4	p.W60G	-2	4.264	Align	0.00	0.7041	4	-4.11	Non-neutral	5	87%	0.06	[Beltran-Valero de Bernabe et al., 1999a]
		p.L61P	-3	2.294	Align	0.00	0.1187	7	-4.87	Non-neutral	1	63%		[Phornphutkul et al., 2002]
		p.Y62C	-2	3.416	Align	0.00	0.7325	4	-4.62	Non-neutral	6	93%	22.5	[Beltran-Valero de Bernabe et al., 1999a]
		p.F73L	0	1.928	Align	0.05	0.3902	2	-1.45	Non-neutral	0	58%		
		p.P92T	-1	2.516	Align	0.02	0.4052	1	-3.19	Non-neutral	3	78%		[Phornphutkul et al., 2002]
	5	p.W97R	-3	4.567	Align	0.00	0.9745	9	-3.82	Non-neutral	6	93%		[Phornphutkul et al., 2002]
	6	p.C120F	-2	2.979	Align	0.00	0.8661	7	-2.66	Non-neutral	2	70%		
		p.C120W	-2	3.05	Align	0.00	0.7364	4	-3.99	Non-neutral	3	78%		[Goicoechea De Jorge et al., 2002]
		p.A122V	0	1.832	Align	0.10	0.6792	3	-3.08	Non-neutral	2	70%		[Ladjouze-Rezig et al., 2006],
		p.G123R	-2	2.52	Align	0.00	0.6783	3	-3.44	Non-neutral	4	82%		
		p.L137P	-3	1.476	Struct	0.20	0.4112	1	-2.59	Neutral	2	69%		
	7	p.N149K	0	2.272	Align	0.00	0.4306	1	-2.36	Non-neutral	4	82%		
	8	p.P158L	-3	2.239	Align	0.02	0.3284	3	-2.68	Non-neutral	2	70%		
		p.G161R	-2	2.711	Align	0.00	0.62	2	-5.06	Non-neutral	7	96%	1	[Gehrig et al., 1997]
		p.E168K	1	2.256	Align	0.00	0.6703	3	-2.39	Non-neutral	6	93%		[Higashino et al., 1998]
		p.E168D	2	2.031	Align	0.00	0.0458	9	-2.56	Non-neutral	3	78%		[Phornphutkul et al., 2002]
		p.Q183R	1	2.097	Align	0.00	0.2819	4	-2.25	Non-neutral	1	63%		
	9	p.R187G	-2	2.527	Align	0.01	0.8225	6	-3.15	Non-neutral	2	70%		
		p.G217W	-2	2.757	Align	0.00	0.814	6	-4.65	Non-neutral	4	82%		
	10	p.R225L	-2	3.11	Align	0.00	0.6989	3	-3.55	Non-neutral	4	82%		[Phornphutkul et al., 2002]
		p.P230S	-1	2.986	Align	0.00	0.3754	2	-3.05	Non-neutral	4	82%	4	[Fernandez-Canon et al., 1996]

