






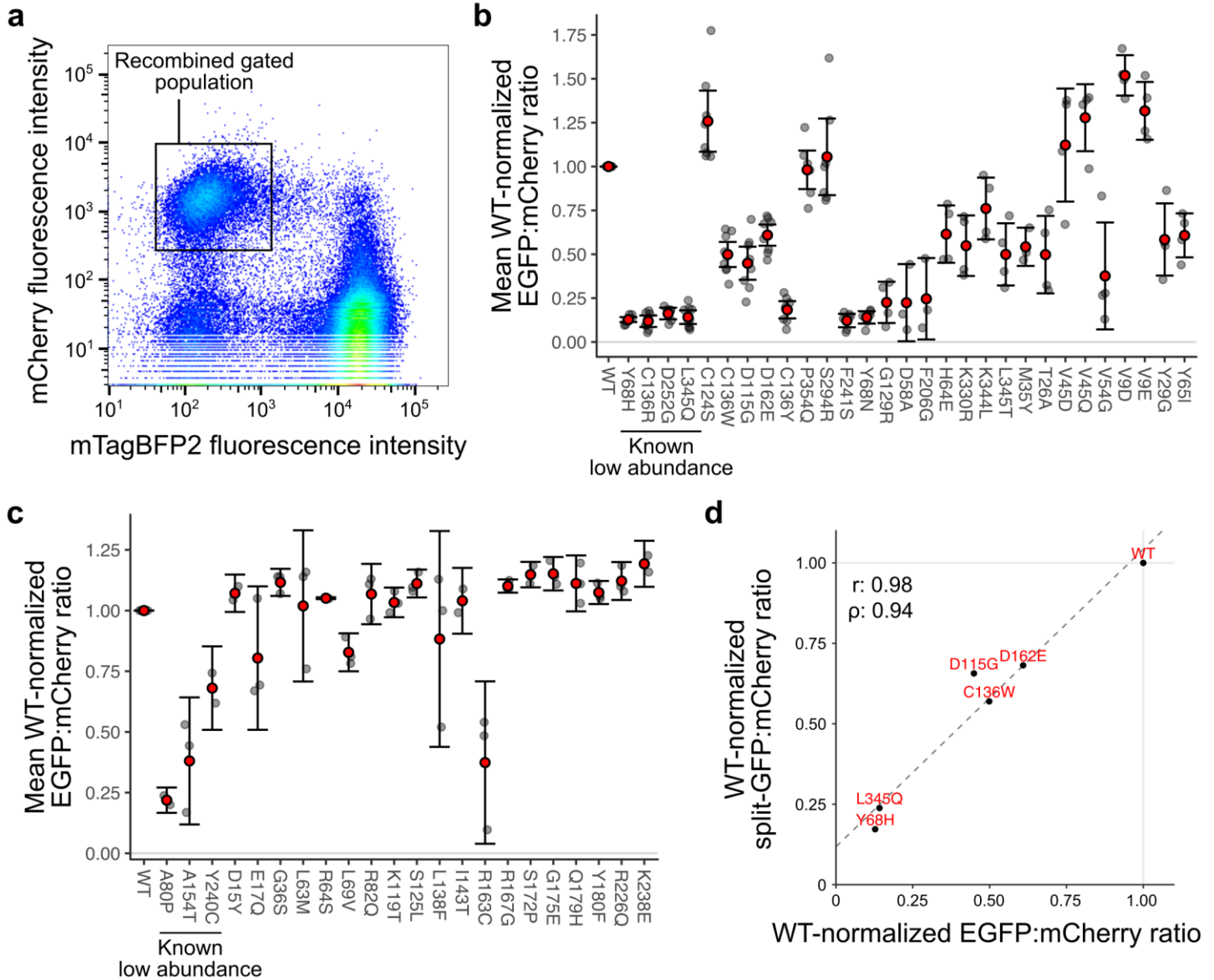
In the format provided by the authors and unedited.

# Multiplex assessment of protein variant abundance by massively parallel sequencing

Kenneth A. Matreyek <sup>1,8</sup>, Lea M. Starita<sup>1,8</sup>, Jason J. Stephany<sup>1</sup>, Beth Martin <sup>1</sup>, Melissa A. Chiasson<sup>1</sup>, Vanessa E. Gray<sup>1</sup>, Martin Kircher <sup>1</sup>, Arineh Khechaduri<sup>1</sup>, Jennifer N. Dines<sup>2</sup>, Ronald J. Hause<sup>1</sup>, Smita Bhatia<sup>3</sup>, William E. Evans <sup>4</sup>, Mary V. Relling<sup>4</sup>, Wenjian Yang<sup>4</sup>, Jay Shendure<sup>1,5\*</sup> and Douglas M. Fowler <sup>1,6,7\*</sup>

---

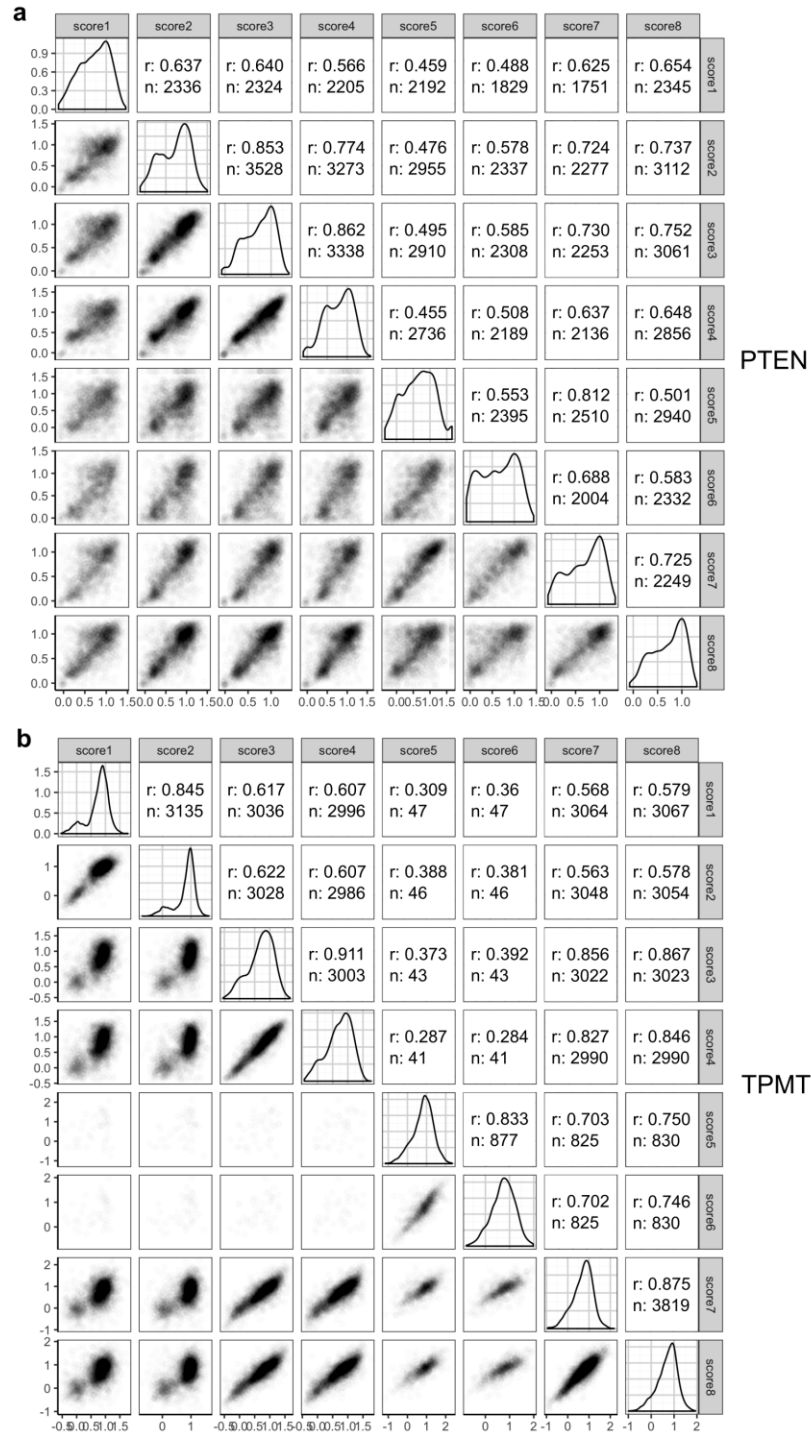
<sup>1</sup>Department of Genome Sciences, University of Washington, Seattle, WA, USA. <sup>2</sup>Department of Medical Genetics, University of Washington, Seattle, WA, USA. <sup>3</sup>School of Medicine, University of Alabama at Birmingham, Birmingham, AL, USA. <sup>4</sup>Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, TN, USA. <sup>5</sup>Howard Hughes Medical Institute, Seattle, WA, USA. <sup>6</sup>Department of Bioengineering, University of Washington, Seattle, WA, USA. <sup>7</sup>Genetic Networks Program, Canadian Institute for Advanced Research, Toronto, Ontario, Canada. <sup>8</sup>These authors contributed equally: Kenneth A. Matreyek, Lea M. Starita. \*e-mail: [shendure@u.washington.edu](mailto:shendure@u.washington.edu); [dfowler@uw.edu](mailto:dfowler@uw.edu)



**Supplementary Figure 1**

**Validation experiments of EGFP-fusions for assessing PTEN and TPMT steady-state abundance.**

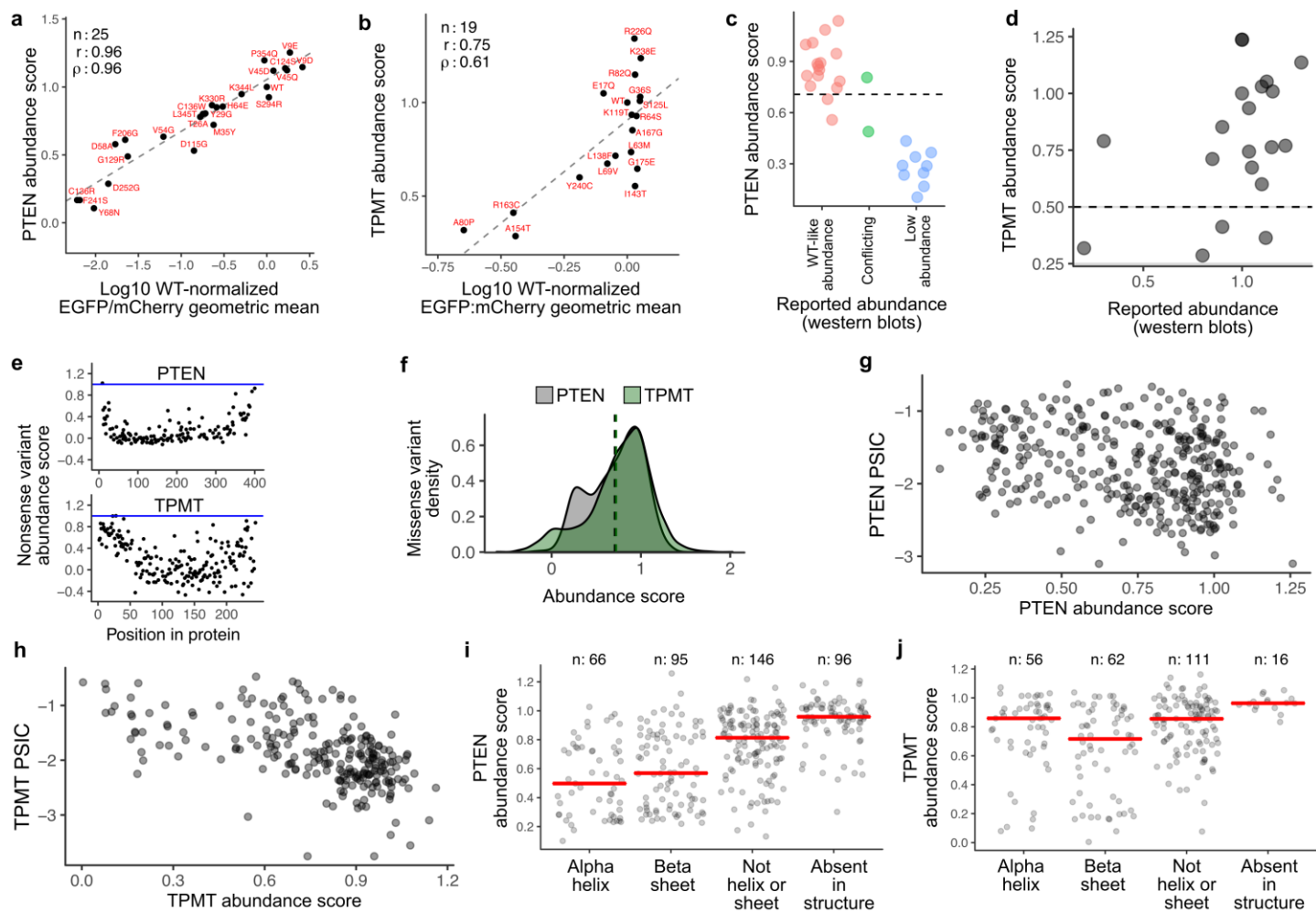
**a**, Representative gating strategy for mTagBFP2 negative, mCherry positive cells containing 15,000 recombined cells. **b**, PTEN variant EGFP:mCherry ratio geometric means as a fraction of WT, for known and previously uncharacterized PTEN low-abundance variants. Error bars denote 95% confidence intervals of the mean (red), with individual data points shown in grey. Each variant was assessed in at least 3 independent experiments. **c**, Similar plot for TPMT, with error bars denoting 95% confidence intervals of the mean (red), with individual data points shown in grey. All variants were independently assessed three times, except variants p.Asp15Tyr, p.Arg64Ser, p.Ala80Pro, p.Ile143Thr, p.Lys238Glu, p.Tyr240Cys, which were assessed twice. **d**, Scatterplot comparison of WT-normalized EGFP:mCherry ratios for EGFP- or 15-aa split-GFP fused PTEN variants. Values are the mean of 3 independently performed experiments.  $n = 6$  samples. “ $r$ ” and “ $p$ ” denote Pearson’s and Spearman’s correlation coefficients, respectively.



