Classifying disease-associated variants using measures of protein activity and stability

Michael Maglegaard Jepsen[a], Douglas M. Fowler[b], Rasmus Hartmann-Petersen[a], Amelie Stein[a,*] & Kresten Lindorff-Larsen[a,*]

[a] Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark.

[b] Departments of Genome Sciences and Bioengineering, University of Washington, Seattle, WA, USA.

Correspondence: amelie.stein@bio.ku.dk (AS) and lindorff@bio.ku.dk (KLL)

Abstract

Decreased cost of human exome and genome sequencing provides new opportunities for diagnosing genetic disorders, but we need better and more robust methods for interpreting sequencing results including determining whether and by which mechanism a specific missense variants may be pathogenic. Using the protein PTEN (phosphatase and tensin homolog) as an example, we show how recent developments in both experiments and computational modelling can be used to determine whether a missense variant is likely to be pathogenic. One approach relies on multiplexed experiments that enable determination of the effect of all possible individual missense variants in a cellular assay. Another approach is to use computational methods to predict variant effects. We compare two different multiplexed experiments and two computational methods to classify variant effects in PTEN. We distinguish between methods that focus on effects on protein stability and protein-specific methods that are more directly related to enzyme activity. Our results on PTEN suggest that ~60% of pathogenic variants cause loss of function because they destabilise the folded protein which is subsequently degraded. Methods that quantify a broader range of effects on PTEN activity perform better at predicting variant effects. Either experimental method performs better than the corresponding computational predictions.

Introduction

Technological advances have made human genome sequencing feasible in routine clinical contexts, revealing around 10,000 changes in the protein-coding regions of each individual[1], of which the majority will be very rare[2] and difficult to interpret[3,4]. We use state-of-the-art high-throughput methods for assessment of protein variants and discuss their performance in discriminating pathogenic from non-detrimental changes. This preprint is a summary of a submitted manuscript.

PTEN (phosphatase and tensin homolog) is a two-domain protein (Fig. 1A and 1B) with dual-specificity phosphatase that acts on both protein and PIP_3 lipid substrates. Its lipid phosphatase function is important for PTEN's role as a tumor suppressor where it counteracts phosphatidylinositol 3-kinase function[5]. A number of germline mutations in the

PTEN gene are associated with various diseases including autism spectrum disorder (ASD) and different tumor-risk syndromes collectively known as PTEN hamartoma tumor syndrome (PHTS)[5,6] (Fig. 1B).

High-throughput genome sequencing combined with assays specific to the function of the protein of interest enable highly parallel assessments of the functionality of each individual variant[7] in so called MAVEs (multiplexed assays of variant effects)[8]. Briefly explained, a MAVE involves creating a large library of variants and subsequently selecting for a property

