Discovering and interpreting genetic variation in neurological disorders

By

Taylor L. Mighell

A DISSERTATION

Presented to the Neuroscience Graduate Program,

Department of Molecular & Medical Genetics,

and Oregon Health & Science University School of Medicine

In partial fulfillment of

The requirements for the degree of:

Doctor of Philosophy

June 2020

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the PhD Dissertation of

Taylor L. Mighell

Has been approved:

Advisor, Brian J. O'Roak

Member and Chair, Kevin M. Wright

Member, Gail Mandel

Member, Philip J. S. Stork

Member, Yoon-Jae Cho

Table of Contents

Acknowledgments

I am fortunate to have had the opportunity to perform the research outlined in this dissertation. It was only possible with the support of a large network.

First and foremost, I would like to thank Brian O'Roak for being my PhD advisor. His creativity, boldness, and extremely high standards have helped me to become a better scientist and thinker. I also thank Andrew Adey, who was effectively a second advisor, for countless pieces of advice on everything from experiments to presentations and papers.

My dissertation advisory committee, including Kevin Wright, Gail Mandel, and Phil Stork, has also been instrumental in shaping my graduate studies and helping to keep me on the right track. I am frequently awed by their collective wisdom.

Additionally, I would like to thank the graduate student body at OHSU. The students of the NGP are among the most inspirational individuals I have had the fortune to meet, and I consider all of them as role models. The members of my cohort, especially, are among the kindest and smartest people I've ever met. I also want to especially thank the members of the O'Roak and Adey labs, who have made graduate school much more fun than it otherwise might have been.

Finally, I want to thank my family for being unrelentingly supportive of my research career. I thank my parents, Greg and Kay, for selflessly providing everything I could have ever wanted or needed. I thank my brothers, Cody and Travis, for keeping things interesting and providing much needed distractions.

Abstract

While recent advances in DNA sequencing have revolutionized clinical genetics, substantial challenges remain. Among the most pressing challenges are reliably and affordably discovering disease-causing variation in patients, and accurately interpreting the functional effects of detected variation. It is now possible to sequence whole human genomes in a matter of days, and at a cost of thousands of dollars. As a result, there has been a staggering accumulation of sequence data from healthy and affected individuals. However, we currently lack generally applicable methods for interpreting the functional consequences of variation. As long as this problem remains unsolved, we will be unable to realize the full potential of constantly advancing DNA sequencing technologies. One solution to this variant interpretation problem is deep mutational scanning (DMS), a novel experimental framework that enables the characterization of thousands of gene variants in parallel. Here, I have developed and implemented a DMS platform for characterizing the phosphatase and tensin homolog (*PTEN*) gene, which is a tumor suppressor among the most commonly somatically mutated genes in several cancers. Germline *PTEN* mutations can lead to a tumor predisposition syndrome called PTEN Hamartoma Tumor Syndrome (PHTS). Curiously, some individuals develop autism spectrum disorder (ASD) with or without accompanying PHTS. I used the DMS platform to measure the enzymatic activity of 7,244 variants, representing ~85% of all possible single amino acid variants. I confirmed a pre-existing hypothesis that hypomorphic *PTEN* variants (i.e. those that retain partial activity) associate with ASD, while completely abrogated variants associate with PHTS. Around the same time, another group developed a platform for

measuring the effects of PTEN variation on protein abundance. In order to maximize the insights from these two complementary datasets, I overlaid them with the largest cohort of well-phenotyped *PTEN* mutation carriers in the world. From this confluence of molecular and human phenotypic data, I establish that DMS data partially explain quantitative clinical traits, identify pathogenic and benign variation, and define subgroups with distinct cancer susceptibility. I also add nuance to the hypothesis that hypomorphic *PTEN* variants lead to ASD. In fact, variants with *any* compromised activity equally increase the chances of developing ASD, while the chances for developing PHTS are strongly related to the magnitude of variant defect.

In addition to interpretation of genetic variation, methods for discovering genetic variation in an affordable and facile way remain necessary. One successful approach has been targeted enrichment, in which sequences of interest are specifically enriched prior to sequencing. Clinical genetics stands to gain greatly from targeted enrichment technologies. This is because, while several disorders have been linked to a relatively small number of genes, it remains challenging to specifically sequence full genes. Current targeted enrichment technologies suffer from several disadvantages, including limited flexibility, GC-content sequence bias, and capturing only the coding portion of the genome. Here, I describe the development of a novel targeted enrichment technology that empowers flexible, low-bias, targeted enrichment of genomic loci of interest. The technology utilizes the CRISPR-Cas12a system, which, similarly to CRISPR-Cas9, enables programmable nuclease activity. In contrast to Cas9, Cas12a cleaves DNA to leave stereotyped single-stranded overhangs. I show that these overhangs can be specifically ligated to sequencing

adapters. This method enables ~50-fold enrichment of sequences of interest, greatly reducing sequencing and associated costs. Overall, these advances can have an immediate impact on clinical discovery and interpretation of human genetic variation, and they represent progress towards fully realizing the potential of human clinical genetics.

Chapter 1. Introduction

1.1 Overview and rationale

One of the greatest promises of the genome sequencing revolution is the ability to predict health outcomes of individuals based on their genome sequence. Whole genome sequencing of newborns may be widespread in the not-too-distant future, and the accurate prediction of health issues could substantially improve the quality of life of patients, save time and resources, and simplify clinicians' decision making. However, until we improve our ability to interpret genetic variation, the clinical usefulness of prospective genome sequencing technology will be severely hampered.

1.1.1 The extent of human genetic variation

The original draft of the human genome, completed in 2001 by the Human Genome Project, came at a cost of billions of dollars and several years of work from labs all over the world[1]. Since then, advances in DNA sequencing technology have enabled the relatively facile sequencing of human genomes in a matter of days and at a cost of thousands of dollars. Spurred by the ease of genome sequencing, multiple consortia have set out to sequence thousands of human genomes with the goal of understanding human genetic variation and to create a framework with which to understand human disease. One of the most ambitious efforts to date is led by the Genome Aggregation Database Consortium (gnomAD). This group has aggregated data from other studies representing 125,748 human exomes (protein-coding portions of the genome) and 15,708 human genomes[2]. An important finding from this

data is that human genetic variation is extensive; 7.9% of all high-quality, protein-coding sites are multi-allelic (that is, different bases occur at these positions in different people). Strikingly, it is estimated that each person carries 100-400 amino acid-altering mutations that have never been seen before in the clinic[3].

A second striking finding from large scale sequencing efforts is that there is extensive misclassification of disease-associated variation. On average, each individual in the gnomAD database harbors 54 alleles that are considered Mendelian disease-causing in either the Human Gene Mutation Database or ClinVar, two of the most widely used clinical genetic databases[4]. While it is possible that some of these alleles have low penetrance, it is likely that many are misclassified. Additionally, when considering all alleles deposited in ClinVar, over half are considered "variants of uncertain significance"[5]. Taken together, this suggests that our classification of alleles as disease-causing is generally inaccurate.

1.1.2 Historical strategies to interpret human genetic variation

Over the years, researchers and clinicians have developed several methods to identify disease-causing variation. Segregation analysis describes a technique in which an allele is traced through a pedigree. If the allele specifically coincides with a positive disease outcome, it can be presumed that the allele is associated with the disease. However, this method is extremely low throughput and retrospective in nature, and limited to situations in which full pedigrees can be traced. Genome-wide association studies (GWAS) and expression quantitative trait loci (EQTL) are higher throughput, and have been effective at identifying the genetic underpinnings of some diseases. These studies rely on large samples of affected and unaffected individuals,

and seek to identify genotypes that occur more frequently in affected individuals. However, these methods are only sensitive to alleles with relatively high population frequency, as they fundamentally compare outcomes between individuals with and without any given allele. Further, due to linkage disequilibrium, it is often challenging to pinpoint the precise variation responsible for the phenotype.

An alternative approach seeks to leverage computation and informatics to predict the effects of mutations *in silico*. An obvious advantage of this approach is that it is widely accessible and extremely high throughput. Laboratory equipment, reagents, and expertise are not needed. Computational predictors rely on sequence features at the DNA and protein level to predict the functional effect of variation. Early algorithms sought to interpret variation in protein coding regions, leverage evolutionary conservation, biophysical and biochemical properties of the substituted amino acids, and the structure of the protein in question.[6,7] As the corpus of functional genomic data burgeoned with efforts such as the ENCODE project[8], newer methods sought to combine many functional annotations, including regulatory information and transcription factor binding sites, along with predictions made by simpler models, to predict variant effects in both coding and non-coding regions.[9] The most recent generation of *in silico* predictors brings to bear the power of deep learning and massive datasets to predict mutation effects.[10] Unfortunately, the predictions made from these models are generally not accurate enough to be useful in the clinic.[11] This stems from several reasons. First, many of these models use evolutionary conservation as a core predictive feature. This may work well for many genes, however it is impossible to know if the human version of a gene has evolved human-

specific function that would confound traditional evolutionary conservation metrics. Further, while our knowledge of protein sequence-function relationships is constantly improving, we have still only sampled a very small fraction of protein sequence space, making the predictions of variant effect challenging. Finally, the majority of the genome is not protein-coding; it likely has some regulatory or structural function. However, our knowledge of the function of non-coding regions of the genome is especially limited, especially considering that different cell types and different contexts display different regulatory activity.

The other major class of methods for interpreting variant pathogenicity is functional assays. Generally, these are laboratory assays designed to measure the functionality of a mutated form of a protein. A simple example would be measuring the catalytic output of a mutated enzyme, compared to the wild-type version. While these assays represent the gold standard for variant interpretation, there are significant challenges in scaling up to large numbers of assayed variants. Further, and as a result of the low throughput, they are typically only performed in a post-hoc manner after a variant has been identified. This can add months or years to the diagnostic odyssey.

There are several factors that limit throughput for traditional functional assays, which generally stem from a requirement to physically compartmentalize individual experiments. The generation and validation of variant sequences requires labor and reagents, and in most cases is impractical for more than hundreds of variants. Further, the actual assay would typically be performed in micro-well plates,

testing a single hypothesis per well, which limits the practical scale to hundreds of variants.

A striking example of traditional functional assays taken to the most extreme is a tour-de-force experiment measuring the functional effects of all possible point-mutation-accessible (2,314) single-amino acid substitutions in the tumor suppressor p53 in a serial, 96-well format[12]. While invaluable, this study was extremely laborious and remains a unique example of the approach nearly 20 years later.

1.1.3 Deep mutational scanning as a solution to the variant interpretation problem

Several technological advances in DNA sequencing and synthesis have led to a new experimental paradigm for variant interpretation, known as deep mutational scanning (DMS). The primary conceptual advantage over "traditional" functional assays is the relaxation of the requirement for physical separation of individual hypotheses. The general approach involves making many mutations *en masse* (at the DNA level) to a gene of interest, expressing the mutated forms of the gene in a population of cells, and then placing the cells in a selective

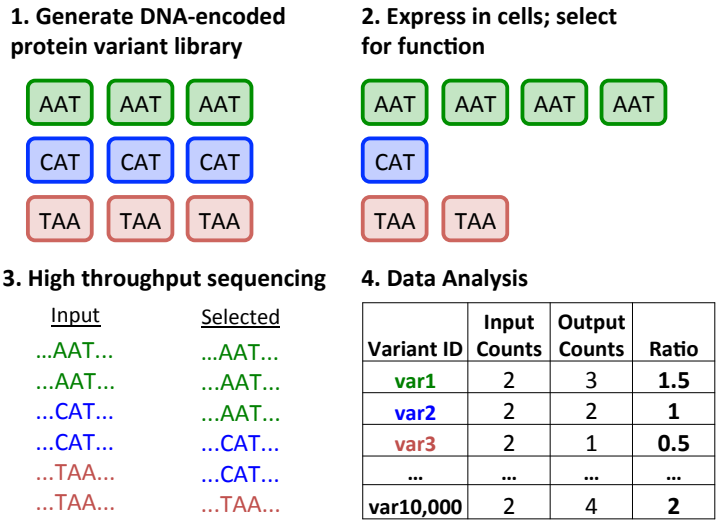


Figure 1. Overview of the deep mutational scanning method.

First, variants are constructed at the DNA level. Then, these variants are expressed in a cellular system and selected for function. DNA is collected from input and selected cells, and then sequenced to tabulate abundance of all variants. The ratio of variant abundance in the selected and input pool is a proxy for variant functionality.

environment which either selects for or against functional alleles[13] (Figure 1). Selection, in this sense, can take many forms. Most simply, the survival of cells could be linked to the function of the gene under study, in which case functional alleles would enrich and non-functional alleles would deplete. Alternatively, functional alleles could drive expression of a fluorescent protein allowing for fluorescence activated cell sorting (FACS). Likewise, if the transcript were a variant-linked barcode, this could be detected directly by RNA sequencing. Regardless of the details of selection, a critical next step uses next generation sequencing techniques to measure the fitness of each protein variant in parallel (Figure 1). The throughput of a DMS experiment is no longer limited by the need for physical compartmentalization. Instead, the limits are imposed by the number of cells that can be practically cultured and the number of sequencing reads that can be practically afforded. In practice, single DMS studies have assayed the effects of hundreds of thousands of unique alleles.[14]

Early studies leveraged DMS to understand fundamental biochemical properties of microbial proteins, such as thermodynamic stability[15] or RNA binding[16]. However, it later became apparent that DMS could be leveraged to measure the functional impact of mutations to human proteins, and potentially offer an empirical and accurate solution to the variant interpretation problem. There are several key advantages of DMS. First and foremost, the parallelization allows a single study to obtain precise measurements for hundreds of thousands of mutations (likely more in the future), which is a large enough number to saturate mutation space for most genes. Second, since the measurements are done in parallel and in the same

conditions, batch effects from different laboratories, cell lines, technicians, or other uncontrolled variables can be avoided. Third, saturating mutation space for a gene provides an extremely information dense dataset that can be used to inform deeper understanding of the form and function of the studied protein. Fourth, since all mutations are prospectively measured, individuals presenting with any possible mutation can be treated in an informed way. Finally, DMS datasets can be aggregated and used as training sets for statistical models that could eventually predict the effects of mutations in other genes with high accuracy[17].

1.1.4 Technical considerations for deep mutational scanning experiments

The first challenge in a DMS experiment is the generation of the library of variant sequences that will be assayed. There are several approaches to make many mutations to a wild-type sequence. The simplest, but generally least effective, is error prone PCR. This approach uses low-fidelity polymerase or PCR additives to increase the error rate. A drawback of this approach is that it generally only introduces single nucleotide changes, limiting the available mutagenesis space. Also, it is challenging to introduce one and only one mutation per allele. Another set of approaches attempts to adapt site-directed mutagenesis methods to introduce mutations at many different sites. The PFunkel approach uses a uracil-containing circular DNA template, to which a mutagenic oligo is annealed and extended. Further processing removes the wild-type strand with exonucleases that specifically act on uracil, yielding a mutagenized plasmid[18]. A common challenge for this type of approach is depleting the wild-type sequence. A different type of approach requires synthesis of short, mutagenic oligonucleotides that are then recombined into an otherwise-wild-type backbone[19].

These methods are more laborious but easier to deplete wild-type sequences and achieve balanced representation of alleles. Once a library of mutated sequences has been generated, these need to be introduced into a cellular model. Appropriate transformation approaches can be used for bacteria or yeast, and lentiviral transduction can be used for human cells. In most cases it is desirable to transduce only one variant per cell; the simplest way to ensure this is to transform or transduce with a relatively low molar ratio of DNA vector to cells. More recently, CRISPR-based systems have been used to introduce saturating mutations[20]. This approach has the key advantage of being able to introduce mutations directly into the genome (instead of a plasmid).

Once a library of variants has been created, the most critical component of a DMS experiment is challenging the variant sequences in a way that stratifies by function and also permits sequencing as a readout. A popular and simple approach is to express the variants in a cellular system whose survival or proliferation is dependent on the function of that gene. Cells that contain a functional variant will increase in abundance when compared to cells without a functional variant, and the abundance of variants can be measured by DNA sequencing. An approach that doesn't depend on selecting for fitness involves using FACS. For example, transcription factor variants could be expressed in cells that also contain a fluorescent reporter gene downstream of that transcription factor's binding site. A functional transcription factor variant would drive higher levels of fluorescent protein production, which could be distinguished from low levels of fluorescent protein by FACS. Finally,

sequencing could be used to detect the relative abundance of all alleles in the high and low fluorescence bins.

Detecting variants by sequencing can become a challenge as the length of mutated sequence exceeds the length of reads supported by current sequencing platforms. One approach to get around this is to incorporate unique DNA "barcodes", that is, a relatively short sequence of unique (often random) nucleotides that are linked with a variant[21], and can thus be used to identify the variant. This allows the researcher to sequence, for example, a 16 nucleotide barcode that uniquely identifies a DNA variant sequence that could be 1,000 nucleotides in length. A simpler but more laborious approach is to split the coding sequence of the gene into fragments that are shorter than the maximum read length of the sequencing technology in use, and then perform mutagenesis, selection, and sequencing in parallel for all fragments.

Finally, once the selection experiment has been performed, sequencing reads must be analyzed in order to attribute functional scores to variants that were present in the experiment. The exact methods employed here will depend on experimental design, but generally either the actual coding sequence or a variant-linked barcode will be sequenced. Either way, variant sequences are identified and tabulated, and relative abundance of the variant can be used to infer functionality. The wild-type sequence is typically included in the pool of variant sequences, and this can act as a positive control. However, it is often desirable to have several different measurements of the wild-type function; to accomplish this, groups often identify all variant sequences that still code for the same amino acid sequence (i.e. variants with synonymous variation). This allows for an estimate of the lower and upper bounds of

wildtype-like functionality. Likewise, practitioners often use the population of early truncating mutations, which are presumed to be null, to estimate the upper and lower bounds of null mutations. All other variants in the experiment can then be interpreted by comparing to the wildtype-like and loss-of-function distributions.

1.1.5 Applications of DMS for interpreting or understanding human genetic variation

One of the first DMS studies to query a human disease gene/protein was an analysis of BRCA1 function[22], variants in which can strongly predispose for breast and ovarian cancer. The authors developed two different DMS assays to measure BRCA1 enzymatic capacity as well as its ability to bind an important interacting protein, BARD1[22]. This study resulted in two important advances for the field. First, the authors showed that predictions of variant effect derived from DMS experiments could perform better than *in silico* predictors of variant effect. Second, the authors demonstrate that probing multiple parameters of protein function results in better prediction accuracy than probing a single parameter in isolation[22]. This is intuitive, as proteins exist in a complex molecular milieu and the full range of a protein's function is not likely to be captured in a single assay. A key obstacle faced by researchers using DMS as a tool to prospectively assess variant effect is that the empirical scores need to be cross-validated with human clinical data, the quality and quantity of which is often lacking. For example, at the time of the BRCA1 study, only 22 relevant variants had been classified for pathogenicity[22].

In a later study, nearly all missense variants in PPARγ were queried[23]. Mutations in this gene predispose for lipodystrophy as well as type 2 diabetes. The

variant library was expressed in cells lacking endogenous PPARγ, and cells were FACS sorted based on expression of CD36, a transcriptional target of PPARγ. A conceptual advance made in this study was the authors' recognition that they could use mutations found in population sequencing studies as putatively benign alleles. This gives practitioners greater power to assess the accuracy of their empirical scores.

Finally, a recent series of yeast-complementation-based DMS experiments were used to demonstrate another critical advance. Due to various biochemical biases, as well as random sampling and bottlenecking during the library generation process, it is difficult to generate a library with 100% of all intended variants represented. Weile e*t al.* showed that machine learning models could be trained on the fraction of variants that were measured, as well as evolutionary, biochemical, and biophysical parameters, to impute the likely score of variants that weren't measured in the study[24]. This means that researchers can design experiments that seek to only saturate a certain fraction of all possible mutations (e.g. 60-80%), saving time and resources.

While these studies set the stage for the expanded use of DMS for clinical variant interpretation, challenges remained. There was little agreement on the best approach for mutagenesis. Likewise, authors generally used custom and inconsistent approaches for analyzing DMS data. Finally, for any given gene, it is critical to select an experimental model that will report on the most clinically significant function of the gene. This can be particularly challenging for genes with multiple functions, or that localize in different cellular compartments, or function in highly specialized physiological niches (i.e. neuronal synapse or muscle fiber). An alternative approach

would probe a general protein feature, such as steady state cellular abundance (see Chapter 3). This approach can be applied to any protein but will not report on specific functional activities.

1.2 PTEN as a critical challenge for clinical genetics

1.2.1 PTEN biochemistry and cell biology

PTEN was discovered and identified as a tumor suppressor mutated in several different types of cancers in 1997[25,26]. Biochemically, PTEN is a dual-specificity phosphatase, with activity towards both protein and lipid substrates. Specifically, PTEN dephosphorylates serine, threonine, and tyrosine residues of acidic protein substrates[27], as well as position 3 of the inositol ring of phosphatidylinositol 3,4,5-trisphosphate (PIP_3)[28]. The early observation that the oncogenic G129E missense substitution specifically ablates lipid phosphatase activity while preserving protein phosphatase activity emphasized the critical importance of the lipid phosphatase activity[29]. Dephosphorylation of PIP_3 is a critical antagonist for the phosphatidylinositol 3-kinase signaling pathway[28,30], activation of which leads to Akt-mediated signaling through several downstream effectors, leading to cell survival, growth, proliferation, and migration[31]. Experimental deletion of *PTEN* in mouse embryonic stem cells[32], mouse brain[33,34], and mouse heart[35] demonstrated that cells lacking *PTEN* exhibit increased PI3K/Akt signaling. This results in deregulated growth and activity, which can eventually result in hypertrophy, hyperplasia, and cancer formation. A later study made the critical finding that even a subtle decrease in PTEN activity can increase the risk of cancer development in mice[36]. This was in

contrast to the two-hit cancer model, which posited that loss of both copies of a tumor suppressor gene was the root cause of cancer. PTEN became the paradigmatic case for the continuum model of cancer, in which disease severity increases as activity level of the tumor suppressor decreases[37].

The partial crystal structure of a PTEN construct was solved in 1999[38]. This construct lacked the 7 residues at the N-terminus, 49 residues at the C-terminus, and 24 residues in an internal loop, all of which represented unstructured regions[38]. The canonical PTEN protein product consists of 403 amino acids, which form two globular domains, a phosphatase and a C2 domain, which closely interact[38] (Figure 2A-B). The catalytic pocket of the phosphatase domain is larger than the pocket of VHR, a prototypical dual-specificity phosphatase, which accommodates the larger lipid head group[38]. The C2 domain is composed mainly of beta sheets, and functions to bring the protein into close proximity to the inner leaflet of the plasma membrane, positioning the phosphatase domain to dephosphorylate PIP_3[38]. C-terminal of the C2 domain is an unstructured domain, commonly known as the C-terminal tail. The function of this domain was discovered to play a role in auto-regulation. Specifically, four serine/threonine residues are phosphorylation targets, and in the

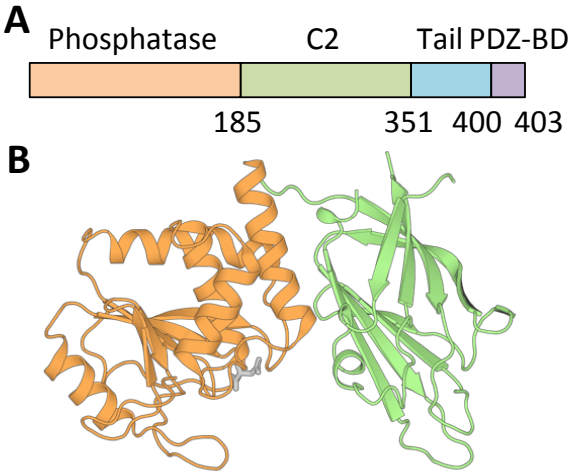


Figure 2. Structure of the PTEN protein. (A) Schematic diagram of PTEN in primary sequence.

(B) Crystal structure of PTEN, colored as in A. The tail and PDZ-BD were not present in the crystallized construct.

phosphorylated state they cause the tail to fold back and inhibit the activity of PTEN[41,42]. This provides a key mechanism for regulating the activity of PTEN. CK2[43] and GSK3[44] have been shown to phosphorylate PTEN's tail, and, intriguingly, the tail can be dephosphorylated, and thus the enzyme de-inhibited, by PTEN itself[45]. Finally, PTEN contains a PDZ-binding motif at the very C-terminus. This motif is critical for several protein-protein interactions, which have been shown to modulate PTEN subcellular localization, as well as forming tightly orchestrated signaling complexes[46].

Early investigations into the subcellular localization of PTEN led to the idea that PTEN exists primarily in the cytoplasmic compartment of cells, and transiently interacts with the inner leaflet of the plasma membrane, where it could interact with the PIP_3 substrate[47,48]. Later studies found evidence of PTEN within several other cellular compartments, including the endoplasmic reticulum and mitochondria[49,50], the nucleus broadly[51], as well as nucleoli[52]. The proposed functions of PTEN at these various subcellular loci will be explored in section 1.2.2, but they are generally poorly understood, and it is therefore challenging to predict how specific disruption of PTEN function at these sites would affect cellular or organismal outcomes.

Two recent discoveries have complicated the view of PTEN form and function. First, biochemical and cell biological evidence suggested that PTEN actually forms dimers and that the dimers are the catalytically active unit (see section 1.2.6 for further discussion)[53,54]. This cell biological and biochemical evidence dovetails with findings in mice that some missense mutations (namely, those that completely abolish lipid phosphatase while preserving stability) actually lead to more severe tumorigenic outcomes than mice lacking a copy of *PTEN*[55]. Therefore, any model that

attempts to predict human clinical phenotype from *PTEN* genotype will need to potentially take dominant negative effects into consideration.

Second, recent studies have shown that PTEN can be translated via alternative, upstream start codons that lead to two N-terminal extended isoforms. [39,40] These isoforms have been reported to localize in the mitochondria and the nucleus, but a full account of their function remains elusive. It is believed that one of these isoforms can be secreted from cells and re-enter other cells[56], and that it may have a role in mitochondrial bioenergetics[39]. Generally speaking, it is poorly understood how mutations differentially affect the canonical PTEN versus this N-terminal extended isoform. This will likely need to be clarified in order to complete an accurate *PTEN* genotype-phenotype map.

1.2.2 PTEN noncanonical functions

Beyond PTEN's vital role in antagonizing the PI3K/Akt signaling pathway through dephosphorylation of PIP_3, several noncanonical functions have been described. First, PTEN has phosphatase activity towards protein substrates. An important substrate is the cluster of phosphorylatable residues in PTEN's C-terminal tail[45]. Additional protein substrates include focal adhesion kinase 1, cAMP-responsive element-binding protein 1, proto-oncogene tyrosine-protein kinase SRC, and insulin receptor substrate 1[57–60]. While these molecules exist in diverse signaling pathways, it is believed that the ultimate effect of PTEN's activity towards these protein substrates is tumor suppression[61].

Understanding the role of PTEN in the nucleus has received substantial attention. This is because some pathogenic mutations affect nuclear localization

without overtly compromising lipid phosphatase activity or steady state stability[62]. In fact, a recent study showed that forcing mislocalizing PTEN missense variants to the nucleus actually rescued the cellular hypertrophy that was observed without forced nuclear localization[63]. These findings suggest that PTEN is playing an important role in the nucleus and disruption of this function could have an effect on human health. However, exactly which roles are most important is as yet unclear. Proposed nuclear functions for PTEN include regulating ribosome biogenesis[52], promoting genome stability[64], regulating DNA replication[65], and controlling chromatin condensation and thereby gene expression[66].

1.2.3 PTEN mutations and human health

PTEN was originally identified as a tumor suppressor upon the karyotypic observation of recurrent deletions of chromosome 10q23 in multiple human cancers[26,67,68]. Further screening demonstrated extremely high mutation frequency in several types of human cancers, especially breast, thyroid, glioblastoma, and prostate (reviewed in Sansal *et al.,* 2004[69]). Besides somatic mutation, it was also discovered that germline *PTEN* mutations led to a variety of tumor-predisposition and overgrowth disorders. Cowden Syndrome described an adulthood presentation characterized by tumors of the breast, thyroid, and skin, and was found to be the result of *PTEN* mutation[70,71]. Lhermitte-Duclos syndrome, characterized by gangliocytomas of the cerebellum, and Bannayan Riley Ruvulcaba Syndrome (BRRS), which is characterized by macrocephaly, lipomatosis, hemangiomas, and speckled penis, typically appearing in childhood[72], were determined to be variable presentations of the same genetic entity as Cowden Syndrom[72]. Another phenotypic

outcome for *PTEN* mutation carriers is neurological disorders including autism spectrum disorder (ASD), developmental delay, or intellectual disability. Screening of individuals with ASD and macrocephaly revealed a subset with *PTEN* mutations[73]. A later study set out to establish the prevalence of *PTEN* mutation in macrocephalic individuals with ASD or mental retardation/developmental delay, and found it to be 8.3% and 12.2%, respectively[74].

Since the discovery that *PTEN* mutations lead to diverse clinical outcomes, researchers have sought to clarify whether certain types of mutations are more strongly associated with certain clinical outcomes. Studies attempting to link the type of mutation (i.e. missense vs. nonsense, or mutations in different exons) with specific cancer-related outcomes have largely been unsuccessful[75–77]. In fact, these and other studies have led to the prevailing belief that *PTEN*-related cancer predisposition syndromes are variable presentations of the same underlying pathobiology. Put another way, there is no meaningful difference amongst the *PTEN* mutations that lead to PHTS outcomes. This is supported by observations of variable PHTS presentations within a single family that share the same *PTEN* missense mutation[78].

However, there has been more progress in elucidating genotype-phenotype relationships with respect to ASD vs. PHTS outcomes. In 2011, a study emerged that leveraged a humanized yeast model to measure the catalytic activity of a series of *PTEN* missense variants[79]. The humanized yeast model will be expanded upon in section 1.2.5, and is used as the experimental model in Chapter 2. The results of the 2011 study suggested that *PTEN* missense mutations associated with ASD tended to retain partial catalytic activity, while those associated with the overgrowth

phenotype of classical PHTS were complete loss-of-function[79]. This study compared functionality of 14 ASD-associated variants and 19 PHTS-associated missense variants[80]. Further biochemical support for this hypothesis came from a study in 2015, in which PTEN missense variants were challenged to antagonize Akt signaling in U87MG breast cancer cell line which is *PTEN*-null[80]. The authors found that, as a group, the 7 ASD-associated PTEN mutations that they tested were destabilized compared to the wildtype. Intriguingly, when the variants were over-expressed to match the protein abundance of PTEN wildtype, the ability of the variants to antagonize Akt signaling was approximately equal to wildtype[80]. In further experiments, it was shown that two of three selected ASD-associated variants, when expressed at wildtype levels, were able to rescue cellular defects in *PTEN*-null, primary cultured mouse neurons. Namely, they rescued soma size, dendritic spine density, and dendritic spine length to comparable levels as wildtype *PTEN*[80]. In a biochemical study published the same year, Johnston and Raines measured the thermostability and enzymatic activity of three ASD-associated *PTEN* missense variants (H93R, E157G, and Y176C). All three mutations were thermodynamically compromised compared to wildtype, however E157G and Y176C retained substantial enzymatic activity (20-30% of wild-type)[81].

In addition to the biochemical and cell biological evidence above, there also exists human genetic data that support less damaging *PTEN* variants leading to ASD while more damaging lead to PHTS. Multiple studies have reported higher rates of missense variants in *PTEN*-ASD individuals than in PHTS individuals (as compared to nonsense, indel, or splice site). Frazier *et al.* report that within their cohort, 51.6% of

PTEN-ASD individuals had missense variants, while only 29.6% of PHTS individuals had missense variants [82]. Spinelli *et al.* aggregated mutations and phenotypes from the literature and came upon similar numbers: 52% of *PTEN*-ASD individuals had missense variants, while 32% of PHTS individuals had missense variants[80]. Since missense variants are generally less damaging than other classes of variation (e.g. nonsense, indel, or splice site), this suggests that hypomorphic alleles may predispose for ASD while highly damaging alleles might predispose for PHTS. However, this data also emphasizes that any genotype-phenotype relationships will not be clear cut, since nearly half of *PTEN*-ASD individuals have a nonsense, indel, or splice site mutation. Further, any mechanistic underpinning for the hypomorphic hypothesis is totally absent.