## Supplementary Figure 2

### Correlations between PTEN and TPMT VAMP-seq replicates.

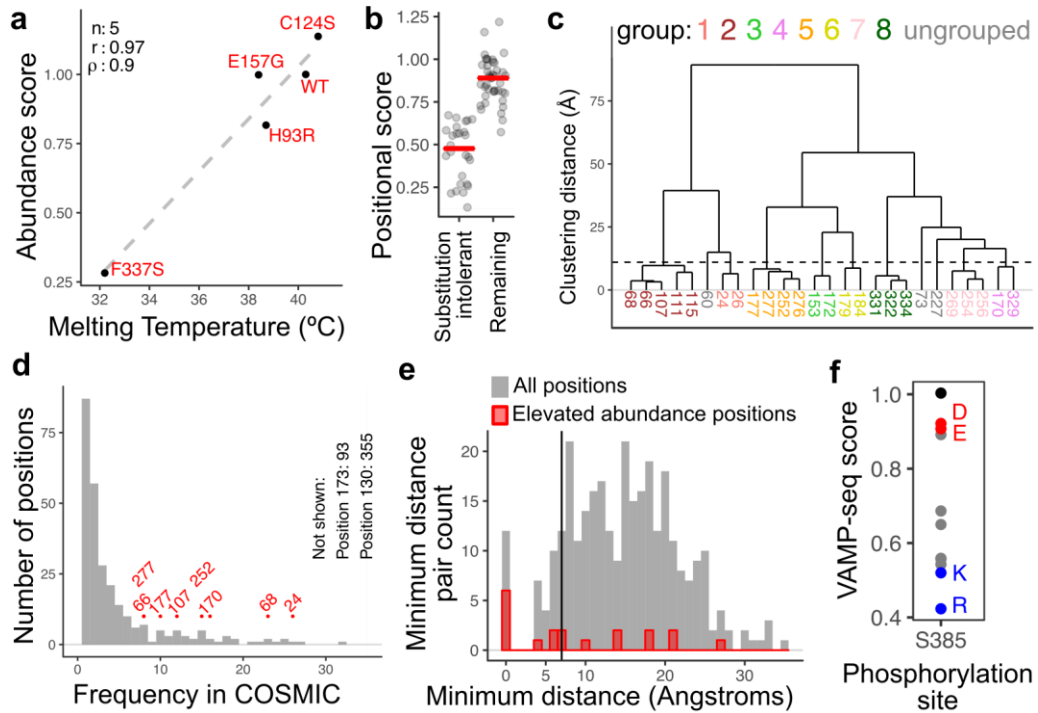
**a, b**, Pairwise VAMP-seq abundance score correlations between replicate sorting experiments for PTEN (a) and TPMT (b). n values are the number of variants scored in both experiments. Replicates 5 and 6 for TPMT contained a subset of mutagenized positions different from those mutagenized in replicates 1 through 4, with both subsets mixed together for Replicates 7 and 8. Pearson's correlation coefficients are shown. Score numbers in this figure correspond to experiment numbers in Supplementary Table 1.



### Supplementary Figure 3

#### Validation analyses for VAMP-seq-derived abundance scores.

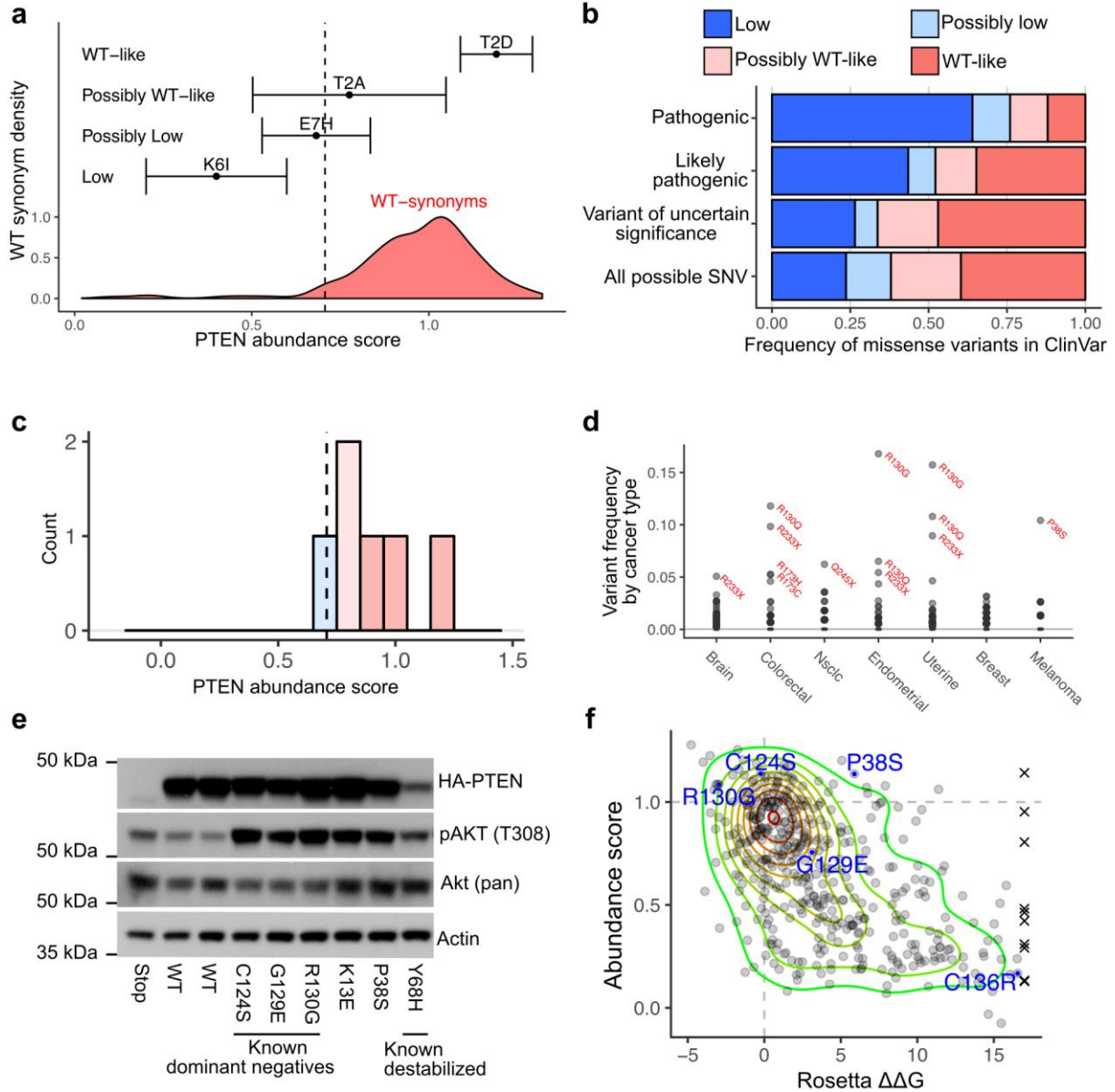
**a, b**, Scatterplot comparison of VAMP-seq abundance scores (x-axis) and individually assessed  $\log_{10}$ -transformed, WT-normalized geometric means of the EGFP:mCherry ratios for various PTEN (**a**) and TPMT (**b**) variants (see also **Supplementary Figure 1b, c**).  $r$  and  $\rho$  denote Pearson's and Spearman's correlation coefficients, respectively. **c**, PTEN VAMP-seq scores for variant steady state expression characterized by western blot analysis in previous publications (See **Supplementary Table 9**). **d**, Scatterplot comparing TPMT VAMP-seq scores (y-axis) and previously published abundance values from western blots (see **Supplementary Table 10**). **e**, Nonsense variant VAMP-seq scores by amino acid position, for PTEN (top) and TPMT (bottom). WT abundance score (1.0) shown as a blue line. N-terminal nonsense variants append a small number of residues to EGFP, which does not affect its abundance. C-terminal nonsense variants remove a small number of residues from PTEN or TPMT, which also does not impact abundance. **f**, Missense variant abundance score density plots for PTEN (gray) and TPMT (green). The thresholds of the 5% lowest synonymous variant scores are shown, for each protein, by the dotted lines. **g, h**, Scatterplot comparing positional median PTEN (**g**) and TPMT (**h**) VAMP-seq scores to PSIC evolutionary conservation scores for each position (Sunyaev *et al.*) **i, j**, Positional median PTEN (**i**) and TPMT (**j**) abundance scores for positions found in various secondary structure types, with the red line denoting the median value for the group.  $n$  values denote the number of positions that fell into each category.



**Supplementary Figure 4**

**Biochemical features associations with VAMP-seq-derived abundance scores.**

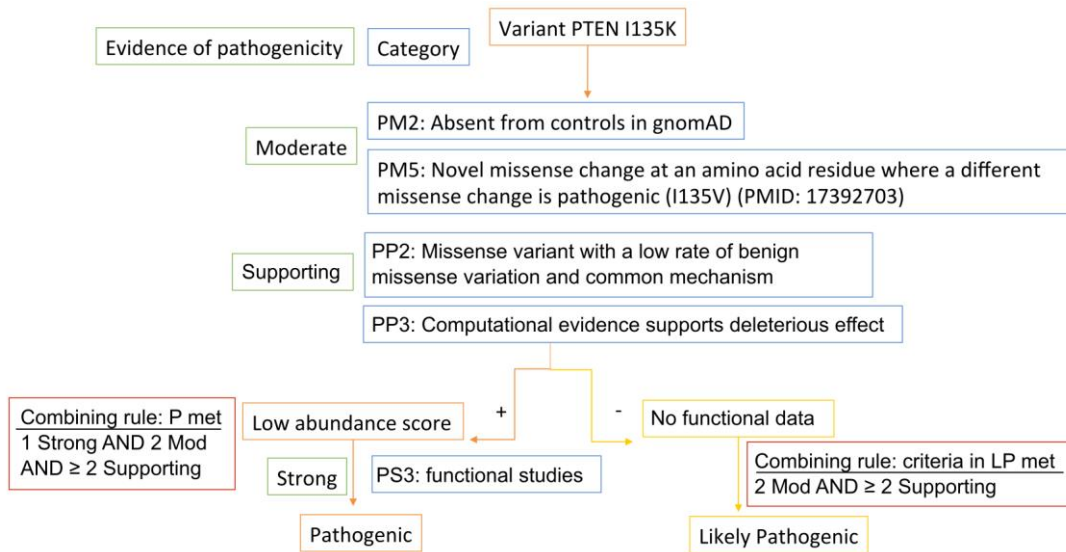
**a**, Scatterplot comparing abundance score (y-axis) to *in vitro* characterized melting temperatures of select PTEN variants (Johnston *et al.*). *r* and *p* denote Pearson's and Spearman's correlation coefficients, respectively. **b**, A plot of positional median scores for PTEN positions with potential hydrogen bonds or salt bridges. A position was considered intolerant only if it had 5 or more variants and more than 90% of the abundance scores were at or below the score threshold containing the lowest 5% of synonymous variants. Red bars denote median abundance score values. *n* = 26 for substitution intolerant, and *n* = 50 for the remaining positions. **c**, Substitution-intolerant PTEN positions with potential polar contacts, clustered by distance based on PDB coordinates (PDB: 1d5r). Positions within 11 Å of each other were considered part of a group. The dashed line shows the 11 Å distance cutoff. **d**, Histogram of the number of PTEN missense variants per position in COSMIC. Substitution-intolerant positions potentially involved in polar contacts with counts in COSMIC greater than 7 are labeled in red. **e**, Minimum distance of all PTEN positions (gray) or elevated-abundance positions (red) from known phospholipid-binding positions. The black line denotes a 7 Å distance. A position was considered elevated in abundance only if it had 5 or more variants and there were more than 5 variants with scores above the median of the synonymous distribution. **f**, VAMP-seq scores for variants at position S385, with a synonymous variant in black, negatively charged variants in red, positively charged variants in blue, and all other variants in gray.