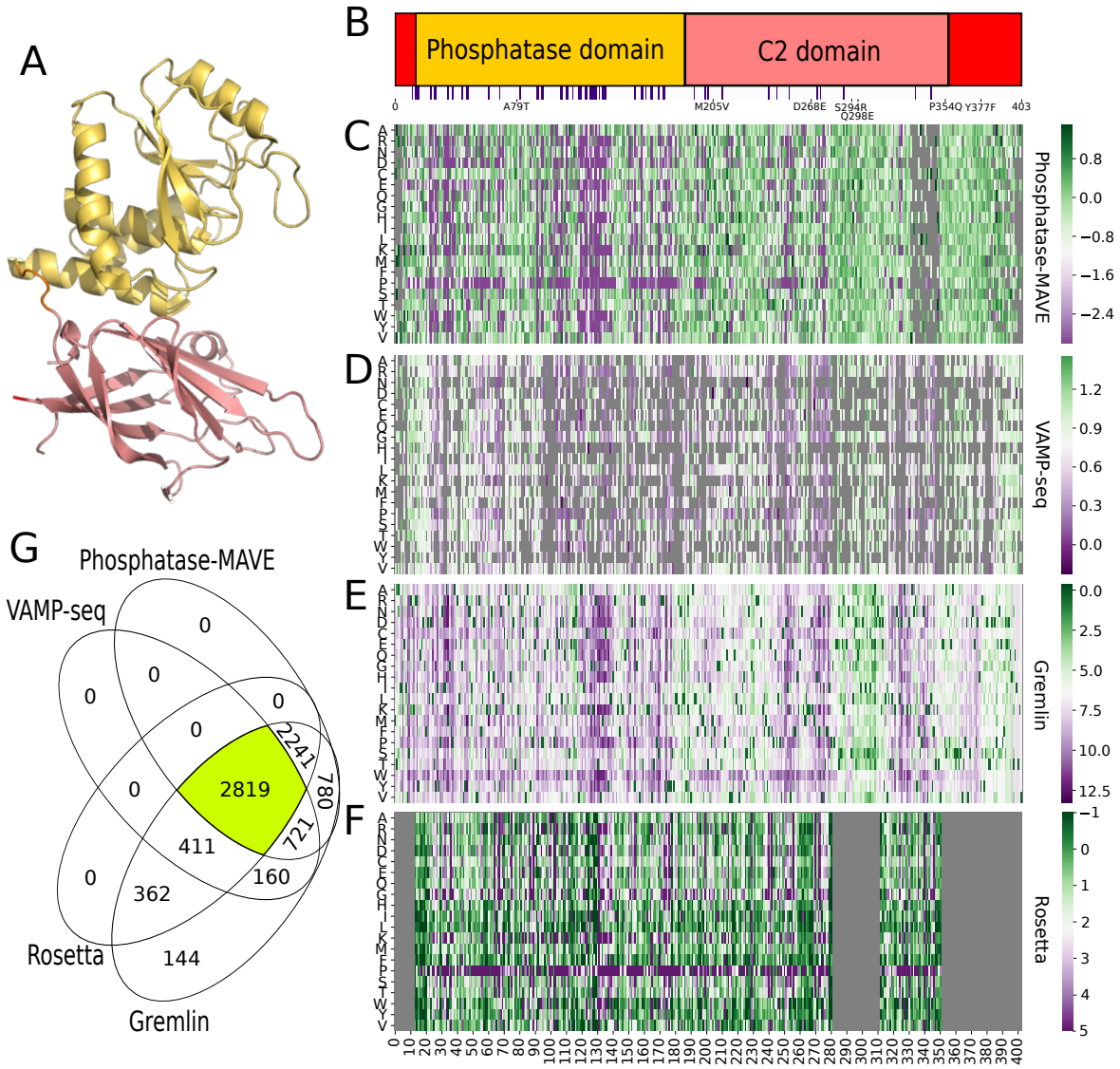

Figure 1: Overview of PTEN and the available single-variant data. A: We show the 3-dimensional structure of PTEN (PDB:1D5R (ref. 35)) with the phosphatase domain in yellow and the C2 domain in pink. B: PTEN domain architecture (top) and labels for gnomAD variants that are not singletons (bottom, details see main text). Purple bars indicate positions where pathogenic variants have been described. C-F: Experimentally and computationally determined scores for single site variants. Green indicates wild-type-like fitness/stability, magenta indicates low fitness/stability, gray indicates missing data. C: Phosphatase-MAVE[28], D: VAMP-seq[21], E: evolutionary sequence energies[24], F: Rosetta cartesian ΔΔG (ref. 25). G: Venn diagram illustrating the availability of scores from each method and their intersections.

of interest[9–12]. MAVEs have been applied to a substantial number of proteins and domains, overall indicating a surprising mutational tolerance as many missense variants retain wild type-like function[13,14].

In the case of PTEN, two different MAVEs have been performed using different selection systems. Mighell *et al*. evaluated the effects of PTEN variants on lipid phosphatase activity in a yeast system[15] (Fig. 1C) and were able to discriminate likely pathogenic from benign alleles[15]. We term this experiment Phosphatase-MAVE. We and others have shown that loss of protein stability and subsequent degradation by the cellular protein quality control machinery is an important factor that underlie pathogenicity of variants in diverse proteins[15–20]. Matreyek *et al.* recently introduced VAMP-seq (variant abundance by massively parallel sequencing), as a MAVE of variant abundance[21]. Application of VAMP-seq to *PTEN* showed that many pathogenic variants were of low cellular abundance and that the resulting data (Fig. 1D) showed separation between pathogenic variants and common variants in the human population.

In the absence of experimental data, computational methods can be used to predict the consequences of missense variants. We here complement and contrast the experimental Phosphatase-MAVE and VAMP-seq experiments with two computational methods, one that assesses sequence conservation through the calculation of an 'evolutionary sequence energy' ($\tilde{E}$) and one that quantifies changes in protein stability ($\Delta\Delta G$). These methods have previously been applied to the identification and classification of pathogenic variants[4,22,23]. Loosely defined, $\tilde{E}$-values are conceptually similar to the outcome of the Phosphatase-MAVE experiment and $\Delta\Delta G$-values are related to the outcome of the VAMP-seq experiment. We use the software Gremlin[24] to calculate evolutionary sequence energies (Fig. 1E) and the Rosetta software to calculate changes in protein stability[25] ($\Delta\Delta G$-values) (Fig. 1F).

