The American Journal of Human Genetics, Volume *102*

Supplemental Data

A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships

Taylor L. Mighell, Sara Evans-Dutson, and Brian J. O'Roak

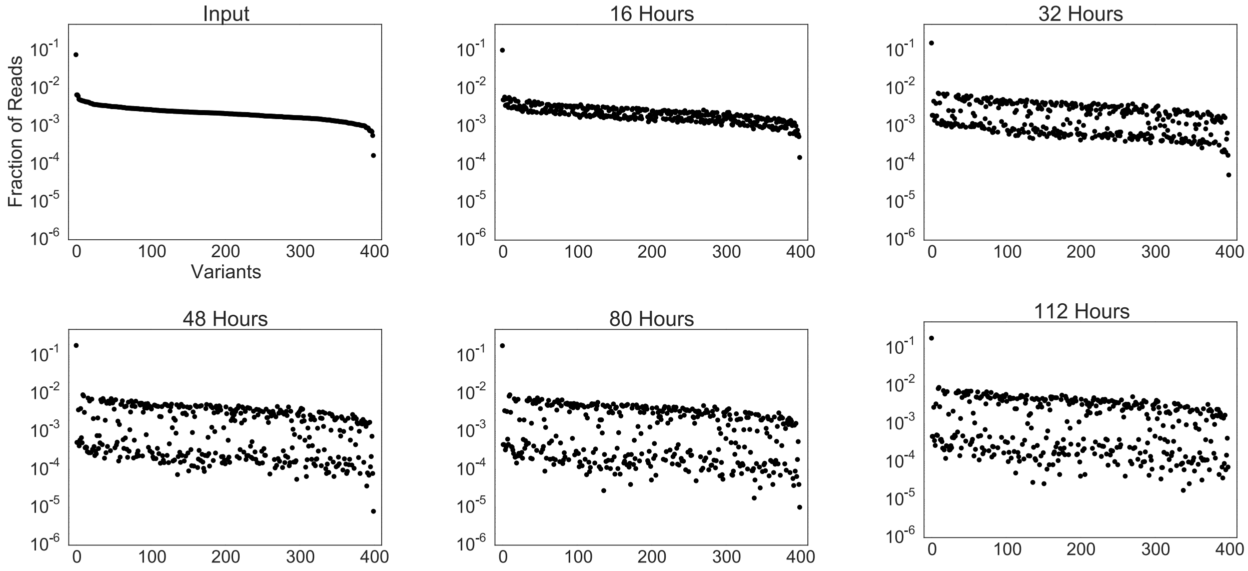
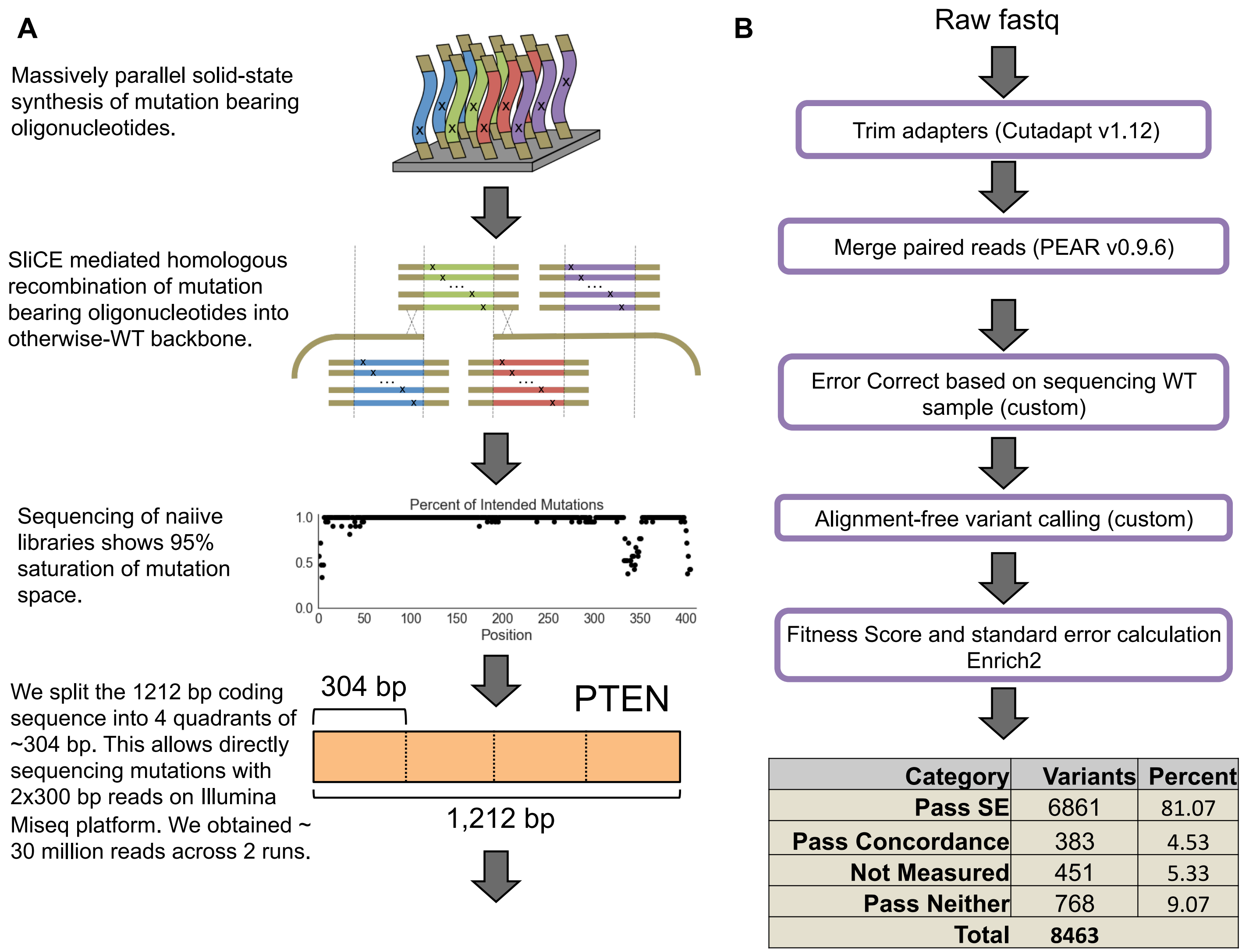


