SUPPLEMENTARY MATERIAL

Blind prediction of deleterious amino acid variations with SNPs&GO

Emidio Capriotti[1], Pier Luigi Martelli[2], Piero Fariselli[3] and Rita Casadio[2].*

[1] BioFolD Unit, Department of Biological, Geological, and Environmental Sciences (BiGeA), University of Bologna, Via F. Selmi 3, Bologna, 40126, Italy.

[2] Biocomputing Group, Department of Biological, Geological and Environmental Sciences (BiGeA), University of Bologna, 40126 Bologna, Italy.

[3] Department of Comparative Biomedicine and Food Science. University of Padova, Viale dell'Università, 16, 35020 Legnaro (PD), Italy.

[*]Correspondence should be addressed to E.C. (emidio.capriotti@unibo.it)

Regression evaluation measures

For all the challenges, we performed regression tests for comparing the experimental and the predicted values associated to each variant. In particular, for each predictor, we calculated the Root Mean Square Error (RMSE) after linear fitting, the Pearson's correlation coefficient ($r_{Pearson}$), the Spearman's rank correlation coefficient ($r_{Spearman}$) and the Kendall's rank correlation coefficient ($r_{KendallTau}$).

Evaluation measures for binary classifiers

In all the performance measures we considered as Positive and Negative classes the *Pathogenic* and *Benign* predictions, respectively. Thus, TP (true positives) are the correctly predicted Pathogenic Single Nucleotide Variants (SNVs), TN (true negatives) are the correctly predicted *Benign* variants, FP (false positives) are the *Benign* SNVs annotated as *Pathogenic*, and FN (false negatives) are the *Pathogenic* variants predicted to be *Benign*.

Predictor performance was evaluated using the following metrics: true positive and negative rates (*TPR, TNR* - also referred as sensitivity and specificity), positive and negative predicted values (*PPV, NPV*) and overall accuracy (Q_2)

$$\begin{aligned} Pathogenic: \; PPV &= \frac{TP}{TP+FP} \quad TPR = \frac{TP}{TP+FN} \\ Benign: \; NPV &= \frac{TN}{TN+FN} \quad TNR = \frac{TN}{TN+FP} \\ Q_2 &= \frac{TP+TN}{TP+FP+TN+FN} \end{aligned} \qquad \text{[Eq. S1]}$$

We computed the Matthew's correlation coefficient *MC* as:

$$MC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad \text{[Eq. S2]}$$

We also calculated the area under the receiver operating characteristic (ROC) curve (AUC), by plotting the True Positive Rate as a function of the False Positive Rate at different probability thresholds of annotating a variant as *Pathogenic* or *Benign*.

CHEK2 and *RAD50* datasets and their predictions

The *CHEK2* gene encodes for a serine/threonine kinase (NP_009125.1), that is part of the signaling pathways activated by DNA damage, particularly DNA double-stranded breaks (Antoni, et al., 2007). Inherited *CHEK2* variations have been found to confer increased risk of breast cancer (Meijers-Heijboer, et al., 2002). Thus, a *CHEK2* mutation-screening on a set including 1,303 female breast cancer patients and 1,109 unaffected female controls, was performed to detect new possible rare risk variants (Le Calvez-Kelm, et al., 2011). In 2010 data-providers released a set of 41 *CHEK2* variants. This dataset is composed of 32 non-synonymous Single Nucleotide Variants (SAVs), 2 double SAVs, 3 truncating variations and 4 deletions.

RAD50 (NP_005723.2) is a component of the *MRN* complex, which plays a central role in double-strand break (DSB) repair, DNA recombination, maintenance of telomere integrity and meiosis (de Jager, et al., 2001). This gene is a candidate gene for breast cancer susceptibility. CAGI data provides a list of 69 *RAD50* variants observed in about 1,400 breast cancer cases and 1,200 ethnically matched controls. The *RAD50* dataset consists of 35 missense, 15 silent, 14 intronic and 5 truncating variants.