Out of the 7,638 possible single amino acids, 2,819 could be assessed by all four methods (Fig. 1G), however below we focus only on the smaller subset of 71 variants that are either known to be pathogenic (42) or that are found in the gnomAD database[26] (29).

Results and Discussion

As a source for pathogenic PTEN variants, we extracted 87 pathogenic missense variants from ClinVar[27] (review criteria provided and no conflicting annotations; accessed August 2018). We turned to gnomAD[26] to examine whether there are variants that are sufficiently common in the population to make it likely that they do not cause disease[21,26,28]. Of the 80 PTEN variants we analyse 29 after removing those absent from the experimental and computational data (Fig. 1). Of these, A79T and M205V are found at allele frequencies of $1{\cdot}10^{-4}$ and $2{\cdot}10^{-5}$, respectively, in gnomAD, that suggest they are likely benign.

We analysed the distribution of scores for the gnomAD and pathogenic variants as obtained by the four different methods (Fig. 2) and find that all four methods show a clear difference in averages and distribution of values across the gnomAD and pathogenic variants. We then performed a 'receiver operating characteristic' (ROC) curve analysis of each of the four methods two quantify their ability to separate pathogenic from the gnomAD dataset (Fig. 3). The 'area under the curve' (AUC) show that two experimental approaches perform better than the corresponding purely computational method (Fig. 3), and that the two methods that

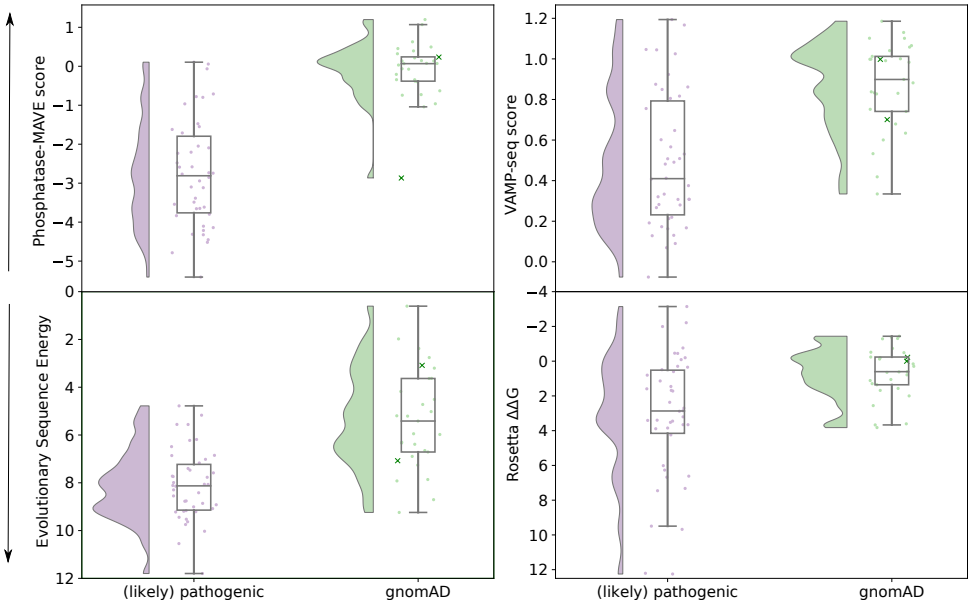

Figure 2: Distributions for pathogenic and gnomAD variants. Distributions of scores from each of the four methods for ClinVar pathogenic (mauve) and gnomAD (green) variants, shown as Raincloud plots[36]. Boxes illustrate the 25th-to-75th percentile, with the median indicated by a horizontal line. Whiskers extend to minimum and maximum value though at most 1.5 times the distance between the 25th and 75th percentile ("interquartile range") away from the box boundaries. As described in the main text, pathogenic variants were removed from the gnomAD set. A79T and M205V are specifically highlighted (crosses), as these are the most common PTEN variants in gnomAD, and thus the most likely to be benign. Only variants for which each of the four methods provide data are included (see also Fig. 1G). Axes are oriented such that values near the bottom correspond to detrimental variants, while values near the top are wild-type-like.

probe a wider range of functionally-relevant properties (Phosphatase-MAVE and evolutionary sequence energies) are able to capture variant effects more accurately than those that focus more directly on protein stability. Nevertheless, the high AUCs for the methods that probe only stability but not other aspects contributing to protein function (VAMP-seq and Rosetta) suggest that loss of stability and resulting decrease in cellular abundance is a key driver for PTEN-associated diseases in line with previous computational[4,29–31] and experimental[17–20,32] studies.