1.2.4 PTEN mutations and the brain

As a result of the recurrence of *PTEN* mutations in ASD, intellectual disability, and developmental delay, there has been a strong push to understand the effects of *PTEN* loss in the brain by use of model organisms. Homozygous germline deletion of *PTEN* in mouse is embryonic lethal[83], and heterozygous germline deletion leads to widespread tumor formation[84]. Therefore, to specifically model PTEN loss in the brain, most groups have used targeted inactivation, either to the brain generally or neurons specifically. An early study used a conditional genetic knockout system, in which Cre recombinase was driven by expression of glial fibrillary acidic protein (GFAP) in mouse to exclusively knock out *PTEN* in the central nervous system. Critical findings from this study include that *PTEN* conditional homozygous knock-out animals had progressively larger brains, and this brain overgrowth led to tonic clonic

seizures, eventually leading to premature death[85]. The larger brains resulted primarily from hypertrophic neurons, which grew increasingly large through time in a cell-autonomous and Akt dependent fashion[85]. The impacts of this cellular growth was clarified over the next decade. A neuron-specific enolase-Cre construct was used to delete *PTEN* from mature neurons in the cerebral cortex[86]. Overgrowth of neuronal soma was recapitulated, and further, the authors demonstrated that neurons lacking PTEN produced dendritic trees that were thicker, longer, and had more dendritic spines than wildtype neurons[86]. It was later discovered that these overgrown dendritic arbors lead to increased excitatory drive[87] that can lead to epileptic seizures[88]. A critical weakness of these studies (and most *PTEN* studies in model organisms) is that the investigators knocked out a copy of *PTEN*, while the human ASD/DD condition is predominantly caused by heterozygous missense mutations (this will be expounded upon in section 2.2.5). One study that did introduce ASD-associated *PTEN* missense mutations into mouse neurons found subtler effects than those induced by *PTEN* knockout[89]. This finding is consistent with the broader hypothesis, supported by several different types of data, that *PTEN* mutations that lead to ASD tend to be hypomorphic.

1.2.5 Humanized yeast assay for measuring PTEN catalytic activity

Because PTEN's substrate, PIP_3, is a lipid that resides within the cell membrane, it has been challenging to develop *in vitro* assays that directly monitor the enzymatic activity of PTEN variants. Additionally, because differences in PTEN activity cause complicated and diverse alterations in mammalian cell signaling, it is

difficult to isolate the direct effects of PTEN variation on enzymatic activity in mammalian cells. One solution to these issues was developed by Rodríguez-Escudero and colleagues in 2005. Their intuition was that the catalytic subunit of PI3K (p110α) could be expressed in yeast, and this enzyme would phosphorylate PIP_2 into PIP_3, which would eventually deplete the PIP_2 pool, leading to growth-inhibiting toxicity. However, PTEN expression could reverse this toxicity and rescue growth. Importantly, the ability of any given *PTEN* variant to rescue growth was related to the enzymatic activity of that variant.[90] Additionally, because PTEN and p110α are heterologous to yeast, and because PIP_3 is generally not present in yeast cells, it is expected that PTEN and p110α biochemical activities will not result in changes to the cellular signaling milieu. Accordingly, survival and growth of cells is directly related to the enzymatic capacity of the PTEN variant. This system has been used to profile enzyme activity of ASD and PHTS-associated *PTEN* variants[79,91,92], variation in p110α[92], and even variation in PI3K regulatory subunits[93]. This model has been used to generate functional data on ~100 *PTEN* variants, but the necessity for individually generating variant sequences and compartmentalizing experiments has precluded a comprehensive assessment of all *PTEN* variants.

1.2.6 Evidence for and implications of PTEN mutations with dominant-negative effects

While PTEN's lipid phosphatase activity at the cell membrane has been well established, it remains unclear exactly what form the catalytic unit takes. In particular, as alluded to earlier, there is evidence that PTEN forms homodimers, and these homodimers are the true catalytic unit. Evidence for dimerization in cells comes

from pulldown experiments, in which unlabeled PTEN can be pulled down by antibodies specific for a tagged PTEN[53,54]. It has also been observed that some *PTEN* variants display dominant negative effects, i.e. that, in the heterozygous state, a *PTEN* variant can actually decrease the activity of the other, wildtype allele. This phenomenon has been observed on the organismal level in mouse, with some missense variants leading to increased tumor burden compared to deletion of one allele.[53,55] It has also been observed in mammalian cells, with some missense variants leading to increased Akt signaling compared to single copy deletion.[53,94] There remain important unanswered questions regarding PTEN dominant negativity. The most parsimonious interpretation of these two observations (that PTEN homodimerizes and that some *PTEN* missense variants are dominant negative) would be that the dominant negativity results from mutant PTEN physically interacting and interfering with wildtype PTEN. However, this remains unproven and there are other possible explanations. Additionally, it is completely unknown how common *PTEN* dominant negative alleles are, and what role they play in determining human clinical outcome.

1.3 Targeted enrichment for sequencing

1.3.1 Targeted enrichment applications

The haploid human genome consists of approximately 3 billion nucleotides. However, in many cases, we are only concerned with the sequence at a small fraction of these sites. For example, there are vast stretches of the genome for which the function is unknown meaning that sequence variation detected at those sites is not

informative. In order to save time, money, and resources, technologies have been developed which allow researchers to enrich for sequences of interest.

In general, our understanding of genomic sequence is much greater for protein-coding as opposed to non-coding regions. Accordingly, one of the most widely used targeted enrichment strategies is exome sequencing, in which the ~1-2% of the genome consisting of protein-coding exons are enriched[95]. Exome sequencing has had tremendous success in establishing links between genes and Mendelian disorders[96] as well as diagnosing disorders in individual patients[97].

Another instance in which targeted enrichment is valuable is disorders for which the genetic etiology is heterogeneous but at least somewhat understood. Several targeted enrichment panels have been developed for cancers, which are commonly driven by mutations in known genes. It is commercially feasible to make custom enrichment panels for cancer, because it is a common disease. However, biotechnology companies have made the calculation that for many other rarer diseases, it is not commercially viable to create custom panels.

1.3.2 Targeted enrichment approaches and technologies

Historically, there have existed two general ways to enrich sequences of interest. The first and simplest method is by using PCR to selectively amplify targets. This approach works well for small numbers of small targets. However, as targets exceed several kilobases in length, it becomes challenging to find amenable PCR conditions. Likewise, as the number of amplicons increases, so too does the likelihood of off-target amplification or interference between primer sets.

The most widely used approach for targeted enrichment currently is hybridization-based[95]. In this type of approach, RNA or DNA probes are designed to have complementarity to regions of interest. Then, through attachment of the probe to a solid state, or by inclusion of a moiety such as biotin on the probe, target-bound probe can be physically separated from non-target sequence. Many contemporary exome or cancer panels use biotinylated probes in solution to capture regions of interest. The main weaknesses with this approach are high cost, GC sequence-content bias, and requirement for microgram scale input DNA. The prohibitive nature of the high cost deserves emphasis: for many uncommon diseases, it is not economically feasible for researchers or companies to create custom capture panels.

1.3.3 Whole-gene sequencing

Human genes are typically on the order of tens of thousands of basepairs long, which makes capturing full genes a major challenge for existing technologies. Identifying pathogenic variation in individuals with Mendelian disorders is a significant clinical challenge, with the diagnostic rate for many of these disorders only ~50%[98]. For example, *PTEN* pathogenic variants are found in only 25-80% of individuals with a Cowden Syndrome diagnosis[99] (Cowden Syndrome represents a subset of PHTS; the diagnosis of PHTS is given only if a pathogenic *PTEN* variant is identified). One explanation for the low diagnostic yield is that technologies focused only on the coding regions are missing pathogenic variation in the non-coding portions of genes (i.e. promoter, introns, or 5' or 3' untranslated regions). Whole gene sequencing can be especially important for genetic diagnoses of recessive disorders,

because a diagnosis requires finding likely gene disrupting variation in both copies of a single gene.

1.3.4 CRISPR-based targeted enrichment technologies

Clustered regularly interspersed short palindromic repeats (CRISPR)-Cas systems have emerged as biotechnological tools of immense value. Cas enzymes are endonucleases that acquire target specificity by complexing with a short guide RNA (gRNA) that is complementary to the targeted DNA sequence. Originally discovered as a form of bacterial adaptive immunity toward bacteriophages, the system was quickly adapted as a means of genome engineering or editing[100].

More recently, several groups have demonstrated that CRISPR can be harnessed as a tool for targeted enrichment for sequencing. A conceptually simple approach involves CRISPR-Cas9 cleavage of one or more regions of interest followed by size selection to isolate the target[101,102]. However, size selection is laborious, requires specialized equipment, and requires large amounts of input DNA. An alternative approach used CRISPR-Cas9 cleavage followed by direct ligation of nanopore sequencing adapters, followed by nanopore sequencing[103]. Ligation of adapters to blunt-ended Cas9 cleavage products (as opposed to overhanging cleavage products) may limit efficiency, and the nanopore sequencing platform has high error rates that make single nucleotide variants hard to call. Another approach involves a catalytically dead version of Cas9, so-called dCas9. Upon RNA-directed binding of the dCas9 to the target, the whole complex can then be pulled down with an oligonucleotide complementary to a tail on the guide RNA.[104]

In a similar vein, I here develop an alternative approach, described in Chapter 4, which leverages CRISPR-Cas12a (expanded upon in 1.3.5), a programmable system that leaves single stranded overhangs at cleavage sites (as opposed to blunt ends). Compared to existing methods, this approach is simple, does not require specialized equipment, flexible, and affordable. It should enable researchers with limited budgets, or those who study rare disorders, to design and implement targeted sequencing experiments.

1.3.5 Cas12a biochemistry

Cas12a, originally known as Cpf1, was discovered and characterized in 2015[105]. This enzyme, like the more well-known Cas9, binds to a gRNA, which then directs the enzyme to a double stranded DNA target by base-pairing between the gRNA and the DNA. Cas12a has several differences from Cas9, though, which make it uniquely suited for certain applications. For example, while Cas9 naturally uses two RNAs, equaling about 100 bases, Cas12a is guided by a single RNA that is only about 45 bases in length. Further, cleavage of target DNA occurs in a different location: Cas9 cleavage occurs proximal to the PAM site and non-homologous end joining often destroys the PAM site. If a researcher was hoping to introduce a genetic edit with homologous recombination, this would be a negative outcome. In contrast, the cleavage site of Cas12a is distal to the PAM site.[106] This suggests that even if a Cas12a cleavage event didn't result in homologous recombination, the enzyme could re-cleave the same target, increasing the odds of successful editing. Additionally, unlike Cas9, Cas12a cleavage results in symmetrical 5' overhangs[105] (Figure 3). This has

been recognized as a valuable characteristic; for example, methods have been developed that take advantage of this feature for molecular cloning[107].

Figure 3. Base-pairing between the gRNA and genomic DNA directs the Cas12a enzyme to the target.

Cleavage of the double stranded DNA target occurs at the 18th and 23rd position downstream of the PAM, resulting in cleavage products with symmetrical 5' overhangs.

Because overhanging ends are better ligation clients than blunt ends, we reasoned that specifically introducing cleavage events at targeted genomic loci would enrich ligatable ends at those loci. Then, a simple ligation reaction could append sequencing adapters to the cleaved ends, resulting in enriched sequencing of the regions of interest. The design and testing of this method are presented in Chapter 4.

Chapter 2. A saturation mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships

Taylor L. Mighell,[1, 2] Sara Evans-Dutson,[2] and Brian J. O'Roak[2]

[1] Neuroscience Graduate Program, Oregon Health & Science University, Portland, Oregon 97239, USA. [2] Department of Molecular & Medical Genetics, Oregon Health & Science University, Portland, Oregon 97239, USA.

Published in The American Journal of Human Genetics, May, 2018

doi: 10.1016/j.ajhg.2018.03.018.

2.1 Abstract

Phosphatase and tensin homolog (PTEN) is a tumor suppressor frequently mutated in diverse cancers. Germline *PTEN* mutations are also associated with a range of clinical outcomes, including PTEN hamartoma tumor syndrome (PHTS) and autism spectrum disorder (ASD). To empower new insights into PTEN function and clinically relevant genotype-phenotype relationships, we systematically evaluated the effect of *PTEN* mutations on lipid phosphatase activity *in vivo*. Using a massively parallel approach that leverages an artificial humanized yeast model, we derived high-confidence estimates of functional impact for 7,244 single amino acid PTEN variants (86% of possible). We identified 2,273 mutations with reduced cellular lipid phosphatase activity, which includes 1,789 missense mutations. These data recapitulated known functional findings but also uncovered new insights into PTEN protein structure, biochemistry, and mutation tolerance. Several residues in the catalytic pocket showed surprising mutational tolerance. We identified that the solvent exposure of wild-type residues is a critical determinant of mutational tolerance. Further, we created a comprehensive functional map by leveraging correlations between amino acid substitutions to impute functional scores all variants, including those not present in the assay. Variant functional scores can reliably discriminate likely pathogenic from benign alleles. Further, 32% of ClinVar unclassified missense variants are phosphatase deficient in our assay, supporting their reclassification. ASD associated mutations generally had less severe fitness scores relative to PHTS associated mutations ($p = 7.16x10^{-5}$) and a higher fraction of hypomorphic mutations, arguing for continued genotype-phenotype studies in larger

clinical datasets that can further leverage these rich functional data.

2.2 Introduction

Recent large-scale exome sequencing studies have highlighted the abundance of protein-coding variation in the human population.[4] It remains challenging to predict variant pathogenicity and clinical outcomes, especially for genes with pleiotropic effects. With most rare variants private to a single family or individual, using traditional approaches to establish pathogenicity such as variant segregation within a pedigree or identification in independent patients is infeasible. Even for well-studied genes, hundreds of variants are currently defined as variants of uncertain significance (VUS). Moreover, purely computational approaches still suffer from high false positive rates[108] and subjective interpretations that limit the clinical utility of these predictions.

To address these challenges for genes of clinical importance, one proposed approach is to prospectively measure the functional effects of all possible mutations, allowing these empirical data to be integrated into the clinical assessment of novel rare variants.[5,11] Historically, these types of functional assays have been conducted in a serial nature, which limits scalability, and often only within a portion of the protein of interest. While there are some notable examples of whole-gene brute force saturation mutagenesis, e.g., *TP53*[109] (MIM: 191170), new more scalable experimental paradigms are being developed that allow the functional dissection of the effects of thousands of genetic mutations in parallel.[13] These approaches leverage recent advances in DNA synthesis and sequencing technologies, and have proven particularly valuable in understanding the effects of mutations in cancer-associated genes.[22,110]

With these issues in mind, we have developed a saturation mutagenesis approach to comprehensively assess the effect of nonsynonymous mutations on the lipid phosphatase activity of phosphatase and tensin homolog (*PTEN* [MIM: 601728]). PTEN antagonizes the phosphoinositide 3-kinase (PI3K) signaling pathway through its lipid phosphatase activity toward the signaling lipid phosphatidylinositol (3,4,5)-trisphosphate (PIP_3).[111,112] In mice, loss of this activity increases tumor susceptibility in a dose dependent manner.[36] This observation led to a continuum model for PTEN's role in cancer development, with the level of phenotypic severity tightly coupled to the level of lipid phosphatase activity.[37]

Germline *PTEN* mutations are associated with a range of clinical outcomes, including autism spectrum disorder (ASD [MIM: 605309])[73,74,113] and tumor predisposition phenotypes collectively known as PTEN hamartoma tumor syndrome (PHTS).[70,114,115] Germline mutation carriers often share the common feature of increased head size or macrocephaly.[116] However, there is substantial variability in the neurological and tumor phenotypes present in these individuals. PHTS is an umbrella term that encompasses Cowden syndrome (MIM: 158350), Bannayan-Riley-Ruvalcaba-syndrome (MIM: 153480), and *PTEN*-related Proteus syndrome (MIM: 176920).[117] PHTS-affected individuals typically present with macrocephaly, hamartomatous polyps, and have an extremely high life-time risk of cancer.[117] *PTEN* mutations have been identified in macrocephaly cohorts of individuals with formal ASD diagnoses or developmental delay (DD)/intellectual disability (ID)[74,118,119] as well as idiopathic ASD.[113,120,121]

It is currently impossible to predict the phenotypic outcome of a given *PTEN*

mutation. Even predicting whether a *PTEN* mutation will have a pathogenic effect is still challenging. This is exemplified by the fact that a majority of missense variants (131/241, 54%) in ClinVar are considered VUS and seven additional variants have inconsistent pathogenicity reported across laboratories. Recent evidence from functional assays on a limited number of mutations and using diverse models, including humanized yeast,[79] cultured human cells,[80] and *in vivo* mouse neurons,[89] suggest that mutations identified in individuals with ASD or DD without obvious PHTS features tend to have hypomorphic lipid phosphatase activity, while PHTS-associated mutations more frequently show complete loss of lipid phosphatase activity. Further supporting this hypomorphic hypothesis, the distributions of mutation types are consistent with ASD associated mutations being generally less severe, with reported missense mutations three to four times as common in ASD compared with PHTS.[80,122] These findings, as well as the established genotype-phenotype relationships for *PTEN* in cancer, led us to hypothesize that, at the population level, ASD-associated PTEN variants are hypomorphic compared to PHTS-associated PTEN variants.

To systematically test this hypothesis and improve our ability to interpret the functional effects of any *PTEN* mutation, we modified a previously validated humanized yeast model for massively parallel functional testing of the effects of *PTEN* mutations on lipid phosphatase activity *in vivo*.[79,90] Given that yeast do not signal through PIP_3 dependent pathways,[123] this model system challenges PTEN protein variants to act on their preferred substrate in a cellular environment, but removes the confounding signaling and regulatory milieu present in mammalian cells.

Accordingly, the model is more sensitive than *in vitro* assays in which PTEN dephosphorylates a water-soluble substrate.[92] The utility of the yeast model for measuring lipid phosphatase activity has been demonstrated through validation of mutation effects on downstream Akt1 activation in mammalian cells, exhibiting complete concordance for the variants tested.[92]

With this system, we analyzed the functional effect of 86% of all possible single amino acid alterations. Overlaying these data onto PTEN secondary and tertiary structures recapitulated many known or predicted structure-function and biochemical relationships but also revealed surprising patterns of mutational tolerance. We discovered that several residues within the catalytic pocket are surprisingly tolerant to mutation and identified residues that are critical for membrane interaction. Moreover, we demonstrate these functional fitness scores have clinical utility by showing that they can outperform *in silico*-based approaches in characterizing likely pathogenic and benign variants. Finally, we provide compelling support for the existence of germline *PTEN* genotype-phenotype relationships that should be further explored in larger longitudinal clinical cohorts.

2.3 Materials and methods

***PTEN* saturation mutagenesis**

We obtained wild-type *PTEN* sequence from GenBank (NM_000314.6). All protein variants are reported relative to the corresponding 403 amino acid protein (GenPept: NP_000305). Our mutagenesis approach was similar to the Mutagenesis by Integrated TilEs (MITE) approach.[19] We designed a series of DNA "tiles" that were complementary to wild-type PTEN except for one codon (Figure S1A). At this single

codon, each molecule bore a substitution to the yeast-optimized codon for each non-wild-type amino acid, the yeast-preferred stop codon, or an in-frame, single codon deletion. Additionally, each set of "tiles" contained unique DNA adapters on either end to allow PCR retrieval of individual tiles from the pool (using primers with prefix: PTEN_sliceprimer, Table S1). These DNA tiles were synthesized as 130-mers (prefix: PTENTile) as part of a 12,000-feature oligo pool by CustomArray (Bothell, WA). For each tile, we designed inverse PCR primers that linearized the pYES2-PTEN wild-type sequence, excluding the portion encoded by the corresponding tile. Following amplification the tile PCR products were incorporated into the appropriate linear pYES2-PTEN by SLiCE mediated recombination.[124] SLiCE reactions were 10 µL and consisted of 100 ng of linearized vector with 15 ng of tile DNA, along with 1x SLiCE buffer and 1x SLiCE extract. SLiCE extract and buffer were prepared as described previously.[125] Reactions were incubated for 60 minutes at 37°C, then diluted 1:10 in water, and 2.5 µL used to electroporate 50 µL of NEB 10-beta electrocompetent *E. coli*. Transformation reactions were plated on LB agar plates with 100mg/mL carbenecillin (GoldBio) and grown overnight at 37°C. Colonies were collected and plasmids isolated with the QIAprep Spin Miniprep Kit (Qiagen).

Yeast selection experiments

Plasmid libraries were normalized and pooled into four mega-pools, each representing saturation mutagenesis for one quadrant (quadrants 1-3 = 100 codons, quadrant 4 = 103 codons). One µg of each mega-pool was transformed into the *S. cerevisiae* strain YPH-499, which already contained YCpLG-p110α-CAAX, using the Li-Ac/SS carrier DNA/PEG method.[126] More than 50,000 colony forming units were

generated per reaction. Colonies for each quadrant were pooled and grown overnight in SC-glucose –leu –ura (synthetic complete medium lacking leucine and uracil, using glucose as carbon source), pelleted and frozen down in 15% glycerol at -80°C.

Selection experiments began with overnight outgrowth of frozen stocks in SC-raffinose –leu –ura (raffinose neither induces nor represses GAL1/10 promoter). Following outgrowth, 25 or 30 million cells (replicate A or B) were pelleted for each quadrant as the "input" sample and frozen at -20°C. Then, 25 or 30 million cells were seeded into three cultures of 50 mL SC-galactose –leu –ura. Cultures were incubated at 30°C with 185 rpm shaking. After 24 and 36 hours of growth, cell concentrations were measured with a TC-20 Automated Cell Counter and 20 million cells (for each replicate) were passaged into fresh medium. At 48 hours, samples of 20 million cells were spun down with 13,000 x g for 30 seconds, medium withdrawn, and frozen at -20°C.

Library prep and sequencing

Plasmid DNA was isolated from pelleted cells (input and 48 hours) with Zymoprep Yeast Plasmid Miniprep II kit (Zymo Research). Stage-one PCR was performed in 25 µL reactions using: 5 ng of plasmid DNA, primers pYES2-PTEN_Q[1-4][F/R]_S1 (containing partial Illumina TruSeq adaptors) at 0.5 µM, 1x KAPA HiFi Hotstart Readymix (KHF), and 1x SYBR Green. Reactions were monitored by qPCR with cycling conditions: [95°C 3 minutes (98°C 20 seconds, 55°C 30 seconds, 72°C 15 seconds, plate read, 72°C 8 seconds) x 28-36 cycles]. Reactions were removed during or immediately following exponential phase of amplification. Stage-two PCR was then performed in 25 µL reactions using: 1 µL of uncleaned stage-one product, custom

Illumina dual index TruSeq primers (prefixes: S2, i7) at 0.5 μM, 1x KHF, and 1x SYBR Green (Table S1). Reactions were monitored by qPCR with cycling conditions: [95°C 3 minutes (98°C 20 seconds, 55°C 15 seconds, 72°C 15 seconds, plate read, 72°C 8 seconds) x 6 cycles]. Reaction products were checked on a 1.5% agarose gel, purified using NucleoSpin PCR Clean-up (Machery-Nagel), and concentrations were measured using a Nanodrop 1000 Spectrophotometer. Samples were normalized and combined into a common pool that was sequenced across multiple runs using paired-end 300 base-pair reads on the Illumina MiSeq platform (v3 reagent kit).

Sequencing data analysis

Paired-end reads were merged with PEAR[127] and common priming sequences were trimmed from the 5' and 3' ends using cutadapt.[128] For each quadrant, a purely wild-type sample was sequenced in order to identify sequencing error profiles. Counts of error reads were normalized to wild-type counts, then this normalized amount of reads were removed from all experimental samples.[110] Sequence variants were identified and counted with custom python scripts. These raw variant counts files were analyzed with Enrich2 v1.2.0[129] to calculate scores and standard errors for each variant. If the 95% confidence interval (based on the standard error) of the fitness score was <= 1, the variant was considered high-confidence. If the 95% confidence interval was > 1 but the measurements from each biological replicate were concordant (both lower or both higher than the 95% bound of the synonymous distribution), the variant was also considered high-confidence.

2.4 Results

Establishing a massively parallel functional assay for PTEN lipid phosphatase activity

We leveraged an artificial humanized yeast model in order to assess the relative phosphatase activity of PTEN variants.[79,90] In this system, the human PI3K catalytic subunit p110α (encoded by *PIK3CA*, [MIM: 171834]) is expressed in *Saccharomyces cerevisiae* and artificially directed to the membrane by a C-terminal prenylation box motif.[90] At the membrane, p110α is able to catalyze the conversion of the essential pool of phosphatidylinositol (4,5)-bisphosphate (PIP_2) to PIP_3, which potently inhibits growth through cytoskeletal disruption.[90] Upon induction of gene expression, cells proliferate at a rate that is proportional to the ability of the PTEN variant to

Table 1. Summary of *PTEN* mutagenesis and high-confidence effect classifications.

	Mutagenesis Summary[a]			**HC Classifications[b]**				
Mut. Type	**Designed**	**Created**	**HC**	**Total < Wt[c]**	**Trunc.-like[d]**	**Hypo.[e]**	**Wt-like[f]**	**> Wt[g]**
Missense	7,657	7,260 (0.95)	6,564 (0.86)	1,789 (0.27)	1,249 (0.19)	540 (0.08)	4,679 (0.71)	96 (0.015)
A.A. del	403	377 (0.94)	340 (0.84)	193 (0.57)	168 (0.49)	25 (0.07)	144 (0.42)	3 (0.007)
Trunc.	403	375 (0.93)	340 (0.84)	291 (0.86)	284 (0.84)	7 (0.02)	49 (0.14)[h]	0 (-)
Total	8,463	8,012 (0.95)	7,244 (0.86)	2,273 (0.31)	1,701 (0.23)	572 (0.08)	4,872 (0.67)	99 (0.014)

[a] Numbers in parentheses represent the fraction of designed variants.
[b] Numbers in parentheses represent the fraction of high-confidence variants.
[c] Total < Wt= less than wild-type; variants with scores less than or equal to -1.11, the lower 95th percentile (two-tailed) for synonymous variants.
[d] Trunc.-like= truncation-like; subset of less than wild-type variants with scores less than or equal to -2.13, the upper 95th percentile (two-tailed) of nonsense mutations at positions 1-349.
[e] Hypo=hypomorphic; subset of less than wild-type variants with scores between -2.13 and -1.11, the upper truncation and lower synonymous 95th percentiles (two-tailed).
[f] Wt-like=wild-type like; variants with scores between -1.11 and 0.89, the 95th percentile (two-tailed) of synonymous variants.
[g] > Wt=greater than wild-type; variants with scores exceeding 0.89, the upper 95th percentile (two-tailed) of synonymous variants.
[h] 48 of these truncating mutations fall within regulatory tail, positions 352-403.
Abbreviations: A.A. del= single amino acid deletion; HC= high-confidence; Hypo.=hypomorphic; Mut.=mutation; Trunc.= truncation; Wt=wild-type.

convert PIP_3 to PIP_2.[92] Co-expression of wild-type PTEN, but not catalytically dead mutants, e.g., p.Cys124Ser, catalyzes the reverse reaction, restoring the PIP_2 pool and allowing the yeast to grow and survive (Figure 1A). Moreover, growth rate provides a quantitative surrogate of lipid phosphatase activity with partial loss of function mutations showing intermediate growth phenotypes.[79]

We made several modifications to this system that allowed for massively parallel testing of preprogrammed mutations. First, to allow for parallel testing, rather than serial plating of single mutations, we modified the assay to support complex populations of PTEN-bearing yeast in liquid culture and sequencing as a readout of growth (Figures 1B-C and S1). We then introduced the yeast-preferred codon for each non-wild-type amino acid, stop codon, and single residue deletion at all *PTEN* codons en masse, utilizing a homologous recombination-based mutagenesis approach (Materials and Methods, Figure 1B and S2A, Table S1).[19,124] To allow direct sequencing of each mutagenized region, mutational space was separated into ~300 base-pair quadrants (Figure S2A).

We transformed two independent yeast populations with our mutagenesis library. Sequencing of naïve yeast libraries indicated that 95% of all intended mutations were present (Figures 1B and S2A). No position had less than 33% mutational coverage. Mutation dropout was largely confined to a single oligo pool in the C2 domain of the protein, which repeatedly performed poorly. We then performed selection experiments on these two independent yeast populations, each with three selection replicates (Figure 1B). We calculated natural log-scaled and wild-type

normalized fitness scores for each variant, along with standard error-based

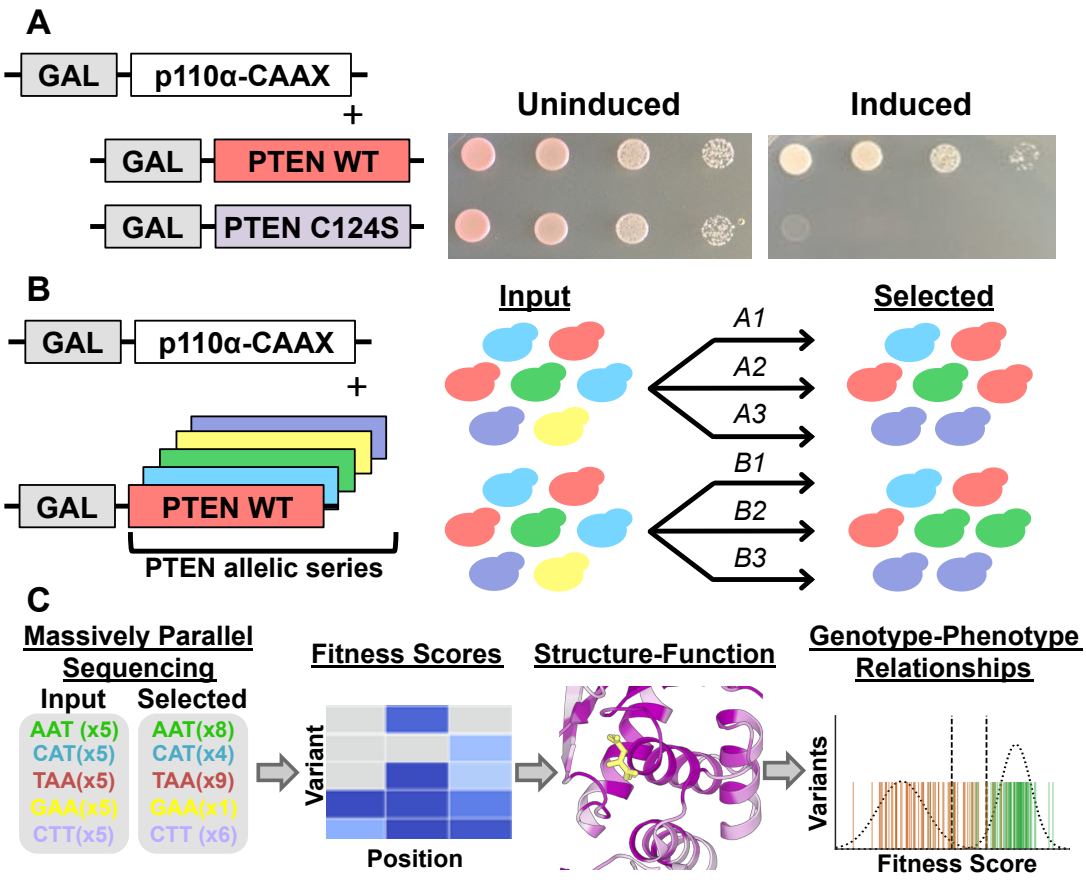


Figure 1. A framework for massively parallel functional testing of PTEN mutations.

(A) Humanized yeast model for evaluating the effect of PTEN mutations on lipid phosphatase activity. Exogenous expression of the catalytic subunit of human PI3K with a membrane-targeting prenylation box motif (p110α-CAAX) in yeast is toxic. However, co-expression of human PTEN wild-type, but not catalytically-dead PTEN p.Cys124Ser, can rescue growth**.** Both genes are under the control of a galactose inducible promoter (GAL).

(B,C) Modifications to allow massively parallel variant assessment.

(B) We generated a comprehensive PTEN allelic series, introduced these variants into yeast en masse, and subjected them to p110α-CAAX-mediated selection in liquid culture. We performed two biological replicates, each consisting of three technical replicates.

(C) We collected input and selected timepoints and subjected these to deep sequencing. We used read counts to calculate fitness scores and used these scores to highlight structure-function insights as well as genotype-phenotype relationships.

confidence intervals (Materials and Methods, Figure S2B).[129] Score estimates were generated for 8,012 (95% of intended) *PTEN* nonsynonymous mutations and

between mutational libraries fitness scores were highly correlated (Pearson's r = 0.76, Tables 1 and S2, Figures S3A and S3B). The distribution of fitness effects illustrates two major populations corresponding to likely damaging and wild-type like mutations (Figure S3A). Based on low standard error or replicate concordance, scores for 7,244 mutations (86% of intended) were classified as high-confidence (Materials and Methods, Tables 1 and S2, Figure S3C). Mutations were classified as wild-type like if their cumulative fitness score was within the 95th percentile (two-tailed) of observed synonymous mutations (Figure S3D). We identified 2,273 likely damaging mutations (31%) and 4,872 wild-type like mutations (67%) (Table 1). We also observed 99 mutations that performed better than wild-type (1%), which was within what was expected due to chance based on the total number of wild-type like variants. Among the likely damaging missense mutations, 1,249/1,789 (70%) fell within the observed distribution for programmed premature truncations (excluding C terminal tail), with the remainder having intermediate phenotypes in this assay.