For each single amino acid variation (SAV) in the *CHEK2* and *RAD50* datasets, we calculated the fraction of carriers in the case set (f_{case}) with respect to the total number of carriers in case (N_{case}) and control (N_{ctrl}) as follows:

$$f_{case} = \frac{N_{case}}{N_{case} + N_{ctrl}} \quad \text{[Eq. S3]}$$

The tested methods (SNPs&GO[09], SIFT and AlignGVGD) are scored performing a grid-like search to find the optimal experimental and predicted f_{case} thresholds. In this test we used an experimental f_{case} (f^{e}_{case}) threshold equal to 0.7, which represents the median of the optimal classification thresholds for the three methods. Thus, we classified SAVs as *Pathogenic* or *Benign* when f^{e}_{case} >0.7 or ≤ 0.7 respectively. As to the predicted f_{case} (f^{p}_{case}), the optimal thresholds were obtained selecting the values resulting in the highest product

among accuracy (Q_2) Area Under the ROC Curve (AUC) and Matthews Correlation coefficient (MC). For *CHEK2* challenge, the optimal f^{p}_{case} for SNPs&GO^{09}, SIFT and AlignGVGD are 0.35, 0.60 and C0 respectively. For the *RAD50* challenge, the selected f^{p}_{case} thresholds for SNPs&GO^{09}, SIFT and AlignGVGD are 0.10, 0.15 and C45 respectively. For SIFT, whose output indicates the probability of benign SAVs, pathogenic variants are selected considering scores below the optimal threshold. For AlignGVGD, the returned classes are ranked according to their level of pathogenicity from the lowest (C0) to the highest (C65).

The *p16* dataset and its prediction.

CDKN2A gene encodes for the *p16INK4A* protein (NP_000068.1), which promotes cell cycle arrest by inhibiting cyclin dependent kinase (*CDK4/6*) (Bockstaele, et al., 2006). Many *p16* variations have been shown to compromise the *CDK4* inhibitory activity of *p16* and/or its ability to block cell cycle and constitute the most common cause of familial malignant melanoma (Kannengiesser, et al., 2009). Recently, Scaini and colleagues characterized the functional consequences of ten *p16* variations (MUT-P16), which were expected to alter in vitro protein cellular functions (Scaini, et al., 2014). They measured the proliferation rate of the mutation-like cells (cases) and compared with the proliferation rate of wild-type cells (controls).

In the *p16* challenge the evaluation of the binary classifier performance was calculated considering an experimental Proliferation Rate threshold equal to 75% for discriminating between pathogenic and benign variants. This value is the threshold suggested by the CAGI data providers and assessors.

The *NAGLU* dataset and its prediction

NAGLU (NP_000254.2) is a lysosomal enzyme that hydrolyzes N-acetyl D-glucosamine from the non-reducing end of heparan sulfate. *NAGLU* deficiency causes the rare neurodegenerative disorder referred as Sanfilippo type B syndrome (Valstar, et al., 2008). To develop a new enzyme replacement therapy, BioMarin (http://www.biomarin.com/) functionally assessed the enzymatic activity of 165 novel SAVs (MUT-NAGLU) from the ExAC database (http://exac.broadinstitute.org/). For each SAV the experimental study determined the change in enzymatic activity of the mutated proteins transfected into HEK293 cells, normalized by the activity in wild-type cells.

The predictions of *NAGLU* challenge were evaluated as a binary classification task, by optimizing the Relative Activity (RelAct) thresholds using a grid-like search. The optimal

thresholds are obtained selecting the values resulting in the highest cube root of the product among accuracy (Q_2), Area Under the ROC Curve (AUC) and Matthews Correlation coefficient (MC). The best performance for MutPred2*, MutPred2, SNPs&GO[13] and SNPs&GO[09] are obtained setting thresholds on the experimental Relative Activity equal to 0.34, 0.55, 0.28 and 0.28 respectively. The classification threshold for the output of all the predictors is set to 0.5.

Supplementary Tables

Table S1. Predictions for the *CHEK2* single amino acid variation (SAV) dataset.

