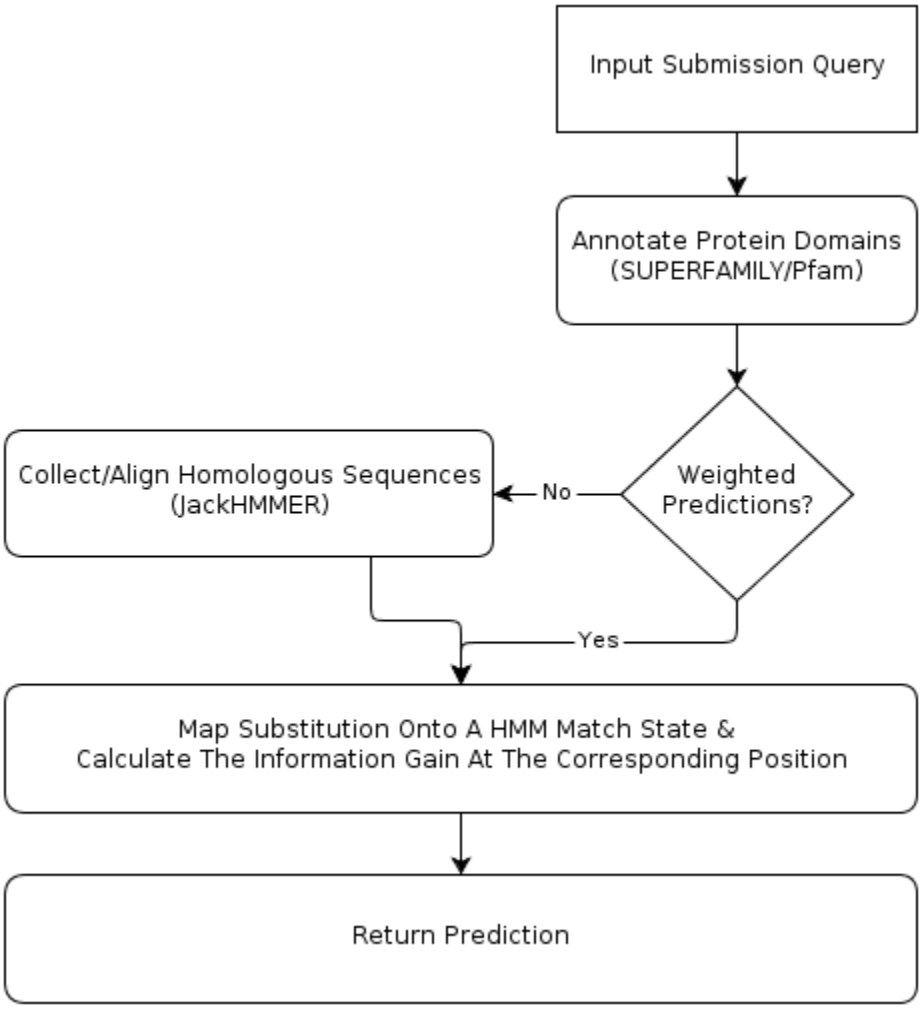
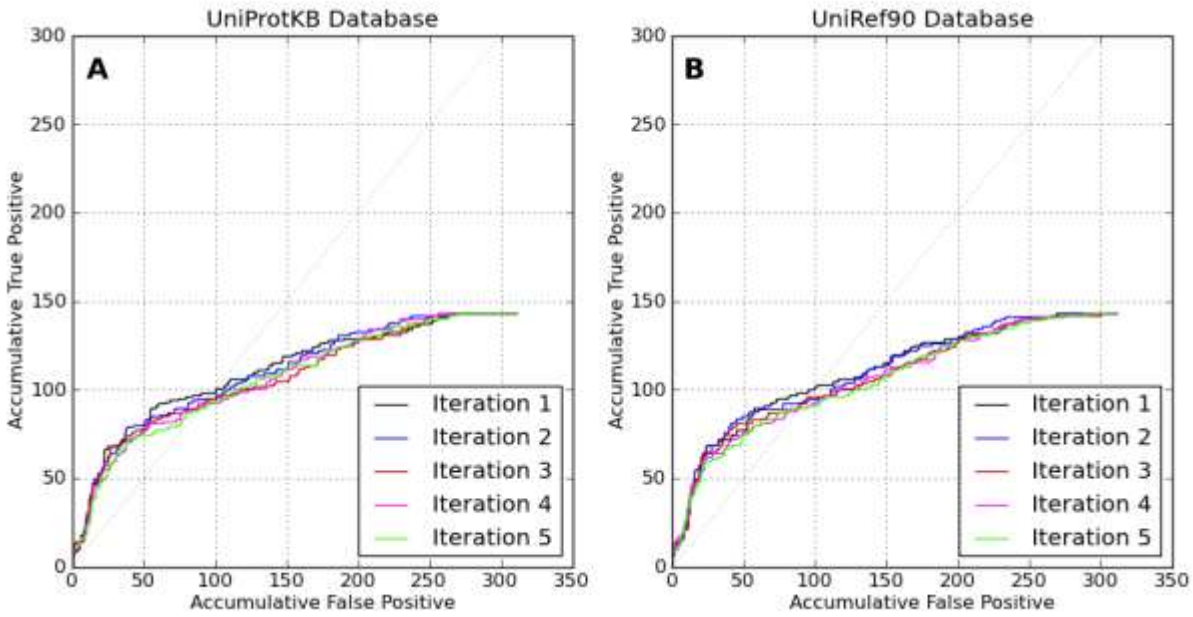