High-resolution mutation data reveal structure-function insights

Using the high-confidence data, we first analyzed structure-function relationships, including known or predicted functional domains. Our complete sequence function map recapitulates many known features of PTEN biochemistry. For example, early truncating mutations are uniformly damaging through the phosphatase and C2 domain, but are tolerated in the regulatory tail (Figure 2A).[92] Overlaying the median fitness score of each position onto the partial crystal structure of PTEN (including residues 7-285 and 310-353) reveals strong intolerance of positions in the phosphatase domain, especially those positions near the catalytic

pocket (Figure 2B). The median fitness scores are also correlated with evolutionary conservation (Spearman, ρ = 0.58, Figure S3E). When compared to positions in alpha helices and beta strands, unstructured positions are very tolerant to mutation (Figure S3F).

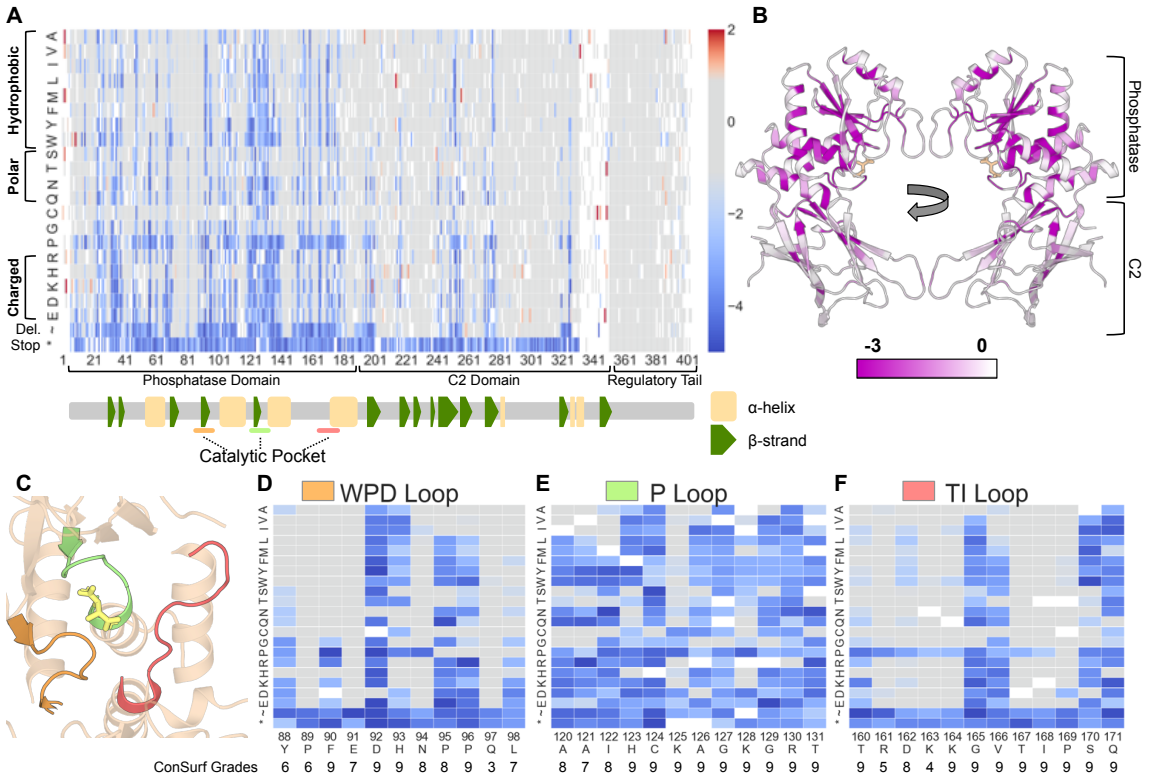


Figure 2. High-resolution map of the functional effects of PTEN mutations.

(A) Heatmap schematic showing high-confidence fitness scores for 7,244 PTEN missense, nonsense, or in-frame deletion mutations (86% of possible). Columns are each protein position and amino acids are listed in rows ordered according to biophysical characteristics. Variants with fitness scores within the 95th percentile (two-sided) of synonymous wild-type like mutations are colored gray. Variants with fitness scores lower than the synonymous distribution are colored blue while variants with higher fitness scores are colored red. The major protein domains, as well as the secondary structure features are indicated in the track below the heatmap (α-helices as yellow rectangles and β-strands as green pentagons).
(B) Ribbon diagram of PTEN crystal structure with residues colored by average fitness score. Darker purple corresponds to more damaging scores.
(C) Ribbon diagram highlighting the crystal structure of the PTEN catalytic pocket, composed of the WPD (orange), P (green), and TI-loops (salmon).

(D-F) The fitness scores of mutations at the residues composing the three catalytic pocket loops. Beneath each position is the Consurf grade (Materials and Methods), which represents the relative evolutionary conservation, with nine being the most conserved and one being the least conserved.

The catalytic pocket of PTEN is composed of the WPD, P, and TI loops (Figure 2C). This motif has sequence homology to dual specificity protein phosphatases, especially within the signature motif (123-HisCysXXGlyXXArg-130).[38] Arg130 is a hot-spot for somatic cancer associated mutations with multiple different missense and truncations frequently reported.[130] We observed this critical position was intolerant to all mutations (Figure 2E). Compared to other phosphatases, PTEN also has unique sequence features in order to accommodate the highly acidic and bulky PIP_3 substrate. Residues His93, Lys125, and Lys128 impart a basic character on the pocket,[38] the importance of which is demonstrated by the mutational intolerance at these positions (Figure 2D-E). Asp92 is a critical residue for PTEN catalysis, but its exact role remains uncertain.[79,131] We find that the only substitution with wild-type like activity is asparagine. Additionally, the PTEN catalytic pocket is larger compared to other dual specificity phosphatases.[38] The Cowden-associated p.Gly129Glu mutation has been shown to abolish lipid phosphatase while preserving protein phosphatase activity.[29] Our data show Gly129 is intolerant to all mutations except to alanine and serine, the two next smallest amino acids (Figure 2E). Unexpectedly, despite their presence in the catalytic pocket, several residues in the WPD and TI loops are highly tolerant to mutations (Figure 2D-F), highlighting the power of functional data to delineate truly functional from non-functional alterations within highly conserved protein domains.

PTEN associates with the plasma membrane through multiple domains. A PIP_2 binding motif in the phosphatase domain (residues 6-15) is rich in positively charged amino acids and allosterically promotes catalysis upon PIP_2 binding.[132,133] An additional positively charged residue, Arg47, contributes to this interaction.[134] Our data suggest that Arg15, Lys13, and Arg47 are the most critical of the positively charged residues in this motif (Figure S4A).[91] Additionally, an intramolecular regulatory interaction between the C-terminal tail and the phosphatase domain is controlled by phosphorylation at four sites in the tail, in mammalian cells.[41] We find that individual phosphomimetic substitutions at these sites are insufficient to decrease activity in our assay (Figure S4B).

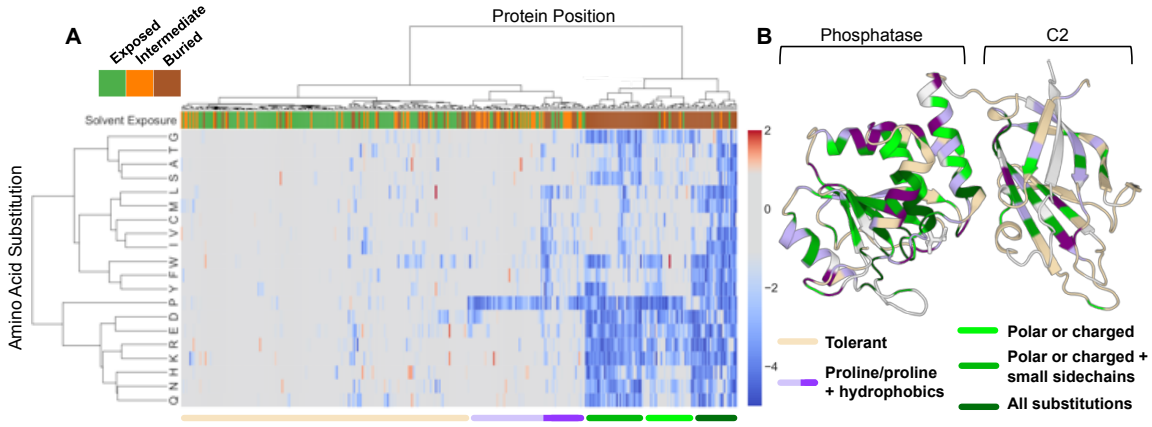


Figure 3. Hierarchical clustering reveals patterns of mutational tolerance among protein positions and amino acid substitutions.
(A) Hierarchical clustering of the 326 sites with all missense mutations measured. Clustering was performed by positions and amino acid substitutions (positions are columns and amino acid positions are rows). Overlaid on this heatmap is a top track showing the solvent exposure of each position in the crystal structure, with solvent exposed positions colored green, intermediate positions orange, and buried positions brown. We identified two major clades, which partitioned into five sub-clades with prevailing characteristics indicated and represented in the bottom track. We further divided the purple clade to reflect major differences in mutational tolerance.
(B) Ribbon diagram of PTEN crystal structure with residues colored according to clade assignment.

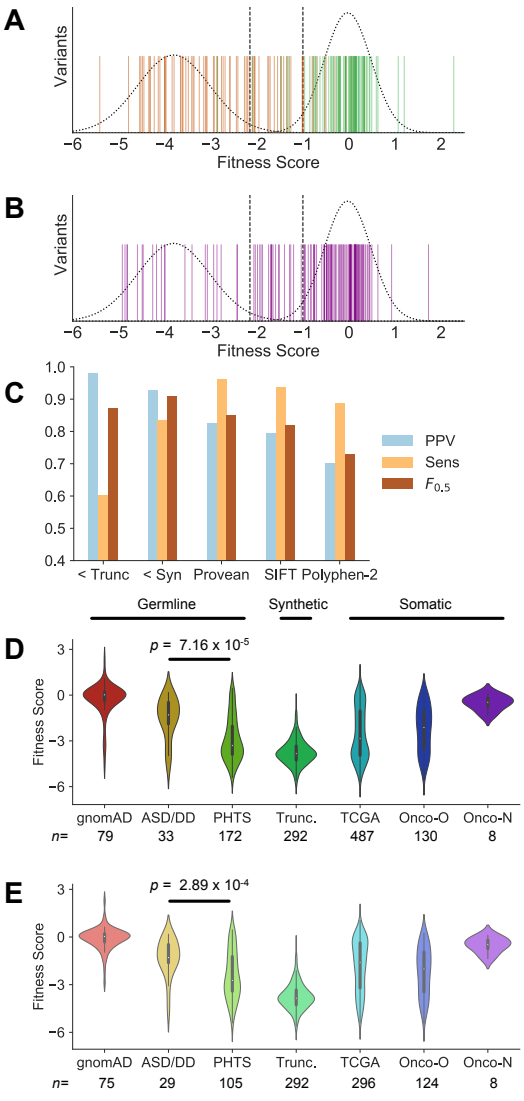
Protein positions cluster into stereotyped patterns of mutational sensitivity

In order to identify patterns of mutational sensitivity among PTEN positions and amino acid substitutions, we performed hierarchical clustering with all positions at which we measured effects of all missense substitutions (including high and low-confidence, n = 326, Figure 3A). We found that positions clustered into two major clades, corresponding to positions broadly tolerant/intolerant to proline or highly sensitive positions. We identified solvent exposure as a highly discriminatory feature between sensitive and tolerant clades, with 80/88 (91%) positions in the sensitive clade being in buried positions, while only 44/170 (26%) are buried in the tolerant clade (Figure 3A). The tolerant clade splits into two major groups with a sub-clade broadly tolerant to all substitutions (beige) and a second sub-clade where positions are either sensitive to proline alone or proline and hydrophobic residues (purple). The proline sensitive positions generally are part of secondary structures that are not buried in the hydrophobic core (Figure 3A). The sensitive clade positions split into three groups (green shaded sub-clades), which differ in their tolerance to charged, polar, or hydrophobic residues. The dark green clade represents the most constrained positions, and includes positions 92, 123, 124, and 130, all of which are in the catalytic pocket and critical for catalysis. Overlaying the sub-clade assignment of each position onto the crystal structure highlights the intolerance of mutations within the hydrophobic core of the phosphatase domain. Many of the solvent exposed positions in the C2 domain are tolerant to mutation (Figure 3B).

Clustering by amino acid substitutions recapitulated known functional relationships with proline correlated poorly with other substitutions (Figure 3A). We

sought to leverage these patterns of correlation to predict the fitness scores of mutations that were not present in our mutagenesis library or that were low-confidence.[24] We developed a heuristic for using only the most closely correlated observed substitutions[135] at the site of interest to compute an "informed position average" (Figure S5A). We combined this with several other prediction based,

Figure 4. Fitness scores discriminate between likely pathogenic and benign variants and support genotype-phenotype relationships.



(A) Fitness scores for missense variants considered pathogenic or likely pathogenic in ClinVar (orange) and putatively benign variants from gnomAD (green). Dashed curves represent truncation (left) and synonymous (right) distributions. Dashed lines at -2.15 and -1 represent the approximate 95% two-tailed distribution of truncations (before regulatory tail) and synonymous mutations, respectively.
(B) Fitness scores of ClinVar VUS (purple), with truncation and synonymous distributions and 95% limits.
(C) To test the ability of fitness scores to discriminate likely pathogenic from likely benign missense mutations, we calculated positive predictive value (PPV), sensitivity, and $F_{0.5}$ scores for our fitness scores ("< Trunc" represents the threshold at -2.15, "< Syn" represents the threshold at -1). In these tests, a true positive represented a ClinVar pathogenic allele having a fitness score less than or equal to the threshold. We compared the performance of the fitness scores at these two thresholds with *in silico* pathogenicity predictors for missense mutations using their default thresholds (Materials and Methods).
(D) Fitness scores of all curated mutations associated with the indicated phenotype (Material and Methods).
(E) Fitness scores of only curated missense mutations associated with the indicated phenotype (programmed truncations also shown for clarity). Abbreviations: ASD/DD, autism spectrum disorder/developmental delay; PHTS, PTEN hamartoma tumor syndrome; TCGA, The Cancer Genome Atlas; Onco-O, OncoKB mutations considered "oncogenic" or "likely oncogenic;" Onco-N, OncoKB mutations considered "likely neutral."

evolutionary, and biophysical features to train and test a random forest regression algorithm on our high-confidence measurements (Materials and Methods, Figures

S5B-C, Table S6)[24]. We used 10-fold cross validation to confirm that this approach can predict unseen data with high confidence (Pearson's r = 0.80, Figure S5E). We further performed a downsampling analysis to assess the expected accuracy of imputing scores at different levels of saturation, finding that reductions of 10-20% (65.8-74% of saturation) achieve similar performance (Figure S5F). Finally, we generated imputations for all variants that were absent from our library or measured with low-confidence (Figure S6 and Table S2).

Fitness scores discriminate between likely pathogenic and benign alleles

To determine if our empirically determined fitness scores were informative for discriminating between germline likely pathogenic and benign alleles, we collected germline missense mutations reported as pathogenic or likely pathogenic from ClinVar[136] and rare variants from gnomAD,[4] excluding p.Arg173His and p.Lys289Glu that are reported pathogenic in ClinVar (Materials and Methods, Tables S3 and S4). Fitness scores alone discriminated pathogenic from benign germline alleles (Figure 4A). We found that the $F_{0.5}$ score, which weights predictive value (PPV) over sensitivity, reaches its maximum at a cutoff based on the synonymous distribution (<= -1, ~95th percentile, PPV = 0.93, sensitivity 0.83), and outperforms several *in silico* mutation effect prediction algorithms (Figure 4C and S7). PPV was maximized (0.98) at a more conservative cutoff based on the 95th percentile of the truncation distribution, but with reduced sensitivity (0.60) (Figures 4A and 4C). Given the high PPV of our scores, we evaluated distribution of fitness scores among ClinVar missense VUS (Figure 4B). We found that 21/127 (17%) VUS with high-confidence data met the strict truncation-based cutoff and 41/127 (32%) met the

synonymous cutoff, suggesting that fitness scores could be used to reclassify a major fraction of VUS.

PTEN mutations are extremely frequent in somatic cancer. We extracted nonsynonymous mutations from The Cancer Genome Atlas (TCGA) and observed a multimodal and wide distribution of fitness scores (Figure 4D-E, Table S5). This is likely due to the presence of both driver and passenger mutations in these data. Similar to the germline analysis, to test if fitness scores could discriminate somatic mutations that are likely pathogenic, we evaluated mutations from Onco-KB, a precision oncology database with expert annotation of somatic mutations (Table S5).[137] We found that fitness scores of *PTEN* mutations considered "oncogenic" or "likely oncogenic" were substantially more negative than those considered "likely neutral." Of the missense likely oncogenic, 86/124 (69%) and 56/124 (45%) were below the synonymous and truncation thresholds, respectively. In contrast, of the 8 variants considered likely neutral (all missense), only one (p.Ala121Val) had a fitness score marginally below the synonymous cutoff (fitness score, -1.3). Taken together, these findings emphasize the ability of empirically determined fitness scores to discriminate between pathogenic and benign human alleles, in both the germline and somatic setting.

Finally, we evaluated potential genotype-phenotype relationships for germline *PTEN* mutations. We first compared the fitness scores of *PTEN* mutations associated with various clinical presentations acquired from multiple sources (Materials and Methods, Figure 4C, Table S5). We found that, as a population, fitness scores of nonsynonymous mutations exclusively reported in ASD/DD cohorts were

less severe than PHTS-associated mutations (Mann-Whitney U-test, two-sided, $p = 7.16 \times 10^{-5}$). Comparing only the missense we found that this significant difference persists (Mann-Whitney U-test, two-sided, $p = 2.89 \times 10^{-4}$), indicating that the mutation type alone does not drive these differences. We found 12/29 (41%) and 21/105 (20%) of the ASD and PHTS missense mutation fell within the hypomorphic activity range, respectively. Overall, these data provide strong support for the hypothesis that ASD/DD associated mutations often retain hypomorphic PTEN phosphatase activity.

2.5 Discussion

Massively multiplexed functional assays represent a promising approach to understanding the effect of mutations on protein function, which can provide immediate insights into structure-function relationships and clinical interpretation. Modifying a humanized yeast assay that uses growth to read out relative phosphatase activity, we were able to assess the functional effects of human *PTEN* mutations on a massive scale. Our approach yielded high-confidence measurements of 86% of the possible single residue nonsynonymous mutations. A limited number of human proteins have been subjected to full length massively multiplexed functional assessment and very few have been assayed at the depth we achieved.[22–24,110,138–141] Similar approaches could be used with this model to the study of various aspects of the PI3K/Akt pathway at scale, including mutations in *PIK3CA/B*[92] (p110α/β (*PIK3CB,* MIM: 602925)), *PIK3R1*[93] (p85α, MIM: 171833), and *AKT1*[142] (MIM: 164730), as well as drug screening for PIK3CA inhibitors.[143]

Several features of the data support the validity of these function estimates and their relevance to human health. We observed high correlation between biological replicates and recapitulated known features of PTEN function. For example, there were no pathogenic mutations within our curated clinical dataset in the C-terminal tail. The set of early terminating mutations confirm that the minimal catalytic unit includes the phosphatase and C2 domains, but not the C-terminal tail.[92] Likewise, we found that position Cys124, which takes part directly in phosphatase catalysis, and position Arg130, which is a hotspot for cancer mutations, are completely mutation intolerant. Additionally, we found that mutations are not well tolerated within the loops forming catalytic pocket or residues mediating interactions with PIP_2. Finally, we found that proline was the most damaging substitution, consistent with a recent meta-analysis of massively multiplexed experiments[135] and decades of biochemistry.[144]

While the humanized yeast system faithfully reports on intrinsic lipid phosphatase activity, mutations that functionally disrupt protein-protein interactions, subcellular localization, post-translational modifications, or function through a dominant negative mechanism[53] in mammalian cells will not be captured. We observed 99 variants with greater than wild-type like activity, none of which were present in curated pathogenic datasets. While it is possible some of these variants increased PTEN activity, the number of variants of this class does not exceed what we would expect under the null assumption of wild-type like activity. PTEN has relatively low thermostability[81] and protein destabilization is a known mechanism for PTEN loss-of-function.[80,145] A concurrent functional screen assaying protein stability found

~1/4th of mutations alter steady state stability.[141] Six mutations that destabilized PTEN in breast cancer cell lines also decreased steady state abundance in this yeast model,[92] suggesting that mutations affecting thermostability will be detected in our screen. However, our sensitivity to detect destabilizing mutations is unknown, as is whether mutations specifically altering the rate of proteasome-mediated degradation[146] will be reported on. We believe that independently assaying these important factors at similar scale would provide useful complementary insights into PTEN function.

We discovered that approximately half of all positions in PTEN were broadly tolerant to substitutions, suggesting that they are not required for lipid phosphatase activity. While there is a degree of correlation between the median fitness score and the evolutionary conservation of each position, we identified positions within the highly conserved catalytic pocket and elsewhere in the protein that are highly tolerant to specific mutations. This is in apparent contradiction with PTEN's high evolutionary conservation (99.75% identity between human and mouse[122]) and constraint in humans.[4] This suggests that many PTEN positions are potentially under selection due to phosphatase independent functions.

Our high-resolution mutation data empowered unique insights into PTEN biochemistry and structure. The substitution p.Gly129Glu is a well-known Cowden-associated mutation that disrupts lipid phosphatase activity while maintaining protein phosphatase activity.[29] We found that substitutions to alanine and serine are tolerated at this position, while mutations to bulkier residues are damaging. This suggests that there is a size limit for the amino acid that occupies this position. Asp92

matches the position of aspartic acid in the WPD loop of PTP1B, which acts as a general acid in the catalytic mechanism.[131] Asp92 is a critical residue in the PTEN catalytic pocket, but its role in the reaction mechanism remains uncertain.[79,131,147] Our data support previous findings that all mutations except p.Asp92Asn are strongly damaging.[79] However, the p.Asp92Asn mutation has been reported in an individual with ASD indicating that it still may have a clinical effect.[148] Similar to our findings, Rodríguez-Escudero and colleagues found in the yeast assay p.Asp92Asn had growth rescue similar to wild-type, but partial activity relative to wild-type using an indirect fluorescence indicator of PIP_3 levels or an *in vitro* phosphatase assay.[79] Combined these data are consistent with the p.Asp92Asn mutation retaining partial activity. We propose that p.Asp92Asn could be showing wild-type like activity in our assay through asparagine deamidation, which is a spontaneous, intramolecular reaction that can result in the conversion of asparagine to aspartic acid.[149] In biochemical systems and mammalian cells, this spontaneous conversion may not be sufficient to fully rescue PTEN activity.

Similar to previous studies,[16,109] we used hierarchical clustering to look for patterns amongst the positions and amino acid substitutions. We found that PTEN positions fall into a few stereotyped patterns of mutational tolerance and that a critical determinant of mutational tolerance is the relative solvent exposure of the position. These findings are consistent with a recent meta-analysis of similar experiments.[17] We leveraged the correlation amongst amino acid substitutions, along with several other features, to generate a random forest regression model that could accurately predict the fitness scores of unseen mutations and create a comprehensive

functional map encompassing the effects of all possible single nonsynonymous mutations. To guide future studies of similar proteins, we performed a downsampling analysis of the training data and found that for similar accuracy, ~70% mutation saturation would likely be sufficient. Moreover, proline substitutions predict poorly and should be directly assayed.

A critical hurdle for the application of massively multiplexed functional assays is bridging the gap between molecular phenotype and human phenotype.[150] We found that fitness scores are able to discriminate between likely pathogenic and benign human alleles in both the germline and somatic condition. On this basis, we expect that these scores will be of tremendous clinical value for reclassifying VUS[11] and also predicting the effects of private alleles that remain to be identified. A major question related to *PTEN* genetics is whether genotype-phenotype relationships can explain the heterogeneity in clinical presentation for carriers of germline mutations. Our comprehensive dataset provides strong evidence that the mutations associated with ASD/DD are hypomorphic for lipid phosphatase activity and are significantly more active than the mutations that lead to PHTS. This suggests that distinct biological mechanisms underlie the differential presentations, and understanding these differences will be critical to the eventual treatment of these disorders. While it is possible that these different mechanisms are the direct result of lipid phosphatase activity at the plasma membrane, ASD-associated mutations may specifically disrupt another of PTEN's cellular functions.[40,66] Supporting this idea, some ASD-associated mutations are excluded from the nucleus and lead to neuronal hypertrophy, but this phenotype can be rescued by artificial direction to the nucleus.[63]

While massively parallel functional data is a significant advance for understanding function-specific mutation effects, further untangling complex genotype-phenotype relationships will require similar advances in clinical genetics databases with standardized descriptors of clinical presentations and symptoms.[122] Our study was limited by both the number of publicly available mutations and associated clinical information. Since there are no coding variants considered benign in ClinVar, we used PTEN variants in the gnomAD database as a proxy for likely benign mutations. While these mutations are on average wild-type like, we recognize that this is an imperfect approach and it is possible that some of the variants in gnomAD are pathogenic. We excluded variants that were only in ClinVar from our genotype-phenotype analysis because of their ambiguous annotation and lack of clinical data. For example, 17% of the pathogenic/likely pathogenic mutation submissions had no indicating condition provided and 36% of all missense entries use the ambiguous term "hereditary cancer-predisposing syndrome." Requiring submitters to provide more information in a consistent way will maximize the utility of massively multiplexed functional data. Finally, it is still unclear if individuals ascertained for neurological phenotypes as children will have a higher risk to develop PHTS like or cancer presentations later in life.[151] Moving forward, large-scale sequencing efforts that permit longitudinal assessment as well as patient re-contact will be instrumental. A new initiative, SPARK, aims to partner with 50,000 individuals with ASD and their families to create the largest genetically characterized ASD cohort to date.[152] It is likely that hundreds of new *PTEN* mutation carriers will be identified in SPARK and would be available for re-contact and detailed prospective study.

We demonstrate that comprehensively assaying the molecular phenotypes of thousands of mutations to a human protein can yield clinically relevant insights, even for proteins with pleiotropic effects. Future efforts that combine multiple functional modalities and rich clinical datasets may allow for the precision needed to fully realize personalized genomic medicine.

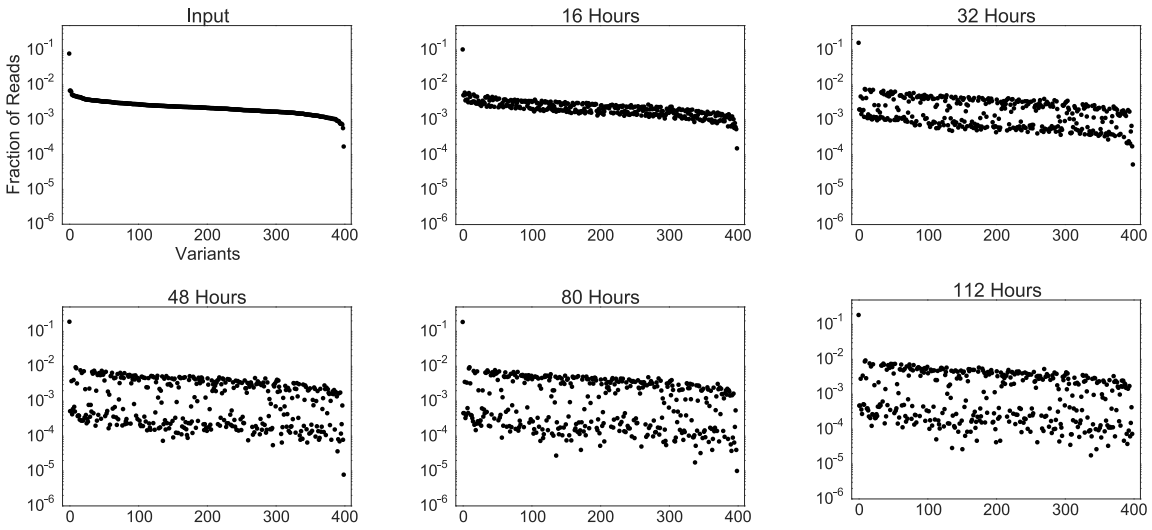
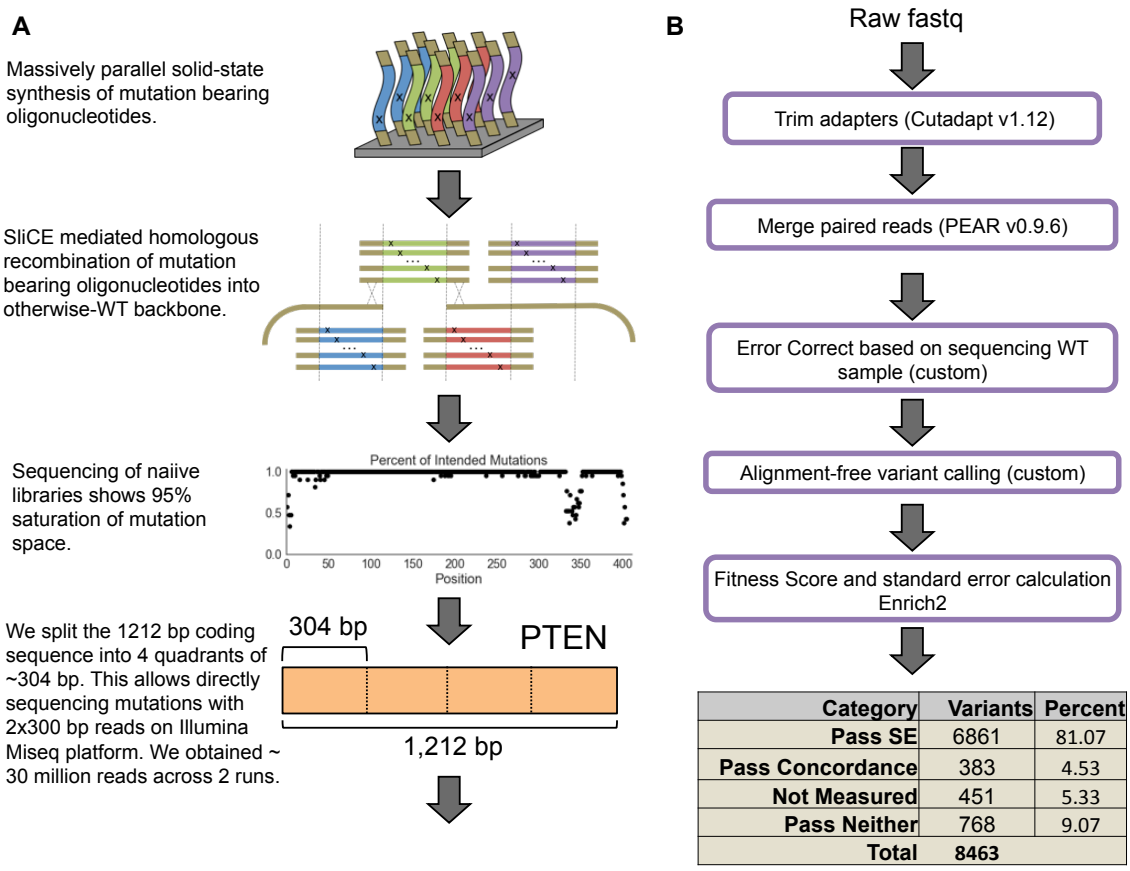


Figure S1. Optimization of humanized yeast assay for liquid culture induction and selection.

