ARTICLES
https://doi.org/10.1038/s41588-018-0122-z
nature genetics

Multiplex assessment of protein variant abundance by massively parallel sequencing

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

Determining the pathogenicity of genetic variants is a critical challenge, and functional assessment is often the only option. Experimentally characterizing millions of possible missense variants in thousands of clinically important genes requires generalizable, scalable assays. We describe variant abundance by massively parallel sequencing (VAMP-seq), which measures the effects of thousands of missense variants of a protein on intracellular abundance simultaneously. We apply VAMP-seq to quantify the abundance of 7,801 single-amino-acid variants of PTEN and TPMT, proteins in which functional variants are clinically actionable. We identify 1,138 PTEN and 777 TPMT variants that result in low protein abundance, and may be pathogenic or alter drug metabolism, respectively. We observe selection for low-abundance PTEN variants in cancer, and show that p.Pro38Ser, which accounts for ~10% of PTEN missense variants in melanoma, functions via a dominant-negative mechanism. Finally, we demonstrate that VAMP-seq is applicable to other genes, highlighting its generalizability.

Every possible nucleotide change that is compatible with life is likely present in the germline of a living human[1]. Some of these variants alter protein activity or abundance, and, consequently, may impact disease risk. However, only ~2% of all presently reported germline missense variants have clinical interpretations[2,3]. Most of the remaining variants, as well as nearly all missense variants not yet observed, are rare and cannot be interpreted using traditional genetic approaches. Computational approaches are insufficiently accurate, and somatic alterations further complicate the picture. These limitations create a major challenge for the clinical use of genomic information.

Deep mutational scans, which enable the simultaneous functional characterization of thousands of missense variants of a protein, offer one potential solution to the variant interpretation problem[4–6]. For example, the effects of nearly all possible single-amino-acid variants of the RING domain of BRCA1 on its E3 ligase and BARD1-binding activity were quantified in a single study[7]. In another example, the effects of all possible single-amino-acid variants of PPARγ on the expression of CD36 in response to different agonists were measured[8]. In both cases, the functional data enabled accurate identification of most known pathogenic variants, suggesting that these data could be useful in interpreting newly observed variants.

So far, deep mutational scans, including of BRCA1 and PPARγ, have relied on assays specific for each protein's molecular function. However, developing specific assays for each of the thousands of disease-related proteins is impractical. To overcome this challenge, we sought to devise a functional assay that was both informative of variant effect and generalizable to many proteins. We based our assay on the fact that most proteins, despite their diversity, must be abundant enough to perform their molecular function. Variants can interfere with steady-state protein abundance in cells via a variety of mechanisms, including by diminishing thermodynamic stability, altering post-transcriptional regulation or interrupting trafficking. In fact, as much as 75% of the pathogenic variation in monogenic disease is thought to disrupt thermodynamic stability and, consequently, alter abundance[9,10]. Furthermore, low-abundance variants of tumor suppressors can lead to cancer[11,12], while low-abundance variants of drug-metabolizing enzymes can alter drug response[13].

Here, we describe VAMP-seq, which measures the steady-state abundance of protein variants in cultured human cells. We applied VAMP-seq to assess 4,112 single-amino-acid variants of the tumor suppressor PTEN and 3,689 variants of the enzyme TPMT. Our results show how changes in protein biophysical properties and interactions within and between proteins alter protein abundance in cells. We identify 1,138 previously uncharacterized, low-abundance single-amino-acid variants of PTEN that are likely to be pathogenic, and 777 TPMT variants that are likely unable to adequately methylate and thereby inactivate thiopurine drugs. We observe selection for low-abundance PTEN variants in cancer and show that LRG_311p1:p.Pro38Ser, which accounts for ~10% of PTEN missense variants observed in melanoma, functions via a dominant-negative mechanism. Finally, we demonstrate that VAMP-seq can be applied to other clinically important proteins including VKORC1, CYP2C9, CYP2C19, MLH1 and PMS2.

Results

Multiplex assessment of PTEN and TPMT variant abundance. Inspired by earlier methods to assess the stability of protein variants in yeast[14] and bacteria[15], and by a microarray-based assay

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

that globally profiled mammalian protein stability[16], we developed VAMP-seq. VAMP-seq is a multiplex assay that uses fluorescent reporters to measure the steady-state abundance of protein variants in cultured human cells (Fig. 1). Each cell expresses a single variant directly fused to EGFP. The stability of the variant dictates the abundance of the EGFP fusion and, accordingly, the green fluorescence signal of the cell. To control for expression, mCherry is either co-transcriptionally or co-translationally expressed.

We first evaluated the ability of VAMP-seq to quantify the abundance of the tumor-suppressor protein PTEN and the enzyme TPMT. Each wild-type (WT) open reading frame was amino-terminally tagged with EGFP and recombined into a single genomic locus of an engineered HEK 293T cell line[17]. We also constructed cell lines expressing known low-abundance variants of each protein. We assessed the EGFP:mCherry ratio by flow cytometry, and found that cells expressing WT PTEN or TPMT had approximately five-fold higher EGFP:mCherry ratios than the known low-abundance variants (Fig. 2a and Supplementary Fig. 1b,c).

We next applied VAMP-seq to measure the steady-state abundance of thousands of PTEN and TPMT single-amino-acid variants in parallel. Barcoded, site-saturation mutagenesis libraries of each protein were separately recombined into our engineered HEK 293T cell line[17,18]. Cells harboring each library had EGFP:mCherry ratios that spanned the range of our WT and known low-abundance variants controls (Fig. 2a). Cells were flow-sorted into bins according to their EGFP:mCherry ratio, and high-throughput DNA sequencing

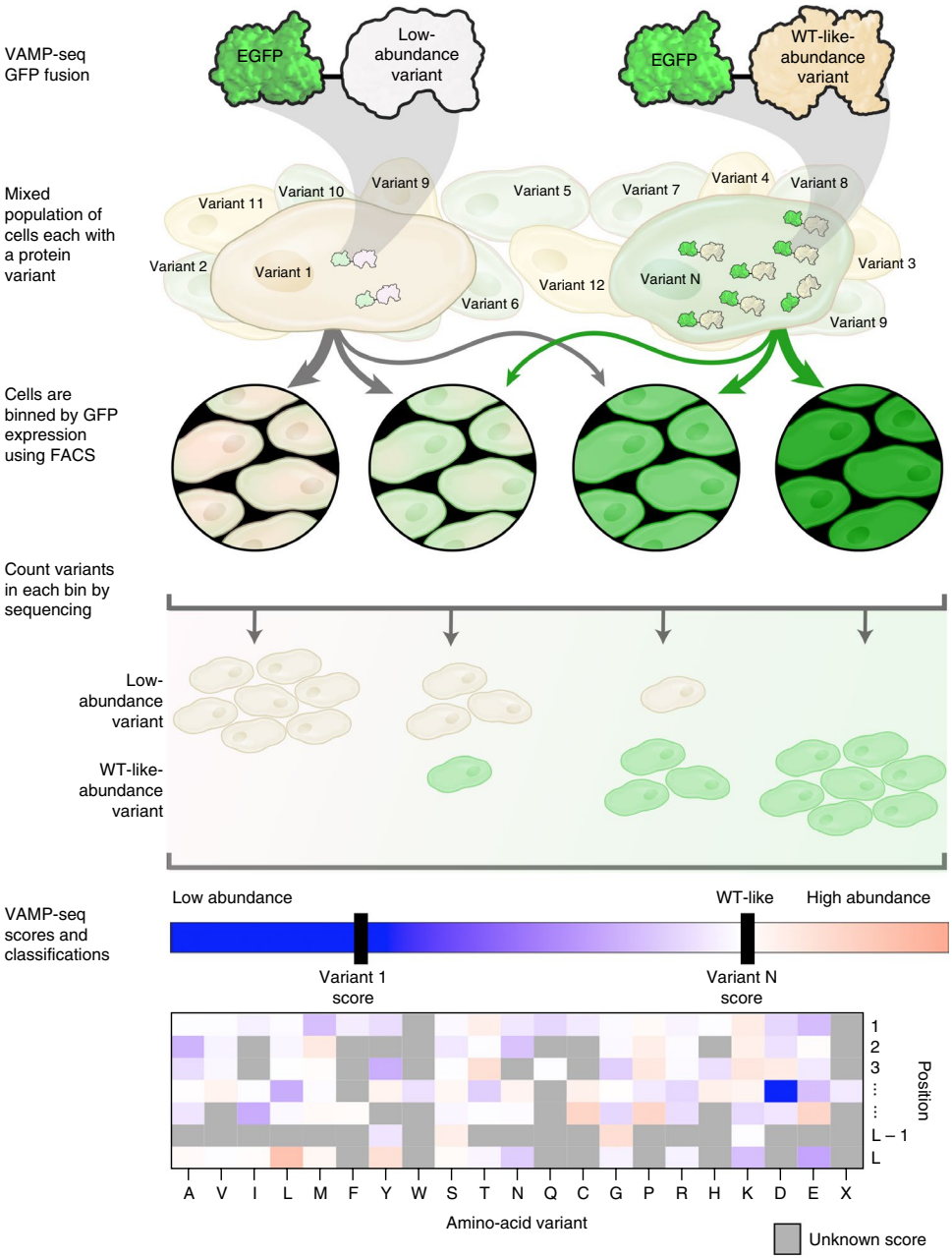

Fig. 1 | Overview of VAMP-seq. A mixed population of cells each expressing one protein variant fused to EGFP is created. The variant dictates the abundance of the variant–EGFP fusion protein, resulting in a range of cellular EGFP fluorescence levels. Cells are then sorted into bins according to their level of fluorescence, and high-throughput sequencing is used to quantify every variant in each bin. VAMP-seq scores are calculated from the scaled, weighted average of variants across bins. The resulting sequence–function maps describe the relative intracellular abundance of thousands of protein variants.