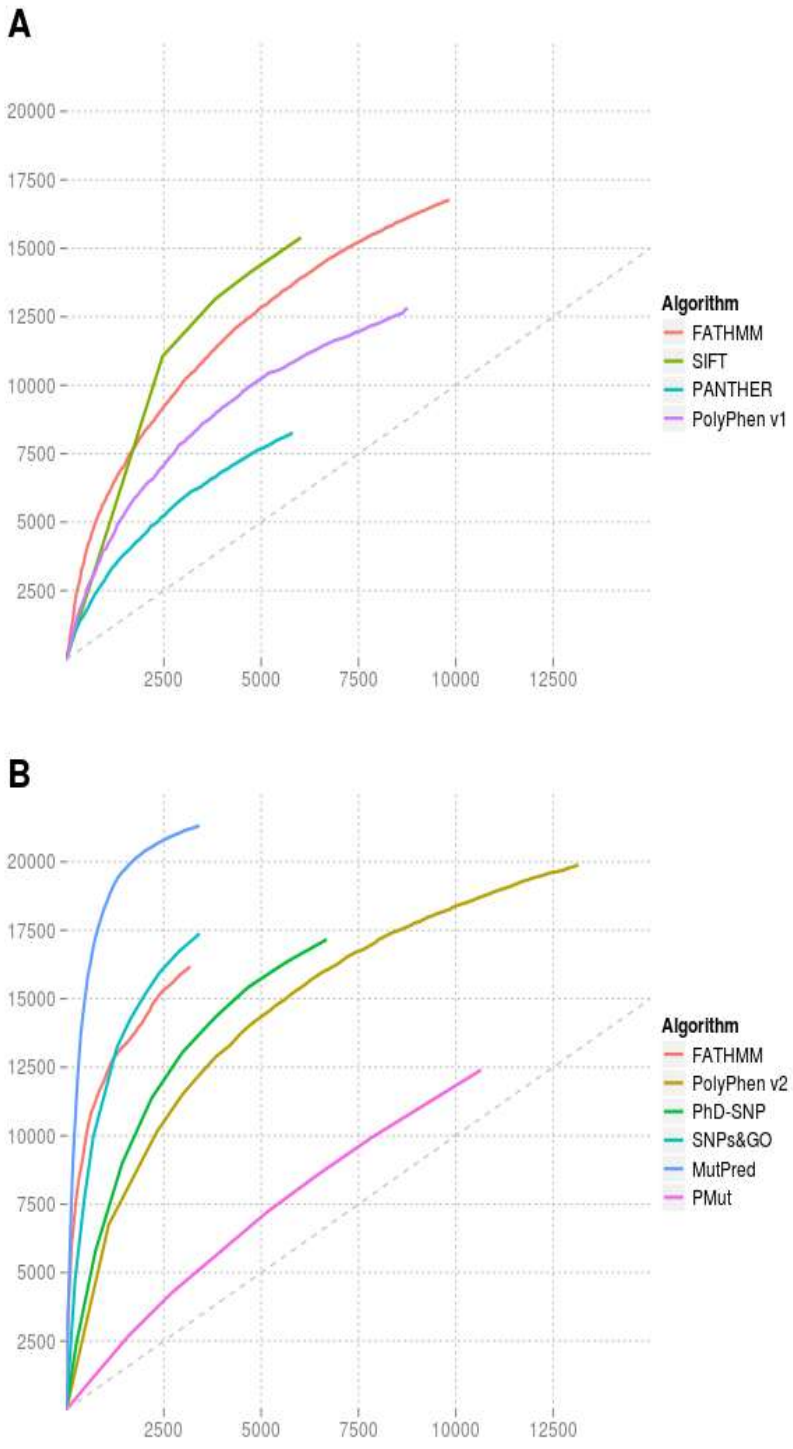