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



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

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

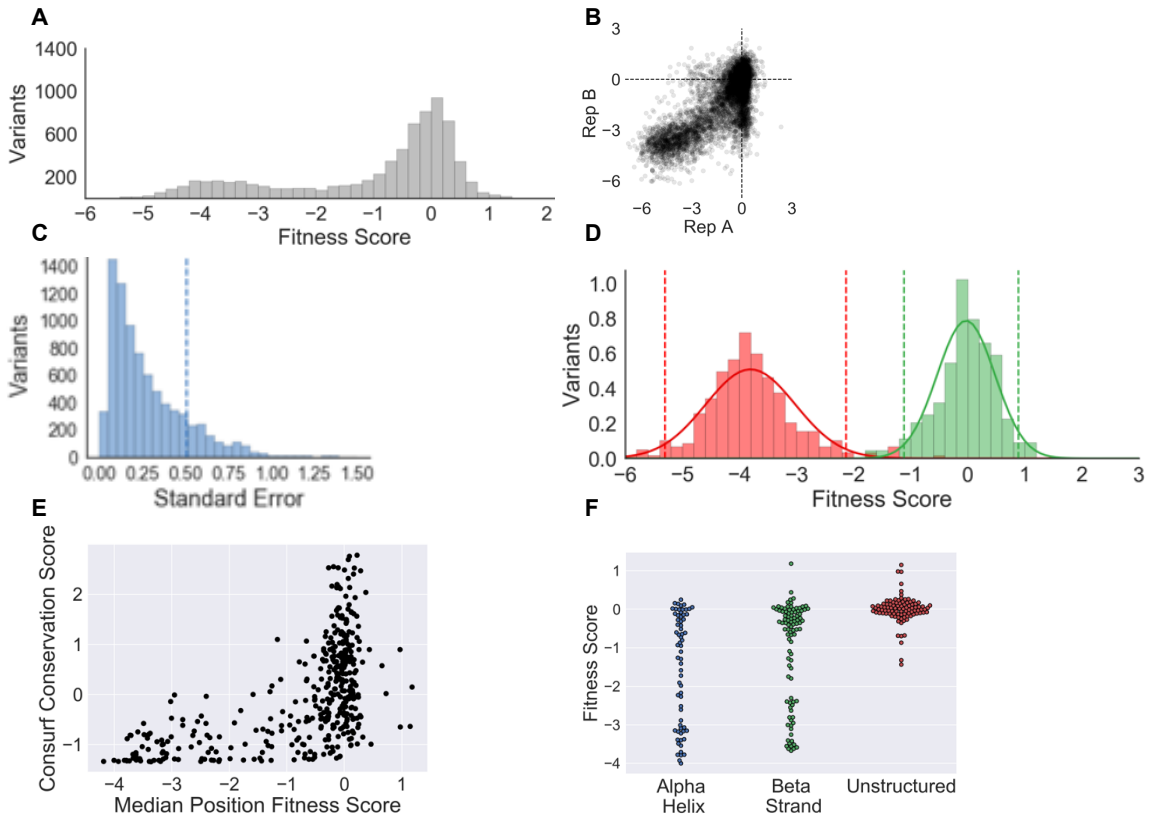


Figure S3. Overview of PTEN saturation mutagenesis dataset and relative fitness scores.

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

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

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

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

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

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

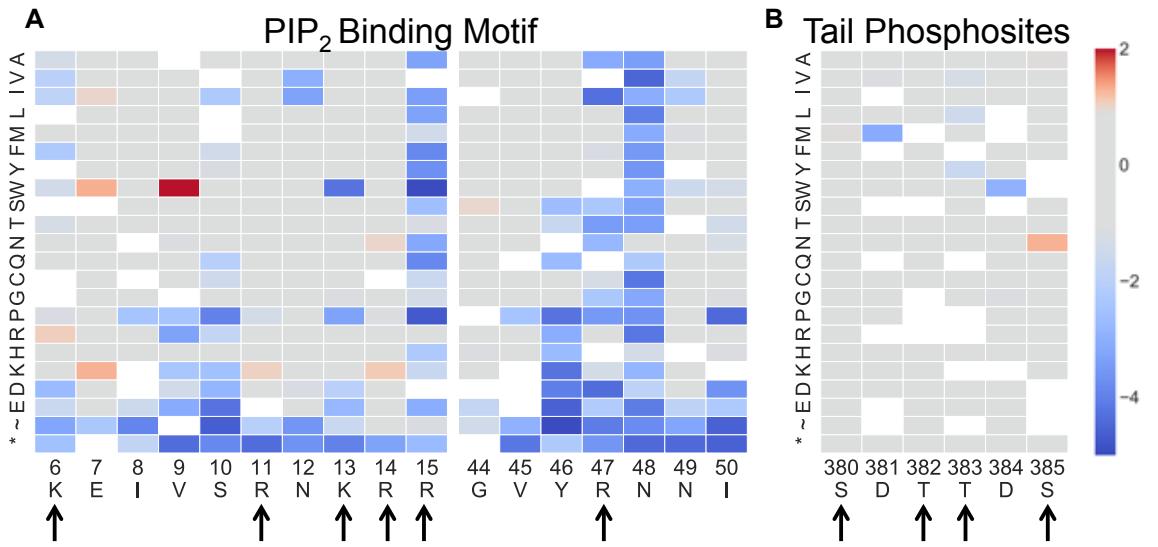


Figure S4. Evaluation of mutation effects within the PTEN predicted PIP_2 binding motif and tail phosphosites.

(A) Fitness scores highlighting positively charged residues in PIP_2 binding domain (Lys6, Arg11, Lys13, Arg14, Arg15) as well as Arg47, with neighboring residues. Lys13, Arg15, and Arg47 are the most critical in our assay.

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

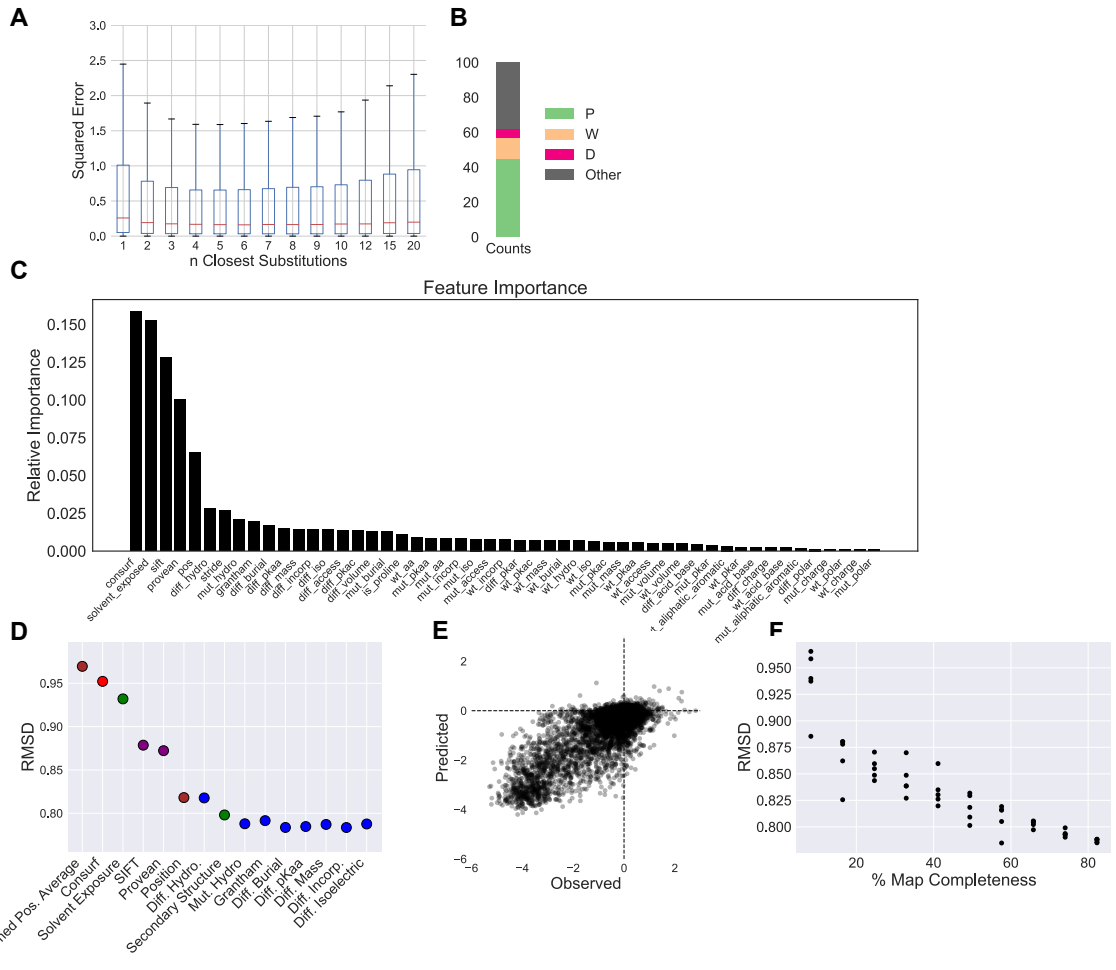


Figure S5. Development of a random forest algorithm to impute relative fitness scores for missing data.

(A) We used correlation coefficients[135] between amino acid substitutions to identify, in aggregate, the number of most closely correlated substitutions that maximized accuracy in the prediction of missing data. To generate each prediction we identified the n most closely correlated substitutions that were measured with high confidence at that positions, and calculated the average weighted by the standard error of each substitution. Box plots represent the squared error between measured value (in our assay) and value predicted from the n closest substitutions for all high-confidence measurements. We chose to use five for subsequent modeling, and define this value as "informed position average".

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

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

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

(E) We used the 15 features in D to train a final model (options same as in C) and performed 10-fold cross validation on the high-confidence variant set. We generated predictions for all high-confidence variants and plotted the observed and predicted values for each variant. Pearson's r = 0.80, options same as above.
(F) RMSD results from downsampling to indicated map completeness. We downsampled from our high-confidence dataset and retrained models at each indicated percent map completeness. The maximum value is 82.3%, which is the percent map completeness that our high-confidence missense dataset represents. 5 replicates were performed at each point % map completeness. Options same as above, except random_state=None.

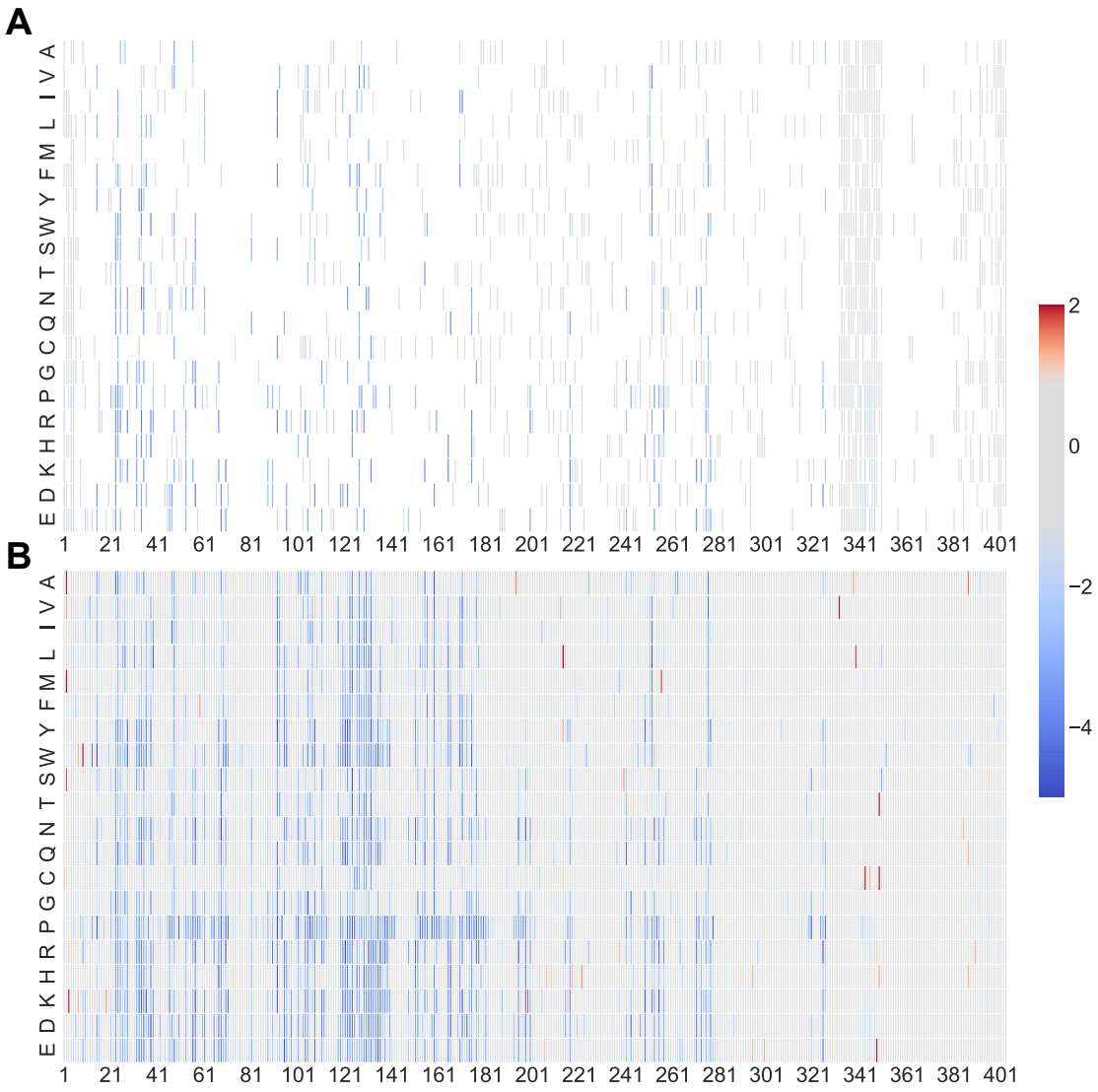


Figure S6. A comprehensive functional map of predicted effects of PTEN mutations using imputed scores.

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

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

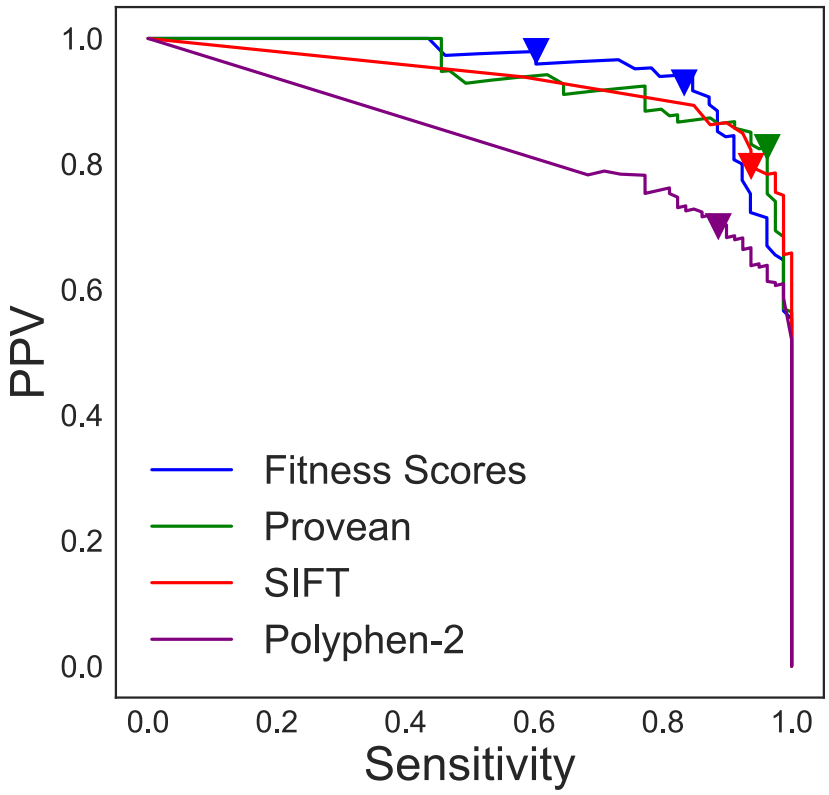


Figure S7. Positive predictive value (PPV) and sensitivity (precision and recall) curves for fitness scores and mutation effect predictors.

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

Chapter 3. An integrated deep mutational scanning approach provides clinical insights on PTEN genotype-phenotype relationships

Taylor L. Mighell[1,8], Stetson Thacker[2,3,8], Eric Fombonne[4,5,6], Charis Eng[2,3,7,*], Brian J. O'Roak[1,**]

[1] Department of Molecular & Medical Genetics, Oregon Health & Science University, Portland, OR 97239; [2] Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA; [3] Cleveland Clinic Lerner College of Medicine, Cleveland, OH 44195, USA; [4] Department of Psychiatry, Oregon Health & Science University, Portland, OR 97239; [5] Department of Pediatrics, Oregon Health & Science University, Portland, OR 97239; [6] Department of Behavioral Neuroscience, Oregon Health & Science University, Portland, OR 97239; [7] Department of Genetics and Genome Sciences, and Germline High Risk Cancer Focus Group, Comprehensive Cancer Center, Case Western Reserve University School of Medicine; Cleveland, OH 44106, USA. [8]These authors contributed equally to this work.

Published in The American Journal of Human Genetics, June, 2020

doi: 10.1016/j.ajhg.2020.04.014

3.1 Prologue: Massively parallel assay for PTEN cellular abundance.

In May of 2018, Matreyek and colleagues published a work describing the effects of approximately half of all *PTEN* single amino acid variants on cellular abundance[94]. This experiment was performed in human cells (HEK293T), and used FACS to sort cells into different bins corresponding to different abundance. Namely, a PTEN-EGFP fusion protein was expressed from the same construct as mCherry. Cells were sorted based on the ratio of EGFP:mCherry. The strength of EGFP signal will be related to the steady state stability of the fused PTEN molecule, and since the mCherry and EGFP are expressed from the same construct, the mCherry acts to normalize the EGFP signal. One strength of this assay is that it is done in human cells, meaning that it should capture reduced abundance due to reductions in thermal stability as well as reduced abundance due to active protein degradation mechanisms, such as the proteasome. A weakness of the study is that PTEN is fused to EGFP, which may affect abundance of certain variants. Further, while the assay is done in human cells, it may not capture activity dependent or cell type dependent abundance effects. Further, the assay only captures about half of all possible variation.

3.2 Abstract

Germline variation in PTEN results in variable clinical presentations, including benign and malignant neoplasia and neurodevelopmental disorders. Despite decades of research, it remains unclear how PTEN genotype is related to clinical outcomes. In this study, we combined two recent deep mutational scanning (DMS) datasets

probing the effects of single-amino acid variation on enzyme activity and steady-state cellular abundance with the largest well-curated clinical cohort of PTEN-variant carriers. We sought to connect variant-specific molecular phenotypes to the clinical outcomes of individuals with PTEN variants. We found that DMS data partially explain quantitative clinical traits, including head circumference and Cleveland Clinic (CC) score, which is a semi-quantitative surrogate of disease burden. We built logistic regression models using DMS and CADD scores to separate clinical PTEN variation from gnomAD control-only variation with high accuracy (AUC = 0.908). Using a survival-like analysis, we identified molecular phenotype groups with differential risk of early onset as well as lifetime risk of cancer. Finally, we identified classes of DMS-defined variants with significantly different risk levels for classical hamartoma-related features (odds ratios range of 4.1-102.9). In stark contrast, the risk for developing autism or developmental delay does not significantly change across variant classes (odds ratios range of 5.4-12.4). Together, these findings highlight the potential impact of combining DMS datasets with rich clinical data, and provide new insights that may guide personalized clinical decisions for PTEN-variant carriers.

3.3 Introduction

Germline mutation of the tumor suppressor gene phosphatase and tensin homolog (*PTEN* [MIM: 601728]) manifests with variable and complex phenotypes, including macrocephaly (with increased occipital-frontal circumferences [OFC]), benign hamartomas affecting all three germ layers, malignant neoplasia across multiple tissues, and neurodevelopmental abnormalities, including autism spectrum disorder (ASD).[153,154] This heterogeneity is reflected clinically with germline *PTEN*

mutations found in variable subsets of defined syndromes, including Cowden syndrome and Bannayan-Riley-Ruvalcaba syndrome (CWS1 and BRRS [MIM: 158350]), as well as macrocephalic ASD (MAS [MIM: 605309]), among others.[70,73,114,154] Collectively, these syndromes have been termed PTEN hamartoma tumor syndrome (PHTS) when a germline *PTEN* variant is identified.[153,154]

The dramatic variability of these clinical presentations has sparked efforts to correlate *PTEN* variants with clinically-relevant phenotypic classes. However, *PTEN* variants resist simple classification approaches based on secondary domain clustering or variant type. Recently, the mapping of a limited subset of germline *PTEN* variants onto the three-dimensional, crystal structure failed to reveal a distinct pattern of distribution between ASD- or cancer-predisposition-associated variants[155]. Classification efforts have also been impacted by limited sample sizes in terms of both functional data and *PTEN* variant cohorts.[80,156] Additionally, the PTEN protein has multiple functional roles in the cell apart from lipid phosphatase activity, which may also play a role in this phenotypic complexity.[154,157,158]

These challenges have prompted recent creative high-throughput methods to functionally measure the molecular phenotypes for thousands of nonsynonymous *PTEN* variants, collectively termed deep mutational scanning (DMS).[94,156] We previously reported the effect of nearly all *PTEN* nonsynonymous variants on lipid phosphatase activity by utilizing a humanized yeast assay where lipid phosphatase activity was linked to cell survival (so called fitness score).[156] These data demonstrated that the solvent exposure of a wild-type residue is a critical determinant of mutational tolerance for lipid phosphatase fitness, with solvent

exposed residues being much more tolerant to mutation. As expected, PTEN lipid phosphatase activity was generally intolerant to mutation in the catalytic pocket and phosphatase domain, though not without exception. Further, in line with suggestions from prior more limited functional studies[80], *PTEN* missense variants associated with ASD tended to retain partial lipid phosphatase activity. [156]

In a second independent study, the effect of ~54% of all *PTEN* nonsynonymous variants on the steady-state cellular protein abundance were estimated using fluorescently tagged PTEN variant proteins (so called abundance score). It was observed that PTEN abundance is, in part, explained by thermodynamic stability and cell-membrane interactions of a given variant. While variant abundance inversely correlates with pathogenicity, notable exceptions are putative dominant-negative PTEN variants, which are highly stable but catalytically inactive.[94]

While these two DMS studies have added essential insights into the effect of *PTEN* variants on protein function, they were limited in their clinical analyses as both relied on previously published clinical reports and ClinVar database[159] variants with varying degrees of validation and phenotypic description. In this study, to further uncover *PTEN* genotype-phenotype relationships and clarify patient risk for these diverse clinical presentations, we integrated these datasets with the largest, prospectively accrued and comprehensively clinically characterized cohort of *PTEN* variant-positive individuals (Cleveland Clinic [CC] cohort). These analyses demonstrate that molecular phenotypes associate with quantitative clinical traits. They also delineate differential lifetime cancer risk and indicate unexpected risk ratio relationships for neurodevelopmental and hamartoma-associated phenotypes.

3.4 Materials and methods

PTEN Variant Function Data and Imputation

We made use of two DMS datasets in this study.[94,156] Briefly, fitness scores were previously determined by assessing a *PTEN* variant's ability to reverse toxicity by means of phosphatidylinositol (3,4,5)-triphosphate (PIP_3) dephosphorylation in a humanized yeast system[90] that expresses a hyperactive kinase.[156] High confidence fitness scores were previously generated for 86% of all variants and a random forest algorithm was used to impute fitness scores for the remaining unmeasured variants.[156] Abundance scores were previously determined by measuring the steady-state level of *PTEN* variants using the VAMP-Seq assay in human cells.[94] Abundance scores were generated for 54% of all variants.

Using a random forest framework similar to what was used to impute fitness scores, here, we imputed abundance scores for the remaining unmeasured variants (Figure S1). Modeling was implemented in Scikit-learn version 0.19.0 (sklearn.ensemble.RandomForestRegressor, n_estimators=500, criterion= "mse", max_features=0.33, random_state=0, oob_score=True). We determined feature importance by training random forest models on the full dataset iteratively, each time randomly permuting a feature. The increase in error upon permutation of a feature is related to the importance of that feature.

Once we had calculated relative feature importance, we iteratively performed 10-fold cross validation, i.e. train the model on 90% of data and test on the remaining 10%. The starting model used feature with the highest importance, position average, i.e. the average score of all other substitution variants at that amino acid position and

the n-1 and n+1 positions. If there were no measured variants at the n-1, n, or n+1 positions, we included the n-2 and n+2 positions. We then iteratively performed 10-fold cross validation with models incorporating features in decreasing order of their importance until the Pearson correlations between predicted and observed scores plateaued. The final model was again assessed using 10-fold cross validation (Figure S1 and Table S1). Finally, we used the final model trained on all measured abundance scores to predict all unmeasured variants.

We classified the full set of missense protein variants (measured and imputed) as wildtype-like, hypomorphic, or truncation-like for fitness and abundance scores (Figure S2, Table S1). For fitness score, we considered variants wildtype-like if they were within the 2.5 and 97.5 percentile of synonymous wildtype fitness scores (Figure S2C-D). We considered variants truncation-like if their fitness scores were within the 2.5 and 97.5 percentile of nonsense variants at positions 1-350, excluding the regulatory tail because nonsense mutations in the tail are not damaging in the yeast assay. We considered variants hypomorphic for fitness score if they were between the wildtype-like and truncation-like bounds.

We classified wildtype-like variants similarly for abundance score with a slight adjustment to the distribution boundaries (Figure S2C-D). Because the abundance score distribution tails were larger than the fitness score distribution tails, we defined the bounds as the 5 and 95 percentile of synonymous wildtype distribution. We considered variants to be truncation-like for abundance score if they were within the 5 and 95 percentile of nonsense variants at positions 30-300, in order to exclude

known experimental artifacts of variants near the protein termini due to the nature of the fusion protein used in the experiments.[94]

PTEN Population Variants from GnomAD

Data from the controls-only subset was downloaded from gnomAD v2.1[2] on January 10, 2019 (Table S2). For the cancer incidence analysis and the clinical outcomes odds ratio analyses, we included all controls-only gnomAD nonsynonymous variants (e.g., missense, nonsense, and indel frameshift). For the pathogenic vs. benign analysis, we considered only gnomAD missense variants (i.e. we excluded frameshifting or truncation variants) with the exceptions of p.Arg173His and p.Lys289Glu, which are classified as pathogenic or likely pathogenic in ClinVar. We also excluded p.Asp268Glu, which occurs at a frequency greater than an order of magnitude over most other variants.

Cleveland Clinic *PTEN* Cohort

This study was performed in accordance with the IRB# 8458 protocol "Molecular Mechanisms Involved in Cancer Predisposition" substudy PTEN, which has been approved by the Cleveland Clinic Institutional Review Board for Human Subjects' Protection, and conducted with informed consent and in accordance with the World Medical Association Declaration of Helsinki. The CC cohort consists of 256 prospectively accrued individuals with germline *PTEN* nonsynonymous variants (145 missense and 111 nonsense variants) (Table 1 and Tables S3-S4). Genotype information concerning each patient's germline *PTEN* variant, demographic, and clinical data were also included. Collection and validation of clinical phenotypes were performed by experienced clinical personnel as detailed in a previous study.[160]

Demographic information includes the age at last follow-up, sex, and age at diagnosis for various clinical phenotypes, including: macrocephaly, neurodevelopmental pathologies including ASD and developmental delay (DD), and several different types of benign and malignant neoplasia. Adult individuals in the cohort were assigned a CC score, which is a derived sum of the weights of specific neurological, breast and gynecological, gastrointestinal, skin, endocrine, and genitourinary clinical features, assessed by clinical specialists. Both benign and malignant clinical figures, including age of onset, are factored into CC score. Moreover, CC score is a validated, individualized estimate of the pretest probability of having a germline *PTEN* mutation. For example, a score of 15 indicates a 10% probability of mutation. Given the methodology for calculating CC score, it also serves as a semi-quantitative measure of burden of disease with larger scores indicating increasing disease burden and/or younger ages of onset. However, the scoring is only applicable/validated for the adult population (individuals 18 years and older).[160] OFC z-scores were calculated using published, age-indexed tables.[161]

An individual was considered ASD/DD positive if they presented with ASD, DD, variable delay, or intellectual disability. An individual was considered PHTS positive if they presented hamartomatous features including any of the following: benign or malignant tumors, mucocutaneous lesions, arteriovenous malformation, lipomas, goiter, or uncommon skin lesions. Individuals with the common skin findings of skin tags, café-au-lait marks, or penile freckling in isolation, meaning without another hamartomatous feature, were not included in the PHTS group. Individuals who

displayed both the neurodevelopmental and hamartomatous features were placed in the ASD/DD & PHTS grouping.

Logistic Regression Modeling for Pathogenic *PTEN* Variation

To test the accuracy of models with all combinations of features (e.g., fitness scores, abundance scores, and CADD scores), the optimal regularization parameters (L1 vs. L2 regularization and regularization strength) for each feature combination were determined using the GridSearchCV function within scikit learn. We assessed the performance of each model with 10-fold cross-validation, i.e. we iteratively trained models on 90% of the data and used that model to make predictions for the outstanding 10%. We repeated this procedure in order to make predictions for all variants. Once we determined that no multivariate model performed better than the univariate fitness score model, we re-trained the fitness score model on the entire set of known pathogenic and benign variants. The optimal model used L1 regularization with strength of 1.0. We then used this model to predict probability of pathogenicity for all single amino acid *PTEN* variants.

Cancer Incidence and Survival Analysis

For fitness and abundance score analyses, we classified all individuals from the CC cohort and gnomAD into wildtype-like missense, hypomorphic missense, truncation-like missense, or true-truncation (i.e. nonsense or frameshifting) groups (Figure S2). For the combined molecular score analysis, we used the hypomorphic cutoff to designate variants as fitness or abundance plus or minus (i.e., -1.11 for fitness score, 0.71 for abundance score). We assumed the gnomAD individuals were cancer free. The observation period for each subject was set from birth to age at last

clinical follow-up/information. For 26 of the 164 gnomAD individuals, we could unambiguously determine their age range by linking to data provided in the full gnomAD v2.1 variant call file. Since these data were provided in five-year increments, we randomly selected a single year from this range for each individual. For the rest of the gnomAD control cohort, we obtained the distribution of ages from the gnomAD FAQ page. We randomly sampled an age range from the weighted distribution of age ranges, and then randomly generated an age within that range. The imputed ages reflected well the original age distribution in the gnomAD database (Goodness-of-fit chi-square= 0.30, df=12, not significant). Differences in cancer incidence between the genotype groups were compared using the Kaplan-Meyer method and log-rank test. Analyses were performed for overall cancer incidence and individuals were right-censored at age at cancer or age at last follow-up. Significant group differences were then examined using pair-wise comparisons. In order to detect potential differences in early onset cancer incidence, survival curves were further compared at age 35 with right-censoring occurring at age of early onset (<35) cancer or age 35 otherwise.

Calculating Odds Ratios for Clinical Outcomes

Although all individuals with an identified germline, pathogenic *PTEN* variant are clinically classified as belonging to the overarching classification of PHTS, we have developed clinical subgroupings with differing presentations in order to enable genotype-phenotype analyses in this study. All individuals with frameshifting and nonsense mutations were treated as true truncations, as all of these variants occurred upstream of both the final exon and the C-terminal tail. We used IBM SPSS statistical

software (version 25) to perform logistic regression modeling on ASD/DD or PHTS outcomes and survival analyses, using molecular phenotypes as exposures.

3.5 Results

Distribution of Missense Variation across the Primary and Crystal Structures of PTEN

In order to examine *PTEN* genotype-phenotype relationships, we prospectively accrued a cohort of individuals with germline nonsynonymous variation in *PTEN* (Table 1). Previously, PHTS has been used as an umbrella term specifically for classically defined *PTEN*-related disorders (e.g., Cowden syndrome and Bannayan-Riley-Ruvalcaba syndrome).[72] Subsequently, as the phenotypic spectrum of *PTEN* mutations expanded, PHTS became a descriptor for all clinical presentations associated with germline *PTEN* variation.[154] In order to explore potential differences between ASD/DD related phenotypes and those associated with hamartoma/cancer phenotypes, we operationally grouped individuals with the classic hamartoma-related PTEN features as PHTS, while individuals with largely neurodevelopmental clinical features were designated ASD/DD (Materials & Methods). Individuals with a combination of neurodevelopmental and hamartoma-related features were designated as a third group, ASD/DD & PHTS.