SAV	N_{case}	N_{ctrl}	f^{e}_{case}	SNPs&GO[09]	SIFT	GV	GD	AlignGVGD
p.Ser5Leu	2	1	0.67	0.15	0.00	129	72	C0
p.Val25Ala	1	0	1.00	0.10	0.24	354	0	C0
p.Pro85Leu	1	3	0.25	0.10	0.04	354	0	C0
p.Arg117Gly	3	1	0.75	0.85	0.00	0	125	C65
p.Arg137Gln	1	0	1.00	0.25	0.17	99	0	C0
p.Ile157Thr	2	1	0.67	0.50	0.07	50	70	C15
p.Arg180Cys	3	0	1.00	0.60	0.06	86	137	C25
p.Arg180His	0	1	0.00	0.40	0.13	86	0	C0
p.Ser192Leu	1	0	1.00	0.40	0.14	99	95	C15
p.Ile221Met	1	0	1.00	0.35	0.22	94	0	C0
p.Ala230Ser	0	1	0.00	0.25	0.81	65	99	C15
p.Glu239Lys	2	0	1.00	0.50	0.50	29	52	C15
p.Cys243Arg	0	1	0.00	0.65	0.22	204	81	C0
p.Ile251Phe	0	1	0.00	0.80	0.02	0	21	C15
p.Met304Thr	1	0	1.00	0.60	0.20	29	70	C25
p.Gly306Ala	1	0	1.00	0.90	0.05	55	53	C0
p.Asp311Asn	1	0	1.00	0.45	0.58	113	0	C0
p.Thr323Pro	1	0	1.00	0.75	0.27	89	29	C0
p.Arg346Cys	3	0	1.00	0.90	0.00	0	180	C65
p.Asn352Tyr	1	0	1.00	0.95	0.00	0	142	C65
p.His371Tyr	2	1	0.67	0.35	0.19	112	25	C0
p.Arg406Cys	1	0	1.00	0.80	0.04	124	109	C15
p.Phe418Cys	1	0	1.00	0.90	0.01	22	192	C55
p.Pro426Ser	1	0	1.00	0.90	0.00	0	73	C65
p.Asp438Gly	1	0	1.00	0.40	0.03	49	68	C15
p.Asp438Tyr	2	2	0.50	0.55	0.01	49	114	C25
p.Asn446Asp	0	1	0.00	0.15	0.62	129	23	C0
p.Ile448Ser	7	2	0.78	0.40	0.88	95	73	C15
p.Thr476Met	1	0	1.00	0.70	0.01	58	81	C15
p.Pro484Leu	1	0	1.00	0.65	0.03	0	98	C65
p.Leu512Val	1	0	1.00	0.05	0.30	245	0	C0
p.Arg519Leu	0	1	0.00	0.30	0.06	99	25	C0

CHEK2 RefSeq ID: NP_009125.1. The amino acid variation numbering system used for the datasets is based on the starting residue as residue 1. N_{case}, N_{ctrl} are the number of carriers in the case and control sets respectively. f_{case} is the fraction of carriers in the case set with respect to the total number of carriers in case and control sets. SNPs&GO[09] and SIFT columns contain the raw probability returned by the methods. AlignGVGD is the output of the method where GV is the Grantham Variation and GD the Grantham Deviation.

Table S2. Predictions for the *RAD50* single amino acid variation (SAV) dataset.