**Supplementary Figure 5**

**PTEN variant abundance classification and relationship to germline and somatic variation.**

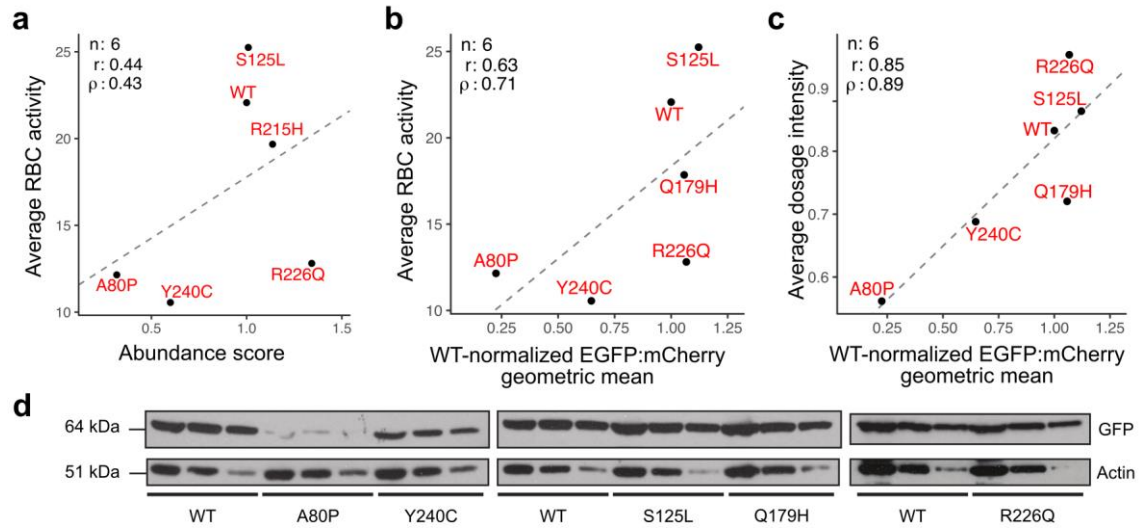
**a**, Illustrative examples of variant abundance classifications, with the dotted line representing the threshold above which 95% of synonymous variants reside. Points represent the VAMP-seq score for each representative variant, with error bars denoting the 95% confidence interval derived from experimental replicates. *n* values are 3, 5, 2, and 4 for p.Thr2Asp, p.Thr5Ala, p.Glu7His, and Lys6Ile, respectively. **b**, Frequencies of each PTEN abundance class for each PTEN ClinVar interpretation, as well as for all possible SNVs with abundance classifications. **c**, Abundance scores and classes for PTEN variants with allele counts highly unlikely to be causal for Cowden's Syndrome. **d**, Frequencies of all observed PTEN variants across different cancer types in the TCGA and AACR GENIE data. Highly recurrent PTEN variants are labeled in red. **e**, Western blot analysis of a clonal line stably expressing WT or missense variants of N-terminally HA-tagged PTEN. This line was derived independently from the line used to generate the data shown in Figure 4f. This experiment was independently performed twice with similar results. **f**, Comparison of PTEN abundance scores with changes in folding energies predicted by Rosetta using the ddg\_monomer protocol. Variants are shown as gray circles, with the exception of those with Rosetta  $\Delta\Delta G$  predictions greater than 17, which are marked by a black "x" at a  $\Delta\Delta G$  value of 17. Contour lines are colored by the regional density of points. Previously or newly identified PTEN dominant negative variants shown as blue points with blue labels.



### Supplementary Figure 6

#### Flow chart of PTEN I135K pathogenicity reclassification using VAMP-seq data.

The ACMG/AMP joint criteria for classifying variants were used, with low abundance classification by VAMP-seq considered strong experimental support of pathogenicity (PS3). Without functional data there is no strong or very strong evidence of pathogenicity for this variant, therefore pathogenic criteria cannot be fulfilled and the variant remains classified as likely pathogenic. With low abundance data, PS3 can be used and pathogenic criteria is met.

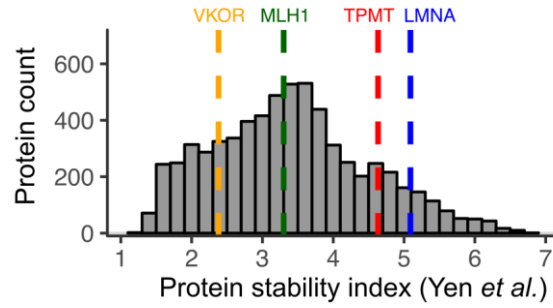


### Supplementary Figure 7

#### Relationship of TPMT variant abundance to drug sensitivity.

**a**, Scatterplot comparing abundance scores and previously characterized red blood cell (RBC) activity from patients. **b**, **c**, Scatterplots comparing individually assessed, WT-normalized EGFP:mCherry geometric means to previously published values of average RBC activity (b), or average patient dosage intensity (c). Dose intensity is the dose where 6-MP becomes toxic to the patient before reaching the 100% protocol dose of 75 mg/m<sup>2</sup>.  $r$  and  $\rho$  denote Pearson's and Spearman's correlation coefficients, respectively.  $n = 6$  samples for each plot. **d**, Western blotting results for individually-expressed TPMT variant GFP fusions. Each variant was blotted with 45, 15, and 5  $\mu$ g of total protein input per lane. This experiment was performed once.

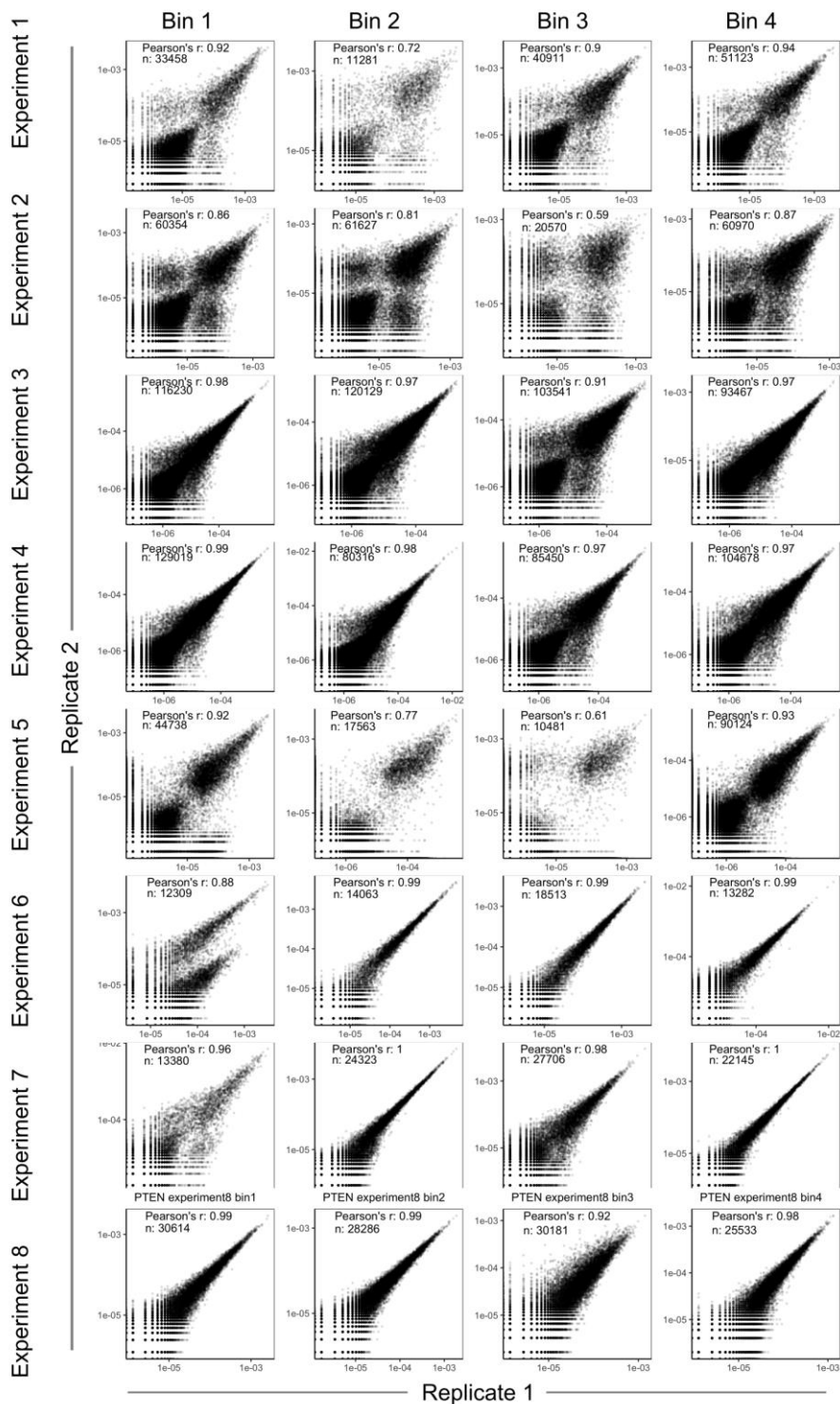




### Supplementary Figure 8

#### Protein stability indices for most human protein N-terminal EGFP fusions.

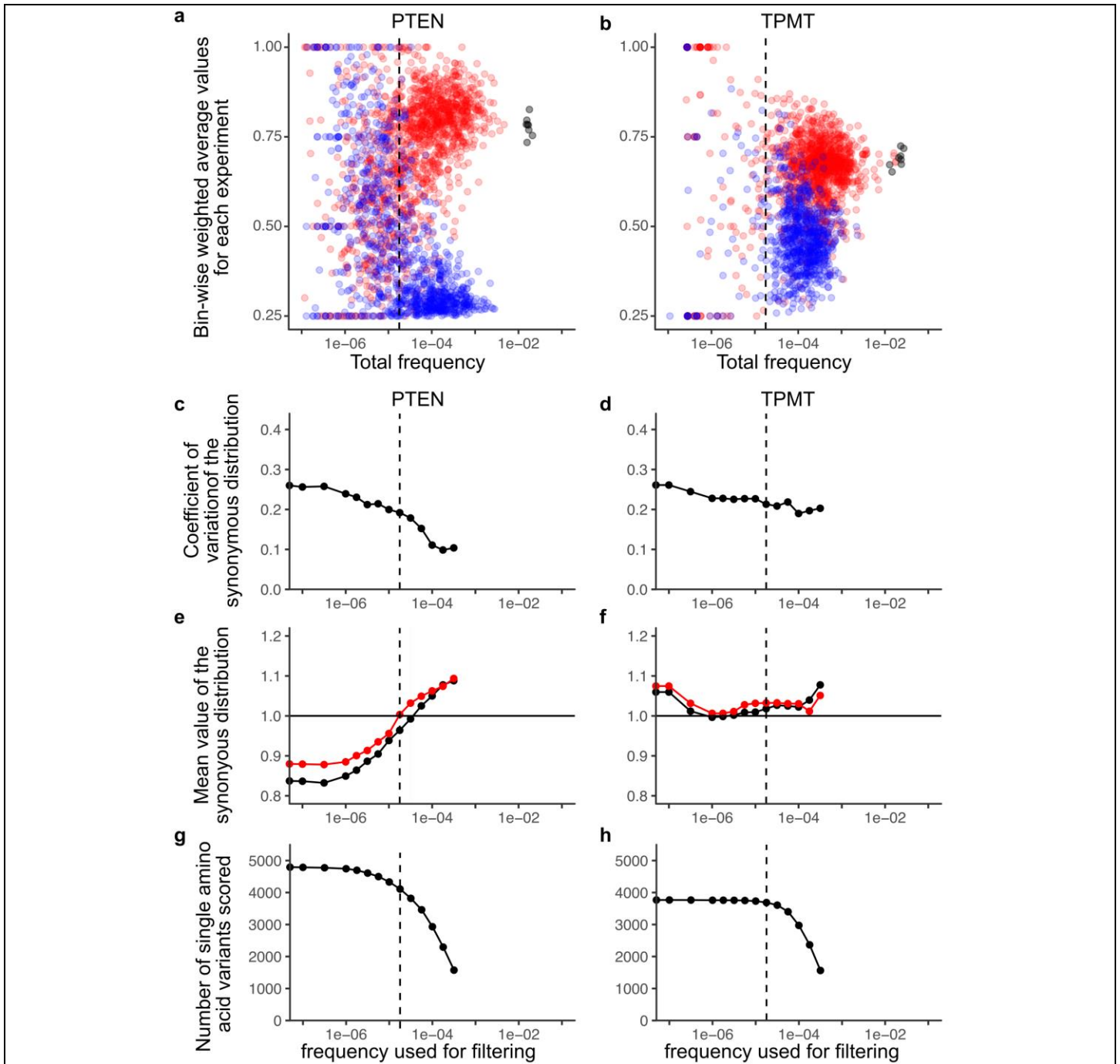
A histogram of protein stability indices from Yen *et al.* Protein stability index values for proteins tested in the VAMP-seq assay are shown as dashed vertical lines. Protein stability indices were not available for PTEN, CYP2C9, CYP2C19, and PMS2.