The cohort recapitulates the previously observed relative enrichment of ASD/DD phenotypes among those with missense as opposed to nonsense variation (16% vs 5%, Table 1).[122,151] As a comparison, we also collated mutation data from control-only individuals in gnomAD, a database that aggregates sequencing studies.[2] As individuals with pediatric disorders are excluded from gnomAD, and since these

Table 1. Cleveland Clinic cohort of individuals with germline nonsynonymous variation in *PTEN*

	Missense		True Truncations		Total		
Phenotype	**All N (%)**	**% Male**	**All N (%)**	**% Male**	**All N (%)**	**% Male**	**% Mis.**
All	145 (100)	40.7	111[a] (100)	49.5	256 (100)	44.5	56.6
ASD/DD	23 (15.9)	78.3	6 (5.41)	100.0	29 (11.3)	82.8	79.3
ASD/DD & PHTS	32 (22.1)	68.8	24 (21.6)	75.0	56 (21.9)	71.4	57.1
PHTS	90 (62.1)	21.1	80 (72.1)	38.8	170 (66.4)	29.4	52.9

[a] One individual did not have qualifying symptoms for any phenotype group.

individuals were specifically accrued as unaffected controls, they were assumed to be free of *PTEN*-related disorders. We categorized missense variants by associated clinical group and then mapped variants to the primary/functional domain and crystal structures of PTEN, including the variants catalogued in gnomAD (Figure 1A-B). The clinical missense variants cluster most heavily in the dual-specificity phosphatase domain (residues 1-178), and are depleted in the C2 domain & tail (residues 179-403), reflecting the importance of the phosphatase domain to *PTEN* function (Figure 1A). Comparing all clinical variants with gnomAD variants demonstrates an enrichment of gnomAD variants in the C2 domain & tail, as compared to the phosphatase domain (odds ratio = 5.13, 95% CI = 2.5-10.8, p = 1.6×10^{-6}, Fisher's exact test, Figure 1B). In contrast, and consistent with similar studies, the distributions of variants were similar across clinical outcomes (p = 0.78 for ASD/DD vs. PHTS, p = 0.71 for ASD/DD vs. ASD/DD & PHTS, p = 0.42 for PHTS vs. ASD/DD & PHTS, Fisher's exact test, Figure 1A). In 3D space, gnomAD variants are significantly more solvent exposed (i.e., exposed to the surface of the protein) than the group of all clinical variants (medians of 87.4% vs 7.8%, p = 1.28×10^{-18}; Mann-

Whitney U-test, Figure 1B). Variation at solvent exposed positions is generally more tolerated because these variants are less likely to disrupt protein structure.[162]

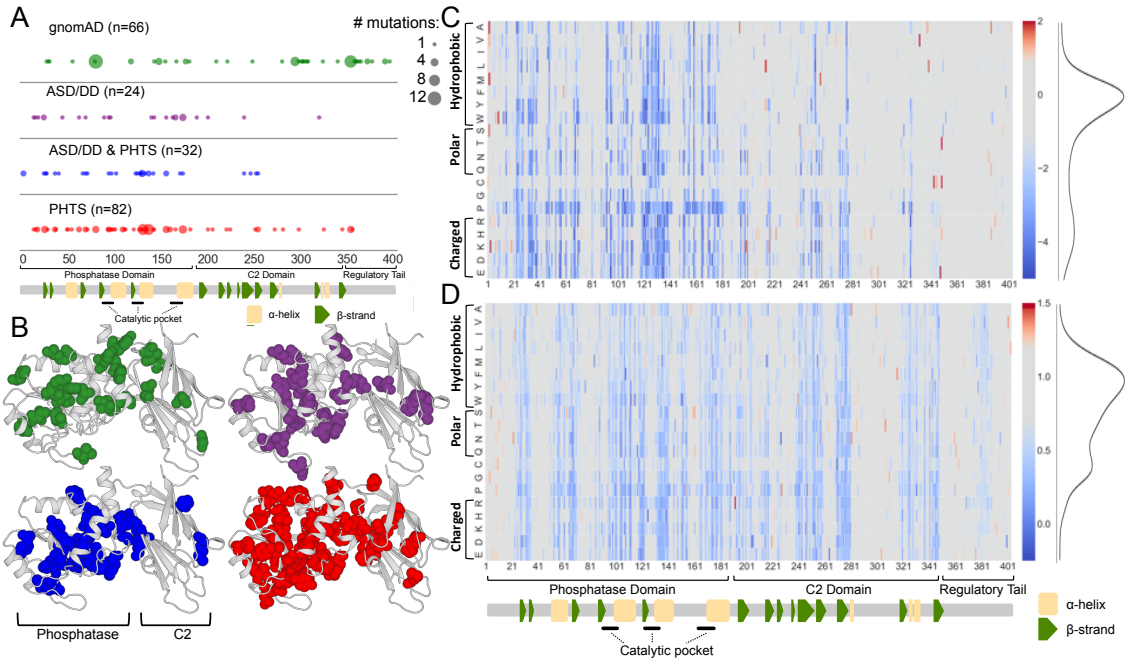


Figure 1. Overview of the datasets used in this study
(A) Diagram of PTEN primary protein structure with locations of PTEN missense variants found in the controls-only gnomAD population or associated with various clinical presentations in the CC cohort. The major protein domains and secondary structure assignments are indicated. Size of circle indicates number of different amino acid variants at that position.
(B) Diagram of PTEN 3D crystal structure with locations of PTEN missense variants found in the controls-only gnomAD population or associated with various clinical presentations in the CC cohort. The C-terminal tail was not solved in the crystal structure and therefore variants falling in this region are not shown. Color of spheres indicates same groups as shown in (A).
(C) Lipid phosphatase fitness scores are displayed as a heatmap, with blue corresponding to damaging variants (i.e. low lipid phosphatase fitness or low abundance), gray corresponding to wildtype-like, and red corresponding to putative fitness or abundance increasing variants (Materials & Methods). The distributions of all missense variants are shown as smoothed histograms on the right.
(D) Cellular abundance scores displayed as in (C).

Visualization and Imputation of Molecular Phenotypes of *PTEN*

We hypothesized that variant level molecular phenotype data might uncover new genotype-phenotype associations, based on the reasoning that protein function data should correlate better with clinical outcome than variant locations in primary or tertiary sequence space. We aggregated molecular phenotype information derived

from recent DMS studies on the effect of thousands of variants on PTEN protein function, including inferred lipid phosphatase activity (i.e., fitness score) and steady-state protein stability (i.e., abundance score).[94,156] Previously, we demonstrated that by using a random forest-based machine learning modeling approach, fitness scores of variants withheld from model training could be imputed with high accuracy. The model incorporated the position average effect of variants missing from a nearly complete DMS dataset (86% saturation) with biophysical, biochemical, and evolutionary data. Therefore, using the imputations from this model, we previously constructed a comprehensive lipid phosphatase functional map of fitness scores (Figure 1C).

We previously showed by down sampling the fitness dataset that a similar strategy could be used for less complete DMS datasets and still result in highly accurate predictions.[156] We developed a similar modeling strategy for the protein abundance DMS dataset, which was at ~54% saturation. Cross validation showed the best performing model could accurately predict withheld abundance scores with an accuracy similar to biologic replicates (Pearson r = 0.75, Figure S1). Therefore, using this approach, we imputed abundance scores for all missing missense variants (Figure 1D and Table S1). Combined, these complete datasets represent estimates of the effect of any given *PTEN* missense variant on the lipid phosphatase activity and steady-state abundance of PTEN protein. All analyses presented here used the

combination of high confidence measured and imputed scores.

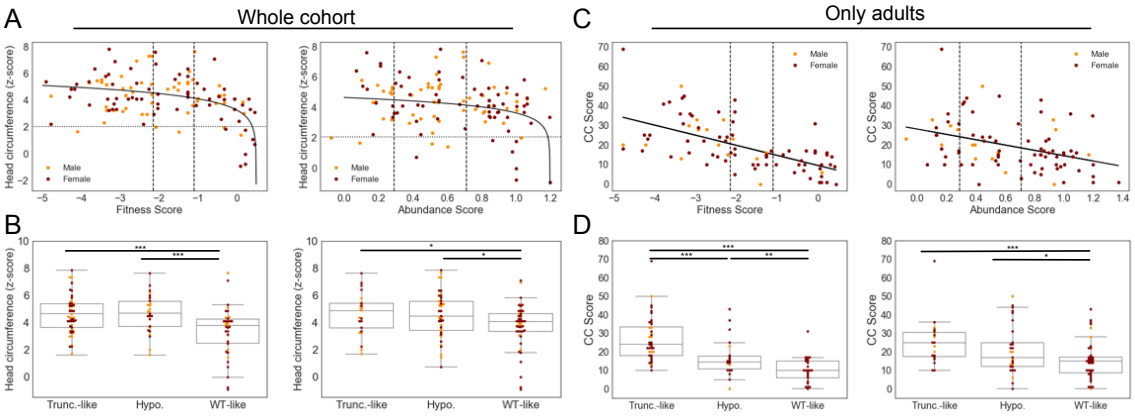


Figure 2. Relationships between molecular phenotype scores and quantitative clinical traits for missense variants

(A) Occipital frontal circumference (OFC, z-score) plotted as a function of continuous fitness score or abundance score for all individuals with missense variants. Males are shown as orange, females are shown as maroon. Vertical dashed lines indicate the hypomorphic and truncation-like cutoffs at -1.11 and -2.15, respectively. Horizontal dashed line indicates threshold for macrocephaly (z-score = 2.054). Solid lines indicate logarithmic curves fit to the data, mean squared error = 2.06 and 2.30 for fitness and abundance, respectively.

(B) Box plot of OFC z-scores for all individuals with missense variants with fitness or abundance scores in the wildtype-like, hypomorphic, or truncation-like ranges. Fitness scores: Trunc.-like vs. WT-like, Cohen's r = 0.35. Hypo. vs. WT-like, Cohen's r = 0.40. Abundance scores: Trunc.-like vs. WT-like, Cohen's r = 0.24. Hypo. vs. WT-like, Cohen's r = 0.22.

(C) CC score for adults with missense variants as a function of continuous fitness or abundance scores. Analyses are restricted to adults as CC score is not valid for individuals under 18. Vertical dashed lines indicate the hypomorphic and truncation-like cutoffs at -1.11 and -2.15, respectively. Solid lines indicate linear curves fit to the data.

(D) Box plot of CC scores for adults with missense variants with fitness or abundance scores in the wildtype-like, hypomorphic, or truncation-like ranges. Fitness scores: Trunc.-like vs. Hypo., Cohen's r = 0.44. Trunc.-like vs. WT-like, Cohen's r = 0.70. Hypo. vs. WT-like, Cohen's r = 0.34. Abundance scores: Trunc.-like vs. WT-like, Cohen's r = 0.42. Hypo. vs. WT-like, Cohen's r = 0.20. *p<0.05; **p<0.01; ***p<0.001.

Fitness scores are modestly correlated with abundance scores (Pearson's r = 0.43, Figure S2E), suggesting some information overlap but that each assay is also capturing (and failing to capture) unique variant effects on protein function. We used the distribution of programmed truncating (nonsense) and synonymous variants in these assays to define truncation-like, hypomorphic, and wildtype-like missense variant categories (Materials and Methods, Figure S2A-D). The missense variants in

both datasets are bimodally distributed, with the majority of variants having wildtype-like scores (Figures 1C-D and S2). We found that for both measures, variation in the phosphatase domain was generally more damaging, i.e. truncation-like or hypomorphic, than variation in the C2 domain or regulatory tail (fitness: 42% vs. 12%, abundance: 50% vs. 38%, Figure 1C-D).

Fitness and Abundance Scores Explain Quantitative Clinical Traits

In an effort to link genotype to quantitative clinical phenotypes, we evaluated whether fitness and abundance scores of individuals' *PTEN* missense variants could explain the degree of macrocephaly or phenotype burden. Burden was assessed by CC score, which takes into account neurological features as well as benign and malignant lesions of the body, for individuals over 18 (Materials and Methods). Molecular phenotype scores were evaluated numerically as well as using the defined functional categories (e.g., wildtype-like, hypomorphic, and truncation-like). We found a logarithmic relationship between fitness score and head size measured by OFC, with z-scores plateauing around the hypomorphic cutoff (Figure 2A, left). Accordingly, we found a significant difference in OFC between the population of wildtype-like variants and truncation-like as well as hypomorphic variants ($p = 4.3x10^{-5}$ and $6x10^{-4}$, respectively, Mann-Whitney U-test), but no difference between hypomorphic and truncation-like variant fitness scores (Figure 2B, left). We also observed a logarithmic relationship between OFC and abundance score (Figure 2A, right). Treating abundance as a categorical variable revealed significant differences between wildtype-like and both truncation-like and hypomorphic variants ($p = 0.02$ and 0.01, respectively, Mann-Whitney U-test). Similar to the fitness score, there was

no difference between the distribution of truncation-like and hypomorphic variants (Figure 2B, right).

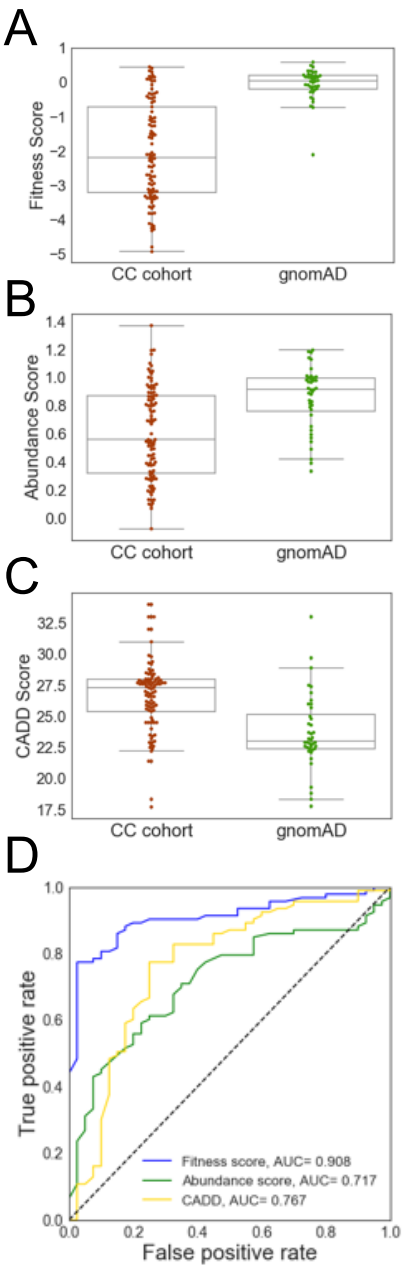


Figure 3. Molecular phenotypes discriminate clinical from gnomAD missense *PTEN* variation

These analyses are done at the variant level to prevent recurrent variants from biasing the results. Two known pathogenic variants were removed from the gnomAD list (Materials & Methods).

(A) Box plots comparing fitness scores between variants found in the CC cohort versus gnomAD. Cohen's r = 0.61.

(B) Box plots comparing abundance scores between variants found in the CC cohort versus gnomAD. Cohen's r = 0.37.

(C) Box plots comparing CADD scores between variants found in the CC cohort versus gnomAD. Cohen's r = 0.45.

(D) Receiver operator characteristic curves for univariate models. Feature weights reported in Table S7.

In our analysis of phenotype burden, we found a significant linear relationship between missense variant fitness score and CC score ($p = 3.7 \times 10^{-10}$), with fitness score explaining 37% of the variation in CC score (Figure 2C, left). Similarly, treating fitness score as a categorical variable, we found that more damaging groups of variants had distributions shifted toward higher (more severe) CC scores (Figure 2D, left). CC scores for truncation-like variants were significantly higher than those of hypomorphic variants ($p = 2.5 \times 10^{-4}$). Additionally, CC scores for hypomorphic variants were in turn significantly higher than those of wildtype-like variants ($p = 9.2 \times 10^{-3}$).

Alternatively, for abundance scores, while a significant linear relationship exists between CC score and abundance score ($p = 3.2x10^{-4}$), it explains only 14% of the variation in CC score (Figure 2C, right). Likewise, when treating abundance score as a categorical variable, more modest trends were observed compared to the trends for fitness score. The abundance scores for truncation-like variants trend toward higher CC scores than hypomorphic variants ($p = 0.08$), while hypomorphic variants are nominally higher than wildtype-like ($p = 0.045$) variants. Truncation-like variants are significantly different from wildtype-like variants ($p = 6.7x10^{-4}$; Figure 2D, right). Combined, these results underscore the potential for molecular phenotypes to partially explain clinical outcomes.

Molecular Phenotype Data Accurately Distinguishes Likely Pathogenic from Benign Variation

We previously showed that fitness scores discriminate ClinVar pathogenic/likely pathogenic *PTEN* variation from gnomAD putatively benign variation.[156] Here, we examined whether molecular phenotype data could identify pathogenic variation in this set of alleles, and whether combining molecular phenotype data could improve performance compared to univariate approaches. Thus, we contrasted the CC cohort of likely pathogenic *PTEN* variants to putatively benign, population *PTEN* variants catalogued in the gnomAD control-only individuals (Materials and Methods, Figure 3, Table S5). We found that variants from the CC cohort were predicted to be significantly more damaging by both fitness score ($p = 6.5x10^{-13}$, Mann-Whitney U-test) and abundance score ($p = 7.6x10^{-6}$, Figure 3A-B). As a comparison, CADD scores[9,163] are also more damaging for the CC cohort group (p =

$6.5x10^{-8}$, Figure 3C). Of these three predictors, fitness scores demonstrate the highest area under the receiver operating characteristic curve (AUC = 0.908, 10-fold cross validation, Figure 3D). While these predictors are correlated, the relationships are modest (Spearman rho = 0.52-0.59, Figure S3A), suggesting that multivariate models could yield improved performance. Therefore, we constructed logistic regression models using various combinations of the molecular phenotype data and CADD scores to find the model that most accurately discriminates the two groups (Figure S3B). We found that no multivariate model performed significantly better than the fitness score univariate model (Figure S3B). Nevertheless, the substantial increase in predictive power of the fitness score model over the CADD model highlights the power of empirical molecular phenotype data to accurately predict pathogenicity.

Molecular Phenotypes Identify Subgroups with Distinct Cancer Susceptibility

While it is known that *PTEN* mutations dramatically increase the lifetime risk of developing specific cancers, we sought to understand whether molecular phenotypes could highlight functional classes of missense variants with differences in cancer susceptibility (Materials and Methods). As a comparison group, we included individuals with variants that are predicted to be truly truncating (e.g., nonsense and frameshifting). Survival functions were first compared between all classes of missense fitness or abundance scores and the true truncations, with pairwise comparisons of survival functions when significant differences were found (Figure 4A-B). In this analysis, cancer free status was considered as the survival criterion.

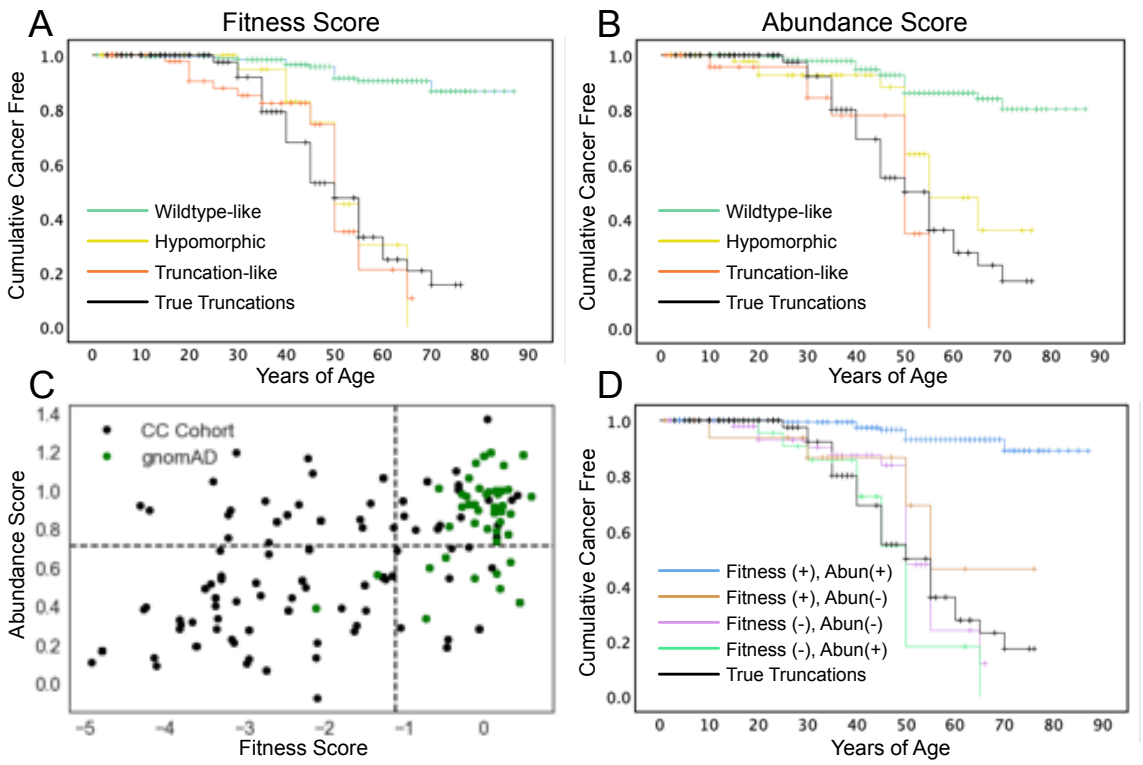


Figure 4. Effects of molecular phenotype on cumulative cancer incidence

(A) Survival-like analysis of individuals with different fitness score classes of *PTEN* missense or true-truncation (e.g. nonsense, frameshifting) variants. "Survival" here is defined as being cancer-free. Ticks represent right-censored individuals, i.e. the age at last follow-up. n=204 wildtype-like; 35 hypomorphic; 68 truncation-like, 114 true truncations.

(B) Survival-like analysis of individuals with different abundance score classes of *PTEN* missense or true truncation (e.g. nonsense, frameshifting) variants. "Survival" here is defined as being cancer-free. Ticks represent right-censored individuals, i.e. the age at last follow-up. n=210 wildtype-like; 65 hypomorphic; 32 truncation-like, 114 true truncations.

(C) All CC cohort and gnomAD individuals plotted as a function of fitness and abundance scores. Dotted lines indicate wildtype-like thresholds (-1.11 and 0.71 for fitness and abundance scores, respectively).

(D) Survival-like analysis for individuals with *PTEN* missense variants falling in fitness/abundance quadrants or true truncating variants. "Survival" here is defined as being cancer-free. Ticks represent right-censored individuals, i.e. the age at last follow-up. n=29 fitness(-), abundance(+); 74 fitness(-), abundance(-); 23 fitness(+), abundance(-); 181 fitness(+), abundance(+), and 114 true truncations.

For fitness scores, survival functions were significantly different ($p = 3.2 \times 10^{-24}$, Log rank, Figure 4A and Table S6). Pairwise comparisons showed that all of the reduced fitness score categories survival functions were similar to each other and significantly different from the wildtype-like survival function (Table S6). Based on the shape of the survival functions, we hypothesized that there may be a difference in early onset risk. Therefore, we conducted a subanalysis with right-censoring at age

35, which again showed significant overall differences ($p = 3.0x10^{-6}$, Log rank). Pair-wise comparisons showed that these differences were driven by truncation-like and true truncations categories, each significantly deviated from wildtype-like variants ($p = 1.0x10^{-5}$ and $2.4x10^{-7}$). The hypomorphic survival function appears visually to be intermediate between the groups. However, across this age range the hypomorphic function did not significantly differ from the wildtype-like function ($p = 0.35$) or either of the truncation-like or true truncation categories ($p = 0.155$ and 0.122, Figure 4A).

Variant classes defined by abundance scores also had significantly different survival functions ($p = 6.2x10^{-18}$, Log rank, Figure 4B and Table S6). In contrast to the fitness scores, pairwise comparisons revealed a step-wise relationship for abundance scores, with hypomorphic missense variants conferring greater lifetime hazard than wildtype-like ($p = 1.3x10^{-4}$), and truncation-like missense conferring greater hazard than the hypomorphic abundance class ($p = 0.024$, Figure 4B). True truncations may confer greater hazard than hypomorphic missense, but this comparison was not significant ($p = 0.07$). Right-censoring at age 35 identified significantly different survival functions for the abundance-defined variant classes as well ($p = 1.1x10^{-5}$). Similar to the fitness score analysis, these differences were driven by truncation-like and true truncations functions which were significantly different from wildtype-like between birth and age of 35 ($p = 2.5x10^{-5}$ and $1.0x10^{-6}$, respectively). The hypomorphic survival function was visually intermediate between wildtype-like ($p = 0.06$) and the truncation groups ($p = 0.15$ and 0.20), but none of the comparisons were significantly different.

We next leveraged the two-dimensional molecular phenotype data to separate missense variants into four categories based on deficiencies in fitness score, abundance score, or both. For this analysis, hypomorphic and truncation-like scoring variants were combined as the negative group for PTEN function for each score in order to keep adequate group sizes. Compared to CC Cohort, gnomAD individuals are enriched in the fitness positive, abundance positive quadrant (Figure 4C). The survival functions for these combined molecular phenotype defined groups were significantly different ($p = 6.3x10^{-24}$, Log rank). Pairwise comparisons, showed variants retaining wildtype-like fitness (+) and abundance (+) have the lowest overall hazard and have a survival function that is significantly different from all other groups (Table S6). The remaining three classes are deficient for either fitness, abundance, or both scores, and are not significantly different from each other or the true truncations (Figure 4D). These data provide high resolution comparisons of cancer risk for different variant classes, which can be further clarified by larger sample sizes.

Molecular Phenotypes Identify Distinct Risk Profiles for ASD/DD and PHTS Subgroupings

Understanding the molecular differences between the variants that associate with ASD/DD versus PHTS (especially cancer occurrence) outcomes is a critical goal for understanding PTEN pathobiology, which ultimately guides clinical management. Consistent with our and others' previous findings,[80,156] fitness scores of individuals in the ASD/DD group are less damaging than the PHTS positive groups ($p = 5.5x10^{-3}$, 0.011, Mann-Whitney U, for ASD/DD vs ASD/DD & PHTS and ASD/DD vs PHTS,

respectively, Figure 5A). However, there is no difference in abundance scores between the clinical phenotype groups (Figure 5B).

Next, we tested whether the severity of variant molecular phenotype, as assessed by fitness or abundance score, affected the odds of developing ASD/DD or PHTS symptoms (regardless of presence or absence of the other qualifying symptoms). We included all members of the CC cohort as well as gnomAD individuals (Materials & Methods). Using a logistic regression model, we calculated odds ratios (OR) for ASD/DD and PHTS as a function of fitness or abundance scores. Wildtype-like variants were used as the reference group. For both molecular phenotypes, more severe missense variants do not significantly increase the odds of an individual developing ASD/DD (OR ranges = 3.9-6.1 and 4.2-7.8 for fitness and abundance, respectively). The odds for true truncation variants are marginally decreased compared to the missense variant classes, though this trend is not significant (Figure S4B). In contrast, for fitness and abundance scores, the odds of an individual developing qualifying symptoms for a PHTS classification increase as mutation severity increases in a stepwise manner, with stronger differences in risk observed for the abundance score (OR ranges = 20.4-51.3 and 5.1-28.7 for fitness and abundance, respectively, Figure S4B).

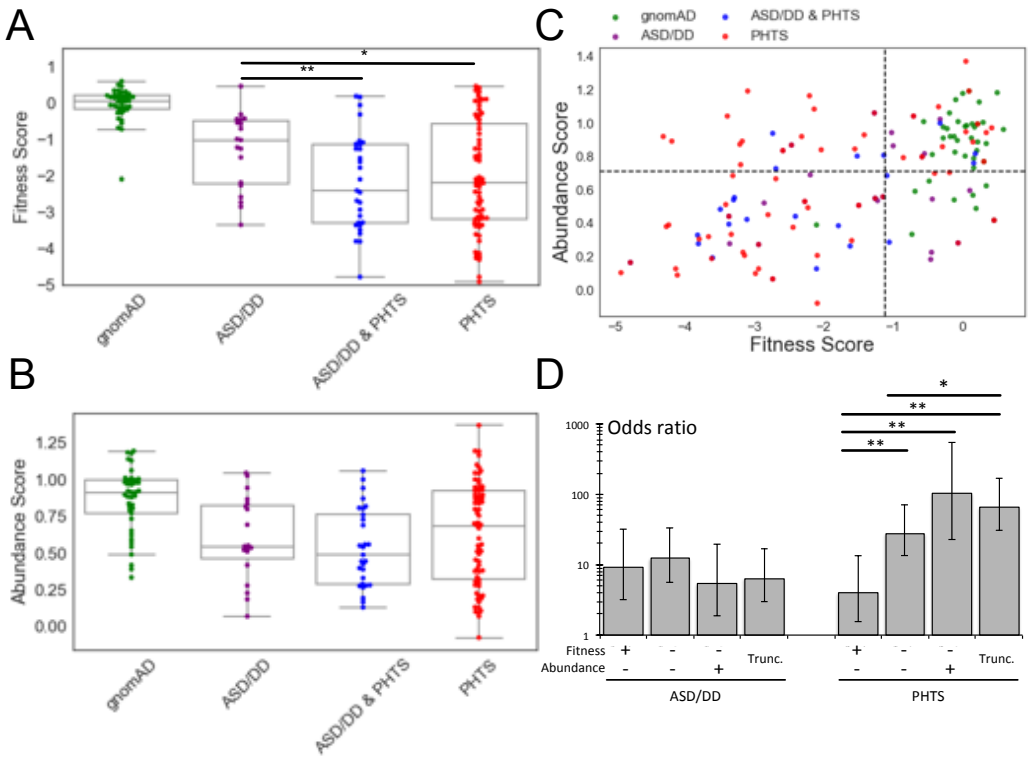


Figure 5. Association of molecular phenotypes with specific clinical outcomes
(A) Box plots showing fitness scores of *PTEN* variants occurring in individuals in gnomAD or different clinical categories. ASD/DD vs. ASD/DD & PHTS, Cohen's r = 0.36. ASD/DD vs. PHTS, Cohen's r = 0.25.
(B) Box plots showing abundance scores of *PTEN* variants occurring in individuals in gnomAD or different clinical categories.
(C) *PTEN* variants plotted according to fitness and abundance scores, colored by clinical group. Dashed lines indicate hypomorphic cutoffs (-1.11 for fitness, 0.71 for abundance).
(D) Odds ratios for developing ASD/DD or PHTS symptoms for different variant classes. Odds ratios represent the comparison of that class with variants in the fitness positive, abundance positive quadrant (top right in C). Error bars represent 95% confidence intervals**.** Odds ratios and 95% confidence intervals reported in Table S8. *p<0.05, **p<0.01.

We next tested whether two-dimensional molecular phenotype data would provide additional insights into the risk for developing ASD/DD or PHTS symptoms. While variants from the control individuals from gnomAD clearly cluster in the fitness positive, abundance positive quadrant, the affected individuals populate the other three quadrants (Figure 5C). Interestingly, compared to the PHTS positive categories, the ASD/DD has a larger fraction of individuals in the fitness positive, abundance

positive (30% vs. 10% and 20% for ASD/DD vs ASD/DD & PHTS and PHTS, respectively), and a smaller fraction in the putatively dominant negative fitness compromised, abundance positive quadrant (5% versus 21% and 25% for ASD/DD vs ASD/DD & PHTS and PHTS, respectively, Figure S4A).

Again, using a logistic regression approach, we generated odds ratios for the combined two-dimensional molecular phenotypes. We again observed no major differences in the odds for developing ASD/DD in any missense groups or the true truncation category (OR range = 5.4-12.4). In contrast, the odds for developing PHTS are highly dependent on the variant grouping (OR range = 4.1-102.9). Missense variants that maintain lipid phosphatase activity but are low abundance show the lowest odds for an individual developing qualifying symptoms for PHTS classification (OR = 4.1, 95% CI = 1.5-10.7, Figure 5D). Missense variants that were fitness and abundance negative showed a significantly different intermediate risk (OR = 27.6, 95% CI = 13.5-56.5). Variants that have wildtype-like abundance but abrogated lipid phosphatase activity (putative dominant negative variants) have the highest odds of an individual developing qualifying symptoms for PHTS classification (OR = 102.9, 95% CI = 22.8-464.0), though not significantly different from variants in the fitness negative/abundance negative or true truncation categories (Figure 5D).

3.6 Discussion

Despite two decades of effort, we still lack a clear understanding of how *PTEN* genotype affects specific clinical phenotypes. Recent advances in DNA synthesis and sequencing technologies allow a new experimental paradigm in which the effects of thousands of variations on protein function can be empirically measured in parallel.

Two such experiments recently explored the effects of *PTEN* variation on lipid phosphatase activity (fitness score) and steady state cellular abundance (abundance score).[94,156] Using imputation, we generated estimated functional scores for all possible *PTEN* missense variants. In order to understand how molecular phenotype data relates to clinical outcomes, we integrated these data with clinical information from the CC cohort of *PTEN* mutation-positive individuals. These analyses have validated the clinical utility of comprehensive multi-dimensional functional scores and have uncovered unexpected insights into the PTEN genotype-phenotype map.