	Exon	Variants	Blossum 62	POLYPHEN		SIFT	PMUT		PANTHER	SNAP			Enzyme Activity [b] (%wt)	Publication
				PSIC score	Exclusion [a]	Score	NN score	Reliability	subPSEC	Prediction	Reliability	Accuracy		
Other variants previously published		p.Q258P	-1	2.509	Align	0.21	0.6444	2	-3.4	Non-neutral	0	58%		[Phornphutkul et al., 2002]
	11	p.H269R	0	3.327	Align	0.00	0.6455	2	-3.29	Non-neutral	4	82%		
		p.G270R	-2	2.79	Align	0.00	0.8246	6	-5.06	Non-neutral	6	93%		[Muller et al., 1999]
	12	p.V300G	-3	2.975	Align	0.00	0.5968	1	-3.47	Non-neutral	5	87%	1.9	[Fernandez-Canon et al., 1996]
		p.S305F	-2	2.575	Align	0.00	0.8493	6	-3.65	Non-neutral	3	78%		[Phornphutkul et al., 2002]
		p.R321P	-2	3.136	Align	0.00	0.7986	5	-3.96	Non-neutral	4	82%		
	13	p.P359L	-3	3.231	Align	0.00	0.4018	1	-3.04	Non-neutral	2	70%		
		p.G360R	-2	2.613	Align	0.00	0.8888	7	-4.94	Non-neutral	5	87%		[Porfirio et al., 2000]
		p.G362E	-2	1.599	Struct	0.01	0.8426	6	-2.54	Non-neutral	2	70%		[Phornphutkul et al., 2002]
		p.M368V	1	2.373	Align	0.01	0.2407	5	-2.76	Non-neutral	5	87%	37	[Beltran-Valero de Bernabe et al., 1998]
		p.P373L	-3	2.074	Align	0.00	0.8929	7	-4.86	Non-neutral	1	63%		
	14	p.E401Q	2	2.031	Align	0.00	0.0877	8	-3.03	Non-neutral	3	78%		[Goicoechea De Jorge et al., 2002]
	2	p.L25P	-3	2.626	Align	0.00	0.1398	7	-1.91	Non-neutral	2	70%		[Felbor et al., 1999]
	3	p.S47L	-2	2.016	Align	0.00	0.6876	3	-2.75	Non-neutral	0	58%		[Zatkova et al., 2000b]
		p.R53W	-3	3.171	Align	0.00	0.9616	9	-5.07	Non-neutral	3	78%		[Rodriguez et al., 2000]
		p.K57N	0	1.506	Align	0.03	0.4212	1	-2.02	Neutral	1	60%		[Grasko et al., 2009]
	5	p.W97G	-2	4.342	Align	0.00	0.7724	5	-4.09	Non-neutral	6	93%	0.1	[Beltran-Valero de Bernabe et al., 1998]
	6	p.A122D	-2	1.676	Align	0.16	0.7101	4	-2.81	Non-neutral	3	78%	33.5	[Beltran-Valero de Bernabe et al., 1999a]
		p.F136Y	3	0.958	Struct	1.00	0.0993	8	-1	Neutral	6	92%		[Phornphutkul et al., 2002]
		p.E143D	2	0.187	Align	0.64	0.0442	9	-1.45	Neutral	8	96%		[Phornphutkul et al., 2002]
	7	p.D153G	-1	2.39	Align	0.00	0.7486	4	-3.31	Non-neutral	3	78%	32.7	[Beltran-Valero de Bernabe et al., 1998]
	8	p.P158R	-2	2.905	Align	0.00	0.6465	2	-3.32	Non-neutral	4	82%		[Phornphutkul et al., 2002]
		p.Q159H	0	1.344	Align	0.00	0.2005	5	-1.7	Non-neutral	1	63%		[Phornphutkul et al., 2002]
		p.K171N	0	0.596	Align	0.31	0.4522	0	-1.88	Neutral	4	85%		[Phornphutkul et al., 2002]
		p.E178D	2	2	Align	0.01	0.0481	9	-1.41	Non-neutral	0	58%		[Phornphutkul et al., 2002]

	Exon	Variants	Blossum 62	POLYPHEN		SIFT	PMUT		PANTHER	SNAP			Enzyme Activity [b] (%wt)	Publication
				PSIC score	Exclusion [a]	Score	NN score	Reliability	subPSEC	Prediction	Reliability	Accuracy		
		p.V181F	-1	2.23	Align	0.00	0.5987	1	-3.35	Non-neutral	2	70%		[Rodriguez et al., 2000]
	9	p.S189I	-2	1.836	Align	0.02	0.375	2	-2.97	Non-neutral	1	63%	3.4	[Beltran-Valero de Bernabe et al., 1998]
		p.G198D	-1	2.51	Align	0.00	0.8432	6	-3.91	Non-neutral	4	82%		[Mannoni et al., 2004]
		p.I216T	-1	1.762	Align	0.00	0.4444	1	-3.02	Non-neutral	4	82%	0	[Beltran-Valero de Bernabe et al., 1998]
	10	p.R225H	0	2.639	Align	0.00	0.5247	0	-3.82	Non-neutral	5	87%	0.1	[Beltran-Valero de Bernabe et al., 1998]
		p.F227S	-2	3.29	Align	0.00	0.5709	1	-3.35	Non-neutral	4	82%	0.1	[Beltran-Valero de Bernabe et al., 1998]
		p.P230T	-1	2.986	Align	0.00	0.4558	0	-3.41	Non-neutral	4	82%	1	[Beltran-Valero de Bernabe et al., 1999a]
		p.K248E	1	2.246	Align	0.00	0.3299	3	-3.69	Non-neutral	5	87%		[Porfirio et al., 2000]
	11	p.D291E	2	2.114	Align	0.00	0.0428	9	-2.76	Non-neutral	5	87%	0	[Beltran-Valero de Bernabe et al., 1999a]
		p.H292R	0	3.419	Align	0.00	0.7496	4	-3.6	Non-neutral	5	87%		[Rodriguez et al., 2000]
	12	p.W322R	-3	4.487	Align	0.00	0.9666	9	-4.31	Non-neutral	4	82%		[Rodriguez et al., 2000]
		p.R330S	-1	2.83	Align	0.00	0.7325	4	-3.07	Non-neutral	4	82%	0.4	[Beltran-Valero de Bernabe et al., 1999b]
	13	p.H371R	0	3.419	Struct	0.00	0.6099	2	-4.15	Non-neutral	6	93%	0	[Beltran-Valero de Bernabe et al., 1999b]

