Supplementary Table S2. Summary of features commonly used used in bioinformatic tools to identify deleterious amino acid substitutions (AAS). Features were obtained using the SNVBox software for all datasets (CPD, DM and SNP sets). The descriptions of features were taken from the SNVBox User Manual (http://karchinlab.org/apps/snvbox/userdoc.pdf); for more details please refer to SNVBox [25].

Feature name	Feature subset	Description
AA Hydrophobicity	Amino acid features	Change in hydrophobicity as a result of the AAS
AA Charge	Amino acid features	Change in formal charge as a result of the AAS
AA Volume	Amino acid features	Change in residue volume as a result of the AAS (cubic Angstroms)
AA Polarity	Amino acid features	Polarity change as a result of the AAS
AA Matrix	Amino acid features	Amino acid substitution scores from BLOSUM 62, PAM250, EX, Venkatarajan and Braun matrix & Miyazawa-Jernigan contact energy matrix
AA Transition	Amino acid features	Frequency of transition between two neighboring amino acids based on all human proteins in SwissProt
AA Grantham score	Amino acid features	The Grantham distance from reference to mutation amino acid residue
AA Frequencies	Amino acid features	Frequency of AAS type (e.g. alanine to glycine) in HGMD (2003), HapMap (dbSNP build 129) and COSMIC (release 38)
Exon Conservation	Exonic features	Entire exon conservation computed from a 46-way genomic vertebrate alignment
Exon SNP Density	Exonic features	Number of HapMap verified SNPs in the exon where the mutation is located divided by the length of the exon
Genomic multiple sequence alignments (MSA)	Genomic MSA	Features calculated from 46-way genomic vertebrate alignments, which includes Shannon entropy and the Kullback-Leibler divergence
Protein multiple sequence alignments (MSA)	Protein MSA	Features calculated from multiple sequence alignment of diverse homologous proteins. Features computed include the Shannon entropy and Kullback-Leibler divergence

Solvent accessibility	Protein structure	Prediction that wild-type residue is buried, partially buried or exposed in terms of solvent accessibility
Secondary structure	Protein structure	Prediction that wild-type residue is helix, loop or strand
Protein stability	Protein structure	Prediction of the degree to which the wild-type residue contributes to protein stability e.g. highly stabilizing
Backbone flexibility	Protein structure	Prediction of the flexibility of the backbone of the wild-type residue
Protein composition	Regional protein composition	Features based on regional amino acid composition in a 15-amino-acid-residue window centred on the AAS
UniProt annotations of human proteins	Annotated functional sites	Includes functional sites annotated by UniProt, including binding sites (e.g. DNA, RNA, lipid, metal, carbohydrate, calcium), catalytic sites, sites of post-translational modification, localization signals, disulphide bonds, protein-protein interaction sites