**Supplementary Figure 9**

**Amplification and sequencing technical replicates for PTEN.**

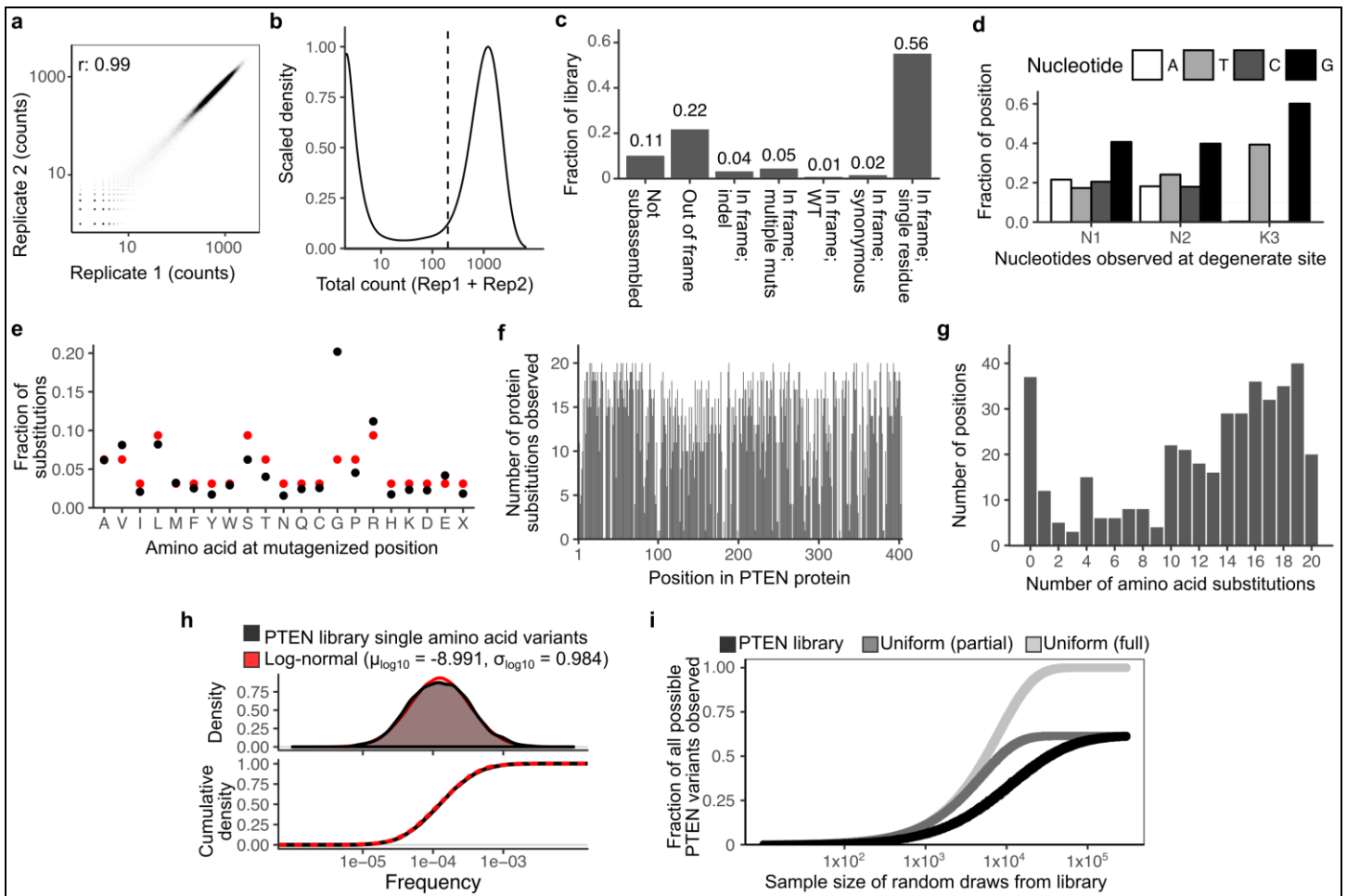
Scatterplots comparing variant frequency derived from replicate PCR amplification and sequencing for each of the four bins in every PTEN experiment are shown.



**Supplementary Figure 10**

**Scheme to determine total frequency filtering threshold value.**

**a, b**, Scatterplots showing the total frequencies and weighted average values of wt (black), synonymous variants (red), or non-terminal nonsense variants (blue) for each experiment, for PTEN and TPMT respectively. A combination of synonymous variant coefficient of variation (**c** and **d**), synonymous variant mean (black) and median (red) (**e** and **f**), and total number of scored missense variants (**g** and **h**) for PTEN (**c**, **e**, and **g**) and TPMT (**d**, **f**, and **h**) were assessed at increasing total frequency filtering threshold values to obtain the threshold value that we required across the four bins for a variant to be included in the analyses we present. The  $1 \times 10^{-4.75}$  total frequency threshold used for the final analysis is displayed as a dotted line in each plot.



### Supplementary Figure 11

#### Statistics for the PTEN library.

**a**, Barcode counts from independent amplifications of the barcoded PTEN library plasmid preparation used for recombination.  $n = 67,162$  data points.  $r$  denotes Pearson's correlation coefficient. **b**, A filter based on a minimum count of 200 was imposed (black dotted line), resulting in 40,560 unique barcodes. **c**, The barcode-variant map was used to determine the frequencies of different types of sequences in the plasmid preparation of the barcoded PTEN library. **d**, Nucleotide biases at the degenerate codon for the single amino acid PTEN variants. **e**, Amino acid biases of the single amino acid variants of the PTEN library, with the frequencies expected from perfect NNK mutagenesis shown in red. **f**, Number of substitutions observed at each position of the PTEN protein amongst the 40,560 barcodes in the PTEN library plasmid preparation. **g**, Distribution of number of substitutions per position in the PTEN protein. **h**, Distribution of single amino acid variant frequencies in the PTEN library (black), along with an illustrative log-normal distribution that closely fits the PTEN data (red), shown as a density plot (top panel), or a cumulative distribution function plot (bottom panel). **i**, Sampling simulations of observed and hypothetical PTEN libraries, displaying the fraction of the 8,040 possible PTEN single amino acid and nonsense variants observed for increasing sampling sizes, with a step size of 1. Results of sampling from the PTEN variant frequency distribution observed in the library plasmid preparation are shown in black. Results of sampling hypothetical, uniformly distributed libraries containing either the subset of single amino acid variants observed in the PTEN library plasmid preparation (dark gray), or all possible PTEN single amino acid variants (light gray) are shown for comparison.

**Supplementary Table 1. PTEN and TPMT library fluorescence activated cell sorts.** The four way sorts in TPMT replicate experiments 7 and 8 were performed after mixing the recombined cell enrichments from transfections B and C.

Protein	Experiment number	Transfection	Cells sorted for recombinants	Cells sorted in Bin1	Cells sorted in Bin2	Cells sorted in Bin3	Cells sorted in Bin4
PTEN	1	1	110,000	396,000	389,000	396,000	488,000
PTEN	2	1	110,000	771,000	881,000	867,000	734,000
PTEN	3	1	110,000	403,000	397,000	407,000	497,000
PTEN	4	1	110,000	820,000	900,000	950,000	790,000
PTEN	5	2	(not enriched)	80,000	60,000	68,000	61,000
PTEN	6	2	135,000	650,000	575,000	680,000	520,000
PTEN	7	2	135,000	485,000	500,000	500,000	485,000
PTEN	8	3	310,000	1,410,000	1,410,000	1,410,000	1,440,000
TPMT	1	A	237,821	500,000	500,000	500,000	500,000
TPMT	2	A	237,821	500,000	500,000	500,000	500,000
TPMT	3	B	186,000	500,000	500,000	500,000	500,000
TPMT	4	B	186,000	500,000	500,000	500,000	500,000
TPMT-fill-in	5	C	87,739	500,000	500,000	500,000	400,000
TPMT-fill-in	6	C	87,739	500,000	500,000	500,000	500,000
TPMT + TPMT-fill-in	7	B + C	186,000 + 87,739	500,000	500,000	500,000	500,000
TPMT + TPMT-fill-in	8	B + C	186,000 + 87,739	462,018	393,779	430,967	500,000

**Supplementary Table 2. Description of columns names for supplementary data tables.**

variant: single letter variant notation for variant.

position: position of mutation.

start: wt residue.

end: mutant residue. X denotes stop codon.

class: whether the variant is WT, or a synonymous, missense, or nonsense variant.

abundance class: whether the variant is "WT-like", "Possibly WT-like", "Possibly low" or "low" abundance.

See methods, and Supplementary Figure 5a for explanation.

score: mean score for a variant across replicate experiments.

sd: standard deviation across replicate experiments.

expts: number of replicates the variant was observed in.

se: standard error across replicate experiments.

lower\_ci, upper\_ci: lower and upper confidence intervals of the abundance score assuming a normal distribution.

score 1-8: Individual VAMP-seq abundance score to the variant in replicate experiments 1 through 8.

median\_w\_ave: The median weighted average value for the variant across replicate experiments.

exp#\_w\_ave (1-8): Individual weighted average values for the variant in replicate experiments 1 through 8.

hgvs: Human Genome Variation Society protein level change designation

snv: value is 1 if the variant is possible through single nucleotide variation

splice\_jxn: value is 1 if the codon overlaps with the exonic A-G nucleotides preceding the splice junction, or the exonic G nucleotide after the junction.

lit\_destabilized: see supplementary Table 9. Conflict denotes that two studies observed differing phenotypes for that variant.

egfp\_geomean: WT-normalized geometric mean of the green:red fluorescence ratio for individually expressed and assessed variants.

egfp\_geomean\_log10: Log base 10 –transformed values of the above geometric mean ratio.

egfp\_geomean\_lower\_ci, egfp\_geomean\_upper\_ci: confidence intervals for the individually assessed variants, assuming a normal distribution.

sgfp\_geomean: WT-normalized geometric mean of the green:red fluorescence ratio for individually expressed and assessed split-GFP fused PTEN variants.

tm: Empirically determined PTEN melting temperature based on Johnston *et al* (PMID 25647146).

ddg:  $\Delta\Delta G$  free energy calculated value by Rosetta (see methods).

xca: mean x-axis location in PTEN protein data bank file 1d5r

yca: y-axis location in PTEN protein data bank file 1d5r

zca: z-axis location in PTEN protein data bank file 1d5r

abs\_tco: absolute value of the cosine of the angle between C=O of the current residue and C=O of previous residue.

kappa: The virtual bond angle defined by the three C-alpha atoms of the residues current - 2, current and current + 2. Used by DSSP to determine bend.

alpha: column indicating the chirality (alpha torsion).

phi: IUPAC peptide backbone torsion angles.

psi: IUPAC peptide backbone torsion angles.

rsa: relative surface area

hbond\_sum: sum of all hydrogen bonds estimated by DSSP

helix: DSSP helix secondary structure prediction (1 if helix deemed present)

sheet: DSSP beta-sheet secondary structure prediction (1 if sheet deemed present)

bfactor: temperature factor (also known as the isotropic B value, Debye-Waller factor) from the pdb.

schbond\_unique\_partners: number of unique side chain hydrogen bonds estimated to be formed.

schbond\_partner\_seq\_dist: maximum primary-sequence distance for the side-chain hydrogen bonds estimated to be formed.

saltbridge\_unique\_partners: number of unique salt-bridge interactions estimated to be formed.

saltbridge\_partner\_seq\_dist: maximum primary-sequence distance for the salt-bridge interactions estimated to be formed.

crowding: Number of alpha carbon atoms in a 6-angstrom distance from the alpha carbon of the residue in question.

substr\_dist: Distance in angstroms of the alpha carbon of the residue in question from the center of the substrate / mimic in the crystal structure.

grantham: Grantham score of the amino acid change

hydro1: Hydrophobicity of the wild-type amino acid

hydro2: Hydrophobicity of the mutant amino acid

hydrodiff: Difference in hydrophobicity between the wild-type and mutant amino acids

vol1: Volume of the wild-type amino acid

vol2: Volume of the mutant amino acid

voldiff: Difference in volume between the wild-type and mutant amino acids

polarity1: Polarity of the wild-type amino acid

polairty2: Polarity of the mutant amino acid  
polaritydiff: Difference in polarity between the wild-type and mutant amino acids  
weight1: Molecular weight of the wild-type amino acid  
weight2: Molecular weight of the mutant amino acid  
weightdiff: Difference in molecular weight between the wild-type and mutant amino acids  
AA1\_PI: Isoelectric point of the wild-type amino acid  
AA2\_PI: Isoelectric point of the mutant amino acid  
deltaPI: Difference in isoelectric point between the wild-type and mutant amino acids  
AA1\_psic: Position specific conservation score of the wild-type amino acid  
AA2\_psic: Position specific conservation score of the mutant amino acid  
Delta\_psic: Difference in position specific conservation score between the wild-type and mutant amino acids  
evolutionary\_coupling\_avg: Average of EV-fold evolutionary coupling scores for the position  
gnomad\_allele\_freq: Minor allele frequency listed in gnomad  
clinvar\_disease: information in ClinVar "Condition(s)" column  
clinvar\_interpretation: information in ClinVar Clinical significance (Last reviewed) column.  
clinvar\_review: information in ClinVar Review status column.  
clinvar\_pathog: 1 means Pathogenic interpretation for variant in ClinVar.  
clinvar\_likely\_pathog: 1 means Likely Pathogenic interpretation for variant in ClinVar.  
clinvar\_uncertain: 1 means Variant of Uncertain Significance (VUS) for variant in ClinVar.  
clinvar\_likely\_benign: 1 means Likely Benign interpretation for variant in ClinVar.  
clinvar\_benign: 1 means Benign interpretation for variant in ClinVar  
cosmic\_count: Number of times the variant was observed in the Catalog of Somatic Mutations (COSMIC). Nonsense variants were not included in this column.  
cancer\_brain\_count: Number of times the variant was observed in Glioma or Glioblastoma in cancer genomics data  
cancer\_uterine\_count: Number of times the variant was observed in Uterine Cancers in cancer genomics data  
cancer\_breast\_count: Number of times the variant was observed in Breast Cancers in cancer genomics data  
cancer\_colorectal\_count: Number of times the variant was observed in Colorectal Cancers in cancer genomics data  
cancer\_nslc\_count: Number of times the variant was observed in Non-Small Cell Lung Cancers in cancer genomics data  
cancer\_endometrial\_count: Number of times the variant was observed in Endometrial Cancers in cancer genomics data  
cancer\_melanoma\_count: Number of times the variant was observed in Melanomas in cancer genomics data  
breast\_cancer\_expected: Frequency that the given variant was expected to occur based on the nucleotide mutational frequencies observed in breast cancer genomic data.  
uterine\_cancer\_expected: Frequency that the given variant was expected to occur based on the nucleotide mutational frequencies observed in uterine/endometrial cancer genomic data.  
endometrial\_cancer\_expected: Frequency that the given variant was expected to occur based on the nucleotide mutational frequencies observed in uterine/endometrial cancer genomic data.  
nslc\_cancer\_expected: Frequency that the given variant was expected to occur based on the nucleotide mutational frequencies observed in lung cancer genomic data.