We show two-dimensional plots of the change in stability (VAMP-seq or Rosetta (ΔΔG) and "activity effects" (Phosphatase-MAVE or evolutionary sequence energies) [18,20] (Fig. 4) and separate the plots into four quadrants (Methods). Through this analysis we find that 27 of the 42 pathogenic variants fall in the unstable-and-non-functional quadrant in the experimental analysis (Fig. 4A) and 24 of the 42 variants fall in the corresponding quadrant in the computational analysis (Fig. 4B), suggesting that about 60% of disease-causing variants in PTEN cause disease via loss of stability and cellular abundance.

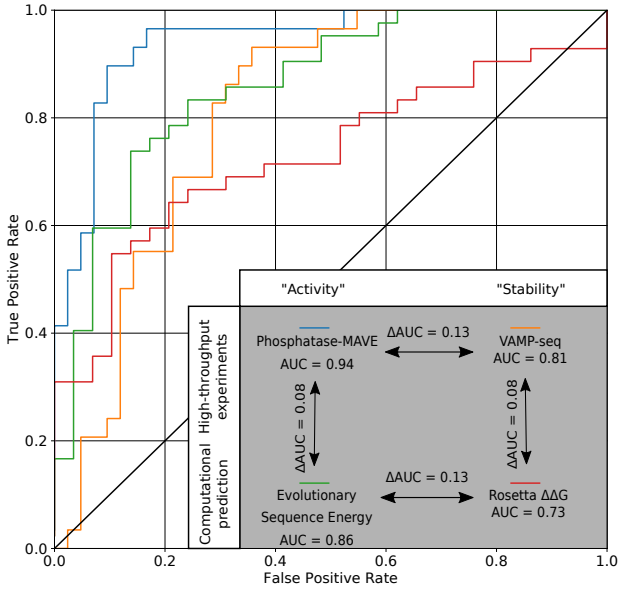

Figure 3: Receiver Operating Characteristic (ROC) curves for each method. ROC curves describe how well each method separates the ClinVar pathogenic variants from the gnomAD variants. As expected, methods assessing overall PTEN function perform better, with Phosphatase-MAVE experimental data (blue, AUC 0.94) exceeding evolutionary sequence energies (green, AUC 0.86) in overall sensitivity and specificity. Assessment of stability alone correctly identifies most pathogenic and gnomAD variants as well, again with experimental data (VAMP-seq, orange, AUC 0.81) exceeding performance of Rosetta ΔΔG calculations (red, AUC 0.73). The insert shows the AUC values and changes in AUC when moving between experimental and computational methods, and between methods that capture activity broadly or methods focused on stability.

Conclusions

We have here taken advantage of the availability of both Phosphatase-MAVE and VAMP-seq data for the human protein PTEN to analyse how well these methods are able to classify known pathogenic variants, and have compared the results to computational methods. Our results show that all four methods we tested are able to provide relatively accurate classifications of variant effects with AUCs ranging from 0.74-0.94 (Fig. 3). Here we remind the reader that we in our selection of pathogenic variants have combined disease-causing variants for different diseases (ASD and PHTS) although it has been suggested that ASD results from variants with milder effects compared to those that give rise to PHTS[28]. Similarly, we note that we have used also rare variants from gnomAD as a proxy for benign variants, and that some of these may indeed give rise to ASD or PHTS. With these caveats in mind we highlight two observations regarding prediction accuracies. First, the experimental methods give rise to AUCs that are about 0.1 unit higher than the corresponding computational method (Phosphatase-MAVE (0.94) vs. evolutionary sequence energies (0.86) and VAMP-seq (0.81) vs. Rosetta (0.73)). Second, the methods that probe function more generally give rise to AUCs that are about 0.1 unit higher compared to those that probe only protein stability (Phosphatase-MAVE (0.94) vs. VAMP-seq (0.81) and evolutionary sequence energies (0.86) vs. Rosetta (0.73)). One advantage of the methods that focus on protein stability is that they provide a more direct mechanistic model for how the variants cause disease. Indeed, our two-dimensional analysis of stability and 'function' reveal that many, but not all, disease-causing variants appear to cause loss of function