Figure S1. Optimization of humanized yeast assay for liquid culture induction and selection. We performed a pilot experiment with ~400 variants (single tile) to determine when effect size was maximized under induction conditions. We sequenced the input library (pre-induction) and p110α and PTEN induced populations at indicated time points. At each time point, five million yeast cells were passaged to fresh induction medium and the remainder used for DNA extraction. Displayed are the relative read counts of each variant, plotted in the same order as input. Effect size reaches a plateau at 48 hours, which we then used as the selected time points for the rest of the experiments in this study.


A
x
...
Percent of Intended Mutations
1.0
0.5
0.0
0
50
100
150
200
250
300
350
400
Position
304 bp
PTEN
1,212 bp
B
Raw fastq
Trim adapters (Cutadapt v1.12)
Merge paired reads (PEAR v0.9.6)
Error Correct based on sequencing WT sample (custom)
Alignment-free variant calling (custom)
Fitness Score and standard error calculation Enrich2

Massively parallel solid-state synthesis of mutation bearing oligonucleotides.

SliCE mediated homologous recombination of mutation bearing oligonucleotides into otherwise-WT backbone.

Sequencing of naiive libraries shows 95% saturation of mutation space.

We split the 1212 bp coding sequence into 4 quadrants of ~304 bp. This allows directly sequencing mutations with 2x300 bp reads on Illumina Miseq platform. We obtained ~ 30 million reads across 2 runs.