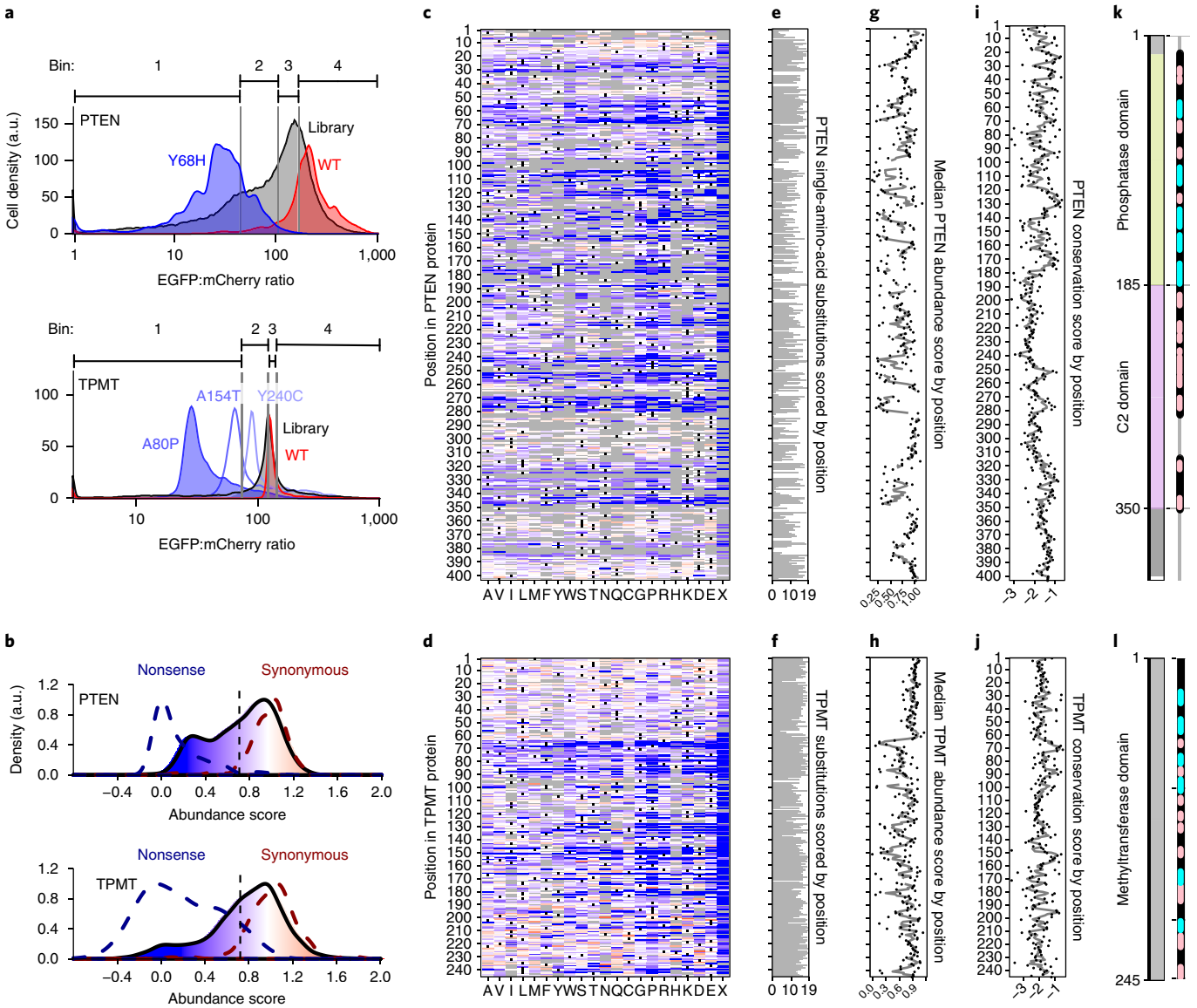

Fig. 2 | VAMP-seq abundance scores for PTEN and TPMT. a, Flow cytometry profiles for PTEN (top) and TPMT (bottom), with WT (red), known low-abundance variant controls (blue) and the variant libraries (gray) overlaid. Bin thresholds used to sort the library are shown above the plots. Each smoothed histogram was generated from at least 1,500 recombined cells from control constructs, and at least 6,000 recombined cells from the library. **b**, VAMP-seq abundance score density plots for PTEN (top) and TPMT (bottom) nonsense variants (blue dashed line), synonymous variants (red dashed line) and missense variants (filled, solid line). The missense variant densities are colored as gradients between the lowest 10% of abundance scores (blue), the WT abundance score (white) and abundance scores above WT (red). **c**,**d**, Heatmap of PTEN (**c**) and TPMT (**d**) abundance scores, colored according to the scale in **b**. Variants that were not scored are colored gray. **e**,**f**, Number of amino-acid substitutions scored at each position for PTEN and TPMT. **g**,**h**, Positional median PTEN and TPMT abundance scores, computed for positions with a minimum of five variants, are shown as dots. The gray line represents the mean abundance score in a three-residue sliding window. **i**,**j**, PTEN and TPMT position-specific PSIC conservation scores are shown as dots, and the gray line represents the mean PSIC score within a three-residue sliding window. **k**,**l**, PTEN and TPMT domain architecture, with positions in alpha helices and beta sheets colored cyan and pink, respectively.

was used to quantify each variant's frequency in each bin. Finally, an abundance score was calculated for each variant based on its distribution across the bins (Fig. 1 and Supplementary Table 1). Abundance scores ranged from about 0, indicating total loss of abundance, to about 1, indicating WT-like abundance (Fig. 2b).

Abundance scores correlated modestly well between replicates (mean Pearson's $r=0.63$ and mean Spearman's $\rho=0.62$ for PTEN; and mean $r=0.73$ and mean $\rho=0.67$ for TPMT; Supplementary Fig. 2). To improve accuracy, final abundance scores and confidence intervals were computed from eight replicate experiments. The resulting data set describes the effects of 4,112 of the 7,638 possible single-amino-acid PTEN variants and 3,689 of the 4,655 possible TPMT variants (Fig. 2c,d and Supplementary Data 1 and 2 and Supplementary Table 2). VAMP-seq-derived abundance scores were highly correlated with the abundances of protein variants assessed in individual experiments ($n=25$, $r=0.96$ and $\rho=0.96$ for PTEN; $n=19$, $r=0.75$ and $\rho=0.61$ for TPMT; Supplementary Fig. 3a,b). Furthermore, PTEN variant abundances measured using full-length EGFP or a 15-amino-acid split-GFP tag[19] were in agreement ($n=6$, $r=0.98$ and $\rho=0.94$; Supplementary Fig. 1d). Finally, our abundance scores were consistent with 41 PTEN and 20 TPMT variant abundance effects assessed by western blotting (Supplementary

Fig. 3c,d). Thus, VAMP-seq accurately quantifies steady-state protein variant abundance.

For both proteins, the distribution of abundance scores was bimodal, with peaks that overlapped WT synonyms and nonsense variants (Fig. 2b). Nonsense variants exhibited consistently low scores, except for those at the extreme N- or C-terminus of each protein (Supplementary Fig. 3e). A larger fraction of PTEN variants had low abundance scores than TPMT variants, possibly reflecting the lower thermostability of PTEN (melting temperature $T_m = 40.3\,°C$) relative to TPMT ($T_m = \sim 60\,°C$) (Supplementary Fig. 3f)[20,21]. This inverse relationship between low-abundance and thermostability is consistent with a deep mutational scan of GFP ($T_m = \sim 78\,°C$) that found relatively few variants with a large effect on fluorescence[22,23]. Median variant abundance scores at each position illustrated tolerance to amino-acid substitution (Fig. 2g,h and Supplementary Data 3 and 4 and Supplementary Table 2), which was inversely related to conservation ($\rho = -0.26$ and -0.59 for PTEN and TPMT, respectively; Fig. 2i,j and Supplementary Fig. 3g,h). In PTEN, alpha helices and beta sheets were less tolerant to substitution, while flexible loops were highly tolerant (Fig. 2k,l and Supplementary Fig. 3i). In TPMT, beta sheets, which comprise the core of protein, were less tolerant of substitution (Supplementary Fig. 3j). The abundance data can be explored using an interactive web interface (see the URLs section).

Thermodynamic stability partly explains variant abundance. Variants can potentially alter protein abundance inside cells via a variety of mechanisms, including by changing thermodynamic stability. We compared our abundance scores to various biochemical and biophysical features and found that hydrophobic packing, which affects thermodynamic stability in vitro[24–26], was a key correlate of abundance. Alteration of WT hydrophobic aromatic, methionine or long nonpolar aliphatic amino acids produced the largest decreases in abundance for both proteins (Fig. 3a). In fact, WT amino-acid hydrophobicity was negatively correlated with abundance (WT hydrophob.; Fig. 3b), whereas mutant amino-acid hydrophobicity was positively correlated with abundance (MT hydrophob.). Conversely, alterations of WT amino acids with high relative solvent accessibility (RSA), polarity (WT polarity) and crystal-structure temperature factor (B-factor), all features associated with polar residues present on the protein surface, were associated with high abundance (Fig. 3b). Consistent with the importance of hydrophobic packing, positions with the lowest average abundance scores were largely in the solvent-inaccessible interiors of each protein (Fig. 3c,d). Finally, PTEN abundance scores correlated strongly with in vitro melting temperatures[20] ($n = 5$, $r = 0.97$, $\rho = 0.90$; Supplementary Fig. 4a). These observations, consistent between PTEN and TPMT, suggest that variant thermodynamic stability is a major driver of variant abundance in vivo.

Next, we explored the role of polar contacts, using the PTEN structure to identify all side chains predicted to form hydrogen bonds and ion pairs. Of the 76 positions potentially participating in these interactions, only 26 were mutationally intolerant (Supplementary Fig. 4b). These 26 intolerant positions largely clustered into discrete groups in three-dimensional space (Fig. 3e and Supplementary Fig. 4c). The groups highlighted regions of PTEN particularly important for abundance, and often included positions distant in primary sequence. For example, group 5 positions, along with p.Ser170, mediate inter-domain contacts between the PTEN phosphatase and C2 domains[27], and we found that alterations at these positions resulted in loss of abundance (Fig. 3e). Alterations at these positions also frequently occur in cancer[27]; our data suggest that they may compromise function by virtue of their low abundance. Similarly, loss of abundance from abrogation of intra-domain polar contacts may account for the high frequency of alterations at p.Lys66, p.Tyr68 or p.Asp107 (group 2) in cancer (Fig. 3e and Supplementary Fig. 4d). TPMT lacked clusters of intolerant, polar-contact positions, possibly because it is a smaller, single-domain protein with a higher melting temperature.