SAV	N_{case}	N_{ctrl}	f^{e}_{case}	SNPs&GO[09]	SIFT	GV	GD	AlignGVGD
p.Asn38Ser*	1	0	1.00	0.70	0.00	0	46	C45
p.Leu84Val*	1	0	1.00	0.35	0.00	0	31	C25
p.Arg87His*	1	0	1.00	0.40	0.04	26	23	C0
p.Ile94Leu*	11	9	0.55	0.10	0.16	31	0	C0
p.Val127Ile*	2	2	0.50	0.05	0.00	84	19	C0
p.Ala171Ser	3	0	1.00	0.05	0.72	91	36	C0
p.Thr191Ile	3	3	0.50	0.20	0.17	58	69	C15
p.Gln199His	1	0	1.00	0.15	0.18	0	24	C15
p.Arg224His	5	4	0.56	0.10	0.00	124	0	C0
p.Glu239Gln	0	1	0.00	0.05	0.41	113	2	C0
p.Leu262His	1	0	1.00	0.15	0.00	145	19	C0
p.Val315Leu	3	3	0.50	0.05	0.43	29	5	C0
p.Arg327His	4	5	0.44	0.05	0.20	125	5	C0
p.Arg365Gln	1	0	1.00	0.05	0.60	124	0	C0
p.Gln426Arg	0	1	0.00	0.10	0.37	0	43	C35
p.Lys446Glu	2	0	1.00	0.15	0.00	0	57	C55
p.Arg486Cys	1	0	1.00	0.10	0.11	105	150	C25
p.Ser557Cys	0	1	0.00	0.10	0.05	170	91	C0
p.Asp637Glu†	0	1	0.00	0.05	0.58	0	45	C35
p.Val683Ile†	1	0	1.00	0.05	0.64	31	0	C0
p.Arg725Trp†	3	0	1.00	0.10	0.05	26	96	C35
p.Ile761Met	0	1	0.00	0.00	0.05	21	10	C0
p.Arg763His	1	1	0.50	0.10	0.18	57	0	C0
p.Gln799His	1	2	0.33	0.10	0.05	116	0	C0
p.Val842Ala	1	0	1.00	0.05	0.07	70	30	C0
p.Thr917Ile	0	2	0.00	0.05	0.11	104	69	C0
p.Asp946Val	1	0	1.00	0.05	0.12	134	36	C0
p.Gly1080Asp	1	0	1.00	0.10	0.30	98	34	C0
p.His1087Arg	0	1	0.00	0.05	0.66	29	0	C0
p.Arg1093Gln	0	1	0.00	0.00	0.41	88	0	C0
p.Tyr1104Cys	0	1	0.00	0.30	0.18	83	155	C25
p.Arg1166Trp*	1	0	1.00	0.15	0.03	180	59	C0
p.Leu1264Phe*	2	1	0.67	0.10	0.00	0	22	C15
p.Arg1279His*	2	0	1.00	0.10	0.05	29	0	C0
p.Lys1301Arg	3	0	1.00	0.00	0.87	82	0	C0

RAD50 RefSeq ID: NP_005723.2. The amino acid variation numbering system used for the datasets is based on the starting residue as residue 1. N_{case}, N_{ctrl} are the number of carriers in the case and control sets respectively. f_{case} is the fraction of carriers in the case set with respect to the total number of carriers in case and control sets. SNPs&GO[09] and SIFT columns contain the raw probability returned by the methods. AlignGVGD is the output of the method where GV is the Grantham Variation and GD the Grantham Deviation. †Amino acid variations in Zn hook domain. *Amino acid variations in P-loop hydrolase domain.

Table S3. Predictions for the *p16INK4A* single amino acid variation (SAV) dataset.

SAV	RelPro	StDev	SNPs&GO[13]	DrCancer	SPARK-LAB
p.Gly23Ala	54	9	0.904	0.873	0.700
p.Gly23Cys	87	13	0.946	0.893	0.707
p.Gly23Arg	92	15	0.932	0.885	0.991
p.Gly23Ser	69	5	0.923	0.821	0.738
p.Gly23Val	90	10	0.940	0.892	0.936
p.Gly35Glu	60	11	0.590	0.805	0.666
p.Gly35Arg	54	2	0.618	0.802	0.642
p.Gly35Trp	86	9	0.841	0.796	0.807
p.Leu65Pro	66	10	0.878	0.740	0.739
p.Leu94Pro	94	14	0.979	0.930	0.913

p16INK4A RefSeq ID: NP_000068.1. The amino acid variation numbering system used for the datasets is based on the starting residue as residue 1. RelPro is the Relative Proliferation rate of mutation-like cells with respect to the wild-type cells and StDev is its standard deviation.

Table S4. Predictions for the *NAGLU* single amino acid variation (SAV) dataset.