Category	Variants	Percent
Pass SE	6861	81.07
Pass Concordance	383	4.53
Not Measured	451	5.33
Pass Neither	768	9.07
Total	8463	

Figure S2. Schematic overview of mutagenesis and computational workflow.
(A) We generated a saturation mutagenesis library by incorporating single-mutation-bearing oligonucleotides into an otherwise wild-type backbone. Oligos were synthesized on solid-state arrays (CustomArray) in 31 individual tiles/pools. Oligo tiles were PCR amplified separately. Long range PCRs of otherwise wild-type plasmid with custom primers for each tile were used as template for SliCE mediated homologous recombination. We divided the protein coding sequence into 4, ~300 bp fragments/quadrants so that we could cover each entire mutation-bearing segment with 2x300 base-pair (bp) paired-end sequencing reads. Mutagenized plasmids were transformed into bacteria. Clones from individual mutagenesis tiles were pooled by quadrant and transformed into yeast for functional assays.
(B) Overview of the computational pipeline for processing reads and obtaining fitness scores. Variant predictions were considered high-confidence if passing a standard error (SE) filter or showing concordant effects between two biologic replicates (Materials and Methods).

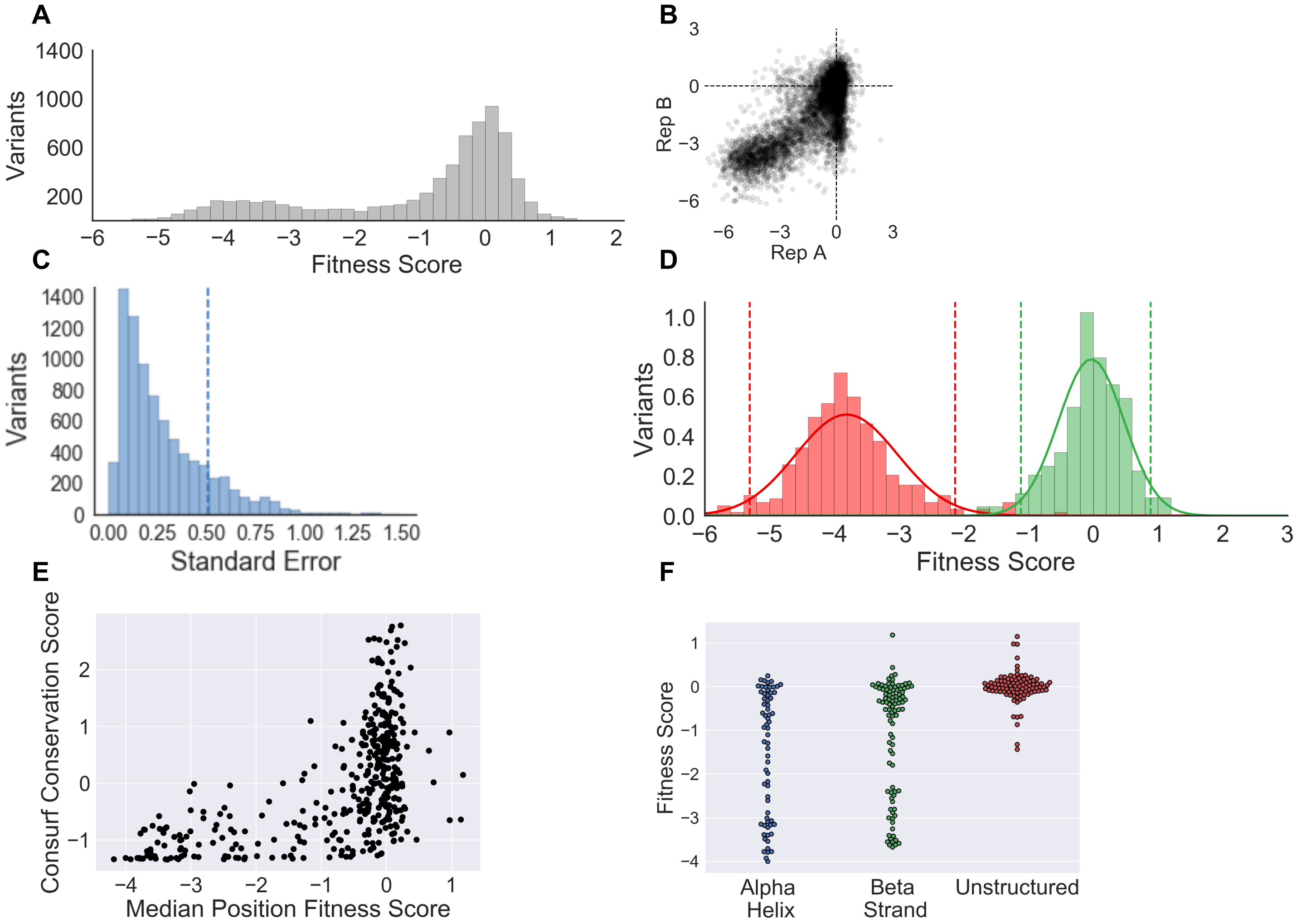


Figure S3. Overview of PTEN saturation mutagenesis dataset and relative fitness scores.

(A) Distribution of fitness effects for all high-confidence variants (7,244) derived from two biologic replicates, with three technical replicates each.

(B) Biological replicates show high correlation (Pearson's r = 0.76).

(C) Distribution of standard errors for measured variants. High-confidence variants to the left of the dashed line have 95% confidence intervals less than or equal to one natural-log fold change.

(D) The distributions of truncating mutations (excluding those in the regulatory tail) (red, left) and synonymous wild-type like mutations (green, right) are shown. Dashed lines indicate the two-tailed 95th percentile limits for synonymous and truncating variants.

(E) The median fitness score of all high-confidence scores at each position is correlated with the evolutionary conservation at that position (Spearman ρ = 0.58). Evolutionary conservation for all positions was obtained with ConSurf, using following options: "Amino-Acids", "No known protein structure", "No MSA", and default homolog search parameters.

(F) Comparison of median fitness scores for positions in alpha helices, beta strands, or unstructured regions. Alpha helix and beta strand assignments obtained through STRIDE for structure PDB: 1D5R. Unstructured positions are those absent from the crystal structure (1-13, 282-312, 352-403).