Cell membrane interactions modulate PTEN variant abundance. Although VAMP-seq does not explicitly query post-translational modification, trafficking or partner binding, each of these can impact abundance. Therefore, we searched for signatures of these properties in our abundance data. PTEN mediates the removal of the 3 phosphate from phosphatidylinositol-3,4,5-triphosphate ($PtdIns(3,4,5)P_3$) to produce phosphatidylinositol-4,5-diphosphate ($PtdIns(4,5)P_2$) at the membrane[28]. Membrane interaction is aided by phospholipid-binding positions present in both PTEN domains (Fig. 3f)[29,30]. Furthermore, PTEN membrane binding and activity is negatively regulated by phosphorylation of its unstructured C-terminal tail[28,31]. Active site or C-terminal regulatory phosphosite variants have been found to decrease activity, reduce membrane binding and increase abundance, hinting at the existence of a negative feedback mechanism that degrades membrane-bound, active PTEN[31,32].

We therefore asked whether any PTEN variants increased abundance, perhaps by altering membrane interaction. We identified 41 positions in PTEN that had mean abundance scores higher than the WT. Nineteen of these enhanced-abundance positions were in structurally resolved regions, and 58% of them were within 7 Å of known phospholipid-binding positions. In comparison, only 13% of all structurally resolved PTEN positions were within 7 Å of phospholipid-binding positions (Supplementary Fig. 4e). Thus, positions with abundance-enhancing variants tended to be near the membrane-proximal face of PTEN, and included those important for binding $PtdIns(3,4,5)P_3$, $PtdIns(4,5)P_2$ or PtdIns(3)P (refs [30,33,34]; Fig. 3f). Furthermore, phosphomimetic substitutions at the p.Ser385 PTEN C-terminal regulatory phosphosite exhibited the highest abundance scores, whereas positively charged substitutions had low scores, supporting the impact of phosphorylation at this site on abundance (Supplementary Fig. 4f). Thus, many of the enhanced-abundance variants we identified likely disrupt PTEN membrane localization or $PtdIns(3,4,5)P_3$ phosphatase function.

New, potentially pathogenic low-abundance PTEN variants. VAMP-seq scores can also be used to identify potentially pathogenic variants. To simplify comparisons to clinical variant effects, we classified PTEN missense single-nucleotide variants (SNVs) as either low abundance, possibly low abundance, possibly WT-like abundance or WT-like abundance on the basis of how each variant's abundance score and confidence interval compared to the distribution of WT synonym scores (Fig. 4a and Supplementary Fig. 5a). Then, we analyzed variants present in public databases of either germline or somatic variation in the light of these abundance classifications.

Heterozygous germline loss of PTEN activity can cause a spectrum of symptoms including multiple hamartomas, carcinoma and macrocephaly, collectively known as PTEN hamartoma tumor syndrome[35], which includes Cowden syndrome. Two hundred and sixteen PTEN germline missense SNVs are in ClinVar, a submission-driven database of variants identified primarily through clinical testing[3]. Forty-one of the 216 PTEN missense variants are annotated as pathogenic, 25 of which had abundance scores. Of these 25, 16 (64%) were classified as low abundance (Fig. 4b), a significantly higher proportion than the 24% of scored missense variants that are low abundance (resampling test, $n = 25$, $P < 0.0001$; Fig. 4a and Supplementary Fig. 5b and Supplementary Table 3). Of the remaining nine variants, three were possibly low abundance. Four were active site variants (p.His93Arg, p.Gly129Glu, p.Arg130Leu and p.Thr131Ile) known to be inactive without loss of abundance. The remaining two variants (p.Asp24Gly and p.Arg234Gln) were distal

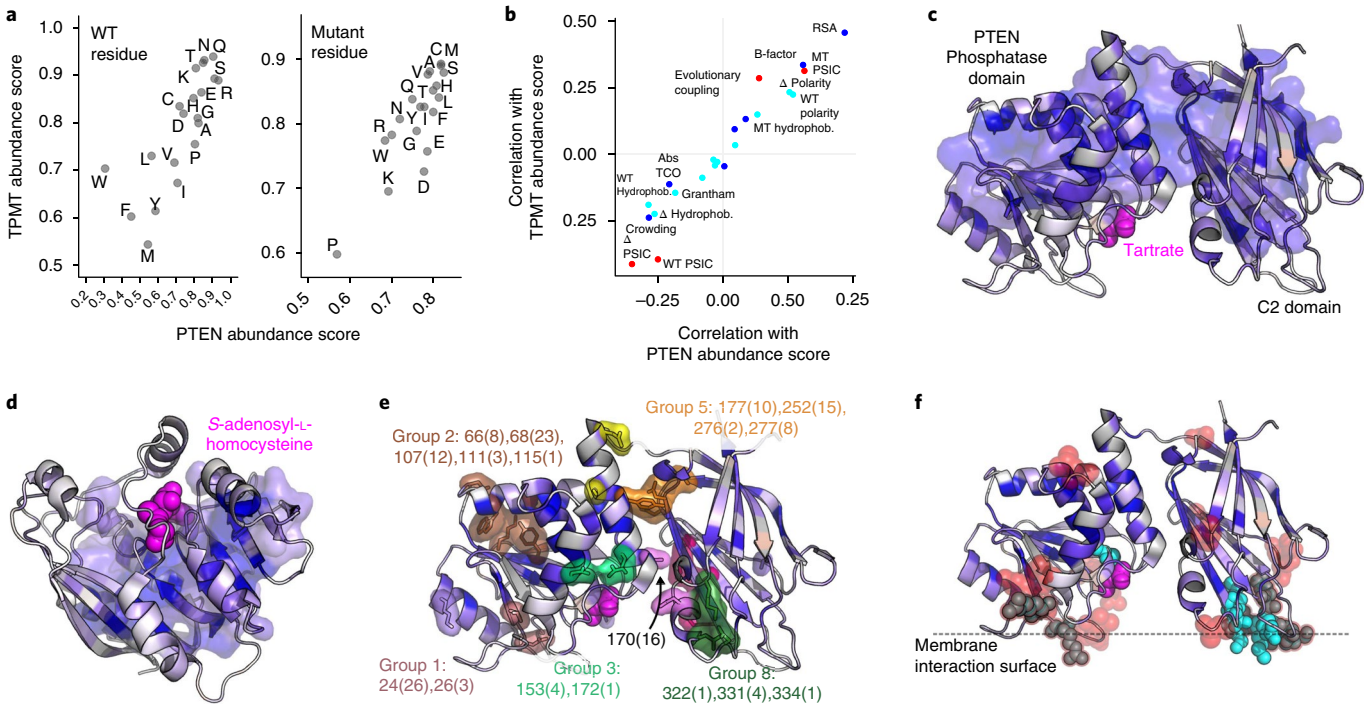

Fig. 3 | Biochemical features influencing intracellular protein abundance. a, Scatterplots of variant abundance scores averaged over all 20 WT residues (left) or mutant residues (right) for PTEN (*x* axis) and TPMT (*y* axis). **b**, A scatterplot of Spearman's rho values for PTEN (*x* axis) or TPMT (*y* axis) abundance score correlations with various evolutionary (red), structural (blue) or primary protein sequence (cyan) features ($n = 3,411$ for PTEN, $n = 3,230$ for TPMT). See the legend of Supplementary Table 2 for information regarding these features. **c**,**d**, PTEN (**c**, PDB: 1d5r) and TPMT (**d**, PDB: 2h11) crystal structures. Chains are colored according to positional median abundance scores using a gradient between the lowest 10% of positional median abundance scores (blue), the WT abundance score (white) and abundance scores above WT (red). The 20% of positions with the lowest scores are shown as a semi-transparent surface. The substrate-mimicking compounds tartrate and *S*-adenosyl-L-homocysteine are displayed as magenta spheres. **e**, Low-abundance PTEN residues with predicted hydrogen bonds or salt bridges are shown as sticks with a semi-transparent surface representation. Residues within 11 Å of each other are clustered and colored as discrete groups. The residues in each group are identified by number, followed, in parentheses, by the number of times any variant at the residue is found in the COSMIC database. **f**, Residues with high abundance scores are shown as semi-transparent red spheres, and known membrane-interacting side chains are shown as opaque cyan spheres. Residues that are both membrane-interacting and have high abundance scores are shown in gray.

to the active site and likely alter PTEN function by an unknown mechanism[36,37]. Thus, VAMP-seq-derived abundance scores, where available and combined with structural knowledge of the PTEN active-site, identify >90% of known PTEN pathogenic variants.

We could not formally assess the VAMP-seq false-positive rate because no PTEN variants are currently classified as benign. However, as has been done before[8], we were able to identify likely non-damaging variants on the basis of their population frequency. Germline PTEN variants cause Cowden syndrome, a high-penetrance, dominantly inherited Mendelian disease, at a rate of at least ~1 per 200,000 individuals[35,38]. We identified PTEN variants occurring at frequencies higher than expected given the prevalence of Cowden's syndrome, strongly suggesting that they are non-damaging[8,39]. Seven variants passed this threshold, and six were in our data set (Supplementary Fig. 5c). None were low abundance. One was possibly low abundance and two were possibly WT-like abundance. The remaining three, p.Ala79Thr, p.Pro354Gln and p.Ser294Arg, were WT-like in abundance and had frequencies higher than 5×10^{-5}, strongly suggesting that they are likely to be benign[2] (Fig. 4a). This analysis suggests that the PTEN abundance score data have a low false-positive rate.

An additional 41 PTEN variants are annotated as likely pathogenic in ClinVar. Of these, 23 had abundance scores, 10 (43%) of which were classified as low abundance (Fig. 4c and Supplementary Fig. 5b). Thus, the likely pathogenic category also had more low-abundance variants than expected (resampling test, $n = 23$, $P = 0.0188$; Supplementary Table 3). The 134 remaining ClinVar variants are of uncertain significance. Eighty-three of these variants had abundance scores, and 22 (27%) were low abundance (Fig. 4d). By providing additional evidence that supports pathogenicity, our abundance data could be used to alter variant clinical interpretations[40] (Supplementary Note and Supplementary Fig. 6). For example, 22 variants of uncertain significance along with 275 possible but not-yet-observed missense variants are low-abundance and could potentially be moved to the likely pathogenic category once observed in the appropriate clinical setting (Supplementary Table 4).