Supp. Figure S1. Scoring the Magnitude of Effect of Amino Acid Substitutions. The expected input for an unweighted prediction is the protein sequence and substitution whereas the expected input for a weighted prediction is the SwissProt/TrEMBL protein ID and substitution. Next, protein domain annotations from the SUPERFAMILY and Pfam databases are made. In addition, if an unweighted prediction is requested, an *ab initio* HMM is built from the alignment of homologous sequences collected as part of the *JackHMMER* algorithm. The amino acid substitution is then mapped onto the corresponding HMM match states where the information gain (as measured by the Kullback-Leibler divergence from the SwissProt/TrEMBL amino acid composition) is then calculated. This is then used to deduce the most informative HMM and a prediction is made accordingly.



Supp. Figure S2. In this study, we interrogate the amino acid conservation within homologous (both orthologous and paralogous) sequences using HMMs. The *JackHMMER* algorithm takes a query sequence and iteratively searches a sequence database for homologous sequences (akin to the PSI-BLAST algorithm). In order to establish the optimal parameters for the *JackHMMER* algorithm, i.e. the sequence database to search (SwissProt/TrEMBL or UniRef90) and the number of iterations required, we randomly sampled 100 proteins from the SwissVar dataset (involving 143 disease-causing and 311 functionally neutral AASs) and compared the performance of FATHMM at each *JackHMMER* iteration using Equation 1 (Figures A & B). In general, we observed no significant improvements in the performance of FATHMM at later iterations in either sequence database, indicating that a single iteration is sufficient when searching for homologous sequences for the computational prediction of the functional effects of amino acid substitutions. Therefore, the final version of FATHMM implements one *JackHMMER* iteration across the UniRef90.



Supp. Figure S3. A Receiver Operating Characteristic (ROC) curve showing the accumulated true positives plotted against the accumulated false positives for all unweighted (A) and weighted (B) computational prediction algorithms evaluated using the SwissVar benchmark dataset. Here, the "HumVar" and "Profile" versions of PolyPhen v2 and PhD-SNP are plotted as they performed best (in terms of performance accuracy).

Supp. Table S1. Mutation Submission Procedure for Computational Prediction Methods

Prediction Method	Methodology
SIFT	Local Installation
PolyPhen 1	Automatic Web Submission/Scraping
PolyPhen 2	Batch Submission
PANTHER	Local Installation
PhD-SNP	Local Installation
PMut	Author Request
SNPs&GO	Automatic Web Submission/Scraping
MutPred	Author Request

For methods without a batch submission/download facility, we developed custom web-scraping scripts in the Python programming language (available upon request) which submitted the mutations, one at a time, and parsed the predictions. For prediction methods where this was not possible, e.g. MutPred, the authors of the method kindly processed our mutation dataset.

Supp. Table S2. Performance of Our Weighted Method using a Leave-One-Out Analysis

	Accuracy	Precision	Specificity	Sensitivity	NVP	MCC
VariBench†	0.86	0.86	0.86	0.85	0.85	0.71
SwissVar†	0.81	0.84	0.85	0.77	0.79	0.63
BRCA1	-	-	0.60	0.47	-	-
MSH2	-	-	0.50	0.74	-	-
MLH1	-	-	0.19	0.95	-	-
TP53	-	-	NA	1.00	-	-

For this analysis, we adjusted our pathogenicity weights, W_d and W_n, if and only when the AAS being evaluated was present in either the HGMD [Stenson et al., 2009] or UniProt [Apweiler et al., 2004] datasets. We observed no significant deviations in the performance measures and concluded that the performances observed in our benchmarks were not biased towards the pathogenicity weights employed.