colorectal\_cancer\_expected: Frequency that the given variant was expected to occur based on the nucleotide mutational frequencies observed in colorectal cancer genomic data.  
brain\_cancer\_expected: Frequency that the given variant was expected to occur based on the nucleotide mutational frequencies observed in glioma genomic data.  
melanoma\_cancer\_expected: Frequency that the given variant was expected to occur based on the nucleotide mutational frequencies observed in melanoma genomic data.  
pph2\_hdiv\_pred: 1 if pph2\_class (polyphen-2 class; HumDiv model) was “deleterious” / 0 if class was “neutral”  
pph2\_hvar\_pred: 1 if pph2\_class (polyphen-2 class; HumVar model) was “deleterious” / 0 if class was “neutral”  
mut\_assess\_pred: 1 if mutation assessor functional impact was “high” or “medium” / 0 if it was “low” or “neutral”  
provean\_pred: Provean prediction score.  
sift\_pred: 1 if Provean prediction was “damaging” / 0 if prediction was “tolerated”  
snap2\_pred: 1 if the Snap2 predicted effect was “effect” / 0 if prediction was “neutral”  
fathmm\_pred: 1 if FATHMM prediction was “damaging” / 0 if prediction was “tolerated”  
ptenpred: Predictor from PMID 27310656. 1 if prediction was “pathogenic” / 0 if prediction was “null”.  
lrt\_pred: 1 if prediction was “D(eleterious)” / 0 if prediction was “N(eutral)”.  
mut\_taster\_pred: 1 if prediction was “D(disease\_causing)” / 0 if prediction was “A(disease\_causing\_automatic)”.  
metasvm\_pred: 1 if prediction was “D(amaging)” / 0 if prediction was “T(olerated)”.  
metalr\_pred: 1 if prediction was “D(amaging)” / 0 if prediction was “T(olerated)”.  
mcap\_pred: 1 if prediction was “D(amaging)” / 0 if prediction was “T(olerated)”.  
cadd\_pred: 1 if CADD phred value was  $\geq 15$ , 0 if CADD phread value was  $< 15$ .  
predictor\_fraction\_deleterious: The fraction of predictions for a given variant that was “Damaging” or “Deleterious”.

rsid: reference SNP identification

tpmt\_allele: allele name based on the star nomenclature system.

single\_WTNorm: Mean of the WT-normalized geometric means of the EGFP:mCherry ratio obtained through testing TPMT variants individually

single\_StErr: Standard error of the WT-normalized geometric means of the EGFP:mCherry ratio obtained through testing TPMT variants individually

published\_western: see references in Supplementary Table 10

catalytic\_efficiency: see references in Supplementary Table 10

rbc\_assay: mean red blood cell activity for patients heterozygous for variant

average\_dose\_intensity: mean dose intensity for patients heterozygous for variant

variants\_scored: The number of variants with VAMP-seq scores at this position

abundance\_class: “intolerant” for positions in which 5 or more variants were scored and greater than 90% of variants at the position were below the threshold value separating the highest 95% of synonymous variants, and “enhanced” for positions in which 5 or more variants were scored, and more than 5 variants at the position had scores above the median of the synonymous distribution.

sc\_xca: the pdb file x-coordinate for the reactive oxygen and nitrogen atoms in the side-chain.

sc\_yca: the pdb file y-coordinate for the reactive oxygen and nitrogen atoms in the side-chain.

sc\_zca: the pdb file z-coordinate for the reactive oxygen and nitrogen atoms in the side-chain.



cosmic\_frequency: sum of all variants observed in COSMIC for that position

**Supplementary Table 3. P-values for enrichment of low abundance PTEN variants in the different ClinVar classes.** The P-values were calculating using a resampling test. Briefly, we drew n = 10,000 random samples, with replacement corresponding to the number of variants scored from each category in ClinVar (pathogenic = 25; likely pathogenic = 23; uncertain significance= 83) from the 1,366 PTEN missense variants (e.g. single nucleotide variants that change an amino acid) with abundance scores. We recorded the frequency of low abundance variants in each round of resampling. Then, we computed the P-value for each category by dividing the number of times the observed frequency of PTEN low-abundance variants fell below the frequencies of low-abundance variants in the resampled sets by 10,000. If the observed frequency of PTEN low-abundance variants never fell below the frequencies of low-abundance variants in the resampled sets, the P-value was listed as < 0.0001. Please see the R Markdown file supplied as Supplementary Data 5 to see the code used to perform this analysis.

Protein	ClinVar interpretation	P-value
PTEN	Pathogenic	< 0.0001
PTEN	Likely Pathogenic	0.0188
PTEN	Uncertain	0.3799

**Supplementary Table 4. Potential ClinVar pathogenicity reclassifications possible with abundance scores considered as ACMG/AMP PS3 criteria.**

**Supplementary Table 5. P-values for observed enrichment of PTEN variants or variant classes in various cancers over samplings based on cancer mutation spectra.** The P-values were calculating using a resampling test. For our statistical analysis of enrichments of low-abundance, dominant negative, or p.Pro38Ser variants in different cancer types, we first used the rates of single nucleotide transitions and transversions observed in TCGA to create mutational probabilities for every possible PTEN missense or nonsense variant. Based on these probabilities we drew n = 10,000 random samples of PTEN variants of size to equal the number of PTEN variants found in each cancer type (n = 337, 192, 153, 186, 77, 113, and 327 for brain, breast, colorectal, endometrial, melanoma, NSCLC, and uterine cancers, respectively). For each cancer type, this created the null distribution of PTEN variant frequencies based on the mutation spectrum alone. Then, for each cancer type, we computed the P-value by dividing the number of times the observed frequency of low-abundance, dominant negative or p.Pro38Ser variants fell below the frequency of the appropriate type of variants in the resampled sets by 10,000. If the observed frequency never fell below the frequency of the resampled set, the P-value was listed as < 0.0001. Please see the R Markdown file supplied as Supplementary Data 5 to see the code used to perform this analysis.

Protein	Cancer Group	Dominant negative P-value	Low abundance P-value	p.Pro38Ser variant P-value
PTEN	Colorectal	< 0.0001	0.0032	0.1978
PTEN	Brain	< 0.0001	< 0.0001	0.4296
PTEN	NSCLC	< 0.0001	< 0.0001	0.0638
PTEN	Endometrial	< 0.0001	0.0002	0.2284
PTEN	Uterine	< 0.0001	0.0032	0.3684
PTEN	Breast	< 0.0001	< 0.0001	0.2058
PTEN	Melanoma	0.0632	< 0.0001	< 0.0001

**Supplementary Table 6. Known or predicted destabilized variants in additional pharmacogenes and disease genes tested.** The potential destabilization conferred by PMS2 V159A was predicted based on protein data bank entry 1h7s.

Gene	Variant	allele	PMID
VKORC1	NP_076869.1:p.Arg98Trp	NA	24963046
CYP2C9	NP_000762.2:p.Arg335Trp	*11	15970795
CYP2C19	NP_000760.1p.Arg433Trp	*5	25001882
MLH1	LRG_216p1:p.Pro640Leu	NA	21404117
PMS2	LRG_161p1:p.Val159Ala	NA	NA
LMNA	LRG_254p2:p.Ile469Thr	NA	12057196;11901143
BRCA1	LRG_292p1:p.Met18Thr	NA	24234437
BRCA1	LRG_292p1:p.Cys44Ser	NA	24234437

**Supplementary Table 7. Oligonucleotide sequences used the study.** For BC-GPS-P7-i#-UMI, # is a specific index sequence per sample and N is random nucleotides.

GPS-landing-f	CGTCAGATCGCCTGGAGCAATTCCAC
BC-GPS-P7-i#-UMI	CAAGCAGAAGACGGCATAACGAGAT#####CANNNNNNNNNNTGGCTGGCAACTAGAAGGCACAGTCG
BC-TPMT-P5-v2	AATGATACGGCGACCACCGAGATCTACACTACAGAAAAGTAACTCGAGCATATGAC
P7	CAAGCAGAAGACGGCATAACGA
TPMT_Read1	CATATGACATGTCCTAGGCTTAAGCTAGC
TPMT_Read2	GAAGGCACAGTCGAGGCTGATCAGTCTAGA
TPMT_Index	GCCTCGACTGTGCCTTCTAGTTGCCAGCCA
eGFP1	GCGGGAGACGTGGAGTCCAACCCAGGGCCCATGGTGAGCAAGGGCGAG
eGFP2	TAGTGGATCCGAGCTCGGTACCAAGCTTAAgCTGTACAGCTCGTCCATGC
Inv_attB_GPS_AscI_R	CGATTGCGGGCGGCCCATAGAGCCACCGCATCC
Inv_attB_GPS_AscI_F	CGATTGCGGGCGGCCGTA AAAAGGCCGCTTGCTGGC
pb_SphI	/5phos/ATCTCTCTCTTTCCCTCCTCCGTTGTTGTTGTTGAGAGAGATCATG
pb_BsrGI	/5phos/GTACGATCTCTCTTTTCCCTCCTCCGTTGTTGTTGTTGAGAGAGATC
KAM499	GAGAACGTATGTCGAGGTAGGC
JJS_501a	GGGTTAGCAAGTGGCAGCCTGATCAGTTATCTAGATCCGGTGGAA
JJS_seq_F	AATGATACGGCGACCACCGAGATCTACACGAATTCACCGGCTGACCTC
JJS_seq_R1a	CAAGCAGAAGACGGCATAACGAGATGATGTACAGGGTTAGCAAGTGGCAGCCT
JJS_seq_R2a	CAAGCAGAAGACGGCATAACGAGATTGCTTTGGGGTTAGCAAGTGGCAGCCT
JJS_seq_R3a	CAAGCAGAAGACGGCATAACGAGATTAACAATACGGGTTAGCAAGTGGCAGCCT
JJS_seq_R4a	CAAGCAGAAGACGGCATAACGAGATATACGTGAGGGTTAGCAAGTGGCAGCCT
JJS_seq_R5a	CAAGCAGAAGACGGCATAACGAGATTCAGTTGGGGTTAGCAAGTGGCAGCCT
JJS_seq_R6a	CAAGCAGAAGACGGCATAACGAGATTTATCCTGGGGTTAGCAAGTGGCAGCCT
JJS_seq_R7a	CAAGCAGAAGACGGCATAACGAGATGACTCATGGGTTAGCAAGTGGCAGCCT
JJS_seq_R8a	CAAGCAGAAGACGGCATAACGAGATCTCTATACGGGTTAGCAAGTGGCAGCCT
JJS_seq_R9a	CAAGCAGAAGACGGCATAACGAGATATTCTCGAGGGTTAGCAAGTGGCAGCCT
JJS_seq_R10a	CAAGCAGAAGACGGCATAACGAGATCAATCTATGGGTTAGCAAGTGGCAGCCT
JJS_seq_R11a	CAAGCAGAAGACGGCATAACGAGATCGAGCGACGGGTTAGCAAGTGGCAGCCT
JJS_seq_R12a	CAAGCAGAAGACGGCATAACGAGATTCGATTATGGGTTAGCAAGTGGCAGCCT
JJS_read_1	GCGTGAGTAGGGTCGACCAAGAACCCTAGATGCGTTCGCTGTACAAATAGTT
JJS_index_1	GGGATCCACCGATCTAGATAACTGATCAGGCTGCCACTTGCTAACCC
JJS_read_2	CGCGGTACCGTCGACGGTTCGAGAAAAGCAAACGACTACTCGC
XbaI_SMRTBell	/5Phos/CTAGCTCTCTCTTTTCCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGAG
SacII_SMRTBell	/5Phos/ATCTCTCTCTTTTCCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGATGC
PTEN_BC_R	GATCAGTTATCTAGATCCGGTGGAT
PTEN_BC_F1.1	CTTAAGAATTCACCGGCTGACCTCCTTCTCCCTCTCTTCAGGTCTGCAATTGCGTGAGTAGGGTCGACCAAGAACCCTA GATGCGTGCCTGTACAAATAGTTNNNNNNNNNNNNNNNNNNNTGCGAGTAGTCGTTTGCTTTCTCGAACCGTCGACGGT ACCGCGGGCCCGGATCCACCGGATCTAGATAACTGATC
JJS_P5_(short)	AATGATACGGCGACCACC