SAV	RelAct	StDev	SNPs&GO[13]	SNPs&GO[09]	MutPred2*	MutPred2
p.Ala16Val	0.42	0.01	0.917	1	0.826	0.633
p.Pro118Ser	0.50	0.07	0.906	1	0.777	0.538
p.Tyr131Ser	0.04	0.02	0.123	0	0.090	0.084
p.Val135Gly	0.21	0.05	0.201	0	0.124	0.181
p.Thr137Met	0.25	0.06	0.068	0	0.203	0.126
p.Ser141Thr	0.34	0.03	0.780	1	0.145	0.124
p.Arg152Gln	0.35	0.07	0.635	1	0.749	0.325
p.Ile154Thr	0.04	0.03	0.199	0	0.098	0.145
p.Ala158Ser	0.10	0.03	0.229	0	0.210	0.202
p.Ser169Asn	0.76	0.09	0.723	1	0.710	0.405
p.Ala181Asp	0.54	0.03	0.775	0	0.585	0.489
p.Gly183Ala	0.11	0.04	0.280	0	0.150	0.241
p.Asn190Ser	0.69	0.08	0.926	1	0.820	0.351
p.Phe192Leu	0.28	0.03	0.357	0	0.340	0.196
p.Phe198Tyr	0.44	0.03	0.681	0	0.250	0.088
p.Gly202Arg	0.40	0.11	0.550	0	0.239	0.333
p.Met204Ile	0.23	0.06	0.120	0	0.244	0.098
p.Asp211Asn	0.70	0.07	0.645	1	0.594	0.492
p.Pro216Arg	0.57	0.03	0.780	1	0.939	0.521
p.Pro216Ser	0.67	0.06	0.843	1	0.917	0.511
p.Gln222Glu	0.47	0.03	0.458	0	0.661	0.123
p.Tyr224Cys	0.34	0.02	0.764	1	0.736	0.379
p.Arg228Gln	0.53	0.06	0.864	1	0.893	0.349
p.Arg228Trp	0.16	0.01	0.435	0	0.779	0.253
p.Gly237Asp	0.06	0.03	0.501	0	0.123	0.084
p.Pro243Ser	0.21	0.05	0.213	0	0.203	0.085
p.Ala244Thr	0.20	0.03	0.670	1	0.466	0.200
p.Glu251Lys	0.18	0.09	0.829	1	0.954	0.537
p.Val253Ile	1.13	0.04	0.946	1	0.966	0.581
p.Thr254Ile	0.67	0.06	0.678	1	0.736	0.513
p.Val260Ile	0.69	0.03	0.765	1	0.893	0.347
p.Asn261Ser	0.72	0.03	0.703	1	0.879	0.371
p.Thr263Met	0.93	0.16	0.839	1	0.612	0.372
p.Gly269Asp	0.89	0.15	0.461	0	0.453	0.374
p.Cys273Arg	0.19	0.14	0.570	0	0.161	0.296
p.Ser276Phe	0.07	0.04	0.436	0	0.118	0.246
p.Leu281Pro	0.06	0.03	0.196	0	0.038	0.196
p.Pro283Leu	1.19	0.06	0.204	0	0.286	0.335
p.Ile290Val	0.21	0.04	0.922	1	0.969	0.604
p.Ile291Leu	1.08	0.20	0.603	1	0.865	0.299
p.Leu294Phe	0.96	0.09	0.894	1	0.757	0.353
p.Arg297Gln	1.05	0.17	0.908	1	0.877	0.401
p.Ile300Thr	0.92	0.24	0.919	1	0.969	0.410