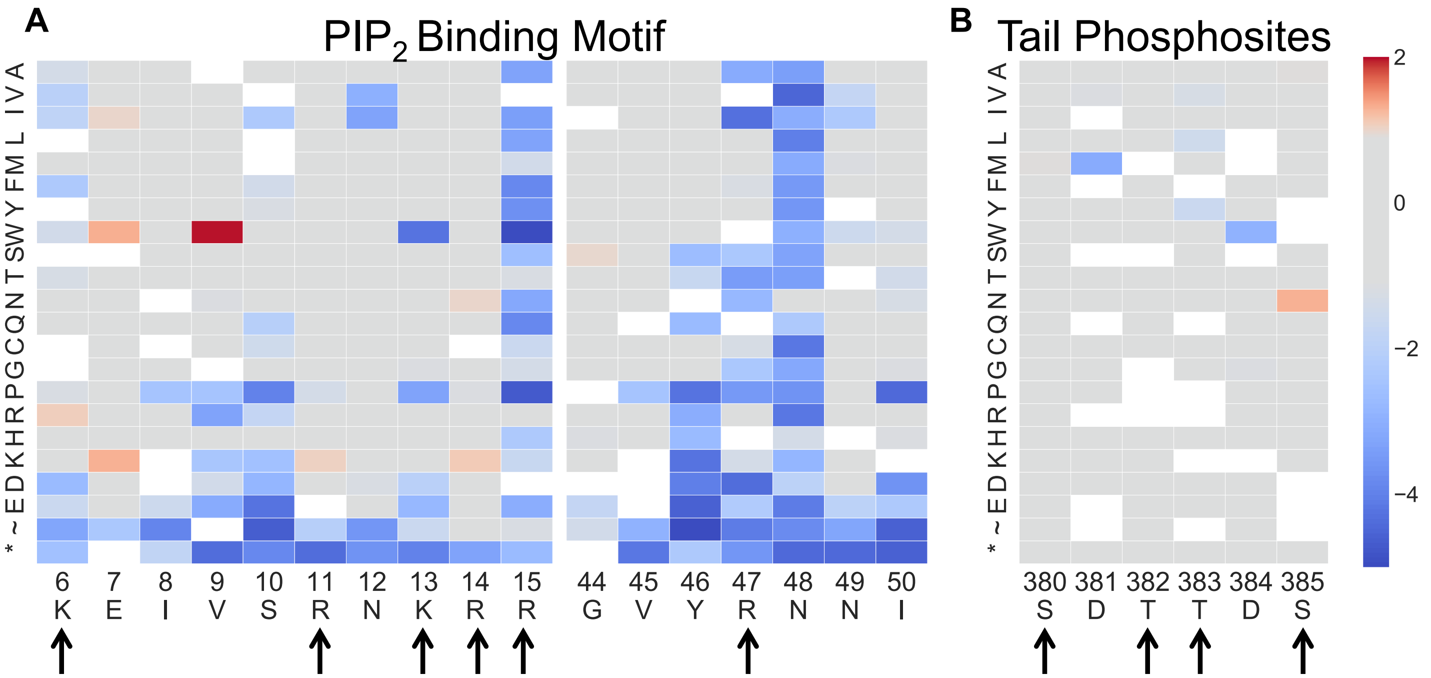


Figure S4. Evaluation of mutation effects within the PTEN predicted PIP$_2$ binding motif and tail phosphosites.

(A) Fitness scores highlighting positively charged residues in PIP$_2$ binding domain (Lys6, Arg11, Lys13, Arg14, Arg15) as well as Arg47, with neighboring residues. Lys13, Arg15, and Arg47 are the most critical in our assay.

(B) Fitness scores for C-terminal regulatory tail phosphosites (Ser380, Thr382, Thr383, Ser385) and neighboring positions.