Abundance data suggest mechanisms of PTEN dysregulation. Somatic inactivation of PTEN by missense variation is an important contributor to multiple types of cancer[41]. We asked whether VAMP-seq derived abundance data could show the contribution of previously reported somatic PTEN variants to tumorigenesis. We collected PTEN missense or nonsense variants found in The Cancer Genome Atlas[42] and the AACR Project GENIE[43], and compared the observed frequencies of PTEN variants of each abundance class to the expected frequencies based on cancer type-specific nucleotide mutation spectra[42]. We observed significantly more low-abundance PTEN variants than expected for every cancer type analyzed (resampling test, all *P* values ≤ 0.0032; Fig. 4e; see Supplementary Table 5 for *P* values). This pattern suggests that selection for low-abundance PTEN variants is a common oncogenic mechanism.

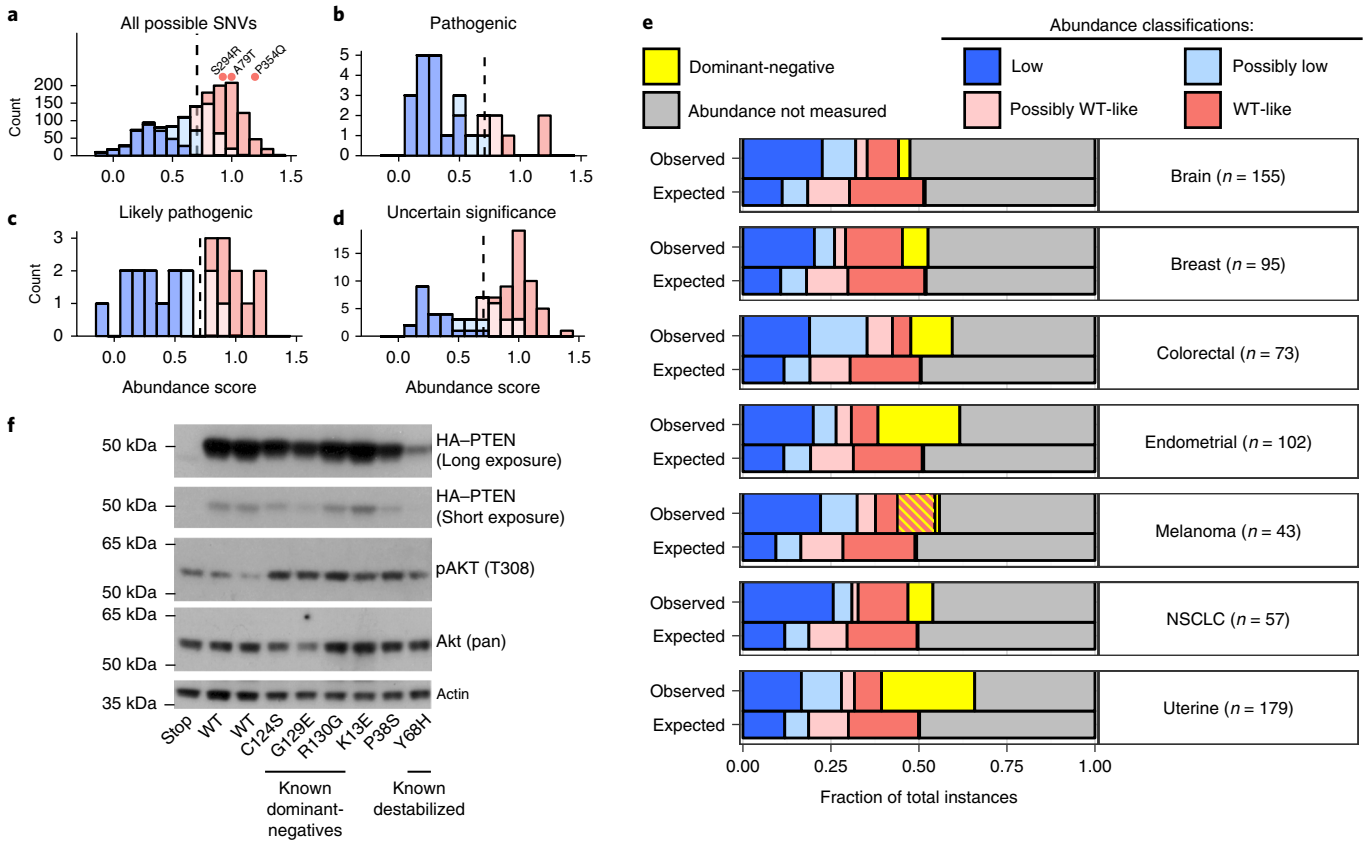

Fig. 4 | PTEN variant abundance classes across PTEN hamartoma tumor syndrome and cancer. a, A histogram of PTEN abundance scores for all missense variants observed in the experiment, with bars colored according to abundance classification. Abundance scores for three possibly benign variants present in the GnomAD database are shown as dots colored by classification. **b**–**d**, Abundance score histograms, colored by abundance classification, for PTEN germline variants listed in ClinVar as known pathogenic (**b**), likely pathogenic (**c**) or variants of uncertain significance (**d**). **e**, PTEN missense and nonsense variants in TCGA and the AACR GENIE project databases are arranged by cancer type. The top bar in each cancer type panel shows the observed frequency of variants in each abundance class as determined using VAMP-seq data. The bottom bar in each cancer type panel shows the expected abundance class frequencies based on cancer type-specific nucleotide substitution rates. Abundance classes are colored blue (low-abundance), light blue (possibly low-abundance), pink (possibly WT-like) or red (WT-like). The p.Pro38Ser variant is additionally colored with yellow stripes. The four known PTEN dominant-negative variants are colored yellow. Variants not scored in the experiment are colored gray. *n* is the number of instances of PTEN variants observed in the indicated cancer type and also scored in our experiments. **f**, A western blot analysis of cells stably expressing WT or missense variants of N-terminally HA-tagged PTEN. This experiment was independently performed twice with similar results (see Supplementary Fig. 5e). NSCLC, non-small cell lung carcinoma.

Some PTEN variants (for example, p.Cys124Ser, p.Gly129Glu, p.Arg130Gly and p.Arg130Gln) are inactive but have WT-like abundance. These inactive variants exert a dominant-negative effect on PTEN activity, leading to enhanced Akt phosphorylation and enhanced tumorigenesis in mouse models[44–46]. As expected, known dominant-negative variants had WT-like or higher abundance scores (p.Cys124Ser = 1.14, p.Arg130Gly = 1.09 and p.Gly129Glu = 0.76). Known dominant-negative variants were also significantly enriched in cancer, largely driven by the high frequencies of p.Arg130Gly and p.Arg130Gln[44,47] (Fig. 4e and Supplementary Fig. 5c; see Supplementary Table 5 for *P* values).

Unlike for every other cancer type we examined, melanoma lacked an enrichment of known dominant-negative variants. However, p.Pro38Ser was significantly enriched, accounting for 10.4% of PTEN missense variants (resampling test, $n = 77$, $P < 0.0001$; Fig. 4e and Supplementary Fig. 5d; see Supplementary Table 5 for *P* values). p.Pro38Ser had been previously observed in melanoma cancer cell lines, yet had never been functionally characterized[48]. p.Pro38Ser had a slightly higher abundance score than the WT (1.14) in our assay. On the basis of its prevalence in melanoma and its WT-like abundance, we hypothesized that it might exert a dominant-negative effect. Indeed, we found that p.Pro38Ser, like known dominant-negative variants, drove increased Akt phosphorylation in the presence of endogenous WT PTEN (Fig. 4f and Supplementary Fig. 5e). In contrast, computational predictors suggested that p.Pro38Ser is thermodynamically unstable, highlighting the utility of VAMP-seq (Supplementary Fig. 5f). Overall, our results show that low-abundance PTEN variants are important cancer drivers and that p.Pro38Ser, over-represented in melanoma, likely acts as a dominant-negative variant.

Implications of TPMT abundance for drug treatment. *TPMT* is 1 of 17 pharmacogenes whose genotype can be used to guide drug dosing[49]. Functional TPMT is required to metabolize thiopurine drugs such as 6-mercaptopurine (6-MP) and its prodrug, azathioprine. Thiopurine drugs are used to treat individuals with leukemia, rheumatic disease, inflammatory bowel disease or rejection in solid-organ transplant. Increased exposure to thiopurines causes treatment interruption or even life-threatening myelosuppression and hepatotoxicity. Three known non-functional variants of TPMT, NP_000358.1: p.Ala80Pro, p.Ala154Thr and p.Tyr240Cys, are found at high allele frequencies (combined minor allele frequency

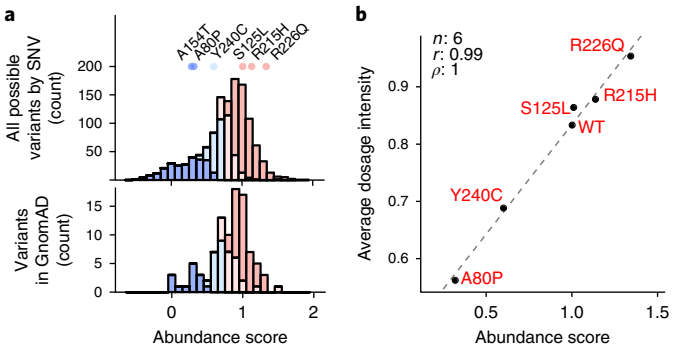

Fig. 5 | TPMT variant abundance classes across pharmacogenomics phenotypes. a, A histogram of TPMT abundance scores for all missense variants observed in the experiment, with bars colored according to abundance classification (top; $n = 1{,}529$ data points). Abundance scores for variants previously identified and characterized in patients are shown as dots colored by classification. Variants found in gnomAD at frequencies higher than 4×10^{-6} are also shown (bottom; $n = 118$ data points). **b**, A scatterplot of abundance score and mean 6-MP dose tolerated by individuals heterozygous for each variant. Dose intensity is the dose at which 6-MP becomes toxic to the patient before the 100% protocol dose of 75 mg m^{-2}. r and ρ denote Pearson's and Spearman's correlation coefficients, respectively.