SAV	RelAct	StDev	SNPs&GO[13]	SNPs&GO[09]	MutPred2*	MutPred2
p.Asp306Gly	0.58	0.06	0.462	0	0.405	0.429
p.Ile308Thr	0.34	0.06	0.343	0	0.489	0.316
p.Gly310Ala	0.68	0.11	0.926	1	0.647	0.384
p.Gly310Val	0.06	0.03	0.719	0	0.553	0.376
p.Ser322Leu	0.74	0.12	0.816	1	0.656	0.512
p.Leu327Phe	0.09	0.05	0.756	1	0.322	0.248
p.Ala329Thr	0.87	0.15	0.874	1	0.816	0.443
p.Thr332Ile	0.81	0.02	0.688	0	0.860	0.506
p.Val334Ile	0.76	0.09	0.976	1	0.870	0.464
p.Met338Ile	0.59	0.05	0.691	1	0.391	0.297
p.Thr339Pro	0.05	0.02	0.427	0	0.667	0.526
p.Ala340Thr	0.95	0.01	0.887	1	0.902	0.580
p.Leu349Phe	0.08	0.03	0.527	1	0.360	0.259
p.Gly351Ala	0.64	0.10	0.747	1	0.161	0.144
p.Phe354Ser	0.07	0.01	0.228	0	0.154	0.129
p.Gln355Lys	0.65	0.11	0.600	1	0.543	0.394
p.Trp361Ser	0.07	0.03	0.128	0	0.159	0.195
p.Ile366Met	0.59	0.10	0.690	1	0.878	0.338
p.Val369Leu	0.14	0.04	0.895	1	0.682	0.426
p.Pro374Ser	0.77	0.13	0.307	0	0.258	0.136
p.Arg375Cys	0.25	0.02	0.537	0	0.719	0.332
p.Arg375His	0.40	0.06	0.637	1	0.879	0.329
p.Arg377Cys	0.15	0.05	0.345	0	0.539	0.286
p.Arg377His	0.06	0.04	0.594	0	0.751	0.279
p.Ala385Val	0.03	0.02	0.438	0	0.423	0.316
p.Glu386Lys	0.03	0.02	0.165	0	0.236	0.257
p.Arg393His	0.85	0.15	0.839	1	0.879	0.408
p.Gln398Lys	0.50	0.02	0.573	0	0.612	0.399
p.Gly399Val	0.05	0.03	0.109	0	0.087	0.110
p.Trp404Cys	0.02	0.02	0.106	0	0.045	0.148
p.His408Tyr	0.04	0.02	0.211	0	0.174	0.126
p.Glu421Lys	0.33	0.06	0.720	1	0.259	0.343
p.Gly426Ala	0.09	0.02	0.666	1	0.378	0.279
p.Arg431Cys	0.05	0.02	0.439	0	0.495	0.348
p.Asn435Thr	0.30	0.05	0.785	1	0.259	0.250
p.Thr437Ser	0.96	0.07	0.903	1	0.721	0.243
p.Thr441Met	0.49	0.05	0.826	1	0.555	0.347
p.Ala444Asp	0.04	0.01	0.425	1	0.317	0.284
p.Ala444Val	0.93	0.07	0.807	1	0.666	0.261
p.Pro445Ser	0.04	0.01	0.551	1	0.203	0.216
p.Ser449Asn	0.21	0.07	0.920	1	0.739	0.553
p.Val454Phe	0.04	0.02	0.446	0	0.254	0.245
p.Tyr455Ser	0.03	0.02	0.126	0	0.079	0.113
p.Ser456Phe	0.23	0.04	0.389	0	0.210	0.285
p.Gly462Arg	0.03	0.03	0.370	0	0.229	0.407

SAV	RelAct	StDev	SNPs&GO[13]	SNPs&GO[09]	MutPred2*	MutPred2
p.Arg464Gln	0.03	0.02	0.737	1	0.603	0.348
p.Ser477Thr	1.01	0.08	0.793	1	0.884	0.273
p.Ala479Thr	0.20	0.04	0.482	1	0.600	0.179
p.Ala480Thr	0.43	0.08	0.865	1	0.903	0.595
p.Arg481Trp	0.09	0.01	0.204	0	0.344	0.251
p.Pro488Leu	0.39	0.08	0.646	1	0.749	0.380
p.Ala490Thr	0.58	0.06	0.698	1	0.788	0.440
p.Gly491Ala	0.67	0.10	0.856	1	0.843	0.621
p.Ala493Val	0.18	0.03	0.323	0	0.173	0.093
p.Leu496Pro	0.05	0.03	0.128	0	0.201	0.197
p.Arg499Gln	0.53	0.11	0.800	0	0.748	0.398
p.Arg499Trp	0.25	0.02	0.394	0	0.518	0.314
p.Asn503Ser	0.33	0.04	0.774	0	0.351	0.215
p.Gly506Glu	0.55	0.06	0.789	0	0.549	0.364
p.Cys509Tyr	0.37	0.03	0.746	0	0.624	0.622
p.Asn513Ser	0.05	0.02	0.923	1	0.524	0.564
p.Arg514Cys	0.51	0.08	0.452	0	0.548	0.400
p.Arg514His	0.96	0.11	0.635	1	0.847	0.422
p.Ser522Phe	0.19	0.07	0.252	0	0.131	0.326
p.Gln524Arg	0.86	0.09	0.793	1	0.618	0.364
p.Arg533Gln	0.56	0.09	0.725	0	0.825	0.564
p.Val536Met	0.45	0.11	0.485	0	0.504	0.265
p.Arg541Gln	0.40	0.06	0.738	1	0.871	0.480
p.Leu542Pro	0.05	0.05	0.185	0	0.338	0.227
p.Leu542Arg	0.47	0.07	0.258	0	0.461	0.224
p.Ala547Thr	0.57	0.09	0.826	1	0.919	0.504
p.Ala555Thr	0.97	0.09	0.972	1	0.904	0.494
p.Arg557Cys	0.48	0.09	0.348	0	0.603	0.232
p.Arg557Leu	0.48	0.05	0.367	0	0.656	0.282
p.Asp559His	0.07	0.01	0.169	0	0.077	0.093
p.Asp559Asn	0.04	0.03	0.194	0	0.140	0.087
p.Leu563His	0.07	0.02	0.324	0	0.270	0.227
p.Val568Gly	0.13	0.03	0.406	0	0.388	0.183
p.Ser573Gly	0.57	0.03	0.772	1	0.762	0.435
p.Tyr575Cys	0.52	0.03	0.854	0	0.574	0.458
p.Gly595Arg	0.82	0.11	0.830	0	0.621	0.408
p.Gly596Cys	1.02	0.07	0.622	0	0.399	0.371
p.Val597Ile	1.02	0.13	0.936	1	0.968	0.597
p.Val597Leu	0.66	0.13	0.910	1	0.904	0.629
p.Glu608Lys	0.44	0.08	0.867	1	0.897	0.547
p.Arg615Cys	0.17	0.05	0.735	1	0.748	0.391
p.Ser620Gly	0.93	0.07	0.831	1	0.849	0.460
p.Arg626Gly	0.26	0.08	0.381	0	0.523	0.214
p.Arg626Gln	0.29	0.05	0.586	1	0.801	0.227
p.Ala627Val	0.08	0.00	0.868	1	0.873	0.401