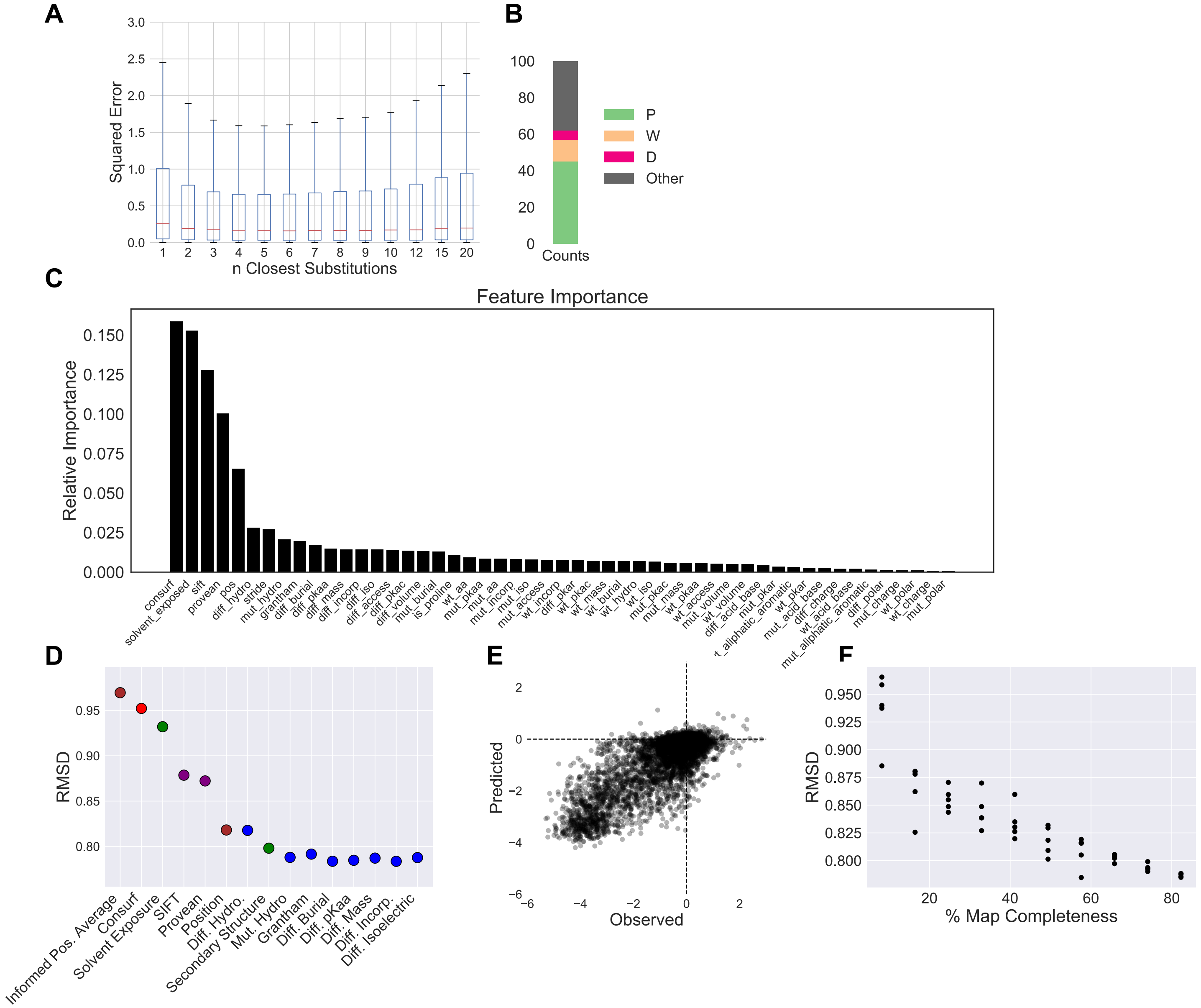


Figure S5. Development of a random forest algorithm to impute relative fitness scores for missing data.

(A) We used correlation coefficients[53] between amino acid substitutions to identify, in aggregate, the number of most closely correlated substitutions that maximized accuracy in the prediction of missing data. To generate each prediction we identified the n most closely correlated substitutions that were measured with high confidence at that positions, and calculated the average weighted by the standard error of each substitution. Box plots represent the squared error between measured value (in our assay) and value predicted from the n closest substitutions for all high-confidence measurements. We chose to use five for subsequent modeling, and define this value as “informed position average”.

(B) The 100 high-confidence substitutions that were predicted most poorly by the five most closely correlated substitutions, which show strong enrichment for proline.

(C) We collected ~50 evolutionary, predictor-based, and biophysical features describing each substitution (as in Weile *et al.*, 2017). Then, we trained a random forest model (Scikit-learn version 0.19.0, sklearn.ensemble.RandomForestRegressor, n_estimators=500, criterion= “mse”, max_features=0.33, random_state=0, oob_score=True) and report here the relative increase in impurity upon random permutation of each feature, which is a surrogate for feature importance.

(D) Then, we trained a model using “informed position average” as the only feature, and iteratively added features, in the order of importance calculated in C. Root mean square deviation (RMSD) of predictions made by iteratively adding indicated features to the model and performing 10-fold cross validation are shown, and we stopped adding features once the decrease in error plateaued. Color of marker indicates the type of feature; brown is intrinsic to the dataset, green is structural, purple is predictor, and blue is biophysical.

