Functional dissection of *SPOP* on the amino acid level reveals a comprehensive functional landscape of variants during tumorigenesis

Seong Kyoon Park,[1*] Jeongha Lee,[2*] Seon Ju Park,[1,3] Ye Na Kim,[1,3] Gi Hyun Shin,[1,3] Kisoon Dan,[4,5] Hee-Jung Choi,[6] Dohyun Han,[4,5] Byung Joon Hwang,[1,3**] and Murim Choi[2**]

[1]Department of Molecular Bioscience, College of Biomedical Science, Kangwon National University, Chuncheon 24341, Republic of Korea

[2]Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul 03080, Republic of Korea

[3]Nabigene Inc., Kangwon Daehakgil, Chuncheon 24341, Republic of Korea

[4]Proteomics and Metabolomics Core Facility, Biomedical Research Institute, Seoul National University Hospital, Seoul 08032, Republic of Korea

[5]Department of Transdisciplinary Medicine, Seoul National University Hospital, Seoul 08032, Republic of Korea

[6]Department of Biological Sciences, Seoul National University, Seoul 08826, Republic of Korea

*. Equal contribution

**. Corresponding authors. Email: bjhwang@kangwon.ac.kr and murimchoi@snu.ac.kr.

Abstract

Numerous proteins display pleiotropic functions in different clinical contexts. However, molecular mechanism underlying such effects is rarely understood. Speckle-type POZ protein (SPOP) is a typical example, exhibiting tumor-suppressing or -promoting effects in different tumor types in accordance with different amino acid changes; specifically, two distinct sets of variants in *SPOP* are commonly found in subsets of prostate cancer and endometrial cancer patients. To comprehensively characterize the functional landscape of *SPOP* alteration, we performed a deep mutational screening (DMS), elucidating the functionality of 7,933 out of 8,228 possible single amino acid changes (96.4% coverage). Leveraging the observation that overexpression of human SPOP leads to yeast growth arrest, we assessed the functionality of each variant using a yeast proliferation assay. In addition, our approach combined long-read and short-read sequencing. While long-read-based scores demonstrated improved concordance with functional classification, both platforms demonstrated a largely consistent classification pattern of external data into the same tiers. Finally, our DMS model enables a clear distinction between loss-of-function and putative gain-of-function changes and reveals their differential characteristics in both protein structure and genetic assessments. These results demonstrate the utility of our approach in high resolution mapping and amino-acid-level interpretation of protein function.

Introduction

As large-scale efforts to sequence broader populations and diverse disease cohorts have become increasingly common, the identification of variants of uncertain significance (VUS) has likewise ballooned, leading to an urgent need for accurate functional interpretation of these variants. Conventional approaches–leveraging evolutionary conservation (1,2), population frequency (3,4), and machine learning models (5-7)–are frequently employed to predict the potential impact of VUS. However, those predictions often lack biological validation and suffer from inconsistent results and mis-interpretation. In that context, deep mutational scanning (DMS) offers a high-throughput experimental framework to assign functional scores to artificially introduced variants within a specific cellular context (8-12). Indeed, a number of genes with disease association have been analyzed through this method to provide insights into the function of pathogenic variants (13-15) and sensitivity to drugs (16,17). Notably, the reliability of DMS depends heavily on the construction of a comprehensive and uniform variant library, as well as the establishment of assay systems that properly report on the gene's function.

Interpreting the functional impact of genetic variants is particularly critical in the context of tumorigenesis, as progressive accumulation of genetic alterations is one of the fundamental mechanisms through which cancer cells gain a survival advantage. Distinguishing driver mutations from passenger mutations among these acquired variants is essential for elucidating tumorigenic mechanisms and informing clinical strategies. Notably, some cancer-associated genes such as *NOTCH1*, *TP53*, and *MYC* exhibit both tumor-suppressive and oncogenic features, whereby both loss-of-function (LoF) and gain-of-function (GoF) mutations can promote tumorigenesis in a context-dependent manner (18-20). It is especially crucial to dissect the molecular mechanisms by which distinct mutations in these bivalent genes contribute to cancer.

Accordingly, DMS has increasingly been applied to cancer genes to systematically identify functionally disruptive variants and elucidate their biological consequences (21-23).

The Speckle-type POZ protein (SPOP) functions as a substrate adapter for the CUL3-RING ubiquitin ligase (CRL3) complex, facilitating the ubiquitination and proteasomal degradation of a broad range of substates, including proto-oncogenes and the bromodomain and extra-terminal domain (BET) proteins. As such, the interaction between SPOP, CUL3, and its substrates is essential for effective proteolysis (24). Beyond its role in the CRL3 complex, the functionality of SPOP is also determined by its structural properties, as SPOP forms a higher-order oligomer structure through its BTB and BACK domains. This oligomerization is crucial for modulating its ubiquitination efficiency through subcellular localization and stability modulation (25,26). Genetic variants altering these properties are known to contribute to tumorigenesis by promoting genomic instability and dysregulating molecular pathways such as the PI3K/mTOR signaling cascade (27,28). According to The Cancer Genome Atlas (TCGA), *SPOP* variants have been identified in a number of cancer types; the most notable are prostate adenocarcinoma (PRAD: 53 carriers/496 patients = 10.7%) and uterine corpus endometrial carcinoma (UCEC: 53/512 = 10.4%), with instances also recorded in cancers of the penis, liver, intestine, and urinary tract (29). While SPOP is widely regarded as a tumor suppressor in that its LoF variants lead to increased oncogenesis, certain mutations found in endometrial cancer have been reported to act as GoF, enhancing ubiquitination activity (30). These divergent effects have been examined in the context of differential mutation hotspots observed between prostate and endometrial cancers (25,31), and numerous studies have been conducted on the mechanism of SPOP-related tumorigenesis and the distinct functional impacts of its mutations; nonetheless, a curated variant profile for *SPOP* is still lacking, as evidenced by the absence of any entry in ClinGen and the presence of only five somatic mutations reported in ClinVar.

To generate a comprehensive functional map of *SPOP*, we applied a DMS approach to systematically assign functional scores to individual variants. Leveraging our observation that overexpression of *SPOP* arrests yeast cell proliferation, we utilized a yeast-based growth assay system to dissect the functionality of individual amino acid changes. We also implemented a long-read sequencing-based analytic system and compared its performance to that of a short-read system. Our approach enables the functional characterization of genetic variants and facilitates the integration of mutation profiles from prostate and endometrial cancers with their corresponding functional impacts.

Materials and Methods

Yeast strain

Yeast assays used *Saccharomyces cerevisiae* strain JC-993 (MATa gal4 gal80 trp1-901 ura3-52 leu2-3 leu2-112 his3 URA3::GAL1-LacZ) with three constructs integrated via CRISPR-Cas9: pADH1(cr)-LexA-ER-B112 (32), activated by ß-estradiol to drive I-Sce I expression through the LexA promoter (top construct in Supplementary Fig. S1A); pADH1(cr)-Zif268-PR-MSN2, activated by progesterone to drive DMS library expression through the pZ promoter (33); and the landing pad construct, integration site for the DMS library (the second construct in Supplementary Fig. S1A). This construct was inserted between TRP5 and CWH41 loci on chromosome VII, and contains a dummy sequence under pZ, I-SceI under LexA, and KanR, conferring G418 resistance.

Library design, cloning, and transformation

For DMS library construction, *SPOP* cDNA (CCDS11551.1) was divided into 16 overlapping tiles (54-122 bp) based on the availability of internal restriction enzyme (RE) sites. As high-quality oligonucleotide synthesis is length-limited (~190 nucleotides), we introduced silent mutations when no suitable RE site was available. Within each tile, saturation mutagenesis was performed at the amino acid level by introducing 19 different substitutions, more than one synonymous changes, one in-frame deletion, and one stop-gain mutation per codon. All designed sequences were synthesized as oligonucleotides (Supplementary Table S1) by Twist Bioscience (South San Francisco, CA). Oligonucleotides from different tiles were separated and amplified using tile specific PCR primers (Supplementary Table S2). PCR was performed with 16 ng of oligonucleotide pool as template using Phusion High Fidelity DNA Polymerase (M0503, New England Biolabs, Ipswich, MA) for 18 cycles. To reconstruct the full-length *SPOP* cDNA, 5’ and 3’ flanking open reading frame (ORF) sequences corresponding to each tile were synthesized and

sequentially ligated via compatible RE sites (Supplementary Fig. S1B). Completed library constructs were cloned into a vector backbone compatible with the landing pad (Supplementary Fig. S1C). To facilitate homologous recombination, the cDNA was flanked by homology arms corresponding to the 5' (pZ promoter) and 3' (KanR) regions of the landing pad cleavage site (the two construct shown at the bottom of Supplementary Fig. S1A).