SAV	RelAct	StDev	SNPs&GO[13]	SNPs&GO[09]	MutPred2*	MutPred2
p.Ala628Val	0.87	0.04	0.929	1	0.862	0.426
p.Val630Gly	1.11	0.09	0.791	0	0.756	0.485
p.Val630Ile	0.77	0.07	0.896	1	0.944	0.558
p.Ala633Gly	0.71	0.12	0.911	1	0.841	0.502
p.Asp636Asn	0.69	0.09	0.920	1	0.884	0.646
p.Gln640Arg	0.79	0.13	0.824	0	0.648	0.472
p.Gln645Pro	0.05	0.04	0.130	0	0.054	0.136
p.Trp649Ser	0.03	0.02	0.081	0	0.037	0.117
p.Pro673Ser	0.25	0.05	0.444	0	0.497	0.304
p.Arg676Trp	0.26	0.05	0.222	0	0.735	0.233
p.Ala681Val	0.86	0.05	0.808	1	0.858	0.508
p.Val686Ala	0.44	0.12	0.907	1	0.811	0.298
p.Gln688Arg	0.67	0.06	0.916	1	0.922	0.676
p.His695Tyr	0.45	0.12	0.814	1	0.839	0.494
p.Gln703His	0.46	0.11	0.901	1	0.813	0.492
p.Val709Ile	0.62	0.20	0.942	1	0.929	0.386
p.Leu710Val	0.94	0.15	0.906	1	0.913	0.668
p.Arg714Met	0.56	0.11	0.889	1	0.853	0.476
p.Gln718His	0.81	0.09	0.776	1	0.839	0.528
p.Asp722Asn	0.93	0.13	0.865	1	0.725	0.333
p.Lys729Gln	0.90	0.12	0.861	1	0.849	0.432
p.Lys729Arg	0.08	0.03	0.876	1	0.943	0.471
p.Phe731Val	0.76	0.12	0.407	0	0.435	0.218
p.Pro736Leu	0.73	0.15	0.377	0	0.688	0.514
p.Arg737Cys	0.46	0.06	0.765	1	0.723	0.573
p.Arg737Gly	1.08	0.04	0.761	1.5	0.690	0.635
p.Arg737His	0.66	0.09	0.791	1	0.862	0.656
p.Arg737Ser	0.69	0.09	0.655	1	0.744	0.591
p.Ala740Pro	0.59	0.14	0.652	0	0.733	0.645
p.Gly741Ser	0.94	0.10	0.606	0	0.836	0.687

NAGLU RefSeq ID: NP_000254.2. The amino acid variation numbering system used for the datasets is based on the starting residue as residue 1. RelAct is the relative enzymatic activity with respect to the wild-type protein and StDev is its standard deviation. SNPs&GO[13] and SNPs&GO[09] are the predictions calculated using the BioFolD server and submitted by the Bologna Biocomputing Group respectively. MutPred2 and MutPred2* are the predictions obtained running the MutPred2 in default mode and without gene-level homology count features respectively.