† The performance measurements reported are calculated from normalised numbers.

Supp. Table S3. Availability of Computational Prediction Methods

	SIFT	PolyPhen v1	PolyPhen v2	PANTHER	PhD-SNP (Profile)	PMut	SNPs&GO	MutPred	FATHMM
Web-Server	✓	*	✓	✓	✓	✓	✓	✓	✓
Average Run-Time (Single Query)	†	-	†	< 1 Minute	2 Minutes	†	†	†	†
Batch Facility Available	✓	-	✓	✗	✗	✗	✗	✗	✓
Batch Facility Limitation	1,000 Proteins	-	150,000 AASs	-	-	-	-	-	Unlimited
Phenotypic Associations	✗	-	✗	✗	✗	✗	✗	✗	✓
Download Available	✓	-	✓	✓	✓	✗	✗	✗	✓
Optional Pre-Computed Database‡	✗	-	✓	✗	✗	-	-	-	✓
Open Source	✗	-	✗	✓	✗	-	-	-	✓

* PolyPhen v1 has now been discontinued and is no longer accepting user submissions
† pre-computed / near-instant predictions available (restrictions may apply)
‡ optional pre-computed database for near-instant predictions while running locally

Supp. Table S4. The Predicted Phenotypic Consequences of Disease-Associated AASs against their Associated Diseases/Abnormalities

Gene & Amino Acid Substitution	Associated Disease & MIM Identifier	Phenotypic Inference
FBN1 C1971Y	Marfan syndrome MIM# 154700	Dilatation of the Ascending Aorta Abnormality of the Aortic Valve Emphysema Mitral Regurgitation
HEXA W485R	GM2-Gangliosidosis type 1 MIM# 272800	Abnormality of Metabolism/Homeostasis Angiokeratoma Mental Deterioration Cardiomegaly Beaking of Vertebral Bodies
PSAP C388F	Atypical Gaucher disease MIM# 610539	Abnormality of the Musculoskeletal System Abnormality of Metabolism/Homeostasis Abnormality of the Immune System Functional Respiratory Abnormality Abnormality of the Lung Respiratory Insufficiency
CHRNG R239C	Escobar syndrome MIM# 265000	Intermittent Episodes of Respiratory Insufficiency Due to Muscle Weakness Prolonged Miniature Endplate Currents Decreased Size of Nerve Terminals Generalized Muscle Weakness due to Defect at the Neuromuscular Junction Muscle Fiber Atrophy Multiple Pterygia Generalized Amyoplasia Decreased Fetal Movement Hypoplastic Heart Cystic Hygroma Poor Feeding due to Muscle Weakness Easy Fatigability Gower Sign Flat Nose Abnormal Cervical Curvature Thin Ribs Vertebral Fusion Weak Cry Ophthalmoparesis Ptosis High-Arched Palate Poor Suck Macrotia Bulbar Palsy Intrauterine Growth Restriction Myopathy Malignant Hyperthermia Umbilical Hernia Joint Dislocation Abnormality of Temperature Regulation Abnormality of Prenatal Development or Birth Abnormality of Muscle Fibers Abnormality of the Nervous System Functional Respiratory Abnormality

Supp. Table S5. Interesting Single Nucleotide Variants (SNVs) Between the "Elite" and "Landrace" Wheat Varieties

Wheat Contig & nsSNP Position	Phenotypic Inference
Contig F0Z7V0F01D2DA5 nsSNP Position 127	1 Main Shoot Growth Seed Development Stages Plant Structure Development Stage Flower Development Stages Corolla Developmental Stages 4 Anthesis 3 Flower Organ Development Stages 4 Leaf Senescence Stage A Vegetative Growth Embryo Development Stages Whole Plant Growth Stage Leaf Production LP.06 Six Leaves Visible LP.12 Twelve Leaves Visible LP.04 Four Leaves Visible F Mature Embryo Stage D Bilateral Stage E Expanded Cotyledon Stage LP.02 Two Leaves Visible LP.10 Ten Leaves Visible C Globular Stage Leaf Development Stages LP.08 Eight Leaves Visible
Contig 09781 nsSNP Position 386	Plant Structure Development Stage