(E) We used the 15 features in D to train a final model (options same as in C) and performed 10-fold cross validation on the high-confidence variant set. We generated predictions for all high-confidence variants and plotted the observed and predicted values for each variant. Pearson’s r = 0.80, options same as above.

(F) RMSD results from downsampling to indicated map completeness. We downsampled from our high-confidence dataset and retrained models at each indicated percent map completeness. The maximum value is 82.3%, which is the percent map completeness that our high-confidence missense dataset represents. 5 replicates were performed at each point % map completeness. Options same as above, except random_state=None.

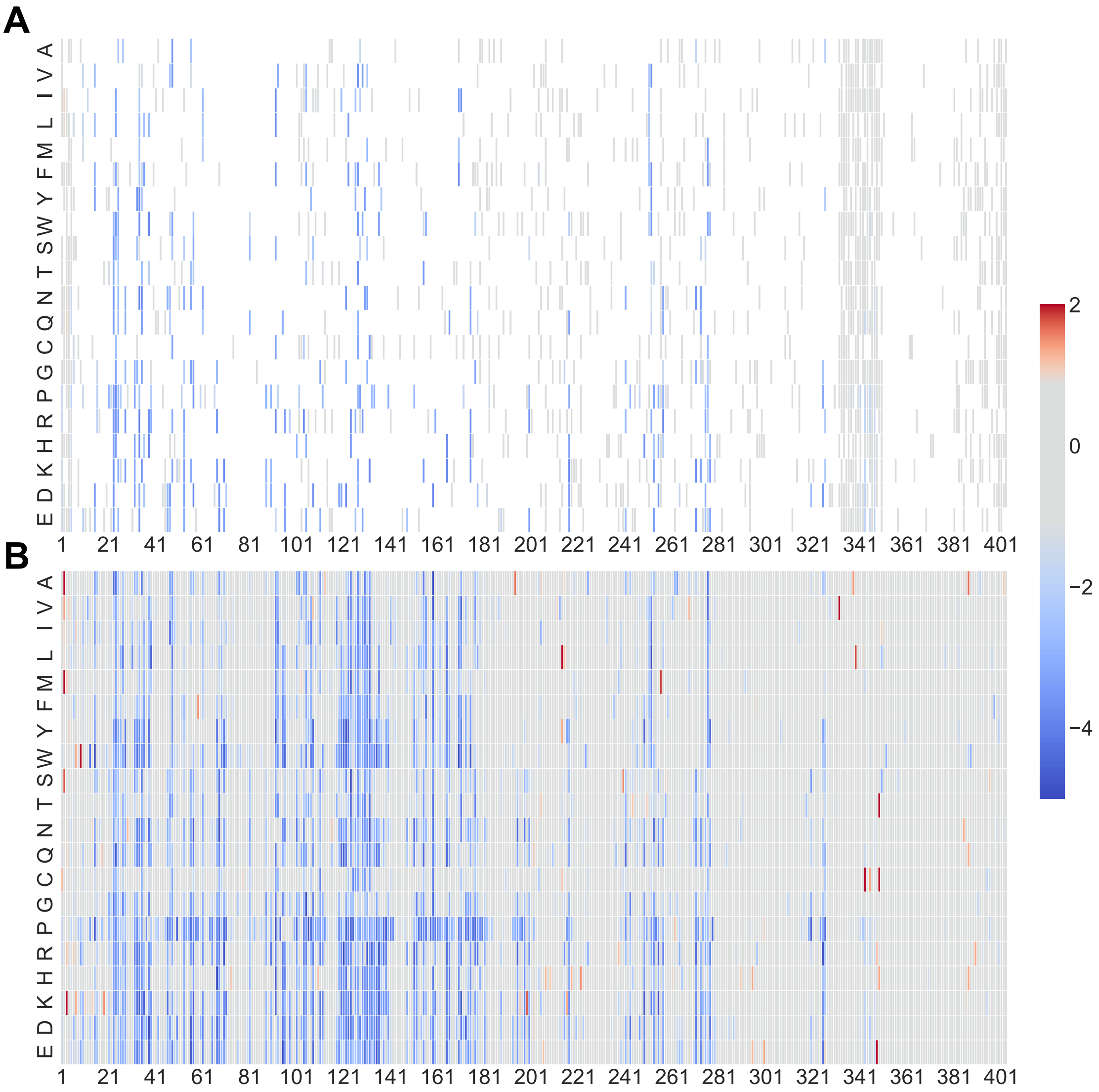


Figure S6. A comprehensive functional map of predicted effects of PTEN mutations using imputed scores.