Yeast cells were transformed through the LiAC/ss carrier DNA/polyethylene glycol method (34). The inducer ß-estradiol (E2758, Sigma-Aldrich, Saint Louis, MO) was applied to activate expression of I-SceI, which recognized and cleaved its RE sites flanking the pZ promoter and KanR, generating double-strand breaks for integration of the transformed library via homologous recombination (Supplementary Fig. S1B) (35-37). Yeast clones harboring the library were selected by supplemention of culturing media with 200 ug/ml of G418 (11-811-023, Thermo Fisher Scientific).

Library screening

To identify differential enrichment of variants in the yeast, two separate pools were prepared containing 2.0 x 10^8 cells each. Both were initially cultured in 50 ml of YPAD media supplemented with 200 ug/ml G418, shaked at 230 rpm for one hour at 30 ºC. Afterwards, the media were replaced with 200 ml YPAD containing 200 ug/ml G418. For the experimental condition (the 'treated', *SPOP*-expressed pool), 500 nM progesterone (P8783, Sigma-Aldrich, Saint Louis, MO) was additionally supplemented; the other, untreated pool served as a control (non-expressed pool). Both cultures were then incubated under 3D culture conditions for 24 hours at 30 ºC (38).

Library sequencing

Genomic DNA was extracted from 1.0 x 10^8 cells via the YeaStar Genomic DNA kit (D2002, ZYMO research, Irvine, CA). Integrated *SPOP* mutant constructs were enriched through 18

cycles of PCR amplification using 500 ng of gDNA as a template, primers flanking the *SPOP* construct (forward: 5'-AGCTGCATAACCACTTTAAC-3', reverse: 5'-GCAAATTAAAGCCTTCGAGC-3'), and Phusion HF polymerase (M0503, New England Biolabs). Amplified products were pooled and purified through ethanol precipitation, agarose gel extraction, and ethanol precipitation. For short-read sequencing, library DNA was prepared using the TruSeq Nano DNA Sample Prep Kit and sequenced 2x300 bp on an Element FreeStyle AVITI. For long-read sequencing, 200 pmole of PCR product was sequenced using Ligation sequencing amplicons V14 (SQK-LSK114, Oxford Nanopore Technologies, Oxford, UK) via the MinION Mk1B platform with a R10.4.1 flowcell (FLO-MIN114, Oxford Nanopore Technologies). Base calling was performed in super-accurate mode with the *SPOP* wild type sequence as reference.

Calculation of z-scores

Scores from short-read and long-read sequencing platforms were calculated separately. Short reads were preprocessed (adapter trimming and contaminant removal) using BBDuk with adapter and PhiX sequences provided, error-corrected with BBMerge, and aligned to the *SPOP* cDNA reference using BBMap. BBDuk, and BBMerge are from BBTools (39). Variants were called using AnalyzeSaturationMutagenesis in GATK (40), and the resulting *variantCounts* files used as per-variant count matrices. These matrices served as input for DESeq2 (41), and the enrichment ratio between *SPOP*-expressed and non-expressed pools was calculated by adapting a pipeline from a previous study (42). Before generating the DESeqDataSet, variants were restricted to only single variants and those with raw count ≥5. Two biological replicates were used and adjusted during DESeqDataSet generation. Since only LoF variants in the non-expressed pool were expected to remain detectable in the *SPOP*-expressed pool, size factors were estimated from stop-gain variants. After differential expression analysis, outlier variants with poor correlation between biological replicates were identified by regularized log

transformation and extracted. Then, a revised input matrix was generated and summary statistics obtained from DESeq. Finally, z-scores were calculated as the log 2 fold-change divided by its standard error.

Long-read data was limited to high-quality ('PASS') reads. To ensure accurate variant calling, hybrid error correction employing AVITI short-reads was performed using the *ratatosk* correct function with options -v -c 16 -G -Q 90 (43). Error-corrected long-reads were subsequently filtered by length using *Seqkit* (44), and regions outside the ORF were trimmed using *cutadapt* (45). Filtered reads were aligned to *SPOP* cDNA using *minimap2* with the -ax map-ont option (46), and variants were called from the aligned bam files by GATK AnalyzeSaturationMutagenesis. The resulting variant count matrix was processed as for short-reads, including the same filtering criteria, application of size factors, and generation of z-scores. Lastly, scores from short-reads and long-reads were combined using inverse-variance weighted averege values. For variants only in short-reads, DESeq2 summary statistics and z-scores were used as-is. Two-tailed *P*-values were calculated with merged scores and adjusted by the Benjamini-Hochberg FDR method.

Serial dilution-spotting assay

Expression plasmids carried the *GAL1* promoter for galactose-inducible expression and a LEU2 auxotrophic marker for plasmid selection. In clone testing, both CEN and 2μ origins were used for clones shown in in Supplementary Fig. S2, but only CEN for the experiment in Fig. 1B. Yeast strains individually transformed with the respective mutations were cultivated in SD/-Leu medium and diluted with fresh SD/–Leu medium to an OD□□□ of 1.0 (~1□×□10□□cells/mL). Serial ten-fold dilutions (10^7-10^3 cells/mL) were prepared for each strain, and 5□μL of each dilution was spotted onto selective induction plates supplemented with 2% galactose.

Yeast proteome analysis

Yeast expressing wild-type *SPOP* were cultured in YPAD media with and without 500 nM progesterone at 30 ºC for one hour. Cells (1.0 x 10^6) were harvested, lysed in SDS buffer via sonication (Branson Ultrasonics, Brookfield, CT), boiled, and lysate protein concentrations measured by tryptophan fluorescence assay (47,48). Preparation and analysis of peptides were described previously (49). Briefly, a total of 100 µg protein per sample was retrieved and resulting peptides were eluted, quantified, and labeled with TMTsixplex (Thermo Fisher Scientific) tag 126-128 for non-expressed or 129-131 for *SPOP*-expressed. The total of 96 fractions were concatenated into 24 and subjected to liquid chromatography-mass spectrometry (LC-MS). The resulting raw MS files were processed using Proteome Discoverer 3.1 (Thermo Scientific) against the *S. cerevisiae* protein sequence database (UniProt, July 2025). The FDR was set to <1%, and a co-isolation threshold of 50% was used for peptide quantification. Differential expressed protein (DEP) analysis utilized the DEP package (50), with criteria of absolute log2 fold-change >0.7 and adjusted *P*-value <0.05. Gene Ontology enrichment analysis utilized the clusterProfiler package (51).

Assigning variant tiers from platform-specific and merged scores

Variants observed in each sequencing platform were independently stratified by score; those with a Log2FC > -1, indicating no depletion in the *SPOP*-expressed pool, were classified as ‘Possible LoF’. Platform-specific tiers were used only for comparing the models derived from each sequencing platform.

Subsequent analyses employed merged scores, which were also stratified independently. Variants with -1 < Log2FC < 1, for which *SPOP* expression did not significantly impact cell number, and variants with Log2FC ≥1, which produced accelerated growth, were collectively classified as ‘Possible LoF’ due to failing to induce growth inhibition. Variants displaying depletion (Log2FCs ≤ -1) but with a non-significant two-sided *P*-value (FDR >0.1) were considered non-differential and classified as ‘Moderate LoF’. Variants with Log2FCs ≤-1 and

FDR ≤0.1 were considered ‘Intermediate’, reflecting significant depletion in the *SPOP*-expressed pool. Finally, variants showing strong depletion (Log2FC ≤-1 and FDR <1 x 10^{-5}) were categorized as ‘Possible hyperactive’.