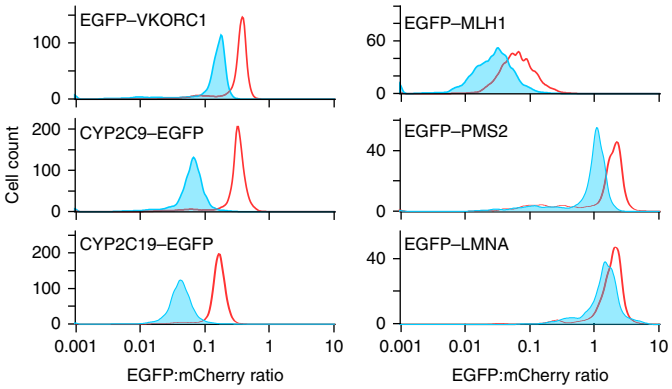

Fig. 6 | Additional drug- and disease-related genes are compatible with VAMP-seq. Representative flow cytometry EGFP:mCherry smoothed histogram plots for WT (red) and known or predicted destabilized variants (blue) for VKORC1, CYP2C9, CYP2C19, MLH1, PMS2 and LMNA. Each smoothed histogram was generated from at least 1,000 recombined cells. This experiment was independently performed three times with similar results.

(MAF) = 0.066) and are responsible for 95% of decreased-function alleles in the population[50]. The drug toxicity to carriers of these variants can be explained, at least in part, by the fact that they result in lower abundance of TPMT relative to the WT[13,21] (Fig. 5a). Accordingly, both abundance scores (Fig. 5a) and individually assessed EGFP:mCherry values (Fig. 2a and Supplementary Fig. 1c) were lower for these non-functional variants compared to the WT allele. Since our abundance scores identified known decreased-function alleles, we analyzed the abundance of rare TPMT variants of unknown function.

In a clinical study of patients with acute lymphoblastic leukemia, 884 patients were analyzed by exome array. Two hundred and seventy-eight of these patients also had exome sequencing data available. Red blood cell (RBC) TPMT activity and 6-MP dose intensity, the dose at which each individual became sensitive to 6-MP, were also measured[51]. The three known, high-frequency drug-sensitivity variants were identified, along with four rare variants: p.Ser125Leu, p.Gln179His, p.Arg215His and p.Arg226Gln (combined MAF < 0.0053). The mean RBC activity of individuals heterozygous for p.Gln179His, p.Arg215His and p.Arg226Gln was lower than the mean activity of individuals without TPMT variants, but higher than the activity of individuals heterozygous for the high-frequency drug sensitivity variants (Supplementary Fig. 7a,b). In contrast, RBC activity for p.Ser125Leu was higher than the WT. Thiopurine dose intensity, which is affected by TPMT activity, is highly correlated with variant abundance ($r = 0.99$, $\rho = 1$, $n = 6$; Fig. 5b and Supplementary Fig. 7c). Although their RBC activity varied over a wide range, the individuals heterozygous for these rare variants tolerated a higher mean dose of 6-MP than individuals heterozygous for the known sensitivity variants. Additionally, the four rare variants are classified as WT-like based on VAMP-seq abundance data. Individual assessment confirmed that these rare alleles do not affect abundance (Supplementary Fig. 7d). Thus, p.Ser125Leu, p.Gln179His, p.Arg215His and p.Arg226Gln may not be decreased-function variants.

Sequencing of the human population[2] and individuals intolerant to thiopurine drugs[52] has identified an additional 118 rare TPMT variants. These variants (MAF range = 0.000004–0.00066) are carried, in aggregate, by 0.2% of the population[2], but the impact of most of these variants on TPMT activity and abundance is unknown[53]. We measured abundance scores for 96 of these variants, classifying 14 (15%) as low abundance and 17 (18%) as possibly low abundance. When these or any of the other 389 missense variants we classified as low or possibly low abundance are identified in the clinic, the risk for thiopurine toxicity may be elevated. Dose reduction or closer monitoring could minimize toxicity and improve outcomes[50].

General utility of VAMP-seq for assessing variant abundance. To demonstrate that VAMP-seq is applicable to diverse proteins, we evaluated WT and known or predicted low-abundance variants for seven additional pharmacogenes or 'clinically actionable' genes[54,55] (Supplementary Table 6). For CYP2C9, CYP2C19 and VKORC1, we found large differences in the EGFP:mCherry ratios of the WT and known or predicted low-abundance missense variants (Fig. 6), whereas MLH1 and PMS2 yielded smaller differences. Thus, VAMP-seq could be applied to these five proteins. Furthermore, ~52% of human proteins yielded at least as much fluorescence as MLH1 when expressed as EGFP fusions[16], suggesting that many human proteins are compatible with VAMP-seq (Supplementary Fig. 8). However, BRCA1 and LMNA resulted in a low EGFP signal or no difference in the EGFP:mCherry ratio between WT and known low-abundance variants (Fig. 6 and data not shown). Thus, VAMP-seq will not be applicable in all cases. In particular, proteins that are marginally stable (such as BRCA1), make large complexes (such as LMNA) or are secreted and therefore break the link between variant genotype and phenotype are not amenable to VAMP-seq.

Discussion

VAMP-seq is a generalizable method for multiplex measurement of steady-state protein variant abundance. Since alterations in abundance may be a general mechanism of pathogenic variation[9,10], an important application of VAMP-seq may be to aid clinical geneticists in understanding the effects of newly discovered missense variants. Indeed, the American College of Medical Genetics suggests that well-established functional assays can provide strong evidence of pathogenicity[40]. Thus, in the context of monogenic diseases where protein inactivation is pathogenic, VAMP-seq-derived abundance data can help to identify pathogenic variants. The utility of VAMP-seq for this purpose is highlighted by the fact that 64% of known PTEN pathogenic missense variants were of low abundance. Furthermore, VAMP-seq identified 1,138 low-abundance PTEN variants that would likely confer an increased risk of PTEN hamartoma tumor syndrome and 777 low-abundance TPMT variants that

would likely require altered drug dosing. If other proteins yielded similar results, VAMP-seq could provide evidence of pathogenicity for greater than half of the pathogenic missense variants we will eventually find as more human genomes are sequenced.

Interpretation of somatic variation is more difficult, but functional data can identify driver variants and, therefore, potential treatments. For example, variation in PTEN, presumably resulting in PTEN loss-of-function, is associated with increased sensitivity to PI(3)K, AKT and mTOR inhibitors, and decreased sensitivity to receptor tyrosine kinase inhibitors[56]. Our PTEN abundance data identify many loss-of-function variants, which could help to clarify the link between PTEN inactivation and altered drug sensitivity, and thus might inform cancer treatment. Furthermore, aided by our abundance data, we identified p.Pro38Ser as a candidate PTEN dominant-negative variant in melanoma. Since the known dominant-negative variants p.Gly129Glu and p.Cys124Ser result in exacerbated oncogenic phenotypes in mice[44,46], p.Pro38Ser status might help to predict tumor aggressiveness.

Despite its utility, VAMP-seq has limitations. Bottlenecks in our library generation method were largely responsible for the ~50% of possible PTEN variants missing from the final data set. In the future, early library validation using deep sequencing along with other well-validated library generation methods[8] could improve coverage. Additionally, as for any assay, VAMP-seq abundance data are subject to uncertainty. To address this concern, we quantified the uncertainty associated with each abundance score. We suggest that abundance score uncertainty should be taken into consideration, as we did when classifying variant abundance. VAMP-seq relies on fusion of the protein of interest to EGFP. We showed a high concordance between VAMP-seq abundance data and abundance as measured by other methods, but this might not always be the case. Furthermore, VAMP-seq cannot yield insight into variants that are pathogenic because of reduced enzymatic activity, altered localization or effects on splicing. Thus, while VAMP-seq abundance data are useful for identifying pathogenic variants, they should not be used to conclude that a variant is benign.

Generalizable assays such as VAMP-seq are a promising way to understand the functional effects of missense variation at scale. In addition to demonstrating its effectiveness for PTEN and TPMT, we provide preliminary evidence that VAMP-seq could be applied to other clinically relevant proteins. Furthermore, repeating VAMP-seq assays in different cell lines could find cell-type specific regulation of variant abundance. Comparing variant abundance data in WT and chaperone knockout cells could identify what makes a protein a chaperone client. Combining VAMP-seq with small-molecule modulators of chaperone or protein degradation machinery may even find variant-specific treatments that could rescue low-abundance variants. Thus, VAMP-seq greatly expands our ability to measure the impact of missense variants on abundance, a fundamental property that underlies protein function.

URLs. VAMP-seq scores are available at http://abundance.gs.washington.edu. Code used for the analyses performed in this work is included as Supplementary Data 5, and is also available at http://github.com/FowlerLab/VAMPseq. Code used for subassembly by PacBio is available at http://github.com/shendurelab/AssemblyByPacBio.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0122-z.