(A) We trained the random forest algorithm on 6,300 missense variants that were measured with low standard error (95% confidence interval < 1 fitness score). We omitted single residue deletions and nonsense mutations. We then predicted the fitness score of the remaining 1,357 variants. Imputed values are colored according to their fitness score. Variants used in the training are white.

(B) Complete sequence function map with high-confidence measurements in addition to imputed values.

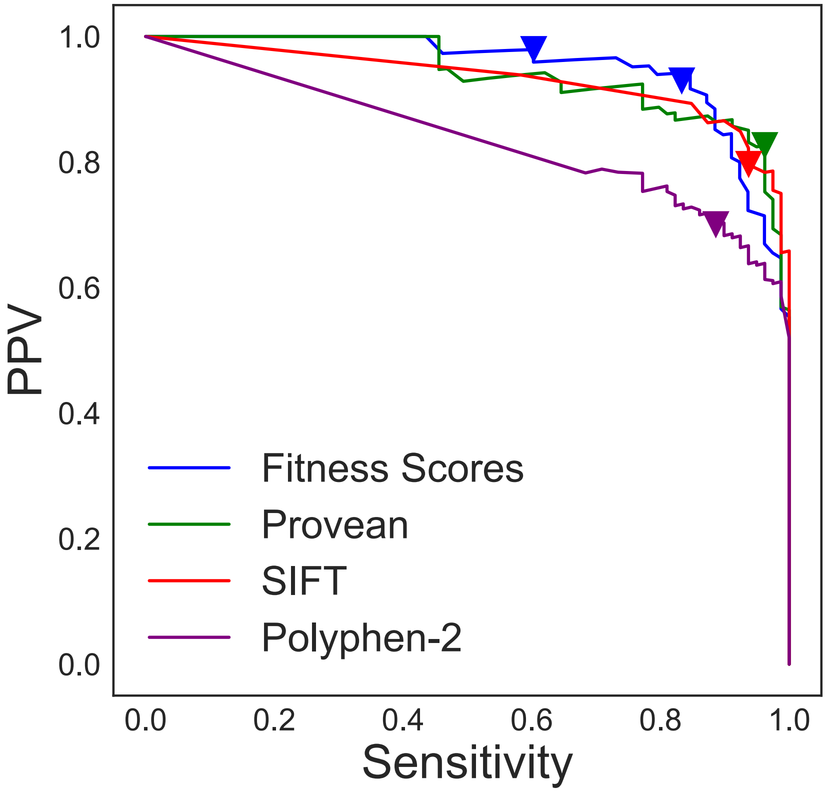


Figure S7. Positive predictive value (PPV) and sensitivity (precision and recall) curves for fitness scores and mutation effect predictors.
PPV and sensitivity were calculated at 200 points between the minimum and maximum of the predictor's output. Triangles represent the cutoff values shown in Figure 4C, based on default setting (Provean=-2.5, SIFT=0.05, Polyphen-2= 0.15). The two blue triangles correspond to the truncation (left) and synonymous (right) thresholds.