**Supplementary Table 8. Statistics for variant-barcode subassemblies.**

Library	TPMT	TPMT fill-in	PTEN
SMRT cells	4	2	5
Reads with more than 3 passes	131498	53471	112076
Passed filters (mapQ, softclipping, correct size barcode)	121306	17073	111639
Unique barcodes	24234	8589	36562
Barcode has one consensus read	8784	3940	9744
Barcode has two consensus reads	6276	2441	8365
Barcode has three consensus reads	4111	1241	6386
Barcode has four or more consensus reads	5063	967	12067
All consensus reads identical	15059	4452	6287
Consensus read assigned by major allele	294	122	8579
Consensus read assigned by highest quality score tiebreaker	97	75	11952
Barcodes associated with WT or single amino acid substitution	19261	7155	22707
Barcodes associated with indel	4479	853	12369
Barcodes associated with > 1 aa substitution	494	290	1486

**Supplementary Table 9. PTEN variants with published abundance phenotypes.**

Variant	Destabilized	PMID
p.His123Tyr	no	10555148
p.Leu345Gln	yes	10555148
p.His93Ala	no	10555148
p.Cys124Gly	no	16645045
p.Gly129Glu	no	24349488
p.Cys124Ser	no	25527629
p.Lys62Arg	no	23475934
p.Lys125Glu	no	23475934
p.Cys136Arg	yes	23475934
p.Gly129Arg	yes	9356475
p.His93Arg	no	11156408
p.Tyr68His	yes	11156408
p.Leu345Gln	yes	11156408
p.Pro204Ser	yes	11156408
p.Leu186Val	yes	11156408
p.Gly251Cys	yes	11156408
p.Lys289Glu	no	11156408
p.Asp331Gly	yes	11156408
p.Ser227Phe	yes	11156408
p.Lys62Arg	no	23475934, 20926450
p.Tyr65Cys	no	20926450
p.Lys125Glu	no	23475934, 20926450
p.Arg233Ter	yes	23475934
p.Arg335Ter	yes	23475934
p.Asp252Gly	yes	25527629
p.Glu157Gly	no	25527629
p.His123Trp	no	25527629
p.Phe241Ala	yes	25527629
p.Asp326Asn	yes	25527629
p.Asn276Ser	yes	25527629
p.Gly129Arg	no	25527629
p.Ser370Ala	yes	25047839
p.Ser380Ala	yes	25047839
p.Ser385Ala	no	25047839
p.Cys124Ser	no	11948419
p.Arg130Gly	no	11948419
p.Tyr240Ala	no	11948419
p.Tyr315Ala	no	11948419
p.Glu388Ter	no	11948419
p.Thr398Ter	no	11948419
p.Tyr68Glu	yes	22891331

p.Tyr155Glu	yes	22891331
p.Tyr240Phe	no	22891331
p.Tyr240Glu	no	22891331
p.Tyr315Phe	no	22891331
p.Tyr315Glu	no	22891331
p.Thr366Ala	no	17444818
p.Ser370Ala	no	17444818
p.Cys124Ser	no	25647146
p.His93Arg	no	25647146

**Supplementary Table 10. TPMT variants with published abundance phenotypes and catalytic efficiencies.** The western blot and catalytic efficiency values are normalized to those of WT, which was set to a value of 1.

Variant	Allele	Western Blot	Catalytic Efficiency	PMID
WT	*1	1	1	NA
p.His227Gln	*7	1.24	0.1	13679074
p.Gly144Arg	*10	1.17	0.31	13679074
p.Ser125Leu	*12	1.41	0.27	13679074
p.Gly28Val	*13	0.92	0.57	13679074
p.Lys122Thr	*19	1	0.69	15652243
p.Arg163His	*16	1	0.32	15652243
p.Ala80Pro	*2	0.2	0.08	18708949
p.[(Ala154Thr; Tyr240Cys)]	*3A	NA	NA	18708949
p.Ala154Thr	*3B	0.8	NA	18708949
p.Tyr240Cys	*3C	1.1	0.52	18708949
p.Leu49Ser	*5	0.3	NA	18708949
p.Tyr180Phe	*6	1.2	0.17	18708949
p.His227Gln	*7	1.2	0.07	18708949
p.Arg215His	*8	1.3	1.13	18708949
p.Lys119Thr	*9	1.2	0.73	18708949
p.Gly144Arg	*10	0.9	0.49	18708949
p.Cys132Tyr	*11	0.9	0.03	18708949
p.Ser125Leu	*12	0.9	0.53	18708949
p.Gly28Val	*13	1	0.16	18708949
p.Arg163His	*16	1.3	0.16	18708949
p.Gln42Glu	*17	1	0.1	18708949
p.Gly71Arg	*18	0.85	NA	18708949
p.Lys122Thr	*19	1.25	0.59	18708949
p.Lys238Glu	*20	1	0.26	18708949
p.Leu69Val	*21	1.15	NA	18708949
p.Ala167Gly	*23	0.9	0.62	18708949
p.Gly36Ser	*30	1.1	0.09	18708949
p.Leu69Val	*21	0.95	0.24	18602085
p.Lys119Thr	*9	0.87	0.76	18602085
p.Gln179His	*24	0.94	0.26	18602085
p.Cys212Arg	*25	1.12	0.68	18602085

## Supplementary Note.

### Extended description of general reagents, DNA oligonucleotides and plasmids.

VAMP-seq evidence that supports pathogenicity can be used to alter variant clinical interpretations<sup>1</sup>. For example, of the 10 low-abundance, likely pathogenic ClinVar variants for PTEN, one variant (p.Ile335Lys) could be reclassified as pathogenic by adding the low-abundance classification to publically available information (**Supplementary Fig. 6**)<sup>1</sup>. Furthermore, 22 PTEN variants of uncertain significance along with 275 possible but not-yet-observed variants are low-abundance and could potentially be moved to the likely pathogenic category once observed in the appropriate clinical setting (**Supplementary Table 4**).

### Extended description of general reagents, DNA oligonucleotides and plasmids.

The PTEN open reading frame was obtained from 1066 pBabe puroL PTEN, which was a gift from William Sellers (Addgene plasmid # 10785), and combined with additional previously-used coding sequences<sup>2</sup> to create attB-EGFP-PTEN-IRES-mCherry-562bgl. This plasmid was modified through splitting of the EGFP coding sequence to create attB-sGFP-PTEN-IRES-mCherry-bGFP, which was used in assessing fluorescence ratios of WT or mutant PTEN using the split-GFP format<sup>3</sup>. The blasticidin resistance gene was obtained from pLenti CMV rtTA3 Blast (w756-1), which was a gift from Eric Campeau (Addgene plasmid # 26429), and fused C-terminally to mCherry to create attB-EGFP-PTEN-IRES-mCherry-BlastR. This construct was used to create the large panel of individually tested PTEN variants. The ampicillin resistance cassette in attB-EGFP-PTEN-IRES-mCherry-562bgl was replaced with a kanamycin resistance cassette to create attB-EGFP-PTEN-IRES-mCherry-562bgl-KanR, which was used to shuttle the mutagenized PTEN open reading frame in the library generation process. The PTEN coding region in attB-EGFP-PTEN-IRES-mCherry-562bgl was replaced to create the constructs used to test VKORC1 (IDT gBlock), MLH1, and LMNA. CYP2C9 and CYP2C19 plasmids were also created using the backbone of attB-EGFP-PTEN-IRES-mCherry-562bgl by replacing the PTEN coding sequence with CYP2C9 or CYP2C19 ORFs (IDT gBlocks) and moving the EGFP tag to the C-terminus of the protein. The MLH1 vector was additionally modified to create attB-EGFP-PMS2-2A-MLH1-IRES-mCherry, as MLH1 co-expression was necessary to observe signal with EGFP-fused PMS2. MLH1 was cloned from pCEP9 MLH1, which was a gift from Bert Vogelstein (Addgene plasmid # 16458)<sup>4</sup>. PMS2 was cloned from pSG5 PMS2-wt, which was a gift from Bert Vogelstein (Addgene plasmid # 16475)<sup>5</sup>. LMNA was cloned from pBABE-puro-GFP-wt-lamin A, which was a gift from Tom Misteli (Addgene plasmid # 17662)<sup>6</sup>. pCAG-NLS-HA-Bxb1 was a gift from Pawel Pelczar (Addgene plasmid # 51271)<sup>7</sup>. The attB\_mCherry\_P2A\_MCS plasmid was built from the pcDNA5/FRT/TO backbone (ThermoFisher). mCherry\_P2A was synthesized (gBlocks, IDT) and EGFP amplified from pHAGE-CMV-eGFP-N (gift from Alejandro Balazs) using primers eGFP1 and 2 was added by Gibson assembly. Wild-type TPMT (NM\_000367.3) was synthesized (gBlocks, IDT) and cloned in-frame with the EGFP by Gibson Assembly. The CMV promoter was replaced with the synthesized AttB sequence (gBlocks, IDT). The final vector was shorted by removing all of the intervening sequence between the E.Coli Ori and the BGH poly-A signal that follows the EGFP-X fusion by inverse PCR with Inv\_attB\_GPS\_AscI\_R and Inv\_attB\_GPS\_AscI\_F, cutting with AscI and religation. Single amino acid mutations were made using the same inverse PCR method described below.



## Extended description of the construction of barcoded, site-saturation mutagenesis libraries for TPMT and PTEN

For TPMT, wild type TPMT was first cloned into pUC19. Next, for each codon, mutagenic primers were ordered with machine-mixed NNK bases at the 5' end of the sense oligonucleotide. Mutagenized TPMT was cloned into the Hind-III/Xho-I sites of attB\_mCherry\_P2A\_MCS. A 15 base, degenerate barcode was then cloned into the XbaI site of the multiple cloning site by Gibson Assembly<sup>8</sup>. Owing to poor coverage in the initial library, a separate "fill-in" library was constructed for TPMT amino acids 192-239 by the same protocol. Colony counts revealed approximately 40,000 and 10,000 barcode clones for the main TPMT and TPMT fill-in plasmid libraries respectively.

For PTEN, eight randomly chosen codons were used to optimized inverse PCR amplification, using attB-EGFP-PTEN-IRES-mCherry-562bgl as the template. Template concentrations between 0.02 pg through 20,000 pg were used to identify the minimum amount of template needed to see bands on an agarose gel after 20 cycles using primer concentrations between 0.25 and 0.5  $\mu$ M. The final concentrations were 250 pg of template plasmid and 0.25  $\mu$ M of forward and reverse primers. Each codon amplification was done in a total volume of 10  $\mu$ L using 20 cycles at the standard conditions recommended for Kapa HiFi (95°C for 3 minutes followed by 20 cycles of 98°C for 20s, 60°C for 15s and 72°C for 30s/kb of template plasmid, followed by a final extension of 5 min). Two  $\mu$ L of each amplified product were run on a 0.7% agarose gel for visual validation of amplification, and the remaining 8  $\mu$ L of product was diluted 1:10 with water. Two  $\mu$ L of this diluted product was quantified using PicoGreen (ThermoFisher) on a BioTek H1 plate reader. PicoGreen measurements were ignored for codons where multiple amplified bands of multiple sizes were observed, and instead replaced by PicoGreen measurements for adjacent codons with amplified bands of the intended size of similar intensity to the amplified band of the intended size for the codon in question. Based on these PicoGreen-derived concentrations, all amplicons were mixed together so that approximately equal amounts of the bands of intended size were present for all amplified codons. This final mixture of the library was cleaned and concentrated by ethanol precipitation. The precipitated product was resuspended in 100  $\mu$ L of ddH<sub>2</sub>O. To phosphorylate the amplified product, 16  $\mu$ L of cleaned product at  $\sim$  11.5 ng/ $\mu$ L was mixed with 2  $\mu$ L of 10x T4 DNA ligase buffer (New England Biolabs) and 2  $\mu$ L of T4 PNK enzyme, and incubated at 37°C for 1 hour. To circularize the amplified product, the entire 20  $\mu$ L reaction was then mixed with 4  $\mu$ L 10x T4 DNA ligase buffer, 14  $\mu$ L of ddH<sub>2</sub>O, and 2  $\mu$ L of T4 DNA ligase, incubated at 16°C for 1 hour, 25°C for 10 min, and heat inactivated at 65°C for 10 min. Residual template plasmid was then removed by adding 1  $\mu$ L of DPN1 enzyme to the tube, and incubated at 37°C for 1 hour. The ligated product was cleaned and concentrated into a final 6  $\mu$ L volume using a Zymo Clean and Concentrate kit, and then transformed into NEB 10-beta electrocompetent *E. coli*. To select against input plasmid and plasmids containing short PCR products, the library was then shuttled into attB-EGFP-PTEN-IRES-mCherry-562bgl-KanR via directional cloning using XbaI and EcoRI. Barcodes were added to the library by filling in a long oligo (PTEN\_BC\_F1.1) supplemented with a short reverse oligo (PTEN\_BC\_R) using Klenow(-exo) polymerase. Here, 0.25  $\mu$ M of PTEN\_BC\_F1.1 and PTEN\_BC\_R were melted and annealed together at 98°C for 3 minutes in Buffer 2.1 (New England Biolabs) and cooled to 25°C at a rate of  $-$  0.1°C/sec. 4000 units of Klenow(-exo) and 0.033  $\mu$ M dNTP's were added, and the mixture was incubated for 15 minutes at 25°C. The polymerase was inactivated by incubating for 20 minutes at 70°C, and the product was cooled to 37°C at a rate of  $-$ 0.1°C/sec. The cooled product was then digested with EcoRI and SacII in Buffer 2.1, purified with a Zymo Clean and Concentrate kit, and eluted in 30  $\mu$ L of ddH<sub>2</sub>O. To digest the mutagenized PTEN library in the attB-EGFP-PTEN-IRES-mCherry-562bgl-

KanR vector, 2 µg of plasmid was mixed with 5 µl of 10x Cutsmart buffer, 1 µl EcoRI-HF, and 1 µl Sac-II in a 50 µl reaction, digested at 37°C for 1 hour, and purified with a Zymo Clean and Concentrate kit. Both purified digestion products were mixed together, ligated with T4 DNA ligase, purified with a Zymo Clean and Concentrate kit, and transformed into NEB 10-beta electrocompetent *E. coli* (New England Biolabs). Colony counts estimated this library to contain roughly 35,200 barcodes.