Table S5. Confusion matrices for the CAGI binary predictions

CAGI Challenge	Method	ET	PT	TN	FP	FN	TP	N
CHEK2	SNPs&GO[09]	0.70	0.35	6	5	4	17	32
	SIFT	0.70	0.60	2	9	1	20	32
	AlignGVGD	0.70	C0	7	4	7	14	32
RAD50	SNPs&GO[09]	0.70	0.10	16	2	10	7	35
	SNPs&GO[09]*	0.70	0.10	4	0	3	4	10
	SIFT	0.70	0.15	11	7	6	11	35
	AlignGVGD	0.70	C35	18	0	15	2	35
p16	SPARK-LAB	75	0.75	16	2	10	7	10
	SNPs&GO13	75	0.75	2	3	0	5	10
	DrCancer	75	0.75	5	0	1	4	10
NAGLU	MutPred2†	0.34	0.50	48	21	11	83	163
	MutPred2	0.55	0.50	86	11	41	25	163
	SNPs&GO[13]	0.28	0.50	43	19	13	8	163
	SNPs&GO[09]	0.28	0.50	46	16	30	71	163

ET: Experimental data threshold for binary classification. PT: Prediction output threshold for binary classification. TP: True Positive. FN: False Negative. FP: False Positive. TN: True Negative. N: Number of amino acid variation in the dataset. *Subset of amino acid variations in Zn hook and P-loop hydrolase domains. † Version of MutPred2 running without gene-level homology features.

References

Antoni L, Sodha N, Collins I, Garrett MD. 2007. CHK2 kinase: cancer susceptibility and cancer therapy - two sides of the same coin? Nat Rev Cancer 7(12):925-36.

Bockstaele L, Kooken H, Libert F, Paternot S, Dumont JE, de Launoit Y, Roger PP, Coulonval K. 2006. Regulated activating Thr172 phosphorylation of cyclin-dependent kinase 4(CDK4): its relationship with cyclins and CDK "inhibitors". Mol Cell Biol 26(13):5070-85.

de Jager M, van Noort J, van Gent DC, Dekker C, Kanaar R, Wyman C. 2001. Human Rad50/Mre11 is a flexible complex that can tether DNA ends. Mol Cell 8(5):1129-35.

Kannengiesser C, Brookes S, del Arroyo AG, Pham D, Bombled J, Barrois M, Mauffret O, Avril MF, Chompret A, Lenoir GM and others. 2009. Functional, structural, and genetic evaluation of 20 CDKN2A germ line mutations identified in melanoma-prone families or patients. Hum Mutat 30(4):564-74.

Le Calvez-Kelm F, Lesueur F, Damiola F, Vallee M, Voegele C, Babikyan D, Durand G, Forey N, McKay-Chopin S, Robinot N and others. 2011. Rare, evolutionarily unlikely missense substitutions in CHEK2 contribute to breast cancer susceptibility: results from a breast cancer family registry case-control mutation-screening study. Breast Cancer Res 13(1):R6.

Meijers-Heijboer H, van den Ouweland A, Klijn J, Wasielewski M, de Snoo A, Oldenburg R, Hollestelle A, Houben M, Crepin E, van Veghel-Plandsoen M and others. 2002. Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. Nat Genet 31(1):55-9.

Scaini MC, Minervini G, Elefanti L, Ghiorzo P, Pastorino L, Tognazzo S, Agata S, Quaggio M, Zullato D, Bianchi-Scarra G and others. 2014. CDKN2A unclassified variants in familial malignant melanoma: combining functional and computational approaches for their assessment. Hum Mutat 35(7):828-40.

Valstar MJ, Ruijter GJ, van Diggelen OP, Poorthuis BJ, Wijburg FA. 2008. Sanfilippo syndrome: a mini-review. J Inherit Metab Dis 31(2):240-52.