Received: 19 September 2017; Accepted: 29 March 2018;
Published online: 21 May 2018

References

1. Shirts, B. H., Pritchard, C. C. & Walsh, T. Family-specific variants and the limits of human genetics. *Trends Mol. Med.* **22**, 925–934 (2016).
2. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
3. Landrum, M. J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, 980–985 (2014).
4. Fowler, D. M., Stephany, J. J. & Fields, S. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat. Protoc.* **9**, 2267–2284 (2014).
5. Gasperini, M., Starita, L. & Shendure, J. The power of multiplexed functional analysis of genetic variants. *Nat. Protoc.* **11**, 1782–1787 (2016).
6. Manolio, T. A. et al. Bedside back to bench: building bridges between basic and clinical genomic research. *Cell* **169**, 6–12 (2017).
7. Starita, L. M. et al. Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* **200**, 413–422 (2015).
8. Majithia, A. R. et al. Prospective functional classification of all possible missense variants in PPARG. *Nat. Genet.* **48**, 1570–1575 (2016).
9. Yue, P., Li, Z. & Moult, J. Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* **353**, 459–473 (2005).
10. Redler, R. L., Das, J., Diaz, J. R. & Dokholyan, N. V. Protein destabilization as a common factor in diverse inherited disorders. *J. Mol. Evol.* **82**, 11–16 (2016).
11. Berger, A. H., Knudson, A. G. & Pandolfi, P. P. A continuum model for tumour suppression. *Nature* **476**, 163–169 (2011).
12. Lee, M. S. et al. Comprehensive analysis of missense variations in the BRCT domain of BRCA1 by structural and functional assays. *Cancer Res.* **70**, 4880–4890 (2010).
13. Tai, H. L., Krynetski, E. Y., Schuetz, E. G., Yanishevski, Y. & Evans, W. E. Enhanced proteolysis of thiopurine S-methyltransferase (TPMT) encoded by mutant alleles in humans (TPMT*3A, TPMT*2): mechanisms for the genetic polymorphism of TPMT activity. *Proc. Natl Acad. Sci. USA* **94**, 6444–6449 (1997).
14. Kim, I., Miller, C. R., Young, D. L. & Fields, S. High-throughput analysis of in vivo protein stability. *Mol. Cell. Proteomics* **12**, 3370–3378 (2013).
15. Klesmith, J. R., Bacik, J.-P., Wrenbeck, E. E., Michalczyk, R. & Whitehead, T. A. Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proc. Natl Acad. Sci. USA* **114**, 2265–2270 (2017).
16. Yen, H.-C. S., Xu, Q., Chou, D. M., Zhao, Z. & Elledge, S. J. Global protein stability profiling in mammalian cells. *Science* **322**, 918–923 (2008).
17. 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).
18. 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**, 90–98 (2014).
19. 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).
20. Johnston, S. B. & Raines, R. T. Conformational stability and catalytic activity of PTEN variants linked to cancers and autism spectrum disorders. *Biochemistry* **54**, 1576–1582 (2015).
21. Wu, H. et al. Structural basis of allele variation of human thiopurine-*S*-methyltransferase. *Proteins* **67**, 198–208 (2007).
22. Ward, W. W., Prentice, H. J., Roth, A. F., Cody, C. W. & Reeves, S. C. Spectral perturbations of the Aequorea green-fluorescent protein. *Photochem. Photobiol.* **35**, 803–808 (1982).
23. Sarkisyan, K. S. et al. Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).
24. Zhou, H. & Zhou, Y. Quantifying the effect of burial of amino acid residues on protein stability. *Proteins* **322**, 315–322 (2004).
25. Kauzmann, W. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **14**, 1–63 (1959).
26. Rocklin, G. J. et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).
27. Lee, J. O. et al. Crystal structure of the PTEN tumor suppressor: implications for its phosphoinositide phosphatase activity and membrane association. *Cell* **99**, 323–334 (1999).
28. Song, M. S., Salmena, L. & Pandolfi, P. P. The functions and regulation of the PTEN tumour suppressor. *Nat. Rev. Mol. Cell Biol.* **13**, 283–296 (2012).
29. Nguyen, H.-N. et al. A new class of cancer-associated PTEN mutations defined by membrane translocation defects. *Oncogene* **34**, 3737–3743 (2015).
30. Walker, S. M., Leslie, N. R., Perera, N. M., Batty, I. H. & Downes, C. P. The tumour-suppressor function of PTEN requires an N-terminal lipid-binding motif. *Biochem. J.* **379**, 301–307 (2004).
31. Das, S., Dixon, J. E. & Cho, W. Membrane-binding and activation mechanism of PTEN. *Proc. Natl Acad. Sci. USA* **100**, 7491–7496 (2003).

32. Vazquez, F., Ramaswamy, S., Nakamura, N. & Sellers, W. R. Phosphorylation of the PTEN tail regulates protein stability and function. *Mol. Cell. Biol.* **20**, 5010–5018 (2000).
33. Wei, Y., Stec, B., Redfield, A. G., Weerapana, E. & Roberts, M. F. Phospholipid-binding sites of phosphatase and tensin homolog (PTEN): Exploring the mechanism of phosphatidylinositol 4,5-bisphosphate activation. *J. Biol. Chem.* **290**, 1592–1606 (2015).
34. Naguib, A. et al. PTEN functions by recruitment to cytoplasmic vesicles. *Mol. Cell* **58**, 255–268 (2015).
35. Hobert, J. A. & Eng, C. PTEN hamartoma tumor syndrome: an overview. *Genet. Med.* **11**, 687–694 (2009).
36. Melbārde-Gorkuša, I. et al. Challenges in the management of a patient with Cowden syndrome: case report and literature review. *Hered. Cancer Clin. Pract.* **10**, 5 (2012).
37. Staal, F. J. T. et al. A novel germline mutation of PTEN associated with brain tumours of multiple lineages. *Br. J. Cancer* **86**, 1586–1591 (2002).
38. Nelen, M. R. et al. Novel PTEN mutations in patients with Cowden disease: Absence of clear genotype–phenotype correlations. *Eur. J. Hum. Genet.* **7**, 267–273 (1999).
39. Whiffin, N. et al. Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet. Med.* **19**, 1151–1158 (2017).
40. Richards, S. et al. 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).
41. Hollander, M. C., Blumenthal, G. M. & Dennis, P. A. PTEN loss in the continuum of common cancers, rare syndromes and mouse models. *Nat. Rev. Cancer* **11**, 289–301 (2011).
42. Kandoth, C. et al. Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
43. AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discov.* **7**, 818–831 (2017).
44. Papa, A. et al. Cancer-associated PTEN mutants act in a dominant-negative manner to suppress PTEN protein function. *Cell* **157**, 595–610 (2014).
45. Leslie, N. R. & Longy, M. Inherited PTEN mutations and the prediction of phenotype. *Semin. Cell Dev. Biol.* **52**, 30–38 (2016).
46. Wang, H. et al. Allele-specific tumor spectrum in *Pten* knockin mice. *Proc. Natl Acad. Sci. USA* **107**, 5142–5147 (2010).
47. Bonneau, D. & Longy, M. Mutations of the human PTEN gene. *Hum. Mutat.* **16**, 109–122 (2000).
48. Aguissa-Touré, A.-H. & Li, G. Genetic alterations of PTEN in human melanoma. *Cell. Mol. Life Sci.* **69**, 1475–1491 (2012).
49. Hodges, L. M. et al. Very important pharmacogene summary. *Pharmacogenet. Genomics* **21**, 152–161 (2011).
50. Relling, M. V. et al. Clinical pharmacogenetics implementation consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing: 2013 update. *Clin. Pharmacol. Ther.* **93**, 324–325 (2013).
51. Liu, C. et al. Genomewide approach validates thiopurine methyltransferase activity is a monogenic pharmacogenomic trait. *Clin. Pharmacol. Ther.* **101**, 373–381 (2017).
52. Appell, M. L. et al. Nomenclature for alleles of the thiopurine methyltransferase gene. *Pharmacogenet. Genomics* **23**, 242–248 (2013).
53. Hamdan-Khalil, R. et al. In vitro characterization of four novel non-functional variants of the thiopurine *S*-methyltransferase. *Biochem. Biophys. Res. Commun.* **309**, 1005–1010 (2003).
54. Kalia, S. S. et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SFv2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19**, 1–7 (2016).
55. Relling, M. et al. New Pharmacogenomics Research network: an open community catalyzing research and translation in precision medicine. *Clin. Pharmacol. Ther.* **102**, 897–902 (2017).
56. Dillon, L. M. & Miller, T. W. Therapeutic targeting of cancers with loss of PTEN function. *Curr. Drug Targets* **15**, 65–79 (2014).

Acknowledgements

We thank J. Underwood and K. Munson of the UW PacBio Sequencing Services for assistance with long-read sequencing; A. Leith of the UW Foege Flow Lab and L. Gitari and D. Prunkard of the UW Pathology Flow Cytometry Core Facility for assistance with cell sorting; and B. Shirts and C. Pritchard in the UW Department of Lab Medicine for advice. The authors would like to acknowledge the American Association for Cancer Research and its financial and material support in the development of the AACR Project GENIE registry, as well as members of the consortium for their commitment to data sharing. Interpretations are the responsibility of study authors. This work was supported by the National Institute of General Medical Sciences (1R01GM109110 and 5R24GM115277 to D.M.F., P50GM115279 to M.V.R. and W.E.E., National Cancer Institute R01CA096670 to S.B. and P30CA21765 to M.V.R.) and an NIH Director's Pioneer Award (DP1HG007811 to J.S.). K.A.M. is an American Cancer Society Fellow (PF-15-221-01), and was supported by a National Cancer Institute Interdisciplinary Training Grant in Cancer (2T32CA080416). M.A.C. and V.E.G. are supported by the National Science Foundation Graduate Research Fellowship. J.N.D. is supported by a National Institute of General Medical Sciences Training Grant (T32GM007454). J.S. is an Investigator of the Howard Hughes Medical Institute. D.M.F. is a Canadian Institute for Advanced Research Azrieli Global Scholar.

Author contributions

D.M.F., J.S., K.A.M. and L.M.S. conceived of, designed and managed the experiments and analyses, and wrote the manuscript; J.J.S. and B.M. cloned expression constructs and libraries and prepared and performed NGS sequencing; K.A.M., M.A.C. and A.K. provided constructs and data for additional disease genes and pharmacogenes; M.K. wrote the scripts to extract barcodes and variable regions from long-read sequences; J.N.D. assisted in using the ACMG guidelines to reclassify PTEN variants; R.J.H. provided constructs for TPMT experiments; V.E.G. designed the website; and S.B., W.E.E., M.V.R. and W.Y. provided clinical data for TPMT comparison.

Competing interests

The authors declare that the variant functional data presented herein are copyrighted, and may be freely used for non-commercial purposes. Licensing for commercial use may benefit the authors. The authors declare no additional competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0122-z.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.S. or D.M.F.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

General reagents, DNA oligonucleotides and plasmids. Unless otherwise noted, all chemicals were obtained from Sigma and all enzymes were obtained from New England Biolabs. *E. coli* were cultured at 37 °C in Luria broth. All cell culture reagents were purchased from ThermoFisher Scientific unless otherwise noted. HEK 293T cells (ATCC CRL-3216) and derivatives thereof were cultured in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum, 100 U ml^{-1} penicillin and 0.1 mg ml^{-1} streptomycin. Induction medium was furthermore supplemented with 2 µg ml^{-1} doxycycline (Sigma-Aldrich). Cells were passaged by detachment with trypsin–EDTA 0.25%. All cell lines tested negative for mycoplasma. All synthetic oligonucleotides were obtained from IDT and can be found in Supplementary Table 7. All non-library-related plasmid modifications were performed with Gibson assembly[57]. See the Supplementary Note for construction of the VAMP-seq expression vectors.