Our analyses demonstrate that molecular phenotype scores are correlated with quantitative clinical traits. Fitness and abundance scores showed a logarithmic relationship with the most penetrant *PTEN* phenotype, macrocephaly (~95% of PTEN patients).[116] In previous work, we designed an algorithm to determine a patient's *a priori* risk for having a germline *PTEN* mutation (CC score). CC score is also a surrogate measure of an adult patient's phenotypic burden accounting for age of onset.[160] CC scores and functional scores have a linear relationship with more severe phenotypic burden associating with worse functional scores. We then demonstrated that molecular phenotype data can be used to model and thus predict likely pathogenic variants with high accuracy, compared to CADD, a completely *in silico* approach.

While broadly predicting pathogenicity has value in a clinical setting and can help resolve *PTEN* variants of uncertain significance (VUS), we were also interested in exploring if these molecular phenotypes could provide additional insights into the diverse clinical outcomes associated with germline PTEN disruption. Our analyses

showed that molecular phenotypes can define subgroups of patients with common or unique age-related cancer hazard. While putative true truncating variants, such as nonsense mutations, showed high lifetime cancer risk, highly damaging missense variants as defined by the molecular phenotypes appear to be at least as impactful. Moreover, our data from single molecular phenotypes show truncation-like and true truncations survival functions separate from wildtype-like functions over an early onset age range. Hypomorphic functions are potentially intermediate over this early onset range but not yet significantly different from the wildtype-like functions. Combining molecular phenotype scores provides further granularity for these cancer risks and identified variants that are lipid phosphatase active (fitness positive) but unstable (abundance negative) as a potentially intermediate class of cancer risk variants.

A growing number of studies have provided important insight into the question of genotypes driving diverse phenotypic outcomes for carriers of germline *PTEN* variants. However, these studies have generally been limited by small sample sizes. For instance, Spinelli et al. (2015) investigated the lipid phosphatase activity and protein stability of seven ASD-associated and five PHTS-associated *PTEN* missense variants using virally transfected U87MG cells. They found ASD-associated variants retained partial phosphatase function but exhibited dramatically decreased stability, whereas PHTS-associated variants lost phosphatase function but exhibited relatively better stability.[80] These findings form the basis for the hypothesis formulated by Leslie and Longy (2016), in which ASD/DD results from hypomorphic *PTEN* variants while traditional PHTS (i.e. hamatomatous and malignant growth)

results from more damaging variants.[122] Our previous work using fitness scores (i.e. inferred lipid phosphatase activity) and ASD/DD- or PHTS-associated variants from the literature lent support to this hypothesis.[156]

Here, using the largest set of clinically annotated variants examined to date, we strengthen these previous findings by showing that ASD/DD-associated *PTEN* variants, on average, retain hypomorphic lipid phosphatase activity, while those associated with either ASD/DD and PHTS or PHTS alone are more damaging. Moreover, the fraction of missense variants and the distribution of variants according to fitness and abundance scores are more similar between the two PHTS associated groups, suggesting that they are in fact molecularly similar. We made the surprising discovery, however, that risk for developing ASD/DD is not dramatically altered across different variant loss-of-function categories, while the risk for PHTS can increase by an order of magnitude. Thus, it appears that while all individuals with pathogenic *PTEN* variants are at substantial risk for developing ASD/DD, the risk, and thus the subsequent penetrance, of PHTS symptoms (i.e. hamartomatous and malignant growth) is significantly greater for true truncations and truncation-like missense variants. These differential risk profiles would then explain the lower fraction of true truncations in cohorts recruited primarily based on an ASD/DD diagnosis.[122,151]

The biologic basis of these differential risk profiles remains unclear. Retaining any lipid phosphatase activity of the variant allele, coupled with the second functional *PTEN* allele, may be sufficient to prevent the formation of hamartomas in some cases. There are numerous PTEN functions that are not described by molecular phenotypes

included in this study. Lipid phosphatase independent functions may also modulate risk. For example, recent studies have shown a potential relationship between altered PTEN subcellular localization and clinical outcomes, where PTEN variants showing aberrant nuclear depletion associated with ASD/DD.[63,82,164,165] Ideally, a comprehensive analysis would include the effect of variation on PTEN's protein phosphatase activity, subcellular localization, nuclear function, and protein-protein interaction. New high-throughput assays may make such datasets available in the near future.

Given that the majority of ASD/DD diagnoses are from children or young adults, an important open question is what will their lifetime risk for neoplasia truly be? Longitudinal tracking to definitively assess neoplasia/cancer risk in this cohort will improve the allocation of clinical resources and guide the precision delivery of care. Our current data suggest that certain subsets of individuals with *PTEN* associated-ASD/DD are likely to have higher cancer risk than other subsets. Longitudinal follow-up with these individuals or new prospective recruitment efforts will be needed to answer this question definitively.

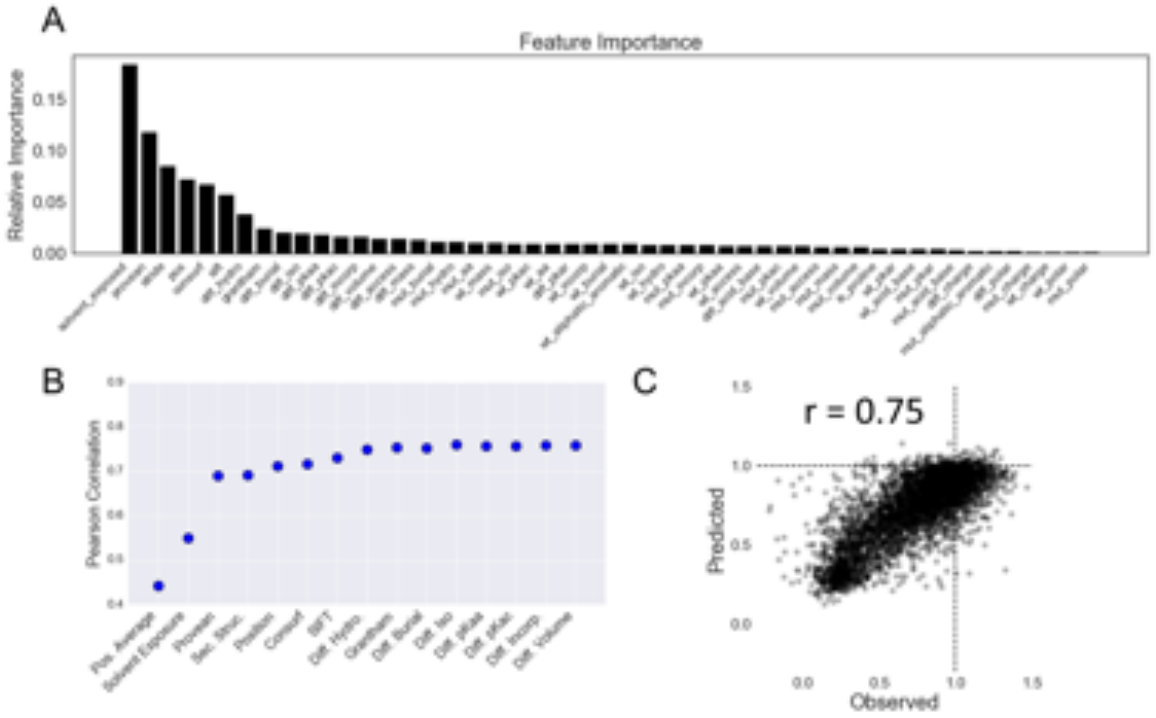


Figure S1. Imputation of missing abundance scores

We used 51 evolutionary, biophysical, biochemical, or *in silico* variant effect predictors to train a random forest machine learning model to predict the effect of unseen variants (See Mighell et al., 2018).

(A) Feature importance of all features (besides position average) from the full dataset, calculated as the relative increase in error upon random permutation of each feature.

(B) Pearson correlations between predicted and observed variant scores. The first model includes only position average, while each successive row includes that feature as well as all features to the left. Features in order include; "Pos. Average": average abundance score of other single amino acid mutations at that position or neighboring positions (see Materials & Methods); "Solvent Exposure": calculated with GETAREA web tool; "Provean": mutation effect predictions; "Sec. Struc.": secondary structure, enumerated with STRIDE; "Position" : position in primary sequence; "Consurf": evolutionary conservation; "SIFT": mutation effect predictor; "Diff. Hydro": difference in hydropathy between wildtype and variant amino acid; "Grantham": Grantham substitution score; "Diff. Burial": difference in burial between wildtype and variant amino acid; "Diff. Iso": difference in isoelectric point between wildtype and variant amino acid; "Diff pKaa": difference in amino pKa between wildtype and variant amino acid; "Diff. PKac": difference in carboxyl pKa between wildtype and variant amino acid; "Diff. Incorp": difference in protein incorporation rate between wildtype and variant amino acid; "Diff. Volume": difference in volume between wildtype and variant amino acid. Note: To ensure that our feature selection approach was consistent across subsets of the data, we iteratively repeated this procedure using 90% subsets of the data. We found that the set of top 12 features were consistent across all folds and that the median ranking of features across the folds was the same as what was originally used for modeling.

(C) 10-fold cross validation demonstrated high accuracy of the final model, yielding 0.75 Pearson correlation between predicted and observed variant scores.

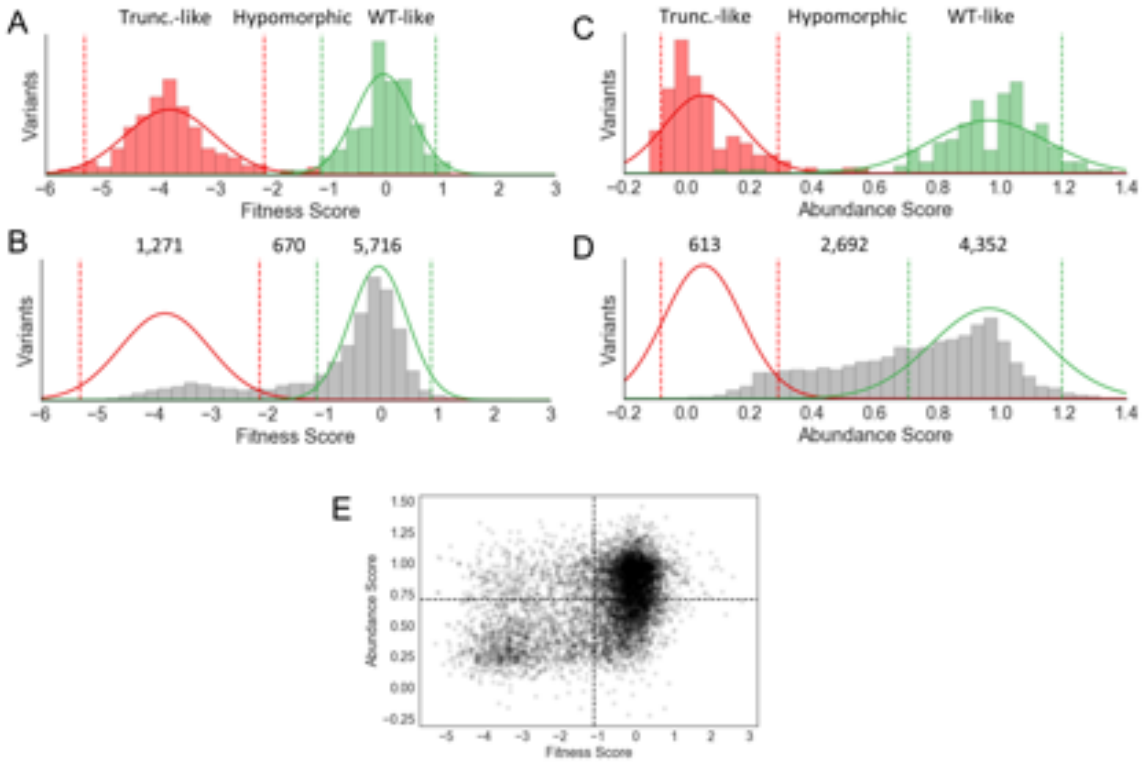


Figure S2. Scaling variant functional scores for integration

(A-B) Histograms showing distributions of fitness scores.

(A) We defined variant fitness scores in relation to the distribution of synonymous-wildtype variants (green) and early truncating nonsense variants (red). We used only truncating variants before the 350th position (i.e. excluding the unstructured tail) to ensure that this distribution represented true loss-of-function. Solid lines are Gaussian fits of the histograms. We drew cutoff lines (dashed lines) corresponding to the 2.5 and 97.5 percentile of these distributions.

(B) Fitness score distribution (gray) for all missense variants, including the high confidence measured and imputed scores, were compared to the synonymous/truncation cutoffs. We considered missense variants to be truncation-like if they fell within the 95 percentile bounds of the truncation distribution, and likewise for missense variants within the synonymous-wildtype distribution. Variants that fell between these two distributions were considered hypomorphic.

(C-D) Histograms showing distributions of abundance scores.

(C) As in (A), variant abundances scores were coded according to their relationship to truncating nonsense (red) and synonymous wild-type variants (green). Only truncating variants between the 30th and 300th position were used in order to exclude measurement artifacts at the termini. Solid lines are Gaussian fits of the histograms. Further, cutoff lines (dashed lines) were drawn at the 5th and 95th percentiles because the tails of the distributions were much longer for abundance scores relative to fitness scores.

(D) As in (B), abundance score distribution (gray) for all missense variants, including the high confidence measured and imputed scores, were compared to the synonymous/truncation cutoffs. For abundances scores, we considered missense variants to be truncation-like if they fell within the 90 percentile bounds of truncation distribution, and likewise for missense variants within the synonymous-wildtype distribution. Variants that fell between these two distributions were considered hypomorphic.

(E) Scatterplot of all missense variants plotted as a function of fitness and abundance scores. Pearson's r = 0.43.

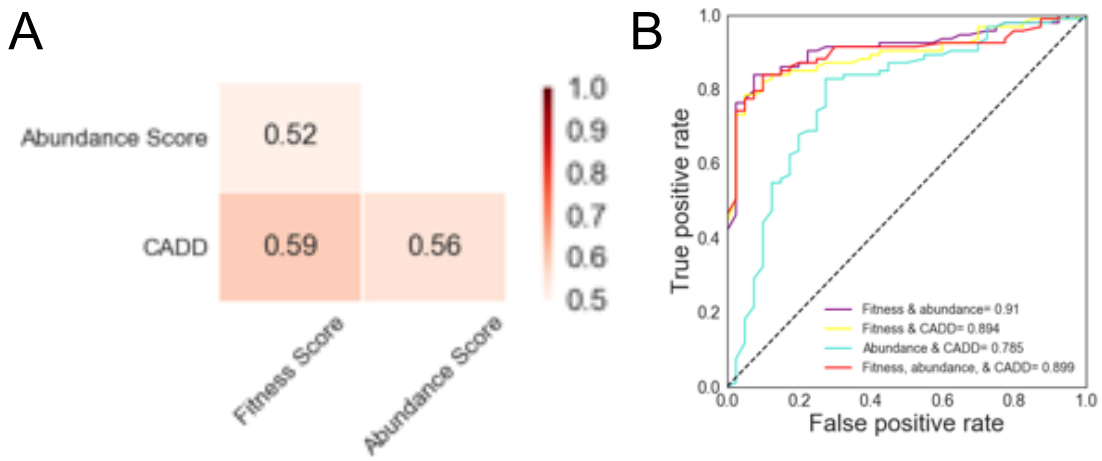


Figure S3. Logistic regression optimization for predicting pathogenic vs. benign PTEN variation

(A) Spearman correlation of features used in the modeling.

(B) Receiver operator characteristic curves with corresponding area under the curve for the various models tested. Model weights reported in Table S7.

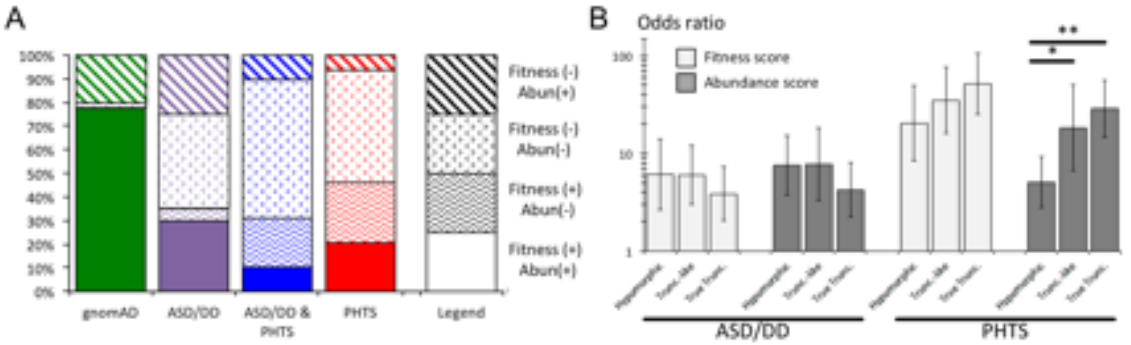


Figure S4. Risk scores for fitness and abundance exposures

(A) Variants in gnomAD or different clinical categories occupy different fitness/abundance quadrants (as shown in Figure 5C).

(B) We calculated logistic regression-based odds ratios for individuals to develop ASD/DD or PHTS features as a function of exposure to hypomorphic scoring, truncation-like scoring, or true truncation variants. Odds ratios are calculated as a comparison between the variant class of interest and the wildtype-like scoring variants. Odds ratios and 95% confidence intervals reported in Table S8. *p<0.05, **p<0.01.

Chapter 4. CRISPR-Capture: a novel, low-cost, and scalable method for targeted sequencing

2.1 Introduction

While advances in next generation sequencing technologies have dramatically reduced the cost and effort required to sequence human genomes, there remains significant clinical and research benefits of targeted enrichment for sequencing. Restricting genomic interrogation to loci of interest minimizes sequencing costs and analytic labor, reduces data processing and storage, and abrogates ethical issues around returning incidental genetic findings to patients and families. An important application of targeted enrichment for sequencing is identifying pathogenic variation in Mendelian disorders. This remains a significant clinical challenge, as the diagnostic rate for many disorders is only ~50%.[98] Exome sequencing or hybridization-based gene-panels have been widely used, but these approaches have critical weaknesses, including sequence bias, limited scalability, high DNA input requirements, and cost (especially for custom applications). Further, in most cases the non-coding regions of gene bodies are not captured. A novel technology that solved these problems could have major benefits for clinical and research applications.

CRISPR-Cas systems have emerged as invaluable biotechnological tools empowering user-programmed nuclease activity.[100,166] These systems leverage target-specific guide RNAs (gRNAs) to direct Cas endonucleases to specific genomic loci. Several methods have taken advantage of CRISPR-based cutting for specific capture and sequencing of genomic regions. However, most of these rely on laborious

size selection steps or require specialized equipment or reagents.[101,167,168] A different set of approaches take advantage of the Cas9-gRNA ribonucleoprotein (RNP) affinity to target DNA by pulling down DNA-bound RNP[104,169]. Recently, a method relying on direct adapter ligation to Cas9 cleavage sites, followed by long-read sequencing with nanopore technology was demonstrated.[103] This method enables detection of structural variation, but suitability for calling single nucleotide variation was not demonstrated.

Here, we set out to design and implement a novel, low-cost, and scalable targeted capture technology around the Cas12a (also known as Cpf1) enzyme, which cleaves target DNA in a staggered fashion, leaving four to five nucleotide overhangs.[105] We reasoned that treating genomic DNA with Cas12a and a pool of gRNAs would result in enrichment of ligatable overhangs specifically at targeted sites. To validate the method, we designed a pilot set of gRNAs targeting 47 known and candidate genes associated with Joubert Syndrome (JS), a genetically heterogeneous, recessive ciliopathy that manifests with hindbrain malformations. We used the performance of the guides in this pilot set to train a linear regression model which learned the sequence determinants of guide performance. We then used this model to design an optimized guide set for a subset of the JS genes. The method is currently compatible with Illumina sequencing platforms, as this is the gold standard for identifying single nucleotide variants and small indels. However, with minor modifications, the method could be adapted to other sequencing technologies.

2.2 Materials and methods

Design of pilot guide set

We obtained RefSeq hg19 genomic coordinates for the 47 genes from UCSC Table Browser as a bed file. Overlapping intervals were merged with Galaxy to obtain a single interval per gene, to which we then padded with 3,000 basepairs upstream and 500 basepairs downstream, in hopes of capturing promoters and 3' untranslated region sequences. Then, we used FlashFry[170] to find all possible Cas12a target sites (i.e. the presence of "TTTN" PAM) within these target regions and to report the copy number of each potential gRNA target sequence. We filtered out guide target sequences that had copy number greater than one, or that had many similar off target sequences (>25 off targets within 1 edit distance, or >100 off targets within 2 edit distance). We also filtered guides that overlapped a common single nucleotide polymorphism (SNP, minor allele fraction > 0.1%, dbSNP, release 153). Then, for each gene, we defined targets by simply enumerating 500 basepair intervals, and selected the gRNA with cut site closest to the target. This resulted in 7,176 guide sequences. We then designed DNA oligo sequences that contained the following in the 5' to 3' direction: dial out priming site, T7 RNA polymerase priming site, crRNA backbone, protospacer sequence, DraI cut-site ("TTTAAA"), and another dial out priming site. We synthesized these gRNA templates as 99-mers across two 12,000-feature oligo pools from CustomArray.

Guide amplification and *in vitro* transcription of pilot guide set

We used PCR to amplify the gRNA templates from the oligo pool. Reactions contained 1x KAPA HiFi Hotstart Readymix, 10 ng of template, 0.5 μM primers, and 1x SYBR Green. Reactions were pulled upon completing exponential amplification, which occurred at 19-22 cycles. Agarose gel electrophoresis confirmed bands of 99

basepairs. We purified reactions with NucleoSpin PCR cleanup columns (Machery Nagel). Then, we treated purified products with DraI restriction enzyme in order to remove the priming site downstream of the gRNA sequence. Reactions contained 500 ng of PCR product, 40 units of DraI (New England BioLabs), and 1x CutSmart buffer. Incubation was done at 37° and proceeded overnight. Reactions were cleaned up with NucleoSpin PCR cleanup columns, and complete digestion was confirmed with agarose gel electrophoresis.

We used MEGAscript T7 Transcription Kit (Thermo Fisher Scientific) to generate gRNAs from the templates. Reactions contained ~60-130 ng DNA (depending on recovery from previous step), and were incubated at 37° overnight. Following incubation, reactions were treated with TURBO DNase and incubated at 37° for 15 minutes. Then, RNA Clean & Concentrator (Zymo Research) columns were used to purify RNA. We quantified RNA with Qubit RNA Broad Range Assay (Thermo Fisher Scientific) and diluted to 10 μM.

CRISPR-Capture workflow

For a detailed protocol, see Supplementary Note 1. Briefly, genomic DNA is treated with phosphatase to enzymatically remove the terminal phosphates from genomic DNA molecules. Then, genomic DNA is treated with gRNA-complexed Cas12a, which creates overhangs specifically at targeted sites. Custom i5 adapters that contain complementary overhangs, a unique molecular identifier (UMI), and 5' biotin modification are added with T4 ligase. Then, the i7 adapter is added through Tn5 tagmentation. A streptavidin-mediated pulldown step purifies those molecules that have an i5 adapter (excluding the molecules with only i7 adapters), and on-bead

PCR (followed by size selection/purification as necessary) generates ready-to-sequence libraries. All libraries were sequenced in paired-end mode on the Illumina NextSeq platform with Mid Output 150 cycle v2.5 kits. Cycles were allocated as follows: 35 cycles for read1, 10 cycles for index1, 6 or 10 cycles for index2 (depending on the presence of unrelated multiplexed libraries), and 113 or 118 cycles for read2.

Sequencing data processing and analysis

Our custom adapter contains a UMI in place of the i5 index. The first step of our informatics pipeline is appending the sequence from the i5 index read to the end of the read name line of both read 1 and 2 fastq files with a custom python script. This is done for compatibility with UMI-tools[171]. Next, adapters are trimmed with cutadapt[128] and paired end reads are aligned to the hg19 reference genome with BWA-MEM[172]. Following paired end read alignment, duplicates are removed with umi_tools dedup.

Modeling sequence determinants of guide performance

We estimated the performance of guides by the number of sequencing reads that aligned to the predicted cut site. Namely, a read was assigned to a guide if the first base of the read was within the 16th to 26th position downstream of a guide's PAM. An additional pseudocount read was added to all guide counts, enabling log transformation of all read counts, which we used as the dependent variable. We then collected 667 sequence-based features as in previous work[173]. Four bases upstream of the PAM and six bases downstream of the protospacer were considered. Position-specific nucleotides and dinucleotides were included (excluding the first three positions of the PAM, which are fixed as "T"), as well as two features relating to GC

content: the GC imbalance of the protospacer (i.e. how far the actual GC content was from 50%), and the GC content of the predicted overhang (positions 26-30). Additionally, we included the estimated minimum free energy of the RNA molecule[174].

Feature selection was done with the elastic net procedure, implemented in scikit-learn version 0.19.0. We found optimal hyperparameters with cross validation (ElasticNetCV) on 90% of the data (6,447 guides). This procedure resulted in 287 features with non-zero coefficients. To further eliminate inconsequential features, we trained ordinary least squares linear regression models with increasing numbers of features (rank ordered by elastic net coefficient absolute value) and made predictions on the 10% (729) fully withheld guides. Prediction performance did not substantially improve once the top ~100 features were added (Supplementary Fig. 2). Therefore, we fit a final ordinary least squares linear regression model to all available data (training and test), with the 100 selected features, which we then used to make predictions for the optimized guide set.

Design of optimized guide set

We used the same procedure as for the pilot guide set for obtaining padded genomic coordinates, identifying all possible Cas12a target sites, and excluded potential guides with copy number >1 and overlapping SNPs >0.001 allele frequency (dbSNP build 153). We used a restricted list of 34 genes, representing high confidence JS risk genes. We also implemented a more sophisticated procedure for picking guides. First, we designed two guide sets, one targeting the forward genomic strand and one targeting the reverse genomic strand, such that consecutive guides

alternated orientation. After picking a guide, we defined the next target as 250 basepairs downstream of the predicted cut site. We established a set of criteria, prioritizing high-scoring guides, guides most proximal to the target, and guides with a low number of predicted off target sites. Predicted off target sites for each guide were found by enumerating all possible single nucleotide deletions from the guide sequence and finding perfect matches for these in the genome. If there were no guides of the correct orientation fulfilling the criteria and within 250 basepairs of the target, we broadened our search to guides in the opposite orientation. If there were still no suitable guides, we moved on without choosing a guide. Once this process had been completed for all genes, we identified all "gaps" (i.e. no guides present) of greater than 600 basepairs. We reasoned that flanking the gaps with guides in the optimal orientation (i.e. forward guides upstream and reverse guides downstream of the gap) would maximize our ability to obtain coverage in the gap regions. So, if correctly oriented guides were present within 100 basepairs of the gap, regardless of predicted performance, we additionally picked those guides. We picked a total of 11,438 guides for the optimized set, and guides were synthesized as two oPools at the picomole per oligo scale (Integrated DNA Technologies, Supplementary Table 7) .

Variant Calling

Base quality scores were recalibrated with GATK (version 4.1.2.0, then variants were called with HaplotypeCaller with a minimum base quality score of 20 (-mbq 20). Sample VCFs were compared to the "Platinum" variant calls[20] with hap.py.[21] Single nucleotide variants were considered separately from insertions or deletions.

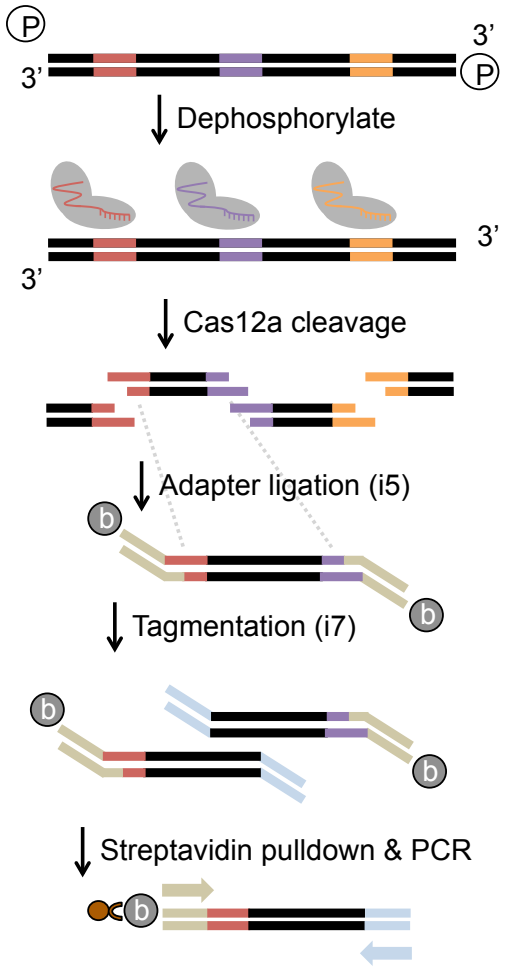


Figure 1. Overview of CRISPR-Capture

Genomic DNA is dephosphorylated and then treated with Cas12a as well as a pool of gRNAs that cleave target sites to leave overhangs. A custom biotinylated adapter with degenerate overhangs is ligated to the cleaved molecules, then the other adapter is added with Tn5 tagmentation. Finally, a streptavidin pulldown isolates library molecules which are amplified by on-bead PCR.

2.3 Results

Overview of the CRISPR-Capture method

The key intuition of the approach is that Cas12a-mediated genomic fragmentation, mediated by a pool of targeted gRNAs, should result in enrichment of ligatable overhanging ends at targeted loci. The approach is made ultra-low cost by synthesizing pools of DNA oligonucleotides containing the gRNA sequence as well as the T7 RNA polymerase priming site, and using *in vitro* transcription to generate pools of gRNAs (Fig. 1A). In order to reduce spurious ligation events, genomic DNA is enzymatically dephosphorylated prior to incubation with the Cas12a-gRNA RNP (Fig. 1B). Cas12a cleavage results in a 5' overhang of four to five nucleotides. Therefore, we designed custom, biotinylated adapters containing the Illumina i5 flow cell and priming sequences, as well as overhangs of four or five degenerate nucleotides (Supplementary Table 1). Following ligation of the i5 adapter, tagmentation with Tn5 transposase adds the i7 sequencing adapter (Fig. 1C). Finally, to enrich for molecules with a ligated i5 adapter (and

deplete molecules with two i7 adapters), a streptavidin-mediated pulldown is performed, followed by PCR directly on the streptavidin beads (Fig. 1D, full protocol can be found in Supplementary Note).

Design and performance of the pilot guide set

To validate the method, a pilot set of guides was designed targeting 47 known and candidate genes associated with Joubert Syndrome (JS), representing 3.5 megabases of DNA. The design tiled 7,176 guides at 500 base-pair intervals and did not use any sequence-based design criteria for picking guides (besides the presence of a "TTTN" protospacer adjacent motif (PAM)). DNA oligonucleotides encoding the T7 RNA polymerase promoter, *Acidaminococcus sp.BV3L6* (*As*) Cas12a constant loop region, and target specific protospacer region were produced with array-based synthesis. Subsequent *in vitro* transcription produced mature gRNAs. Paired-end sequencing of CRISPR-Capture libraries prepared from the well-studied NA12878 reference genome resulted in 5.9% of reads on target, corresponding to a 52.4-fold enrichment. While this enrichment was encouraging, we sought to understand the source of off-target reads. As the primary error modality of array synthesis is single base deletions, we generated a predicted off target list by aggregating all sites in the genome at which gRNAs with a single base deletion aligned (495,299 sites). We observed 12.7% of sequencing reads aligned to these predicted off target sites, which is substantially more than aligned to the same number of size-matched random genomic intervals (1.75%). Since Cas12a cleavage results in symmetrical 5' overhangs, we expected that approximately equal numbers of reads would result from ligation to both overhangs. However, this was not the case: 56% of guides had

greater than 10 times more reads aligning to the enzyme-distal overhang (Supplementary Fig. 1). This bias may be due to Cas12a remaining bound to the enzyme proximal fragment[175] and sterically inhibiting ligation, though treatment with SDS after cleavage did not reduce the bias.

Inspection of the read alignments revealed accumulation of the first read at programmed cut sites, consistent with ligation of the i5 adapter directly to the cut site overhang. In contrast, the second read was scattered across the inter-guide interval, consistent with this adapter being appended by semi-random tagmentation (Supplementary Fig. 1). We reasoned that we should be able to determine which guide led to any given sequencing read. In fact, we found that the first read of 92.6% of on-target read pairs began within 5 bases of a predicted guide cut site. Additionally, we observed that the starting position of the first read corresponds to the expected cut sites of Cas12a (i.e. after the 18^{th} and 23^{rd} bases downstream of the PAM, Supplementary Fig. 1). We used the number of reads assigned to each guide as a proxy for the performance of that guide. Comparing the performance of capture across the full guide set revealed a thousand-fold difference between the best and worst performing guides; however, 49.3% of guides performed within one $\log_{10}$ difference (Fig. 2a).