Quantification of mutant-harboring transcripts

Mutant-harboring transcripts were quantified by bulk RNA sequencing of *SPOP*-expression and non-expressing yeast pools, in three technical replicates. Total RNA was prepared using the TruSeq Stranded Total RNA Sample Prep Kit with Ribo-zero H/M/R (Illumina, San Diego, CA) and sequenced 2x150 bp on a NovaSeq X Plus (Illumina). Raw FASTQ files were aligned by STAR aligner in two-pass mode (52) to the *S. cerevisiae* reference genome (assembly R64), with wild-type *SPOP* appended. Mutation-associated expression was estimated by using pileup to stratify aligned reads by codon along the ORF and counting altered codons. Count matrices for each library were normalized by total read counts, normalized counts averaged within each pool, and differential transcript abundance assessed in terms of the log10-transformed ratio of pool means. Since a majority of assessed variants did not deviate from the mean, variants with normalized count ratio >1 were considered upregulated by the mutation, and those with ratio <-1 downregulated.

Reference set-based evaluation of pathogenicity/benignity prediction

To calculate AUC values and likelihood ratios, we defined reference sets of benign and pathogenic variants. As reported variants are limited, stop-gain mutations and variants with low REVEL scores (5) were incorporated. Reference set 1 and 2 respectively consisted of somatic only and somatic plus germline variants annotated as pathogenic in ClinVar with a least one review star. Reference set 3 and 4 are expanded those sets with stop-gain variants . Finally, the common benign reference set comprised variants with a group maximum allele frequency >0.0001 in gnomAD (V4.1.0) and those with REVEL score <0.183 (53). When calculating

likelihood ratios for pathogenicity and benignity, variants classified as ‘Possible LoF’ and ‘Possible hyperactive’ were considered positive for pathogenicity, whereas variants classified ‘Intermediate’ were considered negative.

Comparison with external datasets

We utilized diverse *in silico* scores and cohort data in assessing our scoring system. Relative conservation scores were obtained from Consurf-DB (54), which provides pre-calculated evolutionary conservation profiles for Chain H of the 8DWV PDB structure. Conservation grades and solvent accessibility annotations (‘B/E’ for buried/exposed and ‘F/S’ for functional/structural) were used to evaluate tier-level enrichment of conserved or buried/exposed residues. *In silico* variant effect scores were obtained from AlphaMissense (6) and ESM1b (7). Variants reported in gnomAD (V4) were used as benign controls, and somatic variants from COSMIC were used to evaluate tumor type-specific impacts (55). Structural and positional visualization utilized PyMOL with the 8DWV (for Fig 2E, Fig S6, Fig S8, and Fig 4C-E) and 3HQI (for Fig 3F) PDB structures as templates (56).

Differential energy prediction for protein stability

To assess mutation impact on the substrate-binding site of the MATH domain, differential interaction energy was quantified by FoldX (57) with the 3HQI PDB structure as template and a repaired structure generated using the *RepairPDB*. Mutations in residues of interest (shown in Fig. 3F) were introduced into Chain B of the repaired structure using *BuildModel* with either substitution or termination and *numberOfRuns* set to 10. The interaction energy between Chain B and Chain D (*i.e.*, PucSBC1 molecule) was calculated for all generated mutant models using *AnalyseComplex*.

Calculating extent of disease clustering (EDC)

EDC for the residues of interest was determined described in a previous study (58) with Chain E of the 8DWV PDB structure as reference. Disease residues comprised those exclusively associated with the ‘Possible hyperactive’ or ‘Possible LoF’ tiers. Randomly selected residues (*n* = 80) served as the negative control. Statistical significance was assessed over 10,000 permutations in which randomly selected sets were taken as disease residues and the remainder as non-disease residues.

Results

Measuring biological effect of SPOP in yeast

To dissect the function of *SPOP* at the amino-acid level, we delivered the DMS library into yeast, induced transgene expression by addition of progesterone, and determined *SPOP* sequences in surviving cells using short-read and long-read sequencing (Fig. 1A). Our library construction and delivery strategy includes insertion of a landing pad into the yeast genome to facilitate insertion of the library, which was made via concatenation of a mutation-containing tile (54-122 bp long) with the remaining 5' and 3' parts of the *SPOP* cDNA (Supplementary Fig. S1 and Methods).

We observed that expression of wild-type *SPOP* inhibits yeast growth (Fig. 1B), whereas depletion of specific functional regions within wild-type *SPOP*, namely the MATH, BTB, and 3-box domains (Supplementary Fig. S2), rescues it. To elucidate the molecular mechanism underlying this SPOP-induced growth inhibition, we performed tandem mass tag-labeled liquid chromatography-mass spectrometry and identified differentially expressed proteins in yeast cells with and without wild-type *SPOP* overexpression. A total of 100 DEPs were detected, 74 up- and 26 down-regulated (Fig. 1C, Supplementary Table S3, Supplementary Fig. S3). To identify direct downstream targets of SPOP's engagement in ubiquitination, we intersected the down-regulated proteins with SPOP-binding consequence (SBC) sequences. This revealed two proteins involved in the glycolysis pathway (FBA1 and CDC19) (Fig. 1C, Supplementary Fig. S4) reduced by the addition of *SPOP*. Notably, FBA1 encodes fructose-bisphosphate aldolase and catalyzes a reversible reaction for both glycolysis and glyconeogenesis (59); its impacts on cell growth and proliferation can be attributed to its role in glycolysis (60,61) and its regulation of the polymerase III RNA complex (62). Meanwhile, CDC19 encodes pyruvate kinase, which converts phosphoenolpyruvate to pyruvate and thereby finishes the glycolysis process in a partially

compensatory manner; its depletion leads to growth arrest by impairing ATP generation (63-65). Based on these observations, it is possible that metabolic burden may have inhibited the cell growth in our *SPOP*-expression system.

Leveraging short- and long-read platforms for generation of a combined score.

Most previous DMS studies based their evaluation of variant functions on short reads (*e.g.*, Illumina). but this approach does not account for the additional inadvertent mutations that can be generated during library construction and sequencing. To accurately identify and address such nucleotide changes in our library, the mutant clones from *SPOP*-expressed and non-expressed pools were sequenced on both short-read and long-read sequencing platforms (Fig. 2A, Supplementary Table S4). The enrichment ratio of variants in the *SPOP*-expressed pool relative to the non-expressed pool was calculated separately for each platform (see Methods, Supplementary Table S5). Two biological replicates were generated, and exhibited high correlation (Supplementary Fig. S5).

To fully leverage the results from both sequencing platforms, we generated combined scores using inverse-variance weighted averege values (Supplementary Table S5). This led to the list of variants, including silent variants (designed = 374, observed = 354), missense-causing variants (designed = 7,106, observed = 6,940), stop-gain variants (designed = 374, observed = 369), single amino acid-deletion variants (designed = 374, observed = 270), and unintentionally occurred frameshift (*n* = 3,439) and in-frame (*n* = 3,099) variants (Fig. 2B, Supplementary Table S6).

Next, we categorized the variants into four tiers based on the log2 fold changes (Log2FC) and Benjamini-Hochberg adjusted *P*-value, as visualized in the volcano plot depicting expression differences of individual clones (left panel of Fig. 2C, Methods). Specifically, clones which

showed depletion in the *SPOP*-expressed pool were classified as ‘Intermediate’ (grey background) or ‘Possible hyperactive’ (blue background) depending on the degree of depletion; clones that were depleted to a non-significant level were assigned as ‘Moderate LoF’ (light red background); and those which did not meet fold-changes and *P*-values significance criteria were classified as ‘Possible LoF’ (red background), as loss of SPOP function will leave the yeast cells intact. This classification strategy demonstrated good correlation with predicted variant impact (right panel of Fig. 2C, Supplementary Fig. S6). For example, stop-gain and frameshift mutations were enriched in the ‘Possible LoF’ and ‘Moderate LoF’ tiers. Missense-causing variants and in-frame deletions exhibited a bimodal distribution with subsets enriched either across the ‘Intermediate’ and ‘Moderate LoF’ tiers or in the ‘Possible LoF’ tier. Only a small subset of missense-causing variants, along with a few silent mutations and in-frame deletions, fell into the ‘Possible hyperactive’ tier. Lastly, silent mutations were enriched across the ‘Intermediate’ and ‘Moderate LoF’ tiers. To account for the possibility that these changes in SPOP function may be due to differences in expression rather than protein function, bulk RNA sequencing was performed to document *SPOP* mRNA expression. This sequencing revealed most variants to exhibit moderate changes in expression, and the correlation of expression with z-score was minimal (adjusted-R^2 = 0.009, Supplementary Fig. S7). Indeed, only 2.3% of the ‘Possible LoF’ variants (2 out of 89) statistically downregulated *SPOP* expression.