Construction of barcoded, site-saturation mutagenesis libraries for TPMT and PTEN. Site-saturation mutagenesis libraries of TPMT and PTEN were constructed using inverse PCR[18]. See the Supplementary Note for a detailed description of construction of the barcoded, site-saturation mutagenesis libraries.

Single-molecule real-time (SMRT) sequencing to link each TPMT and PTEN variant to its barcode. For both PTEN and TPMT, the relationship between variants and barcodes was established using SMRT sequencing (Pacific Biosciences). See the Supplementary Note for a detailed description of variant linking steps using SMRT sequencing.

Integration of single-variant clones or barcoded libraries into the HEK293-landing pad cell line. Barcoded variant libraries or single-variant clones were recombined into the Tet-on landing pad in engineered HEK 293T TetBxb1BFP Clone4 cells that we generated previously[17]. See the Supplementary Note for a detailed description of how variant libraries were integrated into cells.

FACS to bin cells by EGFP:mCherry ratio. Cells harboring variant libraries, prepared as described above, were sorted using a FACSAria III (BD Biosciences) into bins according to the abundance of their expressed, EGFP-tagged variant. First, live, single, recombinant cells were selected using forward and side scatter, mCherry and mTagBFP2 signals. Then, a FITC:PE–Texas Red ratiometric parameter in the BD FACSDIVA software was created. A histogram of the FITC:PE–Texas Red ratio was created and gates dividing the library into four equally populated bins based on the ratio were established. The details of replicate sorts can be found in Supplementary Table 1.

Sorted library genomic DNA preparation, barcode amplification and sequencing. For the TPMT experiments, sorted cells were collected by centrifugation and the FACS sheath buffer was aspirated. Cells were transferred into a microfuge tube, pelleted and stored at −20 °C. Genomic DNA was prepared using the GentraPrep kit (Qiagen). For each bin, all of the purified DNA was spread over eight 25 µl PCR reactions containing Kapa Robust, and the primers GPS-landing-f (in the genome) and BC-GPS-P7-i#-UMI (3 of the barcode) to tag the barcodes with a unique molecular index (UMI) and add a sample index. UMI-tagging PCRs were performed using the following conditions: initial denaturation 95 °C 2 min, followed by three cycles of (95 °C 15 s, 60 °C 20 s, 72 °C 3 min). The eight PCR reactions were pooled and the PCR amplicon was purified using 1 × Ampure XP (Beckman Coulter). To shorten the amplicon and add the p5 and p7 Illumina cluster-generating sequences, the UMI-tagged barcodes were then amplified with the primers BC-TPMT-P5-v2 and Illumina p7. This PCR was performed with Kapa Robust and SYBR Green II on a Bio-Rad mini-opticon qPCR machine, and reactions were monitored and removed before saturation of the SYBR Green II signal, at around 25 cycles. The amplicons were pooled and gel-purified. Barcodes were read twice by the paired-end sequencing primers TPMT_Read1 and TPMT_Read2. The UMI and index were sequenced by the index read and primer TPMT_Index using a NextSeq 500 (Illumina). After converting from the BCL to FASTQ format using Illumina's bcl2fastq version 2.18, the forward, reverse and index reads were concatenated and demultiplexed into a BAM file. Consensus barcodes were called from the forward and reverse reads. To collapse the barcode copies associated with unique UMIs, the UMI (bases 1–10 of the index read) were pasted onto the consensus barcode and unique combinations were identified (sort | uniq -c). The barcode from each unique barcode–UMI pair was used to populate a FASTQ file that could be used by the Enrich 2 software package to count variants.

For the PTEN experiments, sorted cells were replated onto 10 cm plates and allowed to grow for approximately five days. Cells were then collected, pelleted by centrifugation and stored at −20 °C. Genomic DNA was prepared using a DNEasy kit, according to the manufacturer's instructions (Qiagen), with the addition of a 30 min incubation at 37 °C with RNAse in the re-suspension step. Eight 50 µl first-round PCR reactions were each prepared with a final concentration of ~50 ng µl^{-1} input genomic DNA, 1 × Kapa HiFi ReadyMix and 0.25 µM of the KAM499/JJS_501a primers. The reaction conditions were 95 °C for 5 min, 98 °C for 20 s, 60 °C for 15 s, 72 °C for 90 s, repeat 7 times, 72 °C for 2 min, 4 °C hold. Eight 50 µl reactions were combined, bound to AMPure XP (Beckman Coulter), cleaned and eluted with 40 µl water. Forty percent of the eluted volume was mixed with 2 × Kapa Robust ReadyMix; JJS_seq_F and one of the indexed reverse primers, JJS_seq_R1a through JJS_seq_R12a, were added at 0.25 µM each. Reaction conditions for the second-round PCR were 95 °C for 3 min, 95 °C for 15 s, 60 °C for 15 s, 72 °C for 30 s, repeat 14 times, 72 °C for 1 min, 4 °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 for PTEN experiments 1 through 4 were joined using the fastq-join tool within the ea-utils package using the default parameters, whereas only one barcode read was collected for PTEN experiments 5 through 8. Technical amplification and sequencing replicates were conducted for every sample, and compared to assess variability in quantitation stemming from amplification and sequencing. Experiments with poor technical replication across multiple bins were reamplified and resequenced in their entirety, leaving eight replicate experiments with technical replicates shown here (Supplementary Fig. 9). FASTQ files from these technical replicate amplification and sequencing runs were concatenated for analysis with Enrich2[58].

Barcode counting and variant calling. Enrich2 was used to count the barcodes, associate each barcode with a nucleotide variant and then translate and count both the unique-nucleotide and unique-amino acid variants[58]. FASTQ files containing either UMI-collapsed barcodes (TPMT) or total barcodes (PTEN) and the barcode map for each protein were used as input for Enrich2. Enrich2 configuration files for each experiment are available on the GitHub repository (see the URLs section). Barcodes assigned to variants containing insertions, deletions or multiple amino-acid alterations were removed from the analysis.

Calculating VAMP-seq scores and classifications. RStudio v1.0.136 was used for all subsequent analysis of the Enrich2 output. The count for each variant in a bin was divided by the sum of counts recorded in that bin to obtain the frequency of each variant (F_v) within that bin. This calculation was repeated for every bin in each replicate experiment. For each experiment, the total count of each variant across the bins was divided by the total count of all variants across the bins to obtain a total frequency value ($F_{v,total}$) for each variant for each experiment.

$$F_{v,total} = \frac{C_{v,bin1} + C_{v,bin2} + C_{v,bin3} + C_{v,bin4}}{\sum C_{bin1} + \sum C_{bin2} + \sum C_{bin3} + \sum C_{bin4}}$$

This total frequency value was used for filtering low-frequency variants, which we reasoned would be subject to high levels of counting noise, out of the subsequent calculations. We set the $F_{v,total}$ filtering threshold on the basis of the assumption that accurately scored synonymous variants should create a clear, unimodal distribution around the WT. We examined how different minimum $F_{v,total}$ filtering threshold values affected the spread and central tendency of the synonymous distribution (Supplementary Fig. 10). We empirically selected $1 \times 10^{-4.75}$ as the $F_{v,total}$ filtering threshold value as it minimized the skew and coefficient of variation of the synonymous variant abundance score distribution while retaining the majority of missense variants.

Next, for each experiment, a weighted average was calculated for each variant (W_v) passing the $F_{v,total}$ filtering threshold value using the following equation:

$$W_v = \frac{(F_{v,bin1} \times 0.25) + (F_{v,bin2} \times 0.5) + (F_{v,bin3} \times 0.75) + (F_{v,bin4} \times 1)}{(F_{v,bin1} + F_{v,bin2} + F_{v,bin3} + F_{v,bin4})}$$

Thus, all weighted average values ranged from a value of 0.25 to 1.

Finally, for each experiment, an abundance score for each variant (S_v) was obtained by subjecting the weighted average of each variant to min–max normalization, using the weighted average value of WT (W_{WT}), which was given a score of 1, and the median weighted average value for non-terminal nonsense variants ($W_{nonsense}$) at positions 51 through 349 for PTEN, or positions 51 through 219 for TPMT, which was given an abundance score of 0, using the following equation:

$$S_v = \frac{(W_v - W_{nonsense})}{(W_{WT} - W_{nonsense})}$$

The final abundance score for each variant was calculated by taking the mean of the min–max-normalized abundance scores across the eight replicate experiments in which it could have been observed. Only variants that were scored in two or more replicate experiments were retained in the analysis. We implemented this filter because many sources of noise are not captured in count-based estimates of variance and because having replicate-level variance estimates was critical to our abundance classification scheme. A standard error for each abundance score was calculated by dividing the standard deviation of the

min–max-normalized values for each variant by the square root of the number of replicate experiments in which it was observed. Lastly, the lower bound of the 95% confidence interval was calculated by multiplying the standard error by the 97.5th percentile value of a normal distribution and subtracting this product from the abundance score. The upper bound of the 95% confidence interval was calculated by instead adding the product to the abundance score. Positional VAMP-seq scores were calculated by taking the median of all single-amino-acid VAMP-seq scores at each position.

For both TPMT and PTEN, the distribution of WT synonyms was used to create VAMP-seq classifications for every variant (see 'Supplementary Fig. 5a for scheme). First, we established a synonymous score threshold by determining the abundance score that separated the 95% most abundant synonymous variants from the 5% lowest abundance synonymous variants (0.71 for PTEN, and 0.72 for TPMT). Variants whose abundance score and upper confidence interval were both below this synonymous threshold value were classified as 'low-abundance' variants, whereas those with abundance scores below this threshold but upper confidence intervals over this this were classified as 'possibly low abundance'. Variants with scores above this threshold but lower confidence intervals below the threshold were considered as 'possibly WT-like abundance'. Variants with scores and lower confidence intervals above the threshold were classified as 'WT-like abundance'.

For both TPMT and PTEN, substitution-intolerant positions were determined on the basis of the proportion of variants at the position with scores below the synonymous threshold, determined as described above. Positions where five or more variants were scored and greater than 90% of the scores were below the synonymous variant threshold value were considered substitution intolerant. Enhanced abundance positions were determined on the basis of the proportion of variants at the position with scores above the median of the synonymous distribution. Positions where five or more variants were scored and more than five variants had scores above the median of the synonymous distribution were considered enhanced-abundance positions.