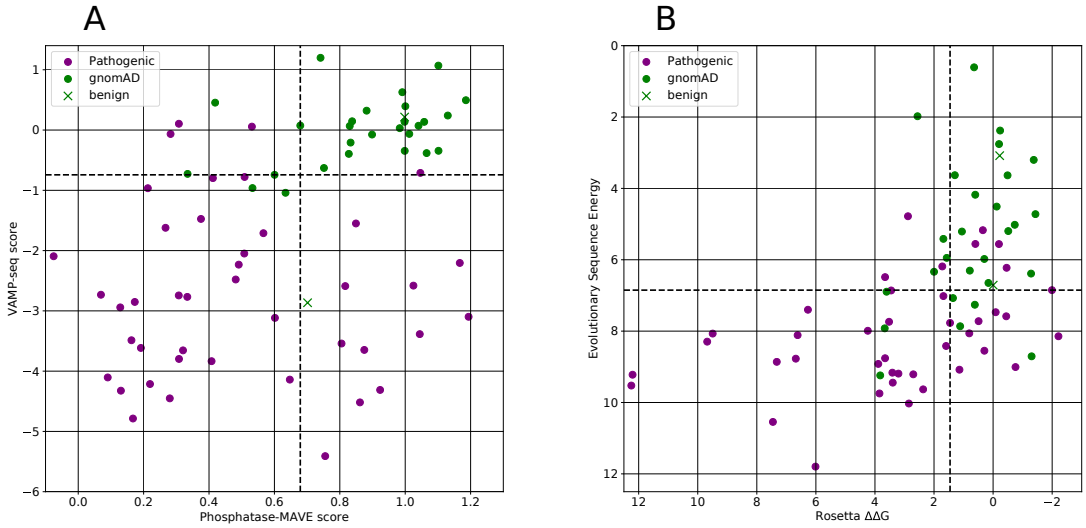

Figure 4: 2-dimensional landscapes integrating changes in stability and fitness. Combining scores from assessment of stability and function shows that most gnomAD variants (green) are wild-type-like, while most pathogenic variants (purple) are correctly identified by one, and often both metrics. Comparing the landscape based on experimental data (left) to that based on predictions (right) shows that availability of experimental data leads to better separation of pathogenic variants, but also that the predictions will provide a good starting point at substantially lower cost. The dashed lines correspond to cut-off values derived from the ROC curves, as those values that give rise to the point on the ROC curve closest to the upper-left corner (TRP=1, FPR=0).

because the variants cause loss of stability and decreased cellular abundance (Fig. 4). We conclude that at the current stage, experimental methods still perform better than the corresponding computational approaches, although the latter perform well on many variants too. Further, we note that many, but not all, pathogenic variants in PTEN appear to give rise to disease via loss of stability and cellular abundance.

Methods

Rosetta ΔΔG Calculations

We used the “cartesian_ddg” application in Rosetta with the ‘beta_nov_16’ variant of the Rosetta energy function to perform the ΔΔG calculations[25] using (PDB ID 1D5R) and Babel[33] to parameterise the ligand. To accomodate for the missing loop in the crystal structure (residues 281 to 313), we added the flag “-missing_density_to_jump” during relaxation of the structure.

When calculating ΔΔG we used three iterations, which were subsequently averaged. The resulting difference in stability was finally divided by 2.9 to bring the ΔΔG values from Rosetta energy units onto a scale corresponding to kcal/mol (Frank DiMaio, University of Washington; personal correspondence).

Evolutionary sequence energies (Ẽ)

We used HHblits[34] to create a multiple sequence alignment based on the UniProt sequence of PTEN (UniProt AC P60484) and used Gremlin[24] to to calculate the log odds score between each single residue variant and the wild type.

Phosphatase-MAVE and VAMP-seq Data
The VAMP-seq data can be accessed and downloaded at: https://abundance.gs.washington.edu/shiny/stability/.

The Phosphatase-MAVE data can be accessed in the supplementary material of the online article from Mighell et al.[28]

ROC-derived thresholds
Thresholds for separating likely pathogenic from likely benign variants are calculated for each metric by determining the point on the ROC curve that is closest to (0,1), the optimum where perfect specificity and sensitivity would be achieved. In this case we give equal weight to detection of pathogenic and benign variants, but note here that, depending on the application, different choices of threshold may be most suitable.