To better characterize each tier, we examined the positional enrichment of variants by tier; this revealed hotspot regions and distinct distributions for variants in the ‘Possible LoF’ and ‘Possible hyperactive’ tiers compared to those in ‘Intermediate’ (Supplementary Fig. S8). We further annotated the variants with AlphaMissesne and ESM1b scores and observed the ‘Possible LoF’ tier to be significantly enriched for variants with those with deleterious impacts (Fig. 2D). This finding suggests that variants in the ‘Possible LoF’ tier are likely to confer functional effects distinct from those of neutral variants (*i.e.*, the ‘Intermediate’ tier). It is also remarkable that,

compared to those in ‘Intermediate’ tier, variants in the ‘Possible hyperactive’ tier displayed a subtle shift to deleterious in both Alphamissense (P = 0.36) and ESM1b (P = 0.0026) scores.

In addition, when we overlaid the evolutionary conservation scores of each amino acid residue from ConSurfDB with the median of missense-causing variant scores, variants in the ‘Moderate LoF’ and ‘Possible LoF’ tiers were found to be enriched in more conserved bins, whereas variants in the ‘Possible hyperactive’ tier displayed a biased distribution toward variable bins (Fig. 2E).

Although we observed a good correlation between short- and long-read data (Fig. 2F, Pearson’s correlation R = 0.85), long-read-based calls assigned more stop-gain variants to the possible LoFs (Log2FC > -1, 35 vs ten variants, Supplementary Fig. S9). Annotation of variants exclusively assigned as ‘Possible LoF’ by each sequencing platform with AlphaMissense scores revealed the long-read-based set to have stronger enrichment for deleterious scores (Fig. 2G). This suggests that long-read-based sequencing provides more accurate assignment of LoF variant function.

Next, we evaluated the performance of our combined score model in distinguishing pathogenic variants from benign variants. Due to the limited number of variants available as ground truth, we generated four reference sets by incorporating additional variants that could be considered credible representatives of either pathogenic or benign categories (see Methods). Of these reference sets, Set 3 yielded the highest predictive likelihood ratio values of 18.325 and 0.0607 for being pathogenic and benign, respectively, leading to an AUC of 0.960 (Fig. 2H, Table 1, Supplementary Table S7). These likelihood ratios closely corresponded to the ‘strong’ evidence levels (PS3 and BS3) defined by the ACMG/AMP 2015 guidelines (66,67). This suggests that our scoring system, which captures the biological mechanism of SPOP-mediated yeast growth inhibition, robustly predicts the pathogenicity of variants reported as somatic variants. All told, our scoring model classified 381 out of 6,261 missense variants as ‘Possible LoF’.

Validation of SPOP fitness scores

To further validate if our z-scores accurately represent the functionality of individual variants, we cross-referenced our scores with external databases and performed experiments to test the pathogenicity of selected variants. Since common variants tend to be benign, we first determined z-scores for the common variants listed in gnomAD; of these, 65.3% (230 out of 352) classified as 'Intermediate' and 30.7% (108 out of 352) as 'Moderate LoF' (Fig. 3A-B). Similarly categorizing variants from tumor primary sites revealed that distributions of z-scores from the prostate and endometrium cancers differ, with variants from prostate cancer showing higher scores than those from endometrium cancer (two-sided *t*-test $P = 1.7 \times 10^{-3}$) (Fig. 3C). This distinction was also observed when plotting the COSMIC variants from each tumor by frequency (Supplementary Fig. S10). Next, we experimentally tested variants found in high frequencies among the prostate cancer patients through assessing yeast survival (Fig. 3D). As expected, variants classified as 'Possible LoF' (*i.e*., p.Y87N, p.F102C, and p.W131G) failed to arrest yeast growth, whereas an 'Intermediate' variant (p.F133L) maintained the growth inhibition phenotype (Fig. 3E). Therefore, while somatic variants associated with prostate cancer were observed across all four tiers, our tier system enables meaningful discrimination of truly pathogenic variants in prostate cancer.

To further illustrate the differences between variants of each tier, we visualized the prostate cancer-mutated residues (*i.e.,* recurrently mutated residues in prostate cancer) on the 3D structure of the MATH domain (Fig. 3F). The five residues in the 'Possible LoF' tier are located near the SPOP-binding consensus (SBC) motif, suggesting that mutation of these residues may disrupt SPOP-SBC interactions. In contrast, variants in other tiers exhibit distinct patterns. Among these residues, residues 42 (in 'Moderate LoF'), 83, 138, 148, and 165 (in 'Intermediate'), which are positioned far from the SBC and are solvent-exposed, were predicted

to have effects similar to silent mutations in terms of interaction energy in SPOP-substrate binding (Supplementary Fig. S11). These findings suggest that for a collection of observed mutations, our z-score-based tiers can provide insight into distinguishing pathogenic from passenger mutations, especially when correlated with structural information.

Utility of in-frame deletion scores

The addition of in-frame deletion variants allowed us to distinguish whether a given amino acid residue functions as LoF or GoF, because we posited that GoF residues would not result in any functional change if removed, whereas LoF residues might. Plotting the z-score values for amino acid residues with in-frame deletion revealed that the half of the residues relevant to prostate cancer (6/13) were enriched in the 'Possible LoF' tier for in-frame deletion (Fisher's exact test $P = 3.3 \times 10^{-3}$) (Fig. 4A). Meanwhile, most residues recurrently observed in endometrium cancer (5/6) were categorized into the 'Intermediate' and 'Moderate LoF' tiers for in-frame deletion-based z-scores.

To elucidate tolerance to in-frame deletion, we aligned solvent accessibility information with residues and observed that the in-frame deletion-based z-score differed significantly depending on whether a residue was buried or exposed (two-sided t-test $P = 0.0024$). Specifically, residues located in exposed sites were more likely to tolerate in-frame deletions (Fig. 4B). Previous studies have reported that residues impacted by variants recurrently observed in prostate and endometrium cancers exhibit distinct structural distributions: those related to prostate cancer are predominantly enriched in the substrate-binding site of the MATH domain, whereas those related to endometrium cancer are more enriched at SPOP domain interfaces that are important for SPOP self-assembly (25,30). This structural distinction has been exploited to explain the divergent pathways by which *SPOP* mutations promote tumorigenesis in those two cancer types

(25,68). In our 3D protein structure annotated with the in-frame deletion tier information (Fig. 4C), residues assigned to ‘Possible LoF’ tier are buried inside by those classified as ‘Moderate LoF’ and ‘Intermediate’. Ultimately, inference from this analysis led us to pinpoint a group of variants likely to confer GoF inferred from the in-frame deletion z-scores being found most frequently in the BTB domain, followed by the MATH domain (Supplementary Fig. S12).

We also investigated the intrinsic importance of the reference amino acids, in which the most pronounced change in z-scores was observed upon alteration of tryptophan (Supplementary Fig. S13A). While most reference amino acids exhibited ‘Intermediate’ mean z-scores upon substitution, replacing tryptophan generally resulted in mean z-scores associated with ‘Possible LoF’, except for substitutions on a few amino acids with nonpolar side chains (Supplementary Fig. S13B).

Disease-causing variants that operate by different molecular mechanisms are known to exhibit distinct tendencies with regard to clustering on 3D protein structure (58). To assess whether our DMS model reflects this phenomenon, we extracted residues exclusively assigned to either ‘Possible hyperactive’ or ‘Possible LoF’ tiers and calculated the extent of disease clustering (EDC) following a previously established method (58). Residues associated with ‘Possible hyperactive’ variants showed an EDC of 1.222, whereas both those associated with ‘Possible LoF’ variants and randomly selected residues exhibited EDC values close to 1 (Fig. 4D). Upon 10,000 permutations of 80 randomly selected residues, the EDC observed for ‘Possible hyperactive’ residues was significantly higher (permutation-based $P = 1.5 \times 10^{-10}$, Fig. 4E). This observation indicates that our categorization of LoF and GoF variants is further supported by their clustering within the 3D structure of SPOP.