Assessment of the PTEN library composition. To better understand the sources of bottlenecking in the PTEN experiments, the composition of the PTEN plasmid library preparation used to generate recombinant cells was assessed by determining barcode frequencies using high-throughput Illumina sequencing. See the Supplementary Note for a description of the steps taken to characterize the PTEN variant library. Metrics regarding the processing of sequencing data for the barcode–variant assignments can be found in Supplementary Table 8.

Variant annotation from online databases. Published western blotting results for PTEN and TPMT variants are listed, along with references, in Supplementary Table 9 and Supplementary Table 10. See the Supplementary Note for a description of the online databases that were accessed to obtain PTEN and TPMT variant annotations.

PTEN ClinVar and cancer genomics analyses. Nine PTEN variants were listed in ClinVar as both likely pathogenic and pathogenic. We examined the evidence for these variants—p.His61Arg, p.Tyr68His, p.Leu108Pro, p.Gly127Arg, p.Arg130Leu, p.Arg130Gln, p.Gly132Val, p.Arg173Cys, and p.Arg173His—and following the ACMG/AMP guidelines[40], all nine were deemed to belong in the likely pathogenic category. An additional two variants—p.Arg15Lys and p.Pro96Ser—had an interpretation of uncertain significance along with another interpretation of likely pathogenic or pathogenic, and thus the clinical significance of the variant was listed as 'Conflicting interpretations of pathogenicity'. As recommended by the ACMG/AMP guidelines[40], variants with conflicting interpretations were considered variants of unknown significance.

Likely non-damaging PTEN variants were identified from the variants observed in gnomAD at allele frequencies rendering them highly unlikely to be causal for Cowden's syndrome, under an autosomal dominant model of inheritance with an estimated prevalence in the population of 1:200,000 (refs [35,38]). For each PTEN variant observed in gnomAD, a binomial distribution of the total number of alleles successfully sequenced at the site was calculated, using a collective pathogenic allele estimate of 1:400,000, genetic and allelic heterogeneity of 1 and a penetrance of 95%, which are all conservative assumptions[8,39]. Each observed PTEN variant was assessed using the following line of code in RStudio: qbinom(0.99, size = (total alleles genotyped at the site), prob = (1/400,000)/0.95). PTEN variants in gnomAD with an observed allele count a full integer above this 99% confidence level of the calculated binomial distribution were considered variants highly unlikely to be causal for Cowden's syndrome.

Statistics and reproducibility. For all figures, r denotes the Pearson's correlation coefficient, whereas ρ denotes Spearman's rho rank correlation coefficient.

For our statistical analysis of the enrichments of low-abundance variants in the pathogenic, likely pathogenic and uncertain significance ClinVar categories, we used a resampling approach. We drew 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 (for example, SNVs 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.

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[42,59] to create mutational probabilities for every possible PTEN missense or nonsense variant. On the basis of these probabilities, we drew 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.

Rosetta $\Delta\Delta G$ predictions. Computational predictions of PTEN variant losses in folding energy (for example, $\Delta\Delta G$s) were performed using the 2017.08 release of Rosetta. The PTEN protein data bank (PDB) file 1d5r was renumbered to accommodate missing residues, and the TLA ligand was removed. Pre-minimization of the ensuing file was performed using Rosetta minimize_with_cst, followed by the convert_to_cst_file shell script. Fine-grain estimations of folding energy changes following PTEN alteration were created with Rosetta ddg_monomer[60] using the talaris2014 scoring function, and the following flags: -ddg:weight_file soft_rep_design, -fa_max_dis 9.0, ddg::iterations 50, -ddg::dump_pdbs true, -ignore_unrecognized_res, -ddg::local_opt_only false, -ddg::min_cst true, -constraints::cst_file input.cst, -ddg::suppress_checkpointing true, -in::file::fullatom, -ddg::mean false, -ddg::min true, -ddg::sc_min_only false, -ddg::ramp_repulsive true, -ddg::output_silent true.

Comparison of TPMT red blood cell activity or dose intensity to abundance scores. Genotypes, TPMT red blood cell activity that was normalized by cohort and dose intensity data for 884 acute lymphoblastic leukemia patients were provided from an earlier study[51]. The mean TPMT red blood cell activity and dose intensity from individuals heterozygous for each unique TPMT variant was calculated. These values were directly compared to abundance scores for that variant from the VAMP-seq assay or the WT-normalized GFP:mCherry ratio from individual flow cytometry experiments (Fig. 5 and Supplementary Fig. 7).

Western blotting. See the Supplementary Note for details of the western blotting procedures.

Code availability. Code used for the analyses performed in this work is included as Supplementary Data 5, and also available at http://github.com/FowlerLab/VAMPseq. Code used for subassembly by PacBio is available at http://github.com/shendurelab/AssemblyByPacBio.

Data availability. All raw sequence data and function scores are freely available for all academic users by non-exclusive license under reasonable terms to commercial entities that have committed to open sharing of *PTEN* and *TPMT* sequence variants and under a free non-exclusive license to non-profit entities. The Illumina and PacBio raw sequencing files and barcode–variant maps can be accessed at the NCBI Gene Expression Omnibus (GEO) repository under accession number GSE108727. VAMP-seq scores are available at http://abundance.gs.washington.edu. The data presented in the manuscript are available as Supplementary Data files.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

References

57. Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
58. Rubin, A. F. et al. A statistical framework for analyzing deep mutational scanning data. *Genome Biol.* **18**, 1–15 (2017).
59. Krauthammer, M. et al. Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. *Nat. Genet.* **44**, 1006–1014 (2012).
60. Kellogg, E. H., Leaver-Fay, A. & Baker, D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* **79**, 830–838 (2011).

nature research

Corresponding author(s): Douglas Fowler, Jay Shendure

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	PacBio Base call files were converted from the bax format to the bam format using bax2bam (https://github.com/PacificBiosciences/pbccs). PacBio consensus sequences for each sequenced molecule in every library were determined using the Circular Consensus Sequencing 2 algorithm with default conditions (ccs, https://github.com/PacificBiosciences/pbccs). Each resulting consensus sequence was then aligned to either the TPMT or PTEN reference sequence using Burrows-Wheeler Aligner60 (http://bio-bwa.sourceforge.net/). Sequencing reads were converted to fastq format and de-multiplexed with bcl2fastq. Barcode paired sequencing reads for PTEN experiments 1 through 4 were joined using the fastq-join tool within the ea-utils package (http://expressionanalysis.github.io/ea-utils/) using the default parameters.
Data analysis	Enrich2 v1.1.0 was used to count the barcodes, associate each barcode with a nucleotide variant, and then translate and count both the unique-nucleotide and unique-amino acid variants. FACSDIVA v8 and FlowJo v10 were used for FACS / flow cytometry collection and analysis, respectively. RStudio v1.0.136 was used for the analysis, with the versions of the relevant packages used listed in Supplementary File 1.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Upon publication, all raw sequence data and function scores will be made freely available for all academic users, by nonexclusive license under reasonable terms to commercial entities that have committed to open sharing of PTEN and TPMT sequence variants, and under a free non-exclusive license to non-profits entities. The data presented in the manuscript are available as Supplementary Tables. The Illumina and PacBio raw sequencing files and barcode-variant maps can be accessed at the NCBI Gene Expression Omnibus (GEO) repository under accession number GSE108727.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences

Study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Preliminary experiments were performed to assess noise and determine the number of samples that could be processed through the experimental work flow. A total of eight replicate experiments were then performed to obtain adequate estimates of variance for the majority of individual variants within the libraries. Sample sizes for downstream analyses were dependent upon the availability of the data present in the publicly accessible databases (e.g. ClinVar, TCGA, etc).
Data exclusions	As described in the methods section, experiments with poor technical replication across multiple bins were reamplified and resequenced in their entirety, resulting in the eight biological replicate experiments we report.
Replication	Confidence intervals were calculated for each abundance measurement. The majority of findings we discuss involve low abundance variants, which had reliable measurements across replicates. For a handful of variants, we also reproduced abundance scores and phenotypes with independent, orthogonal experiments (eg. testing variants for fluorescence individually, comparing variant scores with published western blotting phenotypes, and in-house western blot assays).
Randomization	There were no samples/organisms/participants allocated into experimental groups.
Blinding	There was no group allocation.

Materials & experimental systems

Policy information about availability of materials

n/a	Involved in the study
☐	☒ Unique materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Research animals
☒	☐ Human research participants

Unique materials

Obtaining unique materials	All unique materials are readily available from the authors.

Antibodies

Antibodies used	Western blotting was performed using a 1:2,000 dilution of anti-phospho-AKT (T308; 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).
Validation	These antibodies have been previously validated in our laboratories, and result in the predominant protein bands of the

expected molecular weight. Based on the manufacturer's website, the anti-beta-actin antibody has been used in 176+ publications as a loading control. Control cells lacking HA- or GFP- tagged proteins were first tested in western blotting experiments to validate the correct identification of the correct-sized band that appeared when HA- or GFP-tagged proteins were expressed. Furthermore, specificity of the antibodies were confirmed through observation of expected relative band intensities comparing variants of known effect. Based on the manufacturer's website, the pan- and phospho-AKT antibodies have been used in at least 19 and 104 publications, respectively. Control variants included in our experiments recapitulated observations made by other groups using similar antibodies to test for the same effects.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK 293T (ATCC® CRL-3216™) was the source cell line used, and it was obtained directly from ATCC. These cells were then modified through genome engineering steps performed in our laboratory to create the HEK 293T TetBxb1BFP Clone4 cells used in the experiments.
Authentication	The HEK 293T-based cell lines have not been authenticated.
Mycoplasma contamination	The HEK 293T TetBxb1BFP Clone4 and Clone37 cell lines used in this manuscript have tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Method-specific reporting

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ Magnetic resonance imaging

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	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 µm nylon mesh.
Instrument	Cells were sorted on a BD Aria III FACS machine using an 85 or 100 µm 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. 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.
Software	FACS Diva was used to collect the data. FlowJo_V10 was used for analysis.
Cell population abundance	The entire cell populations sorted were used for downstream analysis.
Gating strategy	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)

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.