Effect on the protein [c]: Benign | Ambiguous | Deleterious

Reliability score [c]: Poor | Medium | Good

Accuracy score [c]: Poor | Medium | Good

[a] Prediction of the effect on the protein base on multiple alignment (Align) or on the structure of the protein (Struct)

[b] From Rodriguez et al., 2000

[c] The color classifications are arbitrary

Supp. Table S2: Splice site effect prediction analysis

	Mutations	Splicing score [a]			Published
		BDGP	NetGene2		
		Score	NN score	Confidence	
Variants found in this study (Novel and previously published)	c.16-1G>A	0.8>0	0.138>0	0.204>0	[Beltran-Valero de Bernabe et al., 1999a]
	c.342+1G>T	0.98>0 0>0.58 [b]	0.465>0	0.833>0	[Beltran-Valero de Bernabe et al., 1998]
	c.342+1G>A	0.98>0 0>0.75 [b]	0.465>0	0.833>0	[Muller et al., 1999]
	c.468+2T>C	0.93>0	0.818>0.042	0.99>0.026	[Phornphutkul et al., 2002]
	c.549+1G>A	0.7>0	0.901>0	0.889>0	
	c.1006+2T>A	0.99>0 0>0.69 [b]	0.282>0	0.747>0	
	c.1007-2A>T	0.82>0	0.639>0	0.415>0	
Other variants previously published	c.177-2G>A	NC [a]	NC [a]	NC [a]	[Phornphutkul et al., 2002]
	c.468+5G>A	0.93>0	0.818>0.079	0.99>0.125	[Porfirio et al., 2000]
	c.550-2A>C	0.91>0	0.776>0	0.39>0	[Phornphutkul et al., 2002]
	c.650-56G>A	0.84>0.84	0.662>0.662	0.871>0.855	[Beltran-Valero de Bernabe et al., 1998]
	c.650-17G>A	0.84>0.84	0.662>0.635	0.871>0.855	[Beltran-Valero de Bernabe et al., 1998]
	c.1188+1G>T	1>0 0>0.52 [b]	1>0	0.997>0	[Rodriguez et al., 2000]

Effect on the splice site [d]		
Benign	Ambiguous	Deleterious

Confidence score variation [d]		
Little	Medium	Large

[a] Wild type score > Variant score
[b] Creation of a new splice site
[c] NC: not calculable, published nucleotide change inconsistent with gene sequence
[d] The color classifications are arbitrary

Supp. Table S3: Other published mutations

	Mutations		Publication
	Nucleotide modification	Protein effect	
Nonsense	c.?	p.S59X	(Phornphutkul et al., 2002)
	c.433A>T	p.R145X	(Phornphutkul et al., 2002)
	c.961C>T	p.R321X	(Rodriguez et al., 2000)
Indel [a]	c.32_33delGGinsATT	p.G11DfsX2	(Beltran-Valero de Bernabe et al., 1998)
	c.357dupG	p.C120VfsX9	(Phornphutkul et al., 2002)
	c.588delC	p.R197GfsX32	(Rodriguez et al., 2000)
	c.1016delT	p.M339RfsX30	(Phornphutkul et al., 2002)

[a] Indel: includes insertions, deletions and insertion-deletions