Discussion

SPOP possesses a remarkable context-dependent functionality in different tumor types. In particular, mutations recurrently observed in prostate cancer cause the gene to act as a tumor-suppressor, but mutations observed in endometrial cancer appear to confer a GoF-like effect. Also, somatic mutations occur widely throughout the gene, making its functional annotation and clinical implication challenging. Despite active studies on the tumorigenesis of SPOP, only variants located in well-characterized structural regions have been reported, and a database categorizing pathogenic variants is lacking.

To evaluate the function of each variant at a higher resolution, we performed a DMS targeting SPOP, covering 96.4% of its all possible variants. Using yeast cell survival as a functional indicator, we were able to resolve variants into four tiers: ‘Possible LoF, ‘Moderate LoF’, ‘Intermediate’, and ‘Possible hyperactive’ tiers. Most stop-gain and frameshift variants fell into the ‘Possible LoF’ tier and silent variants into the ‘Intermediate’ tier; missense variants and in-frame deletions showed wider classification distributions. Notably, our DMS scores classified the majority of gnomAD variants into the ‘Intermediate’ tier and differential score distributions were evident between variants recurrently observed in prostate and in endometrium cancer.

The ‘Moderate LoF’ tier encompassed both neutral-like silent mutations and LoF-like frameshifts, reflecting a mixed composition of both intermediate and LoF features. This characteristic was also evident in their positional distribution, which was similar to that of variants in ‘Intermediate’ tier with a uniform distribution, showed contradictory results when compared with Alphamissense and ESM1b scores, and displayed strong enrichment in highly conserved regions. Therefore, variants in the ‘Moderate LoF’ category require careful interpretation when assessing pathogenicity.

In addition to the functional dissection of *SPOP*, we demonstrate two technical improvements over previous DMS studies. First, we attempted to utilize long reads to fully cover the mutation profile of the DMS library. Albeit the long-read sequencing allowed us to document additional inadvertent mutations in the non-targeted portion of the gene, the fitness scores based on short-read-only and combined sets did not differ substantially in overall concordance rate. Second, we created in-frame deletions across most amino acid residues, which provided indications of whether a given residue would function in a GoF or LoF manner. In particular, integration of the positional enrichments of these variants provided molecular evidence toward the determination of GoF variants, predominantly found in endometrium cancer patients.

Our evaluation of predictive power for *SPOP* variant function yielded robust performance (*i.e.*, likelihood ratio for pathogenic variants = 18.325, approaching the ACMG-recommended threshold of 18.7 for pathogenic variants. Table 1). However, a similar approach using the empirically-detected variants from cancer patients did not achieve such robust prediction, possibly due to contamination of the set with passenger mutations in the gene (Supplementary Table S7).

Multiple DMS studies have been performed in the cancer context to categorize tumorigenesis variants and establish scoring models based on cellular fitness readouts (69-72). The assay strategy often includes cell abundance (11), homology-directed repair reporter gene activity (73), and physical interaction with a partner protein (74). Our DMS library was synthesized through tile-based cloning, whereas several studies introduced mutations via CRISPR-based editing (75,76). While tile-based cloning allows saturation for all kinds of mutations, the CRISPR-based method allows comparison of the scoring system with nucleotide-level cancer signature analysis (70) and impacts on splicing or untranslated regions (71).

Our study has several limitations. The assay system is yeast-based and may not cover all substrate interactions that occur in human cells. It would be also possible to use the library to

assay against a specific substrate, oligomerization status, or formation of phase separation of SPOP (68,77). However, our assay also showed limitations in distinguishing neutral variants from GoF variants, as the ‘Possible hyperactive’ and ‘Intermediate’ tiers exhibited comparable amino acid conservation scores and endometrium-cancer-associated variants were enriched in the ‘Intermediate’ tier.

When we evaluated the performance of our scoring system in predicting pathogenicity, we found that it does not fully capture the malfunctionality of SPOP caused by germline mutations reported in ClinVar. Several factors could contribute to this discrepancy. First, differences in substrate binding to SPOP may lead to variable functional consequences, as germline mutations that disrupt binding of specific substrates involved in diseases may have less impact on substrates whose degradation affects yeast growth. Second, germline mutations in *SPOP* are known to cause Nabais Sa-de Vries syndrome, a neurodevelopmental disorder and inborn genetic disease; hence, germline variants may be more likely to be deleterious compared to somatic mutations, which arise post-developmentally and may be better tolerated. This difference in disease context may also reflect differential tolerance of SPOP function to mutations.

We also observed low resolution in distinguishing between neutral-like and LoF-like variants within the ‘Moderate LoF’ tier, which largely followed the ‘Intermediate’ trend but showed enrichment in conserved regions. Therefore, we excluded this tier from calculation of the likelihood ratio for pathogenicity/ benignity prediction and noted that its interpretation requires caution.

As told, this work presents a nearly complete functional atlas of SPOP single amino acid variants that enables refined interpretation of their potential clinical impact. This resource may provide a foundation for future mechanistic and translational studies through linking variant-specific functional consequences to tumor biology and therapeutic strategies.

Acknowledgment

We thank Theragen Etex and SNUH proteomics facility for technical supports and Eun Ha Kim for administrative support.

Author contributions

Seong Kyoon Park: Data curation, Investigation, Methodology, Resources, Validation, Writing – original draft

Jeongha Lee: Data curation, Investigation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft

Seon Ju Park: Methodology, Writing – review & editing

Ye Na Kim: Methodology, Project administration, Writing – review & editing

Gi Hyun Shin: Methodology, Writing – review & editing

Kisoon Dan: Formal analysis, Writing – review & editing

Hee-Jung Choi: Methodology, Software, Supervision, Writing – review & editing

Dohyun Han: Methodology, Resources, Supervision, Writing – review & editing

Byung Joon Hwang: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – original draft

Murim Choi: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – original draft

Conflicts of interest

The authors declare no conflicts of interest.