Analysis scripts
The scripts used in the analyses and for making the figures of this manuscript are available at https://github.com/KULL-Centre/papers/tree/master/2019/PTEN-variants-Jepsen-et-al.

Acknowledgements

The authors thank Dr. Sofie V. Nielsen and Mustapha C. Ahmed for helpful discussions. This work is supported by a Novo Nordisk Foundation Challenge Grant (PRISM) (D.M.F, R.H-P., A.S., K.L-L.), the Lundbeck Foundation (A.S.), and the National Institute of General Medical Sciences (1R01GM109110 and 1RM1HG010461 to D.M.F.). D.M.F. is a CIFAR Azrieli Global Scholar.

References

1. Kiryluk K, Goldstein DB, Rowe JW, Gharavi AG, Wapner R, Chung WK. Precision Medicine in Internal Medicine. *Ann Intern Med*. 2019;170(9):635. doi:10.7326/M18-0425
2. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285-291. doi:10.1038/nature19057
3. Manolio TA, Fowler DM, Starita LM, et al. Bedside Back to Bench: Building Bridges between Basic and Clinical Genomic Research. *Cell*. 2017;169(1):6-12. doi:10.1016/j.cell.2017.03.005
4. Stein A, Fowler DM, Hartmann-Petersen R, Lindorff-Larsen K. Biophysical and Mechanistic Models for Disease-Causing Protein Variants. *Trends Biochem Sci*. January 2019. doi:10.1016/j.tibs.2019.01.003
5. Yehia L, Ngeow J, Eng C. PTEN-opathies: from biological insights to evidence-based precision medicine. *J Clin Invest*. 2019;129(2):452-464. doi:10.1172/JCI121277
6. Leslie NR, Longy M. Inherited PTEN mutations and the prediction of phenotype. *Semin Cell Dev Biol*. 2016;52:30-38. doi:10.1016/j.semcdb.2016.01.030
7. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. *Nat Methods*. 2014;11(8):801-807. doi:10.1038/nmeth.3027
8. Gasperini M, Starita L, Shendure J. The power of multiplexed functional analysis of genetic variants. *Nat Protoc*. 2016;11(10):1782-1787. doi:10.1038/nprot.2016.135
9. Fowler DM, Araya CL, Fleishman SJ, et al. High-resolution mapping of protein

sequence-function relationships. *Nat Methods*. 2010;7(9):741-746. doi:10.1038/nmeth.1492