Modeling CRISPR-Capture performance

We reasoned that we should be able to use the pilot data as a training set to model the sequence determinants of CRISPR-Capture performance. Toward this end, we collated 667 sequence-based features, representing position-specific nucleotides

and dinucleotides, GC content, and gRNA folding. We modeled CRISPR-Capture performance using linear regression and implemented elastic net regularization to assign hyperparameters and feature coefficients[173]. Hyperparameters were chosen

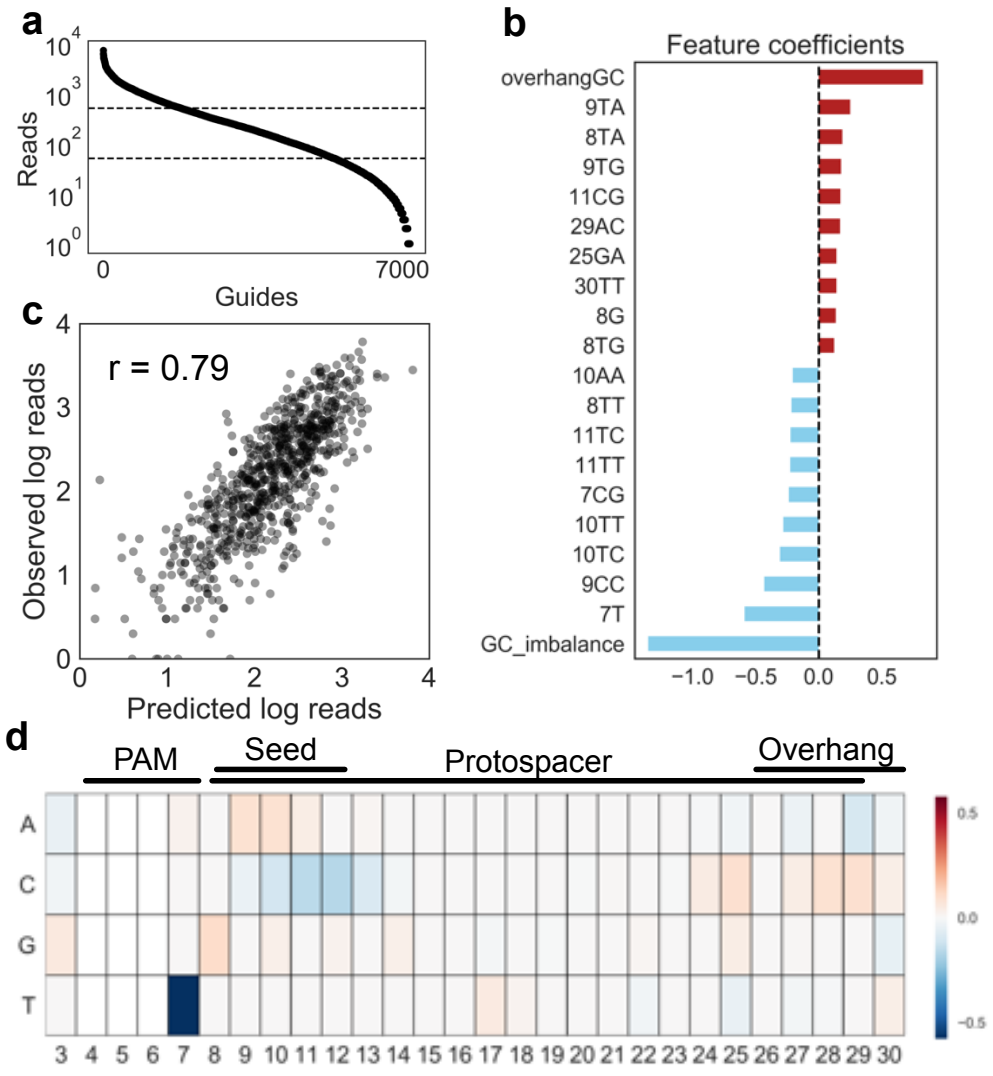


Figure 2. Modeling the sequence determinants of CRISPR-Capture performance

(a) Read uniformity for guides in the pilot experiment. Dashed lines indicate a $\log_{10}$ window within which 49.3% of guides performed.
(b) The twenty features in the linear regression model with the largest positive and negative coefficients.
(c) Performance of the linear regression model on fully withheld test data.
(d) Feature coefficients of individual position-specific nucleotides.

with nested cross-validation, and we tested the resulting model on fully withheld data. The predicted and observed scores were highly correlated (Pearson r = 0.79, Figure 2c). Overall, 287 features were assigned non-zero coefficients, and based on a plateau in predictive performance we used the top 100 (Supplementary Fig. 2). Consistent with previous work, a "T" at the fourth position of the PAM is strongly disfavored. Other important features are related to GC content, with GC imbalance being strongly disfavored and GC content at the overhang positively related to performance, likely due to increased ligation efficiency (Fig. 2b). Inspection of contributions from single position-specific nucleotides suggests the most important positions are within the seed and overhang regions (Fig. 2d).

Design and testing of optimized guide set

We used a combination of predicted performance scores from our model, optimal spacing, and number of predicted off target cut sites to generate optimized guide sets for 34 high confidence JS risk genes using a higher fidelity column-based synthesis platform (Materials & Method, Supplementary Table 6). Due to the observed biased capture efficiency from enzyme distal versus proximal fragments, we designed two interleaved pools, one targeting the forward genomic strand and the other targeting the reverse genomic strand (Supplementary Table 7). Following an initial guide selection step, we identified gaps (>600 basepairs between subsequent guides) and, when possible, picked additional guides flanking the gaps in an attempt to capture the gap sequence.

Captures with the optimized guides achieved an average enrichment of 64-fold (6.3% of reads on target) using NA12878 genomic DNA. Guide uniformity improved

modestly compared to the naïve guide set (54.0% of guides within one $\log_{10}$ difference, Fig. 3a). While cutting at predicted off target sites was present, it made up a relatively smaller fraction of reads compared to the pilot guides (5.5% of reads). Observed guide performance correlated with predictions (Pearson r = 0.38, Supplementary Fig. 3), but this correlation was lower than the cross-validation results. This is likely due to the optimized guides falling within a narrower range of expected performance compared to the cross-validation. For example, when only considering guides above the 2.0 threshold used for picking optimized guides, the cross-validation results in reduced correlation (Pearson r = 0.61). Additionally, the pilot guides were subjected to PCR amplification and restriction enzyme digestion steps prior to

in vitro transcription while the optimized guides were not. These additional steps could introduce biases that are not present for the optimized guide set.

We examined raw coverage of the target region at different levels of downsampling. With 20 million read pairs, 84.4% of bases in the target region are covered by at least 10 reads, and increasing to 40 million read pairs covers 92.8% of

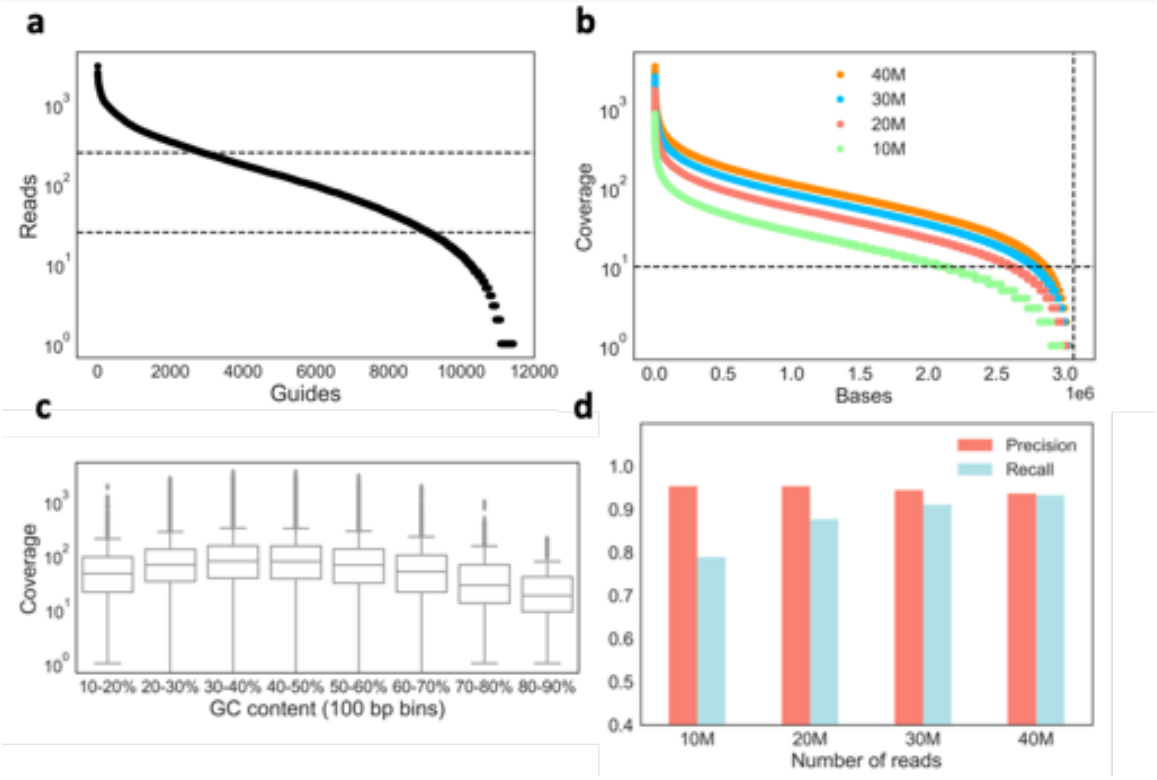


Figure 3. Performance of the optimized guide set

(a) Read uniformity for guides in the optimized experiment. Dashed lines indicate a log10 window within which 54.0% of guides performed.
(b) Per-base read coverage across the full target with downsampled datasets.
(c) Coverage of bases within different 100 basepair GC content bins.
(d) Precision and recall for single nucleotide variant calling of NA12878 compared to the "Platinum" variant calls.

bases by at least 10 reads (Fig. 3b). Considering only those bases outside of repetitive elements (as defined by Repeat Masker), 20 million read pairs cover 86.7% of bases with at least 10 reads, and at 40 million read pairs 94.6% of bases are covered by at least 10 reads (Supplementary Fig. 3). We next examined GC

content coverage bias. 100 basepair bins with extremely low (10-20%) or high (80-90%) GC content have median coverage of 46 and 18, respectively, while the 40-50% bin has median coverage of 78 (Fig. 3c). Finally, we performed single nucleotide variant calling with the downsampled datasets and found that with 20 million read pairs we achieve high precision and recall (0.95 and 0.88, respectively) compared to Illumina platinum calls for this sample. Increasing to 40 million read pairs maintains high precision and boosts recall (0.94 and 0.93, respectively, Fig. 3d). Restricting to bases not Repeat Masked yields slightly improved performance for all conditions (Supplementary Fig. 3).

2.4 Discussion

In this study, we demonstrate and evaluate CRISPR-Capture, a novel method for targeted enrichment of regions of interest for sequencing. This method takes advantage of the unique cleavage pattern of Cas12a to introduce ligatable overhangs specifically at regions of interest. In a pilot experiment we measured capture efficiency for 7,126 gRNAs, chosen without any sequence-specific criteria, and used these data to model the sequence determinants of capture efficiency with high accuracy.

Major strengths of the method are affordability and flexibility. Synthesis of gRNA-encoding DNA oligos is quick and affordable, and standard *in vitro* transcription kits can generate thousands of reactions worth of gRNA from picomolar scale DNA template, effectively making the oligos a one-time cost. For ongoing studies, gRNAs targeting different genes could easily be added to an existing pool. Further, there is no requirement for specialized equipment, and the protocol can be

completed in a single day. Though not implemented here, the method is compatible with liquid handling robots that could dramatically increase throughput. CRISPR-Capture is less susceptible to GC sequence content bias than hybridization-based approaches, and represents a substantial improvement in whole gene sequencing compared to existing technologies.

A drawback of the method is the dependence on a PAM site and a target sequence with high Cas12a cutting efficiency. Efforts are under way, though, to find Cas12a variants with more flexible PAM requirements,[176] increased cutting efficiency,[177] and reduced off-target cutting.[178] Additionally, while we achieve strong enrichment of targeted regions, the majority of sequencing reads originate from off-target loci. We attribute a substantial fraction of off-target reads to synthesis errors in the gRNA encoding DNA oligos. This off-target modality could be reduced by higher fidelity synthesis, or oligo purification schemes that eliminate deletion errors. An improved understanding of the origin of the rest of the off-target reads could lead to substantial improvement in the percent of on-target reads. Finally, we observe an imbalance in the capture efficiency for enzyme proximal versus enzyme distal cleavage product. It remains unclear what the source of this imbalance is, but a resolution of this issue could improve capture performance.

We foresee broad utility of the method in Mendelian disease genetics, where it can be valuable to sequence the full bodies of custom lists of genes. However, the method is compatible with any DNA genome, and could have utility in basic research, diagnostics, or agriculture.

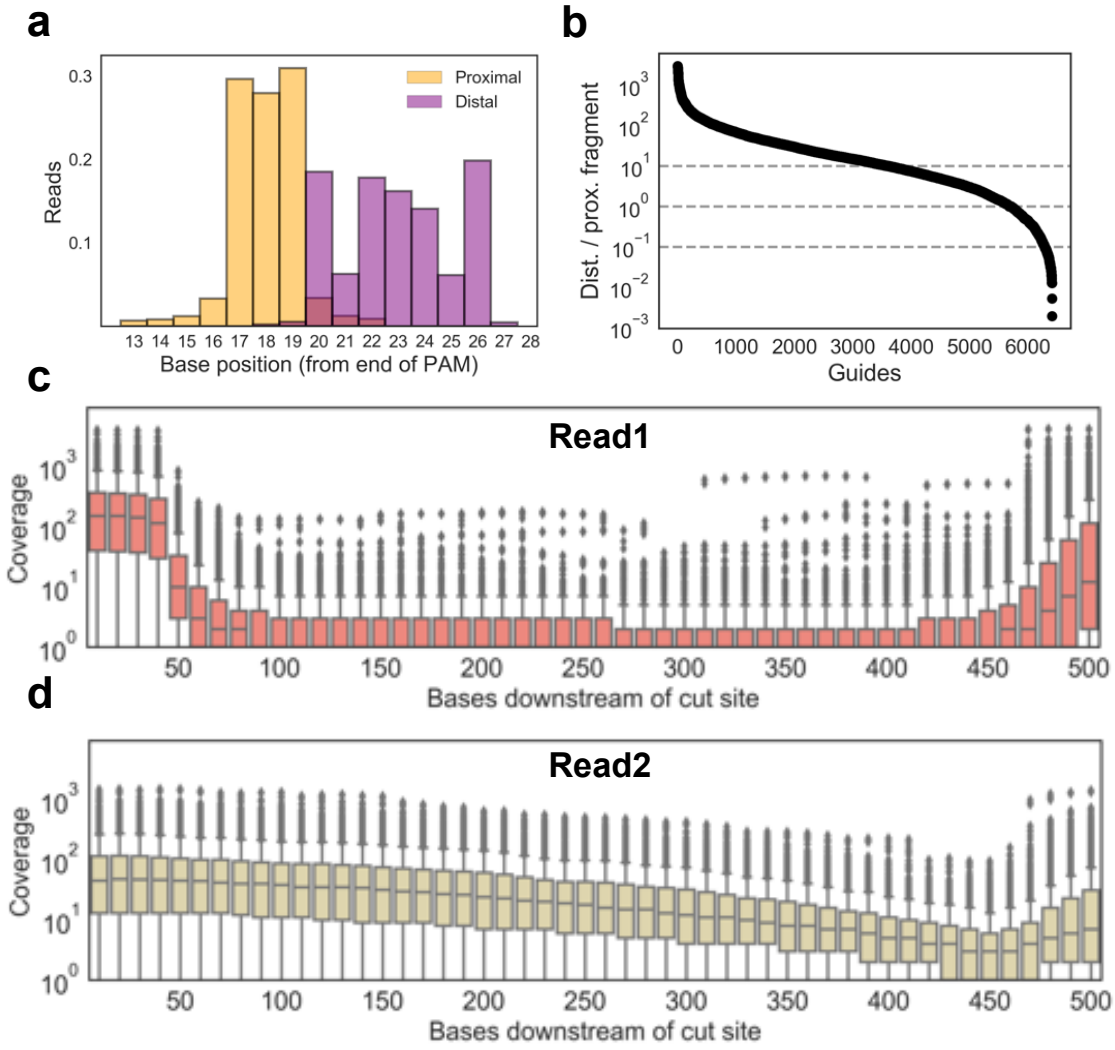


Figure S1. Performance characteristics of CRISPR-Capture

(a) Histogram of position of first base of read1 in relation to the end of PAM (i.e. the start of the protospacer). Reads originating from the Cas12a proximal and distal molecules are colored differently.
(b) Ratio of Cas12a distal to proximal reads for all guides, rank ordered by magnitude of ratio.
(c) Coverage of bases, from read1, as a function of distance downstream from nearest cut site.
(d) Coverage of bases, from read2, as a function of distance downstream from nearest cut site.

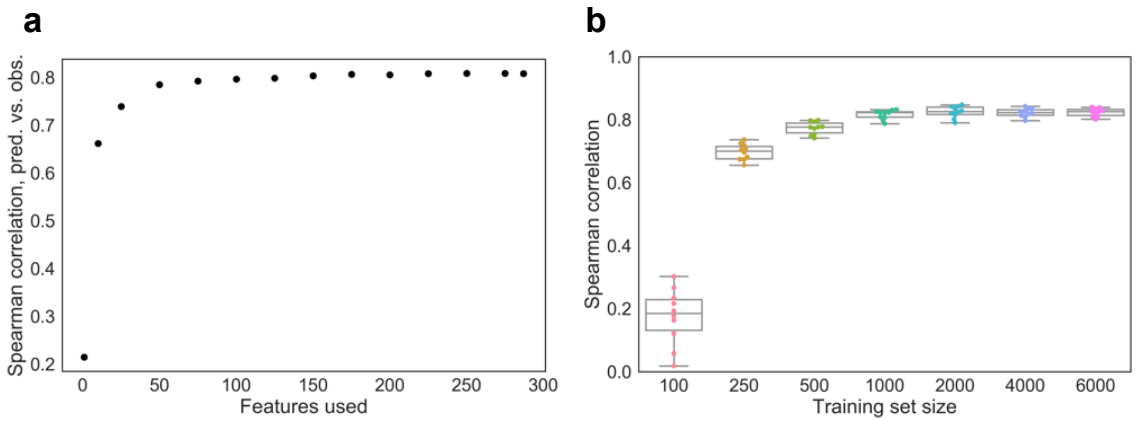


Figure S2. Modeling sequence determinants of CRISPR-Capture performance

(a) Models were iteratively trained with more features, successively adding features with the highest absolute value coefficient.
(b) Models were trained with varying training set sizes.

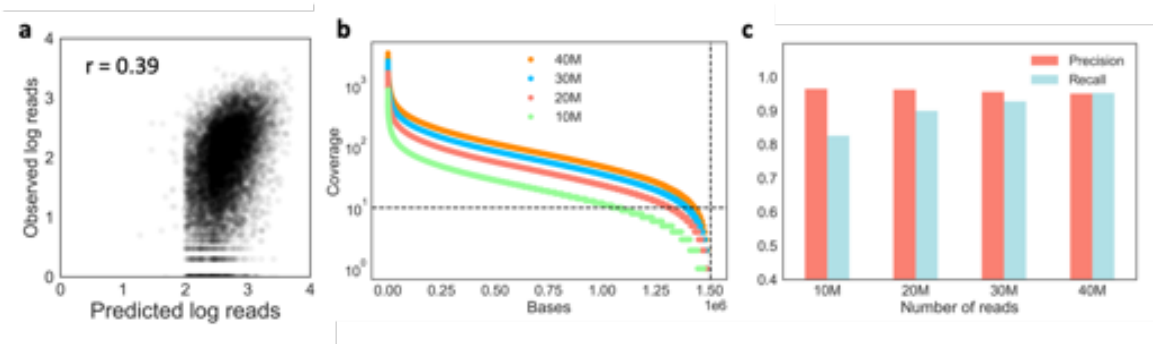


Figure S3. Performance of optimized guide set

(a) Predicted versus observed performance (as defined by assigned reads) for the optimized guide set. Pearson r = 0.39.
(b) Coverage uniformity for all bases outside of repeats (as defined by Repeat Masker) for various downsampled datasets.
(c) Precision and recall for single nucleotide variants at different downsampled read pairs.

Supplementary Note: CRISPR-Capture Protocol

CRISPR-Capture Protocol

Reagents

- Shrimp alkaline phosphatase (rSAP, New England BioLabs, Cat. M0371)
- Phosphate buffered saline (PBS, Thermo Fisher, Cat. 10010023)
- 10x Cas9 reaction buffer
- Alt-R A.s. Cas12a (Cpf1) V3 (IDT, Cat. 1081068)
- NucleoSpin Gel and PCR Clean-Up (Takara, Cat. 740609)
- Custom i5 adapter (Supplemental Table X)
- T4 DNA Ligase (New England BioLabs, Cat. M2622)
- T4 DNA Ligase Buffer (New England BioLabs, Cat. B0202)
- TAPS (Sigma Aldrich, Cat. T5130)
- Potassium Acetate (Sigma Aldrich, Cat. P1190)
- Magnesium Acetate (Sigma Aldrich, Cat. M5661)
- DMF (Sigma Aldrich, Cat. D4551)
- Loaded Tn5 transposase (Picelli et al., 2014. Genome Research)
- Sodium Dodecyl Sulfate (SDS, Sigma Aldrich, Cat. L3771)
- Dynabeads MyOne Streptavidin C1 (Thermo Fisher, Cat. 65002)
- NaCl (Fisher, Cat. M-11624)
- Tris (Fisher, Cat. T1503)
- LiCl (Sigma Aldrich, Cat. L9650)
- EDTA (Sigma Aldrich, Cat. E9884)
- Tween-20 (Sigma Aldrich, Cat. P1379)
- KAPA HiFi Hotstart ReadyMix (Roche, Cat. KK2602)
- Nextera i7 indexed primers (Supplemental Table X)
- Sera-Mag Select SPRI beads (GE Healthcare, 29343045)

Equipment

- Magnetic tube rack
- DNA Engine Tetrad Thermal Cycler (BioRad, or other thermal cycler)
- Thermomixer (Thermo Fisher, Cat. 5382000015, or other heat block)
- Illumina sequencing instrument

Protocol

1. **Dephosphorylate genomic DNA**
    a. Prepare 10x Cas9 reaction buffer (can be done beforehand)
        i. 200 mM HEPES, 1μM NaCl, 50 mM $MgCl_2$, 1 mM EDTA, pH 6.5 @ 25°
    b. Combine 100 ng genomic DNA (quantified with Qubit fluorometer) with 2 μL of 10x Cas9 buffer, 2 μL rSAP, and water to total volume of 20 μL.
    c. Incubate at 37° for 30 minutes.
    d. Incubate at 65° for 5 minutes.

2. **CRISPR cleavage of genomic DNA**
    a. Dilute Cas12a to 1 μM with PBS.

b. Combine 7.5 µL water, 1.5 µL of 10x Cas9 reaction buffer, 2 µL of 10 µM gRNA, and 4 µL of 1 µM Cas12a.
c. Incubate at room temperature for 10 minutes.
d. Combine the gRNA / Cas12a mixture with the reaction from step 1.
e. Incubate at 37° for 30 minutes.
f. Incubate at 65° for 10 minutes.
g. Purify DNA with Nucleospin columns; elute in 20 uL of buffer NE.

3. Ligate i5 adapter

a. Anneal adapter oligos (can be done beforehand and frozen).
i. Combine 10 µL i5_adapter_top, 5 µL i5_adapter_bottom_4N, 5 uL i5_adapter_bottom_5N, and 80 uL TE.
ii. Heat to 95° in thermal cycler, then, cool at a rate of 0.1°/second until reaching 10°.
b. To the eluate from step 2, add 0.5 µL water, 2.5 µL T4 DNA Ligase Buffer, 1 uL T4 DNA Ligase, and 1 µL of 10 µM annealed i5 adapter.
c. Incubate at 25° for 30 minutes.
d. Incubate at 65° for 10 minutes.

4. Tagment DNA

a. Make 1 mL fresh 4x TAPS buffer: 132 µL of 1M TAPS, 52.8 µL of 5M potassium acetate, 40 µL of 1M magnesium acetate, 640 µL of 100% DMF, 135.2 µL water
b. To the reaction from the previous step, add 12.5 µL 4x TAPS buffer, 11.5 µL water, and 1 µL 8 uM loaded indexed Tn5 transposase.
c. Incubate at 55° for 5 minutes.
d. Transfer to ice.
e. Add 5 µL of 2% SDS.
f. Incubate at room temperature for 5 minutes

5. Streptavidin magnetic bead pulldown

a. Prepare following buffers (can be done beforehand):
i. LWB: 10 mM Tris-Cl pH 8.0, 1M LiCl, 1mM EDTA, 0.05% Tween-20, in water.
ii. NWB: 10 mM Tris-Cl pH 8.0, 1M NaCl, 1mM EDTA, 0.05% Tween-20, in water.
iii. TWB: 10 mM Tris-Cl pH 8.0, 0.5mM EDTA, 0.05% Tween-20, in water.
iv. 2x NTB: 10 mM Tris-Cl pH 8.0, 2M NaCl, 1mM EDTA, in water.
b. Warm Dynabeads MyOne Streptavidin C1 beads to room temperature for 30 minutes.
c. For each sample, transfer 5 µL beads to PCR tube in a magnetic rack.
d. Concentrate beads (until supernatant is clear) and remove supernatant.
e. Wash with 200 µL TWB, concentrate, remove supernatant.
f. Resuspend beads in 110 µL 2x NTB.
g. Add samples to beads; shake for 30 minutes at 1,000 rpm at room temp.
h. Wash 1x with 200 µL LWB. Concentrate and remove supernatant.
i. Wash 2x with 200 µL NWB. Concentrate and remove supernatant.
j. Wash 2x with 200 µL TWB. Concentrate and remove supernatant.

k. Resuspend in PCR mix: 12.5 µL KAPA HiFi HotStart ReadyMix, 0.5 µM custom i5 primer, 0.5 µM indexed Nextera i7 primer, water to 25 µL. Ensure that beads are dispersed (i.e. have not settled to the bottom of tube).

6. **On-bead PCR and cleanup**
    - **a.** Thermal cycle: 72° for 3 minutes, 95° for 30 seconds, repeat 15 total times: 98° for 20 seconds, 60° for 15 seconds, 72° for 40 seconds.
    - **b.** Concentrate Dynabeads and transfer supernatant to a new tube.
    - **c.** Cleanup PCR with Sera-Mag Select SPRI beads, at a 0.8x beads to sample ratio. Check size distribution with preferred method.
7. **Perform paired end sequencing on an Illumina instrument**
    a. Read lengths should be:
        i. Read1: 35 cycles
        ii. Index1: 10 cycles
        iii. Index2: 6 cycles
        iv. Read2: 117 cycles

Supplementary Table 1. Primers and oligos

Oligo name	Oligo sequence
i5_adapter_top	/5'biotin/AATGATACGGCGACCACCGAGATCTACACNNNNDDACACTCTTTCCCTACACGACGCTCTTCCGATCT
i5_adapter_bottom_4N	NNNNAGATCGGAAGAGCG
i5_adapter_bottom_5N	NNNNNAGATCGGAAGAGCG
i5_enrichment_primer	AATGATACGGCGACCACCGA
i7_enrichment_primer (Nextera)	CAAGCAGAAGACGGCATACGAGAT[10bp_index]GTCTCGTGGGCTCGGAGATG

Chapter 5. Summary and Conclusions

5.1 A paradigm shift for PTEN clinical genetics

5.1.1 Defining pathogenic and benign PTEN alleles

Prior to the work outlined in this dissertation, there was no accurate method for discriminating pathogenic from benign *PTEN* variation. Based on the data presented in Chapters 2 and 3 of this work, I have demonstrated that the lipid phosphatase fitness scores can discriminate pathogenic variation with high accuracy. The strong performance of the fitness scores on two distinct set of alleles (ClinVar pathogenic alleles in Chapter 2, Cleveland Clinic cohort variant carriers in Chapter 3) strongly supports their broad utility. Further, the ClinGen PTEN Expert Panel has recommended the use of our data for clinical decision making[179].

5.1.2 Different PTEN variant classes confer different cancer risk

Similar to the pathogenic vs. benign question, prior to the work presented here, there was no accurate way to predict risk of neoplasia based on *PTEN* genotype. In fact, many clinicians treated all *PTEN* variation the same. The data presented here is the first to demonstrate that different variant classes confer significantly different risk of both early-onset and life-time risk of cancer. This information will be useful for clinicians as they counsel patients, and it will improve our ability to allocate and steward resources. Additionally, it will provide patients with a clearer picture of what their genotype means. It is reasonable to expect that increased numbers of patients will help clarify and estimate differential cancer risk for carriers of different variants.

5.1.3 Refinement of PTEN genotype-phenotype relationships

While the lipid phosphatase fitness scores can accurately distinguish pathogenic variation, they cannot distinguish variation that leads to ASD vs. PHTS. The data presented here does, though, validate the pre-existing hypothesis that ASD variants are, on average, less damaging than PHTS variants. However, in Chapter 3 I add significant nuance to this model by showing that, in fact, any *PTEN* variation that compromises lipid phosphatase fitness or cellular abundance increases the odds of developing ASD to a similar extent. This is in marked contrast to PHTS, for which the risk of developing symptoms is strongly related to the severity of the variant. This risk pattern suggests that there is a low threshold of activity compromise that confers ASD risk, and severity beyond this threshold does not substantially increase the ASD risk. This reinterpretation clarifies several confusing observations. First, it helps explain why there is such extensive overlap in the distributions of ASD- and PHTS-associated fitness scores (Chapter 2: Figure 4, Chapter 3: Figure 5). According to the original model, less severe variants led to ASD while more severe led to PHTS. According to this model, one might expect individuals with more severe variants to have both ASD and PHTS phenotypes, but this is not always the case.

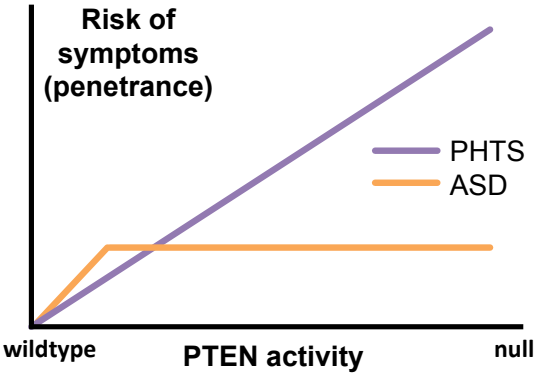


Figure 1. An updated model of PTEN genotype-phenotype relationships. The risk of PHTS increases with decreasing PTEN activity levels. In contrast, there is a threshold which increases risk for ASD, and further decrements in activity do not further increase risk.

The reinterpretation outlined here helps to explain why many PHTS positive individuals are not affected with ASD, by showing that the risk for ASD seems to be somewhat independent from the risk for PHTS. Since a relatively small subset of *PTEN* variant carriers develop ASD, it appears to be the case that the *PTEN* variation sensitizes to ASD, and then some other factor is required to cause symptoms. An important next step will be determining the other factors involved, which may involve PTEN functions not captured in the lipid phosphatase and cellular abundance assays, genetic background, or environmental exposures.

5.1.4 Outlook for deep mutational scanning

The outlook for deep mutational scanning as an approach for interpreting clinical variation is bright. As of 2018, over 200,000 variants in protein-coding or regulatory regions had been assayed[180], and the number continues to grow. Speaking to the promise of this approach, a new center has been established (the Center for the Multiplex Assessment of Phenotype, based at the University of Washington and the University of Toronto) that seeks to define the functional effect of all variants in the human genome. Due to the growth of the community, efforts are being made to harmonize and standardize the design and reporting of deep mutational scanning data[181]. Further, clinicians are weighing in regarding the most informative and responsible use of deep mutational scanning in the clinic[182]. Continued success of the method will depend on freely sharing methods, software code, and data.

5.1.5 A desperate need for improved clinical genetic databases

Deep mutational scanning represents an extremely powerful method for profiling the functional impacts of protein coding variation. However, well annotated, standardized, and publicly available data is an absolute necessity in order to maximize the utility of this data. We were fortunate that the Cleveland Clinic had ascertained a large and well-curated cohort of *PTEN* variant carriers. However, this luxury will not be available for all genes or disorders. In most cases, researchers will have to rely on ClinVar[136] to identify clinically relevant variation. Unfortunately, this database is replete with inconsistent, vague, or conflicting clinical descriptions. Efforts like the UK Biobank, in which genomic, deep phenotypic, and lifestyle data is collected prospectively, will be invaluable resources moving forward.