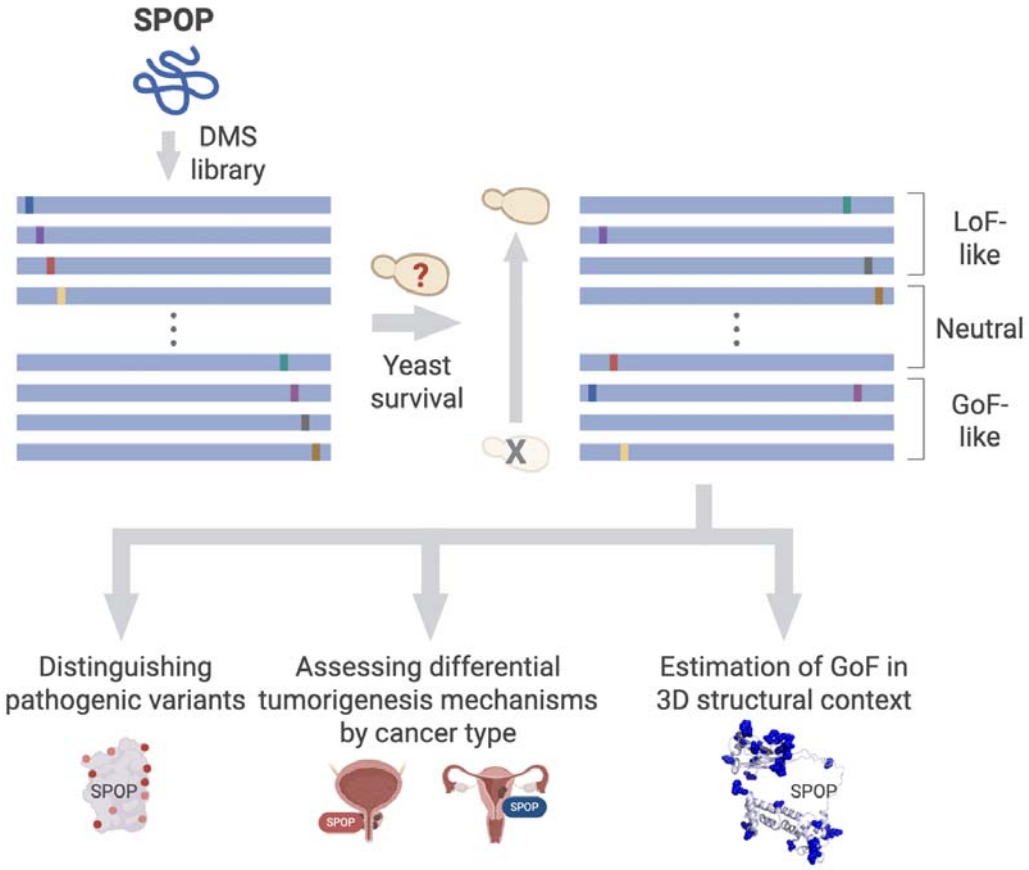
Funding

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1I1A1A01055745, to SKP) and funded by the Ministry of Science and ICT (2019H1A2A1076740 to JHL and NRF-RS-2023-00207857 and NRF-2022M3A9B6082674 to MC). This work is also supported in part by the SNUH Lee Kun-Hee Child Cancer & Rare Disease Project, Republic of Korea (22B-001-0500 to MC).

Data availability

All sequencing data were deposited at Short Read Archives BioProject PRJNA1304161 and Mass spectrum based based proteomics data were examinted at PRIDE (ID to be provided later). All codes for calculating scores in DMS models are available on Zenodo with the DOI 10.5281/zenodo.16787854.

Graphical abstract



Variant set*	Number of variants (P/B)	Pathogenic variants				Benign variants				AUC
		Possible hyperactive	Intermediate	Possible LoF	LR	Possible hyperactive	Intermediate	Possible LoF	LR	
Set 1	3/175	0	0	3	19.444	8	166	1	0.000	0.983
Set 2	11/175	0	10	1	1.768				0.958	0.529
Set 3	278/175	0	16	262	18.325				0.0607	0.960

Table 1. Predictive likelihood ratio for *SPOP* variants called as pathogenic and benign.

Set 4	288/175	0	26	262	17.689				0.0952	0.941

P, pathogenic. B, benign. LR, likelihood ratio. AUC, area under curve.

*. Denotes different combinations of pathogenic variants. Benign variants were commonly defined as listed in gnomAD with group maximum allele frequency >0.0001 and REVEL score <0.183 across sets (*n* = 175). Set 1: Somatic ClinVar variants with review star (*n* = 3). Set 2: Germline and somatic ClinVar variants with review star (*n* = 11). Set 3: Somatic ClinVar variants with review star plus stop-gain variants (*n* = 278). Set 4: Germline and somatic ClinVar variants with review star and stop-gain variants (*n* = 288).

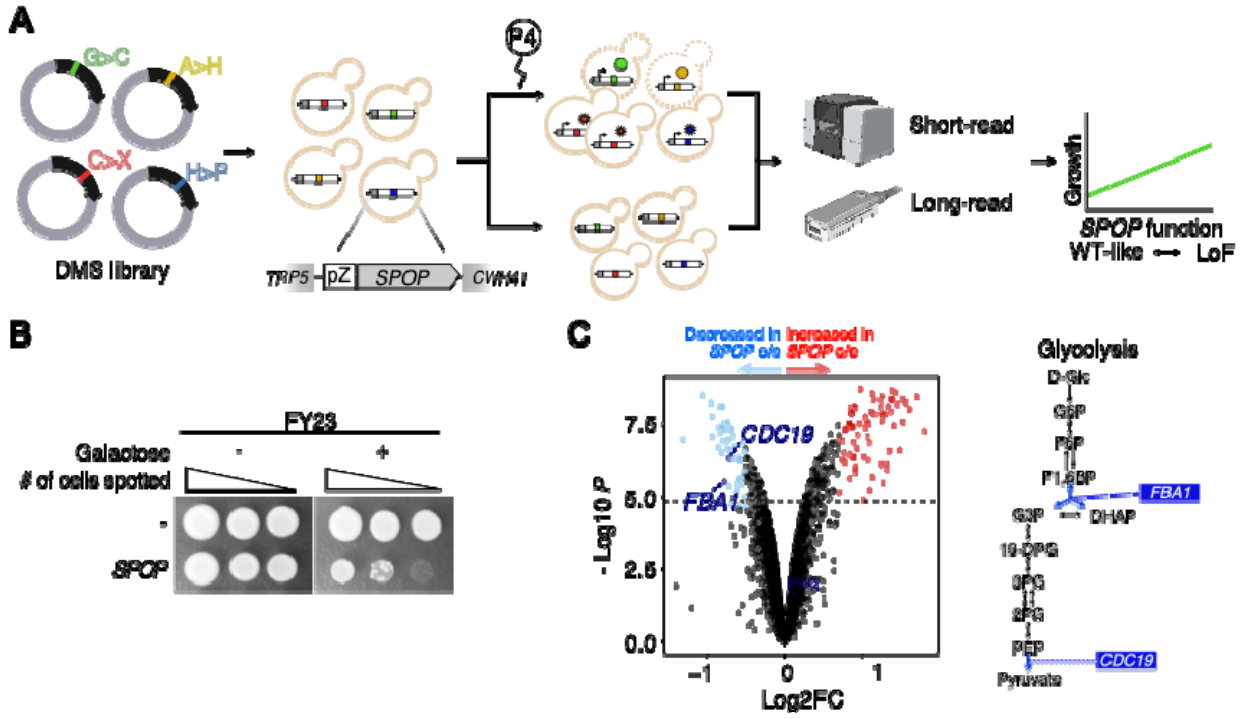


Figure 1. Overall scheme. (A) A DMS library was transfected into yeast. Expression of the introduced *SPOP* variants was induced by progesterone. After induction, genomic DNA was extracted from the surviving yeast cells, and the variants were identified using short- or long-read sequencing. The functional score of each variant was based on the enrichment ratio in the *SPOP*-expressed pool compared to the non-expressed pool. (B) Effect of *SPOP* expression on growth of the yeast strain FY23. (C) Volcano plot of proteins differentially expressed upon overexpression of wild-type *SPOP* (left). Position of highlighted proteins in the simplified glycolysis pathway (right). Abbreviations: P4, progesterone; D-Glc, D-glucopyranose; G6P, D-glucopyranose 6-phosphate; F6P, ß-D-fructofuranose 6-phosphate; F1,6BP, ß-D-fructose 1,6-bisphosphate; G3P, D-glyceraldehyde 3-phophate; DHAP, dihydroxyacetone phosphate; 13-DPG, 3-phospho-D-glyceroyl-phosphate; 3PG, 3-phospho-D-glycerate; 2PG, 2-phospho-D-glycerate; PEP, phosphoenolpyruvate.