10. Starita LM, Young DL, Islam M, et al. Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics*. 2015;200(2):413-422. doi:10.1534/genetics.115.175802
11. Roscoe BP, Thayer KM, Zeldovich KB, Fushman D, Bolon DNA. Analyses of the Effects of All Ubiquitin Point Mutants on Yeast Growth Rate. *J Mol Biol*. 2013;425(8):1363-1377. doi:10.1016/j.jmb.2013.01.032
12. Staller MV, Holehouse AS, Swain-Lenz D, Das RK, Pappu RV, Cohen BA. A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain. *Cell Syst*. 2018;6(4):444-455.e6. doi:10.1016/j.cels.2018.01.015
13. Gray VE, Hause RJ, Fowler DM. Analysis of Large-Scale Mutagenesis Data To Assess the Impact of Single Amino Acid Substitutions. *Genetics*. 2017;207(1):53-61. doi:10.1534/genetics.117.300064
14. Kinney JB, McCandlish DM. Massively Parallel Assays and Quantitative Sequence–Function Relationships. *Annu Rev Genomics Hum Genet*. May 2019. doi:10.1146/annurev-genom-083118-014845
15. Clausen L, Abildgaard AB, Gersing SK, Stein A, Lindorff-Larsen K, Hartmann-Petersen R. Protein stability and degradation in health and disease. In: Donev R, ed. *Advances in Protein Chemistry and Structural Biology*. Vol 114. Molecular Chaperones in Human Disorders. Academic Press; 2019:61-83. doi:10.1016/bs.apcsb.2018.09.002
16. Callebaut I, Hoffmann B, Mornon J-P. The implications of CFTR structural studies for cystic fibrosis drug development. *Curr Opin Pharmacol*. 2017;34:112-118. doi:10.1016/j.coph.2017.09.006
17. Pey AL, Stricher F, Serrano L, Martinez A. Predicted Effects of Missense Mutations on Native-State Stability Account for Phenotypic Outcome in Phenylketonuria, a Paradigm of Misfolding Diseases. *Am J Hum Genet*. 2007;81(5):1006-1024. doi:10.1086/521879
18. Scheller R, Stein A, Nielsen SV, et al. Toward mechanistic models for genotype–phenotype correlations in phenylketonuria using protein stability calculations. *Hum Mutat*. 2019;40(4):444-457. doi:10.1002/humu.23707
19. Nielsen SV, Stein A, Dinitzen AB, et al. Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. *PLOS Genet*. 2017;13(4):e1006739. doi:10.1371/journal.pgen.1006739
20. Abildgaard AB, Stein A, Schultz-Knudsen K, et al. Structural destabilization and chaperone-assisted proteasomal degradation of MLH1 as a mechanism for Lynch syndrome. *bioRxiv*. April 2019:622266. doi:10.1101/622266
21. Matreyek KA, Starita LM, Stephany JJ, et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat Genet*. 2018;50(6):874. doi:10.1038/s41588-018-0122-z
22. Hopf TA, Ingraham JB, Poelwijk FJ, et al. Mutation effects predicted from sequence co-variation. *Nat Biotechnol*. 2017;35(2):128-135. doi:10.1038/nbt.3769
23. Feinauer C, Weigt M. Context-Aware Prediction of Pathogenicity of Missense Mutations Involved in Human Disease. *bioRxiv*. January 2017:103051. doi:10.1101/103051
24. Balakrishnan S, Kamisetty H, Carbonell JG, Lee S-I, Langmead CJ. Learning generative models for protein fold families. *Proteins Struct Funct Bioinforma*. 2011;79(4):1061-1078. doi:10.1002/prot.22934
25. Park H, Bradley P, Greisen P, et al. Simultaneous Optimization of Biomolecular Energy

Functions on Features from Small Molecules and Macromolecules. *J Chem Theory Comput*. 2016;12(12):6201-6212. doi:10.1021/acs.jctc.6b00819

26. Karczewski KJ, Francioli LC, Tiao G, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*. January 2019:531210. doi:10.1101/531210
27. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018;46(D1):D1062-D1067. doi:10.1093/nar/gkx1153
28. Mighell TL, Evans-Dutson S, O'Roak BJ. A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. *Am J Hum Genet*. 2018;102(5):943-955. doi:10.1016/j.ajhg.2018.03.018
29. Casadio R, Vassura M, Tiwari S, Fariselli P, Martelli PL. Correlating disease-related mutations to their effect on protein stability: A large-scale analysis of the human proteome. *Hum Mutat*. 2011;32(10):1161-1170. doi:10.1002/humu.21555
30. Pal LR, Moult J. Genetic Basis of Common Human Disease: Insight into the Role of Missense SNPs from Genome-Wide Association Studies. *J Mol Biol*. 2015;427(13):2271-2289. doi:10.1016/j.jmb.2015.04.014
31. Redler RL, Das J, Diaz JR, Dokholyan NV. Protein Destabilization as a Common Factor in Diverse Inherited Disorders. *J Mol Evol*. 2016;82(1):11-16. doi:10.1007/s00239-015-9717-5
32. Bullock AN, Henckel J, Fersht AR. Quantitative analysis of residual folding and DNA binding in mutant p53 core domain: definition of mutant states for rescue in cancer therapy. *Oncogene*. 2000;19(10):1245. doi:10.1038/sj.onc.1203434
33. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. *J Cheminformatics*. 2011;3(1):33. doi:10.1186/1758-2946-3-33
34. Zimmermann L, Stephens A, Nam S-Z, et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J Mol Biol*. 2018;430(15):2237-2243. doi:10.1016/j.jmb.2017.12.007
35. Lee J-O, Yang H, Georgescu M-M, et al. Crystal Structure of the PTEN Tumor Suppressor: Implications for Its Phosphoinositide Phosphatase Activity and Membrane Association. *Cell*. 1999;99(3):323-334. doi:10.1016/S0092-8674(00)81663-3
36. Allen M, Poggiali D, Whitaker K, Marshall TR, Kievit R. Raincloud plots: a multi-platform tool for robust data visualization [PeerJ Preprints]. https://peerj.com/preprints/27137/. Accessed June 7, 2019.