5.2 Programmable, whole-gene sequencing

5.2.1 CRISPR-Capture is a novel method for programmable, whole-gene sequencing

Due to the length of human genes, which is often greater than 10 kilobases, existing technologies are not well suited to sequence full genes. A widely used technology relies on hybridization of biotinylated probes to sequences of interest. While it is technically possible to capture whole genes with this method, the cost of generating the number of biotinylated probes necessary is prohibitive. CRISPR-Capture enables affordable whole-gene sequencing by leveraging the CRISPR-Cas12a system. In particular, the method is made ultra-low-cost by synthesizing DNA oligos and then using *in vitro* transcription to convert these DNAs into guide RNAs. This

process enables the synthesis of microgram quantities of RNAs, which corresponds to hundreds of CRISPR-Capture reactions.

5.2.2 CRISPR-Capture applications and use cases

Due to the modularity of CRISPR technology, the same approach described here could be modified for use on any animal, plant, fungus, or other DNA genome. It could be valuable, for instance, to profile certain gene families important in crop abundance. Also due to the modularity of the approach, individual, optimized guide sets could be developed and commercialized for genes. This could allow researchers to assemble their own targeted enrichment panel very quickly and easily.

5.2.3 Limitations of CRISPR-Capture

One limitation of CRISPR-Capture, as described here, is that current versions of the Cas12a enzyme are limited to targets with a TTTV protospacer adjacent motif. Therefore, certain regions of the genome will be challenging to target, particularly regions depleted of T's. However, efforts are underway to isolate Cas12a sequences from different strains, which may have different PAM requirements[183]. Also, efforts are underway to generate synthetic variants of Cas12a that have relaxed PAM requirements[177].

While CRISPR-Capture achieves enrichment of targeted sequence on the order of 50-fold, there is still room for improvement. Especially for panels targeting smaller sequence space, there is substantial room for improvement. There are multiple potential contributors to the limited enrichment. First, as demonstrated, a substantial fraction of off target reads result from synthesis-error related cutting. As DNA

synthesis quality improves, this off target contribution could be mitigated. Another potential contributor to lack of enrichment is the fact that, for most guides, ligation of sequencing adapter seems to only occur on the enzyme-distal fragment (See Chapter 4, Supplemental Figure 2). This may be due to Cas12a remaining bound to the enzyme proximal side[175] and sterically inhibiting adapter ligation. If this were truly occurring, then developing a method to dissociate the enzyme from the target DNA could improve enrichment. Additionally, multiple studies have demonstrated that Cas12a exhibits indiscriminate exonuclease activity towards *trans* DNA substrates while bound to gRNA and target DNA[184,185]. It is possible that upon binding and cleavage to target sites, Cas12a is actually introducing spurrious overhanging ends by degrading molecules of off target DNA in the reaction.

References

1. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* 531210 (2019). doi:10.1101/531210
3. Consortium, T. 1000 G. P. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
4. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–91 (2016).
5. Starita, L. M. *et al.* Variant Interpretation: Functional Assays to the Rescue. *Am. J. Hum. Genet.* **101**, 315–325 (2017).
6. Ramensky, V. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* **30**, 3894–3900 (2002).
7. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–4 (2003).
8. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
9. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
10. Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018).
11. 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).
12. Kato, S. *et al.* Understanding the function–structure and function– mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis.
13. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
14. Salinas, V. H. & Ranganathan, R. Coevolution-based inference of amino acid interactions underlying protein function. *Elife* **7**, (2018).
15. Araya, C. L. *et al.* A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Sci.* **109**, 16858–16863 (2012).
16. Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R. & Fields, S. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA* **19**, 1537–51 (2013).
17. Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J. & Fowler, D. M. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Syst.* **6**, 116–124 (2017).
18. Firnberg, E. & Ostermeier, M. PFunkel: Efficient, Expansive, User-Defined Mutagenesis. *PLoS One* **7**, e52031 (2012).

19. Melnikov, A., Rogov, P., Wang, L., Gnirke, A. & Mikkelsen, T. S. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Res.* **42**, e112 (2014).
20. Findlay, G. M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
21. Hiatt, J. B., Patwardhan, R. P., Turner, E. H., Lee, C. & Shendure, J. Parallel, tag-directed assembly of locally derived short sequence reads. *Nat. Methods* **7**, 119–122 (2010).
22. Starita, L. M. *et al.* Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* **200**, 413–422 (2015).
23. Majithia, A. R. *et al.* Prospective functional classification of all possible missense variants in PPARG. *Nat. Genet.* **48**, 1570–1575 (2016).
24. Weile, J. *et al.* A framework for exhaustively mapping functional missense variants. *Mol. Syst. Biol.* **13**, 957 (2017).
25. Steck, P. A. *et al.* Identification of a candidate tumour suppressor gene, MMAC1, at chromosome 10q23.3 that is mutated in multiple advanced cancers. *Nat. Genet.* **15**, 356–362 (1997).
26. Li, J. *et al.* PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* **275**, 1943–7 (1997).
27. Myers, M. P. *et al.* P-TEN, the tumor suppressor from human chromosome 10q23, is a dual-specificity phosphatase. *Proc. Natl. Acad. Sci.* **94**, 9052–9057 (1997).
28. Maehama, T. & Dixon, J. E. The tumor suppressor, PTEN/MMAC1, dephosphorylates the lipid second messenger, phosphatidylinositol 3,4,5-trisphosphate. *J. Biol. Chem.* **273**, 13375–8 (1998).
29. Myers, M. P. *et al.* The lipid phosphatase activity of PTEN is critical for its tumor supressor function. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 13513–8 (1998).
30. Stambolic, V. *et al.* Negative Regulation of PKB/Akt-Dependent Cell Survival by the Tumor Suppressor PTEN. *Cell* **95**, 29–39 (1998).
31. Engelman, J. A., Luo, J. & Cantley, L. C. The evolution of phosphatidylinositol 3-kinases as regulators of growth and metabolism. *Nat. Rev. Genet.* **7**, 606–619 (2006).
32. Sun, H. *et al.* PTEN modulates cell cycle progression and cell survival by regulating phosphatidylinositol 3,4,5,-trisphosphate and Akt/protein kinase B signaling pathway. *Proc. Natl. Acad. Sci.* **96**, 6199–6204 (1999).
33. Backman, S. A. *et al.* Deletion of Pten in mouse brain causes seizures, ataxia and defects in soma size resembling Lhermitte-Duclos disease. *Nat. Genet.* **29**, 396–403 (2001).
34. Groszer, M. *et al.* Negative regulation of neural stem/progenitor cell proliferation by the Pten tumor suppressor gene in vivo. *Science* **294**, 2186–9 (2001).
35. Crackower, M. A. *et al.* Regulation of Myocardial Contractility and Cell Size by Distinct PI3K-PTEN Signaling Pathways. *Cell* **110**, 737–749 (2002).
36. Alimonti, A. *et al.* Subtle variations in Pten dose determine cancer susceptibility. *Nat. Genet.* **42**, 454–8 (2010).
37. Berger, A. H., Knudson, A. G. & Pandolfi, P. P. A continuum model for tumour

suppression. *Nature* **476**, 163–169 (2011).
38. Lee, J. O. *et al.* Crystal structure of the PTEN tumor suppressor: implications for its phosphoinositide phosphatase activity and membrane association. *Cell* **99**, 323–34 (1999).
39. Liang, H. *et al.* PTENα, a PTEN Isoform Translated through Alternative Initiation, Regulates Mitochondrial Function and Energy Metabolism. *Cell Metab.* **19**, 836–848 (2014).
40. Liang, H. *et al.* PTENβ is an alternatively translated isoform of PTEN that regulates rDNA transcription. *Nat. Commun.* **8**, 14771 (2017).
41. Vazquez, F., Ramaswamy, S., Nakamura, N. & Sellers, W. R. Phosphorylation of the PTEN tail regulates protein stability and function. *Mol. Cell. Biol.* **20**, 5010–8 (2000).
42. Rahdar, M. *et al.* A phosphorylation-dependent intramolecular interaction regulates the membrane association and activity of the tumor suppressor PTEN. *Proc. Natl. Acad. Sci.* **106**, 480–485 (2009).
43. Torres, J. & Pulido, R. The Tumor Suppressor PTEN Is Phosphorylated by the Protein Kinase CK2 at Its C Terminus. *J. Biol. Chem.* **276**, 993–998 (2001).
44. Al-Khouri, A. M., Ma, Y., Togo, S. H., Williams, S. & Mustelin, T. Cooperative Phosphorylation of the Tumor Suppressor Phosphatase and Tensin Homologue (PTEN) by Casein Kinases and Glycogen Synthase Kinase 3β. *J. Biol. Chem.* **280**, 35195–35202 (2005).
45. Tibarewal, P. *et al.* PTEN Protein Phosphatase Activity Correlates with Control of Gene Expression and Invasion, a Tumor-Suppressing Phenotype, But Not with AKT Activity. *Sci. Signal.* **5**, ra18–ra18 (2012).
46. Sotelo, N. S., Schepens, J. T. G., Valiente, M. & Hendriks, W. J. A. J. PTEN–PDZ domain interactions: Binding of PTEN to PDZ domains of PTPN13. *Methods* **77–78**, 147–156 (2015).
47. Vazquez, F. *et al.* Tumor suppressor PTEN acts through dynamic interaction with the plasma membrane. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 3633 (2006).
48. Iijima, M., Huang, Y. E., Luo, H. R., Vazquez, F. & Devreotes, P. N. Novel Mechanism of PTEN Regulation by Its Phosphatidylinositol 4,5-Bisphosphate Binding Motif Is Critical for Chemotaxis. *J. Biol. Chem.* **279**, 16606–16613 (2004).
49. Bononi, A. *et al.* Identification of PTEN at the ER and MAMs and its regulation of Ca(2+) signaling and apoptosis in a protein phosphatase-dependent manner. *Cell Death Differ.* **20**, 1631–43 (2013).
50. Zhu, Y., Hoell, P., Ahlemeyer, B. & Krieglstein, J. PTEN: A crucial mediator of mitochondria-dependent apoptosis. *Apoptosis* **11**, 197–207 (2006).
51. Bassi, C. *et al.* Nuclear PTEN controls DNA repair and sensitivity to genotoxic stress. *Science* **341**, 395–9 (2013).
52. Li, P. *et al.* Identification of nucleolus-localized PTEN and its function in regulating ribosome biogenesis. *Mol. Biol. Rep.* **41**, 6383–6390 (2014).
53. Papa, A. *et al.* Cancer-Associated PTEN Mutants Act in a Dominant-Negative Manner to Suppress PTEN Protein Function. *Cell* **157**, 595–610 (2014).
54. Heinrich, F. *et al.* The PTEN Tumor Suppressor Forms Homodimers in Solution. *Structure* **23**, 1952–1957 (2015).

55. Wang, H. *et al.* Allele-specific tumor spectrum in pten knockin mice. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 5142–7 (2010).
56. Hopkins, B. D. *et al.* A Secreted PTEN Phosphatase That Enters Cells to Alter Signaling and Survival. *Science (80-.).* **341**, 399–402 (2013).
57. Gu, T. *et al.* CREB Is a Novel Nuclear Target of PTEN Phosphatase. *Cancer Res.* **71**, 2821–2825 (2011).
58. Shi, Y. *et al.* PTEN is a protein tyrosine phosphatase for IRS1. *Nat. Struct. Mol. Biol.* **21**, 522–527 (2014).
59. Tamura, M. *et al.* Inhibition of Cell Migration, Spreading, and Focal Adhesions by Tumor Suppressor PTEN. *Science (80-.).* **280**, 1614–1617 (1998).
60. Zhang, S. *et al.* Combating trastuzumab resistance by targeting SRC, a common node downstream of multiple resistance pathways. *Nat. Med.* **17**, 461–469 (2011).
61. Lee, Y.-R., Chen, M. & Pandolfi, P. P. The functions and regulation of the PTEN tumour suppressor: new modes and prospects. *Nat. Rev. Mol. Cell Biol.* **19**, 547–562 (2018).
62. Mingo, J. *et al.* A pathogenic role for germline PTEN variants which accumulate into the nucleus. *Eur. J. Hum. Genet.* **26**, 1180–1187 (2018).
63. Fricano-Kugler, C. J. *et al.* Nuclear-Excluded Autism-Associated PTEN Mutations Dysregulate Neuronal Growth. *Biol. Psychiatry* 32243–32246 (2017). doi:10.1016/j.biopsych.2017.11.025
64. Sun, Z. *et al.* PTEN C-Terminal Deletion Causes Genomic Instability and Tumor Development. *Cell Rep.* **6**, 844–854 (2014).
65. He, J., Kang, X., Yin, Y., Chao, K. S. C. & Shen, W. H. PTEN regulates DNA replication progression and stalled fork recovery. *Nat. Commun.* **6**, 7620 (2015).
66. Chen, Z. H. H. *et al.* PTEN interacts with histone H1 and controls chromatin condensation. *Cell Rep.* **8**, 2003–2014 (2014).
67. Gray, I. C. *et al.* Loss of the chromosomal region 10q23-25 in prostate cancer. *Cancer Res.* **55**, 4800–3 (1995).
68. Bigner, S. H. *et al.* Specific chromosomal abnormalities in malignant human gliomas. *Cancer Res.* **48**, 405–11 (1988).
69. Sansal, I. & Sellers, W. R. The biology and clinical relevance of the PTEN tumor suppressor pathway. *J. Clin. Oncol.* **22**, 2954–63 (2004).
70. Liaw, D. *et al.* Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. *Nat. Genet.* **16**, 64–67 (1997).
71. Lynch, E. D. *et al.* Inherited Mutations in PTEN That Are Associated with Breast Cancer, Cowden Disease, and Juvenile Polyposis. *Am. J. Hum. Genet.* **61**, 1254–1260 (1997).
72. Marsh, D. J. *et al.* PTEN Mutation Spectrum and Genotype-Phenotype Correlations in Bannayan-Riley-Ruvalcaba Syndrome Suggest a Single Entity With Cowden Syndrome. *Hum. Mol. Genet.* **8**, 1461–1472 (1999).
73. Butler, M. G. *et al.* Subset of individuals with autism spectrum disorders and extreme macrocephaly associated with germline PTEN tumour suppressor gene mutations. *J. Med. Genet.* **42**, 318–321 (2005).
74. Varga, E. A., Pastore, M., Prior, T., Herman, G. E. & McBride, K. L. The

prevalence of PTEN mutations in a clinical pediatric cohort with autism spectrum disorders, developmental delay, and macrocephaly. *Genet. Med.* **11**, 111–117 (2009).
75. Nieuwenhuis, M. H. *et al.* Cancer risk and genotype–phenotype correlations in PTEN hamartoma tumor syndrome. *Fam. Cancer* **13**, 57–63 (2014).
76. Marsh, D. *et al.* Mutation spectrum and genotype-phenotype analyses in Cowden disease and Bannayan-Zonana syndrome, two hamartoma syndromes with germline PTEN mutation. *Hum. Mol. Genet.* **7**, 507–515 (1998).
77. Lachlan, K. L., Lucassen, A. M., Bunyan, D. & Temple, I. K. Cowden syndrome and Bannayan Riley Ruvalcaba syndrome represent one condition with variable expression and age-related penetrance: results of a clinical study of PTEN mutation carriers. *J. Med. Genet.* **44**, 579–85 (2007).
78. Celebi, J. T. *et al.* Phenotypic findings of Cowden syndrome and Bannayan-Zonana syndrome in a family associated with a single germline mutation in PTEN. *J. Med. Genet.* **36**, 360–4 (1999).
79. Rodriguez-Escudero, I. *et al.* A comprehensive functional analysis of PTEN mutations: implications in tumor- and autism-related syndromes. *Hum. Mol. Genet.* **20**, 4132–4142 (2011).
80. Spinelli, L., Black, F. M., Berg, J. N., Eickholt, B. J. & Leslie, N. R. Functionally distinct groups of inherited PTEN mutations in autism and tumour syndromes. *J. Med. Genet.* **52**, 128–134 (2015).
81. 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).
82. Frazier, T. W. *et al.* Molecular and phenotypic abnormalities in individuals with germline heterozygous PTEN mutations and autism. *Mol. Psychiatry* **20**, 1132–1138 (2015).
83. Cristofano, A. D., Pesce, B., Cordon-Cardo, C. & Pandolfi, P. P. Pten is essential for embryonic development and tumour suppression. *Nat. Genet.* **19**, 348–355 (1998).
84. Podsypanina, K. *et al.* Mutation of Pten/Mmac1 in mice causes neoplasia in multiple organ systems. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 1563–8 (1999).
85. Kwon, C.-H. *et al.* Pten regulates neuronal soma size: a mouse model of Lhermitte-Duclos disease. *Nat. Genet.* **29**, 404–411 (2001).
86. Kwon, C.-H. *et al.* Pten Regulates Neuronal Arborization and Social Interaction in Mice. *Neuron* **50**, 377–388 (2006).
87. Luikart, B. W. *et al.* Pten Knockdown In Vivo Increases Excitatory Drive onto Dentate Granule Cells. *J. Neurosci.* **31**, 4345–4354 (2011).
88. Pun, R. Y. K. *et al.* Excessive Activation of mTOR in Postnatally Generated Granule Cells Is Sufficient to Cause Epilepsy. *Neuron* **75**, 1022–1034 (2012).
89. Vogt, D., Cho, K. K. A., Lee, A. T., Sohal, V. S. & Rubenstein, J. L. R. Parvalbumin/Somatostatin Ratio Is Increased in Pten mutant Mice and by Human PTEN ASD alleles. *Cell Rep.* **11**, 944–956 (2015).
90. Rodríguez-Escudero, I. *et al.* Reconstitution of the mammalian PI3K/PTEN/Akt pathway in yeast. *Biochem. J.* **390**, 613–23 (2005).

91. Gil, A. *et al.* A Functional Dissection of PTEN N-Terminus: Implications in PTEN Subcellular Targeting and Tumor Suppressor Activity. *PLoS One* **10**, e0119287 (2015).
92. Rodrı, I. *et al.* In vivo Functional Analysis of the Counterbalance of Hyperactive Phosphatidylinositol 3-Kinase p110 Catalytic Oncoproteins Amparo Andre Functional definition of relevant epitopes on the tumor suppressor PTEN protein Amparo Andre. *Cancer Res.* **67**, 9731–9739 (2007).
93. Oliver, M. D. *et al.* Insights into the pathological mechanisms of p85α mutations using a yeast-based phosphatidylinositol 3-kinase model. *Biosci. Rep.* **37**, (2017).
94. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).
95. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
96. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *American Journal of Human Genetics* **97**, 199–215 (2015).
97. Suwinski, P. *et al.* Advancing personalized medicine through the application of whole exome sequencing and big data analytics. *Frontiers in Genetics* **10**, 49 (2019).
98. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.* **97**, 199–215 (2015).
99. Eng, C. *PTEN Hamartoma Tumor Syndrome. 2001 Nov 29 [updated 2016 Jun 2]*. *GeneReviews(®)* (University of Washington, Seattle; 1993-2019., 2016).
100. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science (80-.).* **339**, 819–823 (2013).
101. Bennett-Baker, P. E. & Mueller, J. L. CRISPR-mediated isolation of specific megabase segments of genomic DNA. *Nucleic Acids Res.* **45**, e165–e165 (2017).
102. Nachmanson, D. *et al.* Targeted genome fragmentation with CRISPR/Cas9 enables fast and efficient enrichment of small genomic regions and ultra-accurate sequencing with low DNA input (CRISPR-DS). *Genome Res.* **28**, 1589–1599 (2018).
103. Gilpatrick, T. *et al.* Targeted Nanopore Sequencing with Cas9 for studies of methylation, structural variants, and mutations. *bioRxiv* 604173 (2019). doi:10.1101/604173
104. Xu, X. *et al.* CRISPR-assisted targeted enrichment-sequencing (CATE-seq). *bioRxiv* 672816 (2019). doi:10.1101/672816
105. Zetsche, B. *et al.* Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759–71 (2015).
106. Paul, B. & Montoya, G. CRISPR-Cas12a: Functional overview and applications. *Biomedical Journal* **43**, 8–17 (2020).
107. Lei, C. *et al.* The CCTL (Cpf1-assisted Cutting and Taq DNA ligase-assisted Ligation) method for efficient editing of large DNA constructs in vitro. *Nucleic Acids Res.* **45**, e74 (2017).
108. Sun, S. *et al.* An extended set of yeast-based functional assays accurately

identifies human disease mutations. *Genome Res.* **26**, 670–80 (2016).
109. Kato, S. *et al.* Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 8424–8429 (2003).
110. Brenan, L. *et al.* Phenotypic Characterization of a Comprehensive Set of MAPK1 /ERK2 Missense Mutants. *Cell Rep.* **17**, 1171–1183 (2016).
111. Worby, C. A. & Dixon, J. E. PTEN. *Annu. Rev. Biochem.* **83**, 641–669 (2014).
112. Song, M. S., Salmena, L. & Pandolfi, P. P. The functions and regulation of the PTEN tumour suppressor. *Nature Reviews Molecular Cell Biology* **13**, 283–296 (2012).
113. O'Roak, B. J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science (80-.).* **338**, 1619–1622 (2012).
114. Marsh, D. J. *et al.* Germline mutations in PTEN are present in Bannayan-Zonana syndrome. *Nat. Genet.* **16**, 333–334 (1997).
115. Padberg, G. W., Schot, J. D. L., Vielvoye, G. J., Bots, G. T. A. M. & De Beer, F. C. Lhermitte-duclos disease and cowden disease: A single phakomatosis. *Ann. Neurol.* **29**, 517–523 (1991).
116. Mester, J. L. *et al.* Analysis of prevalence and degree of macrocephaly in patients with germline PTEN mutations and of brain weight in Pten knock-in murine model. *Eur. J. Hum. Genet.* **19**, 763–8 (2011).
117. Eng, C. *PTEN Hamartoma Tumor Syndrome*. *GeneReviews*® (University of Washington, Seattle, 1993).
118. Buxbaum, J. D. *et al.* Mutation screening of thePTEN gene in patients with autism spectrum disorders and macrocephaly. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **144B**, 484–491 (2007).
119. McBride, K. L. *et al.* Confirmation study of PTEN mutations among individuals with autism or developmental delays/mental retardation and macrocephaly. *Autism Res.* **3**, 137–141 (2010).
120. C Yuen, R. K. *et al.* Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat. Neurosci.* **20**, 602–611 (2017).
121. O'Roak, B. J. *et al.* Recurrent de novo mutations implicate novel genes underlying simplex autism risk. *Nat. Commun.* **5**, 5595 (2014).
122. Leslie, N. R. & Longy, M. Inherited PTEN mutations and the prediction of phenotype. *Semin. Cell Dev. Biol.* **52**, 30–38 (2016).
123. Cid, V. J. *et al.* Assessment of PTEN tumor suppressor activity in nonmammalian models: the year of the yeast. *Oncogene* **27**, 5431–5442 (2008).
124. Zhang, Y., Werling, U. & Edelmann, W. SLiCE: a novel bacterial cell extract-based DNA cloning method. *Nucleic Acids Res.* **40**, e55 (2012).
125. Zhang, Y., Werling, U. & Edelmann, W. Seamless Ligation Cloning Extract (SLiCE) cloning method. *Methods Mol. Biol.* **1116**, 235–44 (2014).
126. Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2**, 31–34 (2007).
127. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate

Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).
128. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
129. Rubin, A. F. *et al.* A statistical framework for analyzing deep mutational scanning data. *Genome Biol.* **18**, 150 (2017).
130. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
131. Xiao, Y. *et al.* PTEN catalysis of phospholipid dephosphorylation reaction follows a two-step mechanism in which the conserved aspartate-92 does not function as the general acid — Mechanistic analysis of a familial Cowden disease-associated PTEN mutation. *Cell. Signal.* **19**, 1434–1445 (2007).
132. Das, S., Dixon, J. E. & Cho, W. Membrane-binding and activation mechanism of PTEN. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 7491–6 (2003).
133. Campbell, R. B., Liu, F. & Ross, A. H. Allosteric activation of PTEN phosphatase by phosphatidylinositol 4,5-bisphosphate. *J. Biol. Chem.* **278**, 33617–20 (2003).
134. Wei, Y., Stec, B., Redfield, A. G., Weerapana, E. & Roberts, M. F. Phospholipid Binding Sites of PTEN: Exploring the Mechanism of PIP 2 Activation. *J. Biol. Chem.* (2014). doi:10.1074/jbc.M114.588590
135. Gray, V. E., Hause, R. J. & Fowler, D. M. Analysis of Large-Scale Mutagenesis Data To Assess the Impact of Single Amino Acid Substitutions. *Genetics* **207**, 53–61 (2017).
136. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).
137. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis. Oncol.* 1–16 (2017). doi:10.1200/PO.17.00011
138. Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nat. Methods* **12**, 203–206 (2015).
139. Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C. & Shendure, J. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* **513**, 120–123 (2014).
140. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7**, 741–6 (2010).
141. Matreyek, K. A. *et al.* Multiplex Assessment of Protein Variant Abundance by Massively Parallel Sequencing. *bioRxiv* **50**, 211011 (2018).
142. Rodríguez-Escudero, I., Andrés-Pons, A., Pulido, R., Molina, M. & Cid, V. J. Phosphatidylinositol 3-kinase-dependent activation of mammalian protein kinase B/Akt in Saccharomyces cerevisiae, an in vivo model for the functional study of Akt mutations. *J. Biol. Chem.* **284**, 13373–83 (2009).
143. Fernández-Acero, T. *et al.* A Yeast-Based In Vivo Bioassay to Screen for Class I Phosphatidylinositol 3-Kinase Specific Inhibitors. *J. Biomol. Screen.* **17**, 1018–1029 (2012).
144. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Biochemistry* **89**, 10915–10919 (1992).
145. Redfern, R. E. *et al.* A mutant form of PTEN linked to autism. *Protein Sci.* **19**, 1948–1956 (2010).

146. Yang, J.-M. *et al.* Characterization of PTEN mutations in brain cancer reveals that pten mono-ubiquitination promotes protein stability and nuclear localization. *Oncogene* **36**, 3673–3685 (2017).
147. Chia, J. Y.-C., Gajewski, J. E., Xiao, Y., Zhu, H.-J. & Cheng, H.-C. Unique biochemical properties of the protein tyrosine phosphatase activity of PTEN—Demonstration of different active site structural requirements for phosphopeptide and phospholipid phosphatase activities of PTEN. *Biochim. Biophys. Acta - Proteins Proteomics* **1804**, 1785–1795 (2010).
148. Krumm, N. *et al.* Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* **47**, 582–588 (2015).
149. Geiger, T. & Clarke, S. Deamidation, isomerization, and racemization at asparaginyl and aspartyl residues in peptides. Succinimide-linked reactions that contribute to protein degradation. *J. Biol. Chem.* **262**, 785–94 (1987).
150. Shendure, J. & Fields, S. Massively Parallel Genetics. *Genetics* **203**, 617–619 (2016).
151. Tan, M.-H. *et al.* Lifetime cancer risks in individuals with germline PTEN mutations. *Clin. Cancer Res.* **18**, 400–7 (2012).
152. Feliciano, P. *et al.* SPARK (Simons Foundation Powering Autism Research for Knowledge): a US cohort of 50,000 families to accelerate autism research. *Neuron* **97**, 488–493 (2018).
153. Yehia, L. & Eng, C. One gene, many endocrine and metabolic syndromes: PTEN-opathies and precision medicine. *Endocrine-Related Cancer* **25**, T121–T140 (2018).
154. Yehia, L., Ngeow, J. & Eng, C. PTEN-opathies: from biological insights to evidence-based precision medicine. *J. Clin. Invest.* **129**, 452–464 (2019).
155. Smith, I. N., Thacker, S., Jaini, R. & Eng, C. Dynamics and structural stability effects of germline PTEN mutations associated with cancer versus autism phenotypes. *J. Biomol. Struct. Dyn.* **37**, 1766–1782 (2019).
156. Mighell, T. L., Evans-Dutson, S. & O'Roak, B. J. A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. *Am. J. Hum. Genet.* **102**, 943–955 (2018).
157. Orloff, M. S. & Eng, C. Genetic and phenotypic heterogeneity in the PTEN hamartoma tumour syndrome. *Oncogene* **27**, 5387–5397 (2008).
158. Mester, J. L. *et al.* Gene-specific criteria for PTEN variant curation: Recommendations from the ClinGen PTEN Expert Panel. *Hum. Mutat.* **39**, 1581–1592 (2018).
159. Landrum, M. J. *et al.* ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, (2014).
160. Tan, M. H. *et al.* A clinical scoring system for selection of patients for pten mutation testing is proposed on the basis of a prospective study of 3042 probands. *Am. J. Hum. Genet.* **88**, 42–56 (2011).
161. Roche, A. F., Mukherjee, D., Guo, S. & Moore, W. M. Head Circumference Reference Data: Birth to 18 Years. *Pediatrics* **79**, (1987).
162. Saunders, C. T. & Baker, D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J. Mol. Biol.* **322**, 891–901 (2002).

163. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
164. Tilot, A. K. *et al.* Germline disruption of Pten localization causes enhanced sex-dependent social motivation and increased glial production. *Hum. Mol. Genet.* **23**, 3212–27 (2014).
165. Tilot, A. K. *et al.* Neural transcriptome of constitutional Pten dysfunction in mice and its relevance to human idiopathic autism spectrum disorder. *Mol. Psychiatry* **21**, 118–125 (2016).
166. Jinek, M. *et al.* A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science (80-.).* **337**, 816–821 (2012).
167. Nachmanson, D. *et al.* Targeted genome fragmentation with CRISPR/Cas9 improves hybridization capture, reduces PCR bias, and enables efficient high-accuracy sequencing of small targets. *bioRxiv* 207027 (2018). doi:10.1101/207027
168. Shin, G. *et al.* CRISPR–Cas9-targeted fragmentation and selective sequencing enable massively parallel microsatellite analysis. *Nat. Commun.* **8**, 14291 (2017).
169. Lee, J. *et al.* CRISPR-Cap: multiplexed double-stranded DNA enrichment based on the CRISPR system. *Nucleic Acids Res.* **47**, (2019).
170. McKenna, A. & Shendure, J. FlashFry: a fast and flexible tool for large-scale CRISPR target design. *BMC Biol.* **16**, 74 (2018).
171. Smith, T., Heger, A. & Sudbery, I. UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).
172. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* (2013).
173. Kim, H. K. *et al.* In vivo high-throughput profiling of CRISPR–Cpf1 activity. *Nat. Methods* **14**, 153–159 (2017).
174. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
175. Singh, D. *et al.* Real-time observation of DNA target interrogation and product release by the RNA-guided endonuclease CRISPR Cpf1 (Cas12a). *Proc. Natl. Acad. Sci. U. S. A.* **115**, 5444–5449 (2018).
176. Varga, E. *et al.* Improved LbCas12a variants with altered PAM specificities further broaden the genome targeting range of Cas12a nucleases. *Nucleic Acids Res.* **48**, 3722–3733 (2020).
177. Kleinstiver, B. P. *et al.* Engineered CRISPR–Cas12a variants with increased activities and improved targeting ranges for gene, epigenetic and base editing. *Nature Biotechnology* **37**, 276–282 (2019).
178. Chen, P. *et al.* A Cas12a ortholog with stringent PAM recognition followed by low off-target editing rates for genome editing. *Genome Biol.* **21**, (2020).
179. Mester, J. L. *et al.* Gene-specific criteria for *PTEN* variant curation: Recommendations from the ClinGen PTEN Expert Panel. *Hum. Mutat.* **39**, 1581–1592 (2018).
180. Weile, J. & Roth, F. P. Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. *Human Genetics* **137**, 665–678 (2018).

181. Esposito, D. *et al.* MaveDB: An open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol.* **20**, 1–11 (2019).
182. Gelman, H. *et al.* Recommendations for the collection and use of multiplexed functional data for clinical variant interpretation. *Genome Med.* **11**, 85 (2019).
183. Teng, F. *et al.* Enhanced mammalian genome editing by new Cas12a orthologs with optimized crRNA scaffolds. *Genome Biol.* **20**, 15 (2019).
184. Chen, J. S. *et al.* CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science (80-.).* **360**, 436–439 (2018).
185. Swarts, D. C. & Jinek, M. Mechanistic Insights into the cis- and trans-Acting DNase Activities of Cas12a. *Mol. Cell* **73**, 589-600.e4 (2019).