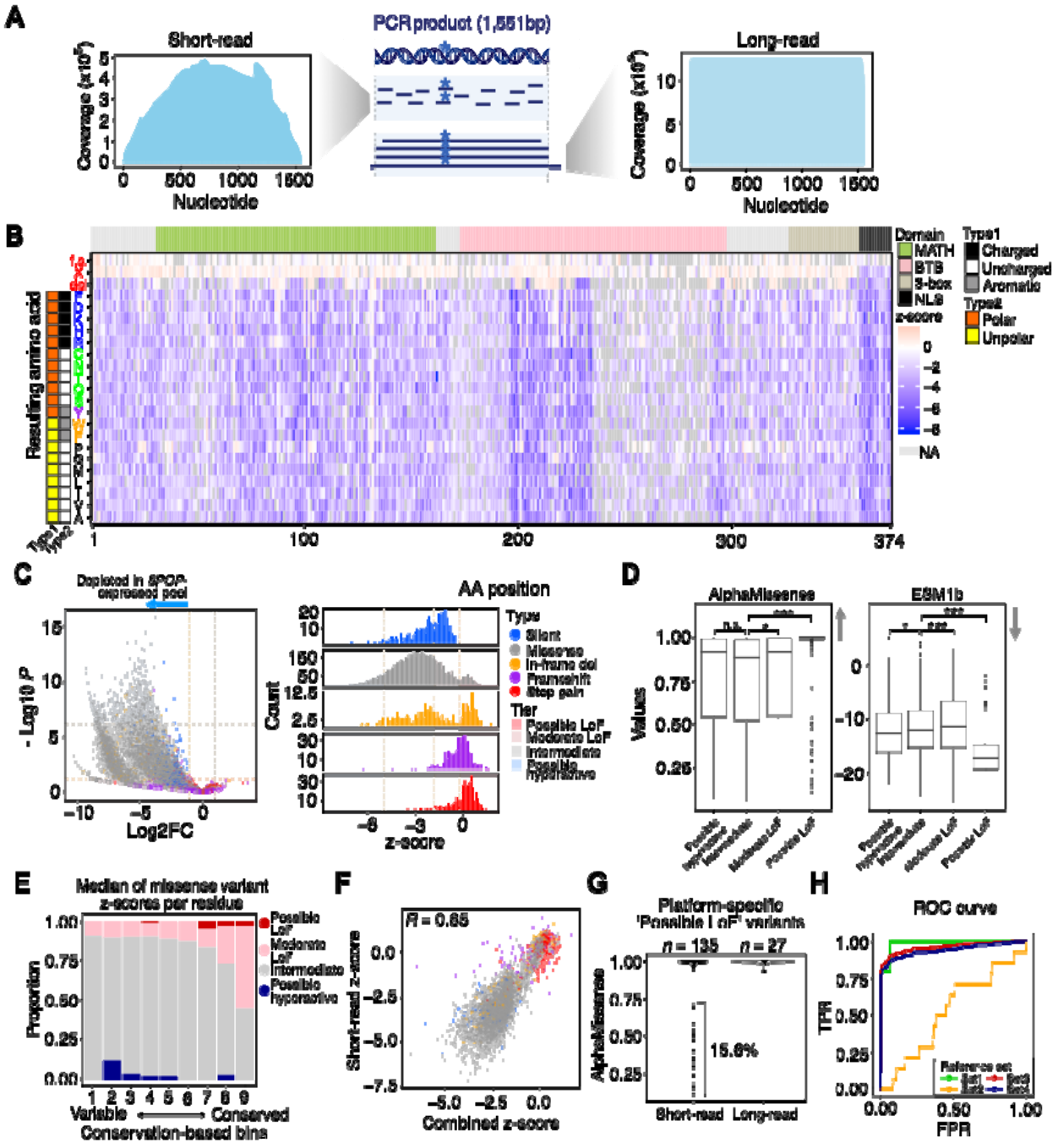


Figure 2. Generation of combined z-scores for the variant functional assessment. (A) Read coverage distributions of the reference template for short-read (left) and long-read (right) sequencing, and schematic illustration of read coverage patterns from each platform (center). (B) Heatmap showing combined z-scores for the covered *SPOP* variants. The x-axis represents the residue number and the y-axis the alternative amino acid. SPOP functional domains are

indicated along the top. Amino acids are color-coded based on their biochemical properties. (C) Volcano plot depicting variants by Log2FC and *P*-value (left). Vertical dashed lines correspond to log2FC values of -1 and 1, and the horizontal dashed lines mark adjusted *P*-values of 0.1 and 1.0×10^{-5}. Histogram displaying z-score distribution by variant impact and assigned tier (right). Variant impacts are distinguished by color, and tiers by background color. (D) Boxplot of AlphaMissesne and ESM1b scores for each tier; values significantly different between tiers were determined by two-sided Student's *t*-test. Asterisks indicate *P*-value magnitude: *** for *P*-value <0.0005, ** for <0.005, * for <0.05, and n.s. for non-significant; *P*-values based on AlphaMissense (Possible hyperactive vs Intermediate = 0.36, Moderate LoF vs Intermediate = 0.035, Possible LoF vs Intermediate = 2.1×10^{-84}), ESM1b (Possible hyperactive vs Intermediate = 2.6×10^{-3}, Moderate LoF vs Intermediate = 1.6×10^{-07}, Possible LoF vs Intermediate = 1.6×10^{-90}). Arrows beside panels indicate the direction of increasing pathogenicity. (E) Functionality score vs. evolutionary conservation. Median missense variant z-score values per each residue were used to determine tiers. (F) Scatter plot showing correlation of z-scores from short-read and combined long- and short-read data. Each point is colored according to variant consequence as defined in (C). (G) AlphaMissense scores of variants classified as 'Possible LoF' by a single sequencing platform. (H) Performance of our scoring model in distinguishing pathogenic variants from benign variants. ROC curves were generated using the four reference sets from Table 1.