In NNK mutagenesis schemes like the one we employed, synonymous variants can be generated at 50 of the 61 amino acid-coding codons that may exist in the template sequence. Notably, the following codons in the template sequence preclude generation of a synonymous variant at that position: ATG (M), ATT (I), TTT (F), GAG (E), GAT (D), AAG (K), AAT (N), CAG (Q), CAT (H), TAT (Y), and TGT (C). Thus, synonymous variants were theoretically possible at 272 and 167 codons for the PTEN and TPMT proteins, respectively. Of these, synonymous variants were observed at 151 PTEN and 138 TPMT codons in our final data set.

### **Extended description of Single Molecule Real Time (SMRT) sequencing to link each TPMT and PTEN variants to its barcode.**

To prepare the circular SMRT-bell templates<sup>9</sup>, library plasmids were digested with restriction enzymes to release the barcode and open reading frame. Hairpin SMRT-bell oligonucleotides with complementary sticky ends and SMRT priming sequences were ligated to the fragments. TPMT libraries were digested using BsrGI and SphI. The correct fragment was size-selected on 1% agarose and gel-purified with NEB Monarch DNA Gel Extraction kit (New England Biolabs). Custom SMRT bell adapters pb\_SphI and pb\_BsrGI were sticky-end ligated to the purified fragment. To make a working stock of 20 µM SMRT bell adaptors in 10 mM Tris, 0.1 mM EDTA, 100 mM NaCl, they were heated to 85°C and snap cooled on ice. The ligation reaction contained 500 ng purified fragment, 2.5 µM of each adaptor, 1 µL of BsrGI, 1 µL of SphI, 1X ligase buffer, and 2 µL of T4 ligase in a 40 µL reaction. The ligation was performed at room temperature for 2 hours, then heat inactivated at 65°C for 20 minutes. 1 µL each of ExoIII and ExoVII were added and incubated at 37°C for 1 hour. The final SMRT bell fragments were purified via AmpurePB (Pacific Biosciences) at 1.8X concentration, washed in 70% ethanol, eluted in 15 µL 10mM Tris and quantified by BioAnalyzer (Agilent). The PTEN library was digested using SacII and XbaI. The correct fragment was size-selected on 1% agarose and gel-purified with a Qiagen Gel Extraction kit (Qiagen). Custom SMRT bell adapters XbaI\_SMRTBell and SacII\_SMRTBell were sticky-end ligated to ~150 ng of the purified fragment in a 50 µL reaction using 1x T4 DNA ligase buffer, 1 µM of each oligo, 800 units of T4 DNA ligase, 5 units of SacII, and 5 units of XbaI. The ligation was performed at room temperature for 30 minutes, then heat inactivated at 65°C for 10 minutes. Ten units of Exonuclease VII (ThermoFisher) and 100 units of Exonuclease III (Enzymatics) were added to the mixture, incubated for 30 mins at 37°C. The final SMRT bell fragments were purified with AmpurePB (Pacific Biosciences) at 1.8X concentration, washed twice in 70% ethanol, eluted in 20 µL 10mM Tris, and quantified using a QuBit (ThermoFisher) and BioAnalyzer (Agilent).

The TPMT and PTEN constructs were sequenced on a Pacific Biosciences RS II sequencer. The main TPMT library was sequenced using four SMRT cells and the fill-in TPMT library was sequenced using two. The PTEN library was sequenced using five SMRT cells. Base call files were converted from the bax format to the bam format using bax2bam (version 0.0.2) and then bam files for each library from

separate lanes were concatenated. Consensus sequences for each sequenced molecule in every library were determined using the Circular Consensus Sequencing 2 algorithm (version 2.0.0) with default parameters (bax2bam and ccs can be found on Github, <https://github.com/PacificBiosciences/unanimity/blob/master/doc/PBCCS.md>). Each resulting consensus sequence was then aligned to either the TPMT or PTEN reference sequence using Burrows-Wheeler Aligner<sup>10</sup> (<http://bio-bwa.sourceforge.net/>). Barcodes and insert sequences were extracted from each alignment using custom scripts that parsed the CIGAR and MD strings. For barcodes sequenced more than once, if barcode-variant sequences differed, the barcode was assigned to the variant that represented more than 50% of the sequences. Barcodes lacking a majority variant sequence were assigned the variant sequence with the highest average quality score as determined by the ccs2 algorithm. The barcode-variant extraction and barcode unification scripts can be found at <https://github.com/shendurelab/AssemblyByPacBio/>. Metrics regarding the processing of sequencing data for the barcode-variant assignments can be found in **Supplementary Table 8**. The final TPMT libraries have 26,416 barcodes associated with 6,251 full-length nucleotide sequence variants that encoded 3,994 unique protein sequences with zero or one amino acid change. The final PTEN library had 22,707 barcodes associated with 7,756 full-length nucleotide sequence variants that encoded 5,043 unique protein sequences with zero or one amino acid change. For both TPMT and PTEN a barcode-variant map file was created that contains each barcode and its nucleotide sequence.

### **Extended description of the integration of single variant clones or barcoded libraries into the HEK293-landing pad cell line.**

These cells harbor exactly one copy of a tet-inducible promoter followed by a Bxb1 recombinase site. Integration of a promoterless plasmid containing a Bxb1 recombinase site results in expression of one variant per cell. First, FuGENE 6 (Promega) was used to transfect the Bxb1 recombinase-expressing pCAG-NLS-HA-Bxb1 plasmid, followed 24-48 hours later by the single variant or library plasmid. Two days after transfection, variant expression was induced by adding 0.5-2  $\mu\text{g}/\text{mL}$  doxycycline to the media (DMEM + 10% FBS). Then, cells were prepared for sorting by lifting from 10 cm plates with Versene solution (0.48 mM EDTA in PBS), washing 1X in PBS, resuspending in sort buffer (1X PBS + 1% heat-inactivated FBS, 1 mM EDTA and 25 mM HEPES pH 7.0) and filtering through 35  $\mu\text{m}$  nylon mesh. Cells were sorted on a BD Aria III FACS machine using an 85 or 100  $\mu\text{m}$  nozzle. mTagBFP2, expressed from the unrecombined landing pad, was excited with a 405 nm laser, and emitted light was collected after passing through a 450/50 nm band pass filter. EGFP, expressed after successful recombination of the variant or library plasmid, was excited with a 488 nm laser, and emitted light was collected after passing through 505 nm long pass and 530/30 nm band pass filters. mCherry, also expressed after successful recombination of the variant or library plasmid was excited with a 561 nm laser, and emission was detected using 600 nm long pass and 610/20 band pass filters. Before analysis of fluorescence, live, single cells were gated using FSC-A and SSC-A (for live cells) or FSC-A and FSC-H (for single cells) signals. Recombinant mTagBFP2 negative, mCherry positive cells were isolated, with mCherry fluorescence values at least 10 times higher than the median fluorescence value of negative or control cells, and mTagBFP2 fluorescence at least 10 times lower than the median of the unrecombined mTagBFP2 positive cells (**See Supplementary Fig. 1a for gating example**). Multiple replicate integrations were conducted and sorted for recombinants (**Supplementary Table 1**). After sorting, the libraries were uniformly mTag2BFP negative and mCherry positive. Analytical flow cytometry was performed with a BD

LSR II flow cytometer, equipped with filter sets identical to those described for the Aria III, with the exception of mCherry emission which was detected using 595nm long pass and 610/20 band pass filters.

### **Extended description of the assessment of the PTEN library composition.**

Two reactions were independently performed from the same plasmid preparation and served as technical replicates. Each 50  $\mu$ L first-round PCR reaction was prepared with a final concentration of  $\sim$ 50 ng/ $\mu$ L input plasmid DNA, 1x Kapa HiFi ReadyMix, and 0.25  $\mu$ M each of the JJS\_seq\_F/JJS\_501a primers. The reaction conditions were 95  $^{\circ}$ C for 3 minutes, 98  $^{\circ}$ C for 20 seconds, 60  $^{\circ}$ C for 15 seconds, 72  $^{\circ}$ C for 15 seconds, repeat 5 times, 72  $^{\circ}$ C for 2 minutes, 4  $^{\circ}$ C hold. The reaction was bound to AMPure XP beads (Beckman Coulter), cleaned, and eluted with 16  $\mu$ L water. 15  $\mu$ L of the eluted volume was mixed with 2x Kapa Robust ReadyMix; JJS\_P5\_(short) and either JJS\_seq\_R1a for technical replicate 1 or JJS\_seq\_R2a for technical replicate 2 were added at 0.25  $\mu$ M each. Reaction conditions for the second round PCR were 95  $^{\circ}$ C for 3 minutes, 95  $^{\circ}$ C for 15 seconds, 60  $^{\circ}$ C for 15 seconds, 72  $^{\circ}$ C for 30 seconds, repeat 19 times, 72  $^{\circ}$ C for 1 minutes, 4  $^{\circ}$ C hold. Amplicons were extracted after separation on a 1.5% TBE/agarose gel using a Quantum Prep Freeze 'N Squeeze DNA Gel Extraction Kit (Bio-Rad). Extracted amplicons were quantified using a KAPA Library Quantification Kit (Kapa Biosystems) and sequenced on a NextSeq 500 using a NextSeq 500/550 High Output v2 75 cycle kit (Illumina), using primers JJS\_read\_1, JJS\_index\_1, and JJS\_read\_2. Sequencing reads were converted to FASTQ format and de-multiplexed with bcl2fastq. Barcode paired sequencing reads were joined using the fastq-join tool within the ea-utils package. Enrich2 was used to count the barcodes in the reads, using a minimum quality filter of 20.

High correlation (Pearson's  $r = 0.99$ ) of barcode counts was observed between technical replicate amplifications (**Supplementary Fig. 11a**). After barcode counts in both replicates were combined, a minimum count filter of 200 was imposed to remove barcodes arising from sequencing error (**Supplementary Fig. 11b**). Each barcode's count was divided by the total number of barcode reads passing this filter to obtain frequencies for each barcode. Using the barcode-variant map generated by PacBio subassembly, a protein sequence was assigned to each barcode. Barcodes missing from the barcode-variant map were categorized as "Not subassembled". The frequency of each type of sequence was determined (**Supplementary Fig. 11c**). The composition of the single amino acid variants in the library were next analyzed to determine sources of potential library bottlenecks. The nucleotide frequencies at each mutated codon were determined (**Supplementary Fig. 11d**), and relative frequencies of each amino acid variant observed in the library were calculated (**Supplementary Fig. 11e**). Single amino acid substitution coverage was determined for each position along the protein (**Supplementary Fig. 11f and 11g**). Lastly, the distribution of single amino acid variants within the library was determined (**Supplementary Fig. 11h**), and simulations of sample sizes required to observe each PTEN single amino acid variant were performed (**Supplementary Fig. 11i**).

### **Detailed analysis of the sources of PTEN variant loss in the VAMP-seq pipeline.**

**Loss at the site-saturation mutagenesis step:** To create the site-saturation mutagenesis library we employed the inverse PCR-based method<sup>11</sup>. There are two major sources of loss of library uniformity in this protocol. First, each inverse PCR reaction is performed separately using individually manufactured oligonucleotides (unlike a method like PALS<sup>12</sup>). Therefore, failed individual reactions and uneven mixing

of the PCR products causes the complete or nearly complete loss of mutations at some positions (**Supplementary Fig. 11f and g**). Secondly, a strong G bias at each degenerate nucleotide position occurred during synthesis, which led to a bias in the amino acids that were created (**Supplementary Fig. 11d and e**, we believe machine mixing by IDT, the oligo supplier, was likely the culprit here). High throughput sequencing of the final barcoded plasmid library (used for cell integration) showed that only ~56% of the plasmid molecules contained single-amino acid substitutions that would be scored in the VAMP-seq experimental pipeline (**Supplementary Fig. 11a, b, and c**). This is analogous to the PALS method where only ~50% (at best) of variants are indel free or not contaminating WT sequences<sup>13</sup>. We note that tile-based library generation methods can lead to higher coverage but have their own drawbacks.

**Loss at the barcoding step:** Given the limitations of experiments based on FACS sorting, which largely arise from machine time vs. number of cells sorted, we decided to limit the number of barcodes/variants that went into the VAMP-seq experiments so that each barcode would be adequately represented in each experiment. Therefore, we bottlenecked the site-saturation mutagenesis library to contain ~40k barcodes, based on colony counts. This bottlenecking step likely prevented low abundance variants from being represented in the final library.

**Loss at the cell integration step and VAMP-seq experiments:** There was also likely loss at the cellular integration step. The distribution of the single amino acid variants in the final barcoded library was far from uniform, and instead exhibited a log-normal distribution with a  $\log_{10}$  standard deviation of ~1 (**Supplementary Fig. 11h**). Given this distribution of variant frequencies in the library, ~100,000 recombinants would theoretically have been needed for full coverage of the ~5,000 protein variants present in the library prep used in the transfections (**Supplementary Fig. 11i**). We actually obtained between an estimated 108,000 and 250,000 recombinant cells following transfection, from which 110,000 cells were collected by FACS to create a relatively pure set of recombined cells preceding each VAMP-seq experiment. Lastly, with the exception of one experiment, greater than 400,000 cells were collected in each bin of each VAMP-seq experiment. Thus, while most experiments exceeded the 100,000 or more cells needed to observe most of the ~5,000 variants confirmed in our library prep, another ~1,000 variants were either lost or represented too infrequently in the sorted pools of cells to yield reproducible scores, and were removed by the frequency filter.

In summary, ~3,000 of the ~8,000 possible PTEN variants appear to have been completely lost at the library generation and barcoding step, while an additional ~1,000 variants were lost at minor bottlenecks occurring at the subsequent steps, including the frequency filter employed during the analysis.

### **Extended description of variant annotations obtained from online databases.**

We collected structural feature information, including absolute solvent accessibilities, using DSSP<sup>14,15</sup> based on PDB structure 1d5r for PTEN and 2h11 for TPMT. For each amino acid in both proteins, we divided the absolute solvent accessibility derived from DSSP by the empirically determined maximum accessibility of that amino acid to yield relative solvent accessibility<sup>16</sup>. The COSMIC (Catalogue of Somatic Mutations in Cancer) release v81 was used for the analyses we presented<sup>17</sup>. Cancer genomics data including those from The Cancer Genome Atlas and AACR Project GENIE<sup>18</sup> data was accessed from

cBioPortal<sup>19</sup> on 2/15/2017 and 2/21/2017, respectively. PTEN variants observed in the GBM, LGG-GBM, and Glioma cancer categories were combined into a single brain cancer category for the analysis. ClinVar<sup>20</sup> data was accessed on 6/29/2017 and filtered to exclude everything except germline missense and nonsense variants. Average evolutionary coupling<sup>21</sup> values by position were calculated using data from <http://evfold.org/>. Mutational spectra from the six transition or transversion categories for breast adenocarcinoma, lung squamous cell carcinoma, uterine corpus endometrial carcinoma, glioblastoma multiforme, colon and rectal carcinoma, ovarian serous carcinoma<sup>22</sup>, and melanoma<sup>23</sup> were used to create expected PTEN variant frequency distributions. Minor allele frequencies were extracted from the GnomAD database (Feb. 2017 release)<sup>24</sup>. TPMT allele names and RSID numbers were taken from <http://www.imh.liu.se/tpmtalleles/tabell-over-tpmt-alleler?l=en>. The PTEN variant effect predictions were obtained from Polyphen-2 (<http://genetics.bwh.harvard.edu/pph2/>)<sup>25</sup>, Provean (<http://provean.jcvi.org/>)<sup>26</sup>, SIFT (<http://sift.jcvi.org/>)<sup>27</sup>, Snap2 (<https://roslab.org/services/snap2web/>)<sup>28</sup>, Mutation assessor (<http://mutationassessor.org/r3/>)<sup>29</sup>, and FATHMM (<http://fathmm.biocompute.org.uk/>)<sup>30</sup> by querying their respective websites. PSIC scores<sup>31</sup> were obtained from the Polyphen-2 output. PTENpred<sup>32</sup> was downloaded and all predictions were run locally. The predictions for LRT<sup>33</sup>, Mutation Taster<sup>34</sup>, MetaSVM<sup>35</sup>, MetaLR<sup>35</sup>, MCap<sup>36</sup>, and CADD<sup>37</sup> were collected with dbNSFP<sup>38</sup>, which was downloaded and run locally.

### **Extended description of western blotting procedures.**

HEK 293T TetBxb1BFP Clone4 or Clone37 cells<sup>2</sup> were transfected with the pCAG-NLS-HA-Bxb1 expression vector and either an attB-PTEN-HA-IRES-mCherry plasmid encoding a PTEN variant or an attB-mCherry\_2A\_GFP plasmid encoding a TPMT variant. Two days after transfection, cells were switched to media containing 2 µg/mL doxycycline. For each variant, approximately 8,000 mTagBFP2 negative, mCherry positive cells were sorted using a FACSAriaIII sorter (BD Biosciences), and allowed to grow to confluence in 6-well plates with Dox-containing media. Cells expressing PTEN variants were then collected with Trypsin-EDTA, washed in PBS, and incubated with lysis buffer (20 mM Tris pH 8.0, 150 mM NaCl, 1% Triton X-100, and Protease Inhibitor Cocktail (Sigma-Aldrich)) for 10 minutes at 4 °C. The tubes were centrifuged at 21,000 x g for 5 minutes, the supernatant was collected, and protein concentration was determined by the DC Protein assay (Bio-Rad) against a standard curve of bovine serum albumin. 40 µg of protein was loaded per well of a NuPage 4-12% Bis-Tris gel (Invitrogen) in MOPS buffer, using Spectra Multicolor Broad Range Protein Ladder (ThermoFisher Scientific) for size comparison. Proteins were transferred to a PVDF membrane using a GenieBlotter (Idea Scientific). Western blotting was performed using a 1:2,000 dilution of anti-phospho-AKT (p.Thr308; 13038; Cell Signaling Technology) followed by detection with a 1:10,000 dilution of anti-rabbit-HRP (NA934V; GE Healthcare); a 1:2,000 dilution of anti-pan-AKT (2920; Cell Signaling Technology) followed by detection with a 1:10,000 dilution of anti-mouse-HRP (NA931V; GE Healthcare); a 1:4,000 dilution of anti-GFP antibody (11814460001;Roche), followed by detection with a 1:10,000 dilution of anti-mouse-HRP; 1:5,000 dilution of anti-HA-HRP (3F10; Roche); or a 1:5,000 dilution of anti-beta-actin-HRP (ab8224; Abcam), using the SuperSignal™ West Dura extended duration substrate (ThermoFisher Scientific).

TPMT expressing cells were removed from the plate with cold PBS, pelleted and resuspended in lysis buffer (50 mM Tris pH 8.0, 150 mM NaCl, 1% NP-40, and Protease Inhibitor Cocktail (Roche)). Protein concentration was determined by Bradford Assay (Bio-Rad). 45, 15 and 5 µg of lysate was loaded per well of a NuPage 4-12% Bis-Tris gel (Invitrogen) in MOPS buffer, using SeeBlue Plus2 Protein Ladder

(ThermoFisher Scientific) for size comparison. Proteins were transferred to a PVDF membrane using a GenieBlotter (Idea Scientific). Western blotting was performed using a 1:3,000 dilution of anti-GFP antibody (11814460001; Roche) followed by detection with a 1:10,000 dilution of anti-mouse-HRP (NA934V; GE Healthcare) or a 1:5,000 dilution of anti-beta-actin-HRP (ab8224; Abcam), using the SuperSignal™ West Dura extended duration substrate (ThermoFisher Scientific).

## Supplementary References

1. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K. & Rehm, H. L. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–423 (2015).
2. Matreyek, K. A., Stephany, J. J. & Fowler, D. M. A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Res.* **45**, e102 (2017).
3. Cabantous, S., Terwilliger, T. C. & Waldo, G. S. Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. *Nat. Biotechnol.* **23**, 102–107 (2005).
4. Papadopoulos, N., Nicolaides, N. C., Wei, Y. F., Ruben, S. M., Carter, K. C., Rosen, C. A., Haseltine, W. A., Fleischmann, R. D., Fraser, C. M. & Adams, M. D. Mutation of a mutL homolog in hereditary colon cancer. *Science* **263**, 1625–9 (1994).
5. Nicolaides, N. C., Littman, S. J., Modrich, P., Kinzler, K. W. & Vogelstein, B. A naturally occurring hPMS2 mutation can confer a dominant negative mutator phenotype. *Mol. Cell. Biol.* **18**, 1635–1641 (1998).
6. Scaffidi, P. & Misteli, T. Lamin A-dependent misregulation of adult stem cells associated with accelerated ageing. *Nat. Cell Biol.* **10**, 452–459 (2008).
7. Hermann, M., Stillhard, P., Wildner, H., Seruggia, D., Kapp, V., Sánchez-Iranzo, H., Mercader, N., Montoliu, L., Zeilhofer, H. U. & Pelczar, P. Binary recombinase systems for high-resolution conditional mutagenesis. *Nucleic Acids Res.* **42**, 3894–3907 (2014).
8. Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. a, Smith, H. O., Iii, C. A. H. & America, N. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–5 (2009).
9. Travers, K. J., Chin, C. S., Rank, D. R., Eid, J. S., Turner, S. W., Biosciences, P. & Park, M. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* **38**, e159 (2010).
10. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
11. Jain, P. C. & Varadarajan, R. A rapid, efficient, and economical inverse polymerase chain reaction-based method for generating a site saturation mutant library. *Anal. Biochem.* **449**, 90–8 (2014).
12. Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively parallel single- amino-acid mutagenesis. (2015). doi:10.1038/nmeth.3223
13. Starita, L. M., Young, D. L., Islam, M., Kitzman, J. O., Gullingsrud, J., Hause, R. J., Fowler, D. M., Parvin, J. D., Shendure, J. & Fields, S. Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* **200**, 413–422 (2015).
14. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–637 (1983).
15. Touw, W. G., Baakman, C., Black, J., Beek, T. A. H., Krieger, E., Joosten, P., Vriend, G., Te Beek, T. A. H., Krieger, E., Joosten, R. P. & Vriend, G. A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.* **43**, D364–D368 (2015).
16. Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J. & Wilke, C. O. Maximum allowed solvent accessibilities of residues in proteins. *PLoS One* **8**, (2013).
17. Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C. G., Ward, S., Dawson, E., Ponting, L., Stefancsik, R., Harsha, B., YinKok, C., Jia, M., Jubb, H., Sondka, Z., Thompson, S., De, T.



- & Campbell, P. J. COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
18. AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discov.* **7**, 818–831 (2017).
  19. Gao, J., Aksoy, B., Dogrusoz, U. & Dresdner, G. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, 1–20 (2013).
  20. Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M. & Maglott, D. R. ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, 980–985 (2014).
  21. Marks, D. S., Hopf, T. a & Sander, C. Protein structure prediction from sequence variation. *Nat. Biotechnol.* **30**, 1072–80 (2012).
  22. Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. a, Leiserson, M. D. M., Miller, C. a, Welch, J. S., Walter, M. J., Wendl, M. C., Ley, T. J., Wilson, R. K., Raphael, B. J. & Ding, L. Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–9 (2013).
  23. Krauthammer, M., Kong, Y., Ha, B. H., Evans, P., Bacchicchi, A., McCusker, J. P., Cheng, E., Davis, M. J., Goh, G., Choi, M., Ariyan, S., Narayan, D., Dutton-Regester, K., Capatana, A., Holman, E. C., Bosenberg, M., Sznol, M., Kluger, H. M., Brash, D. E., *et al.* Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. *Nat. Genet.* **44**, 1006–14 (2012).
  24. Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
  25. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* 7.20.1-7.20.41 (2013). doi:10.1002/0471142905.hg0720s76
  26. Choi, Y. & Chan, A. P. PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**, 2745–2747 (2015).
  27. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
  28. Hecht, M., Bromberg, Y. & Rost, B. Better prediction of functional effects for sequence variants. *BMC Genomics* **16**, S1 (2015).
  29. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Res.* **39**, (2011).
  30. Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., Day, I. N. M. & Gaunt, T. R. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Hum. Mutat.* **34**, 57–65 (2013).
  31. Sunyaev, S. R., Eisenhaber, F., Rodchenkov, I. V., Eisenhaber, B., Tumanyan, V. G. & Kuznetsov, E. N. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. **12**, 387–394 (1999).
  32. Johnston, S. B. & Raines, R. T. PTENpred: A Designer Protein Impact Predictor for PTEN-related Disorders. *J. Comput. Biol.* **23**, 1–7 (2016).
  33. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Identif. deleterious Mutat. within three Hum. genomes.* **19**, 1553–1561 (2009).
  34. Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* **11**, 361–2 (2014).

35. Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K. & Liu, X. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **24**, 2125–2137 (2015).
36. Jagadeesh, K. A., Wenger, A. M., Berger, M. J., Guturu, H., Stenson, P. D., Cooper, D. N., Bernstein, J. A. & Bejerano, G. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.* **48**, 1581–1586 (2016).
37. Kircher, M., Witten, D. M., Jain, P., O’Roak, B. J., Cooper, G. M. & Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–5 (2014).
38. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum. Mutat.* **37**, 235–241 (2016).