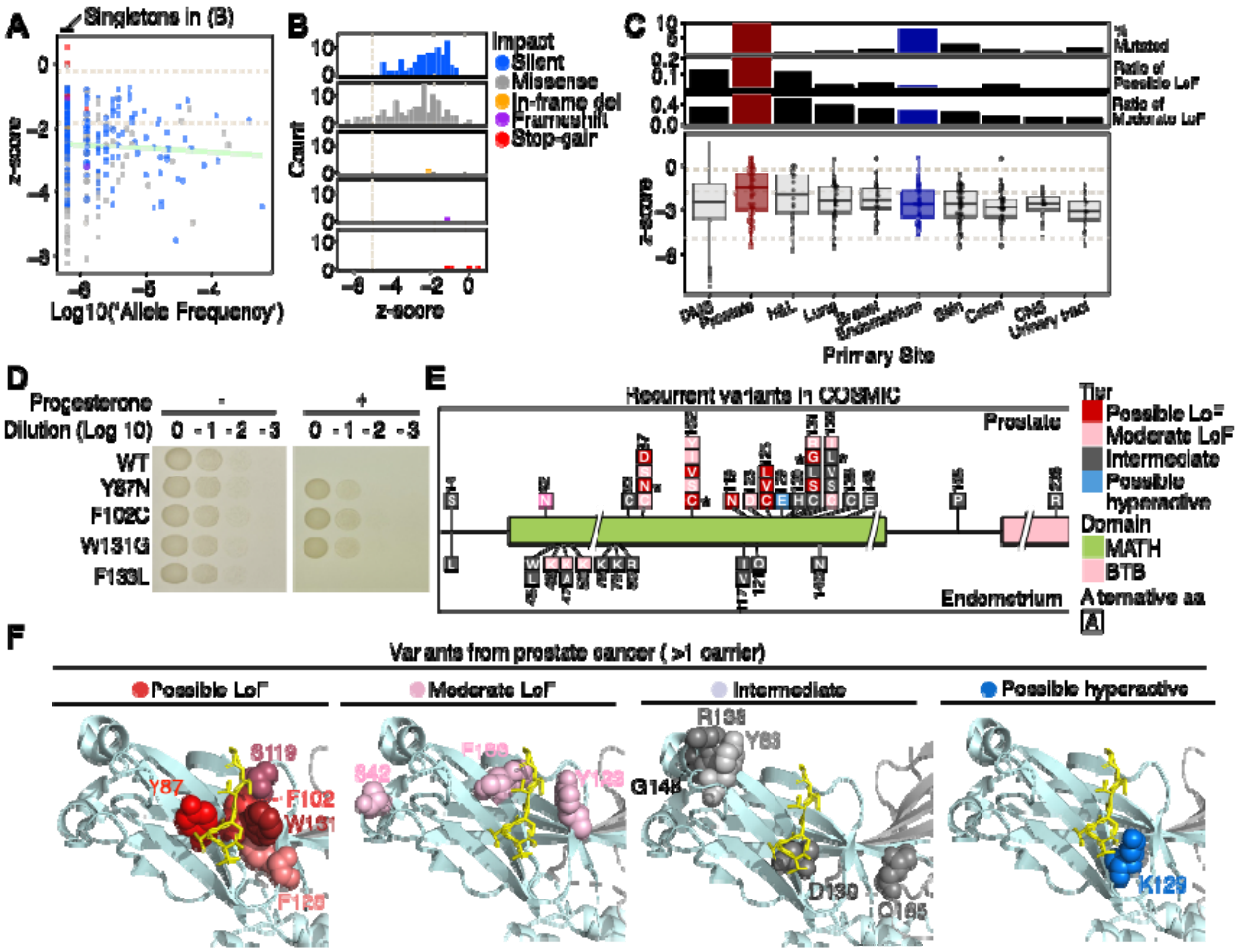


Figure 3. Validation of SPOP functional scores. (A) Correlation of z-scores with allele frequency in gnomAD. Variants on the far left part of the plot are singletons. Variant impacts are distinguished by dot color; the linear regression line based on Log10 transformed allele frequency and z-score is indicated in light green. (B) Z-score distribution of gnomAD singleton variants according to impact on *SPOP*. (C) Z-score distribution of COSMIC variants according to primary tumor sites. The top panel indicates the percentage of patients with *SPOP* somatic mutations; the middle two panels the proportion of variants with positive z-scores per tier; and the bottom panel the distribution of z-scores per tumor site. H&L, Hematopoietic and lymphoid. (D) Yeast growth assay of variants associated with prostate cancer. (E) Lollipop plot of recurrently observed somatic variants in prostate (top) and endometrial (bottom) tumors. Each variant is represented as a box, colored according to its tier. Letters within boxes indicate

alternative amino acids. Asterisks next to boxes denote variants tested in (D). (F) Structural placement diagrams of variant residues found in prostate cancers from panel (E) according to variant tier, shown as color-coded spheres by variant tier (PDB ID: 3HQI). Only the MATH domain and the SBC motf are displayed, represented as a cyan cartoon and yellow sticks, respectively.

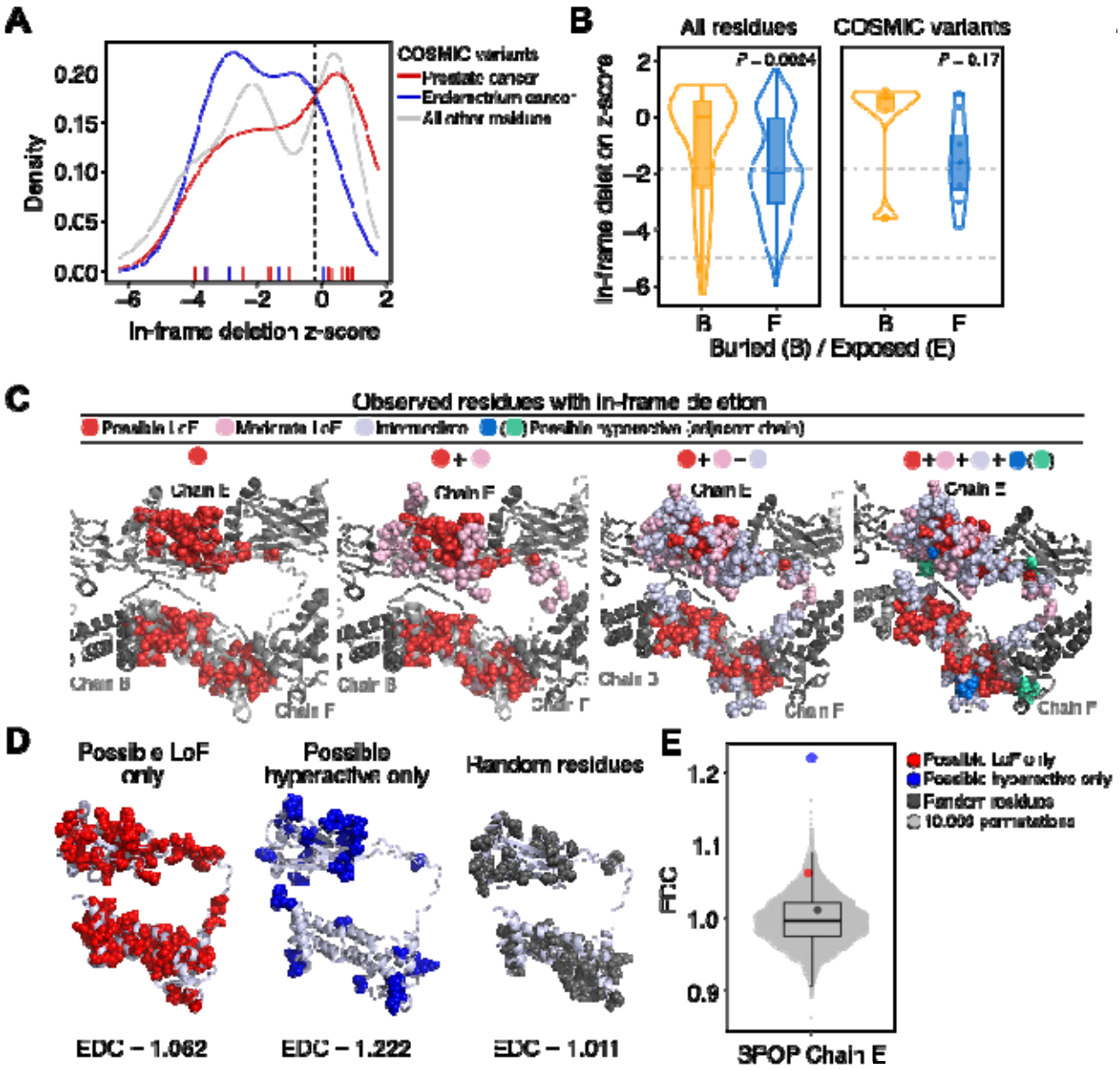


Figure 4. Interpretation of variant impact based on amino acid biochemical properties and structural context. (A) Dissection of COSMIC-overlapping residues for their z-score of the corresponding in-frame deletion on that position. (B) Stratification of in-frame deletion z-scores by residue solvent accessibility. (C) Residues with in-frame deletions are displayed as color-coded spheres on the oligomeric structure of SPOP. Colors indicate variant tier and “Chain” refers to SPOP monomer. (D) Residues exclusively harboring variants with z-score >0.5 (left), <-6 (middle), and randomly selected residues (right) were depicted in red, navy, and grey

spheres, respectively. These sets of residues were used to calculate the EDC values. (E) EDC values from 10,000 permutations were shown as light dots with overlaid boxplot. EDC values corresponding to the residue groups in (D) were highlighted as larger colored dots (red, navy, and dark grey). Protein structure in (C-E) is based on the PBD ID: 8DWV and only Chain E was used for (D-E).

References

1. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. and Siepel, A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*, **20**, 110-121.
2. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A. and Batzoglou, S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol*, **6**, e1001025.
3. Chen, S., Francioli, L.C., Goodrich, J.K., Collins, R.L., Kanai, M., Wang, Q., Alfoldi, J., Watts, N.A., Vittal, C., Gauthier, L.D. *et al.* (2024) A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*, **625**, 92-100.
4. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M. and Maglott, D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*, **42**, D980-985.
5. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D. *et al.* (2016) REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet*, **99**, 877-885.
6. Cheng, J., Novati, G., Pan, J., Bycroft, C., Zemgulyte, A., Applebaum, T., Pritzel, A., Wong, L.H., Zielinski, M., Sargeant, T. *et al.* (2023) Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, **381**, eadg7492.
7. Brandes, N., Goldman, G., Wang, C.H., Ye, C.J. and Ntranos, V. (2023) Genome-wide prediction of disease variant effects with a deep protein language model. *Nat Genet*, **55**, 1512-1522.
8. Rubin, A.F., Stone, J., Bianchi, A.H., Capodanno, B.J., Da, E.Y., Dias, M., Esposito, D., Frazer, J., Fu, Y., Grindstaff, S.B. *et al.* (2025) MaveDB 2024: a curated community database with over seven million variant effects from multiplexed functional assays. *Genome Biol*, **26**, 13.
9. Meitlis, I., Allenspach, E.J., Bauman, B.M., Phan, I.Q., Dabbah, G., Schmitt, E.G., Camp, N.D., Torgerson, T.R., Nickerson, D.A., Bamshad, M.J. *et al.* (2020) Multiplexed Functional Assessment of Genetic Variants in CARD11. *Am J Hum Genet*, **107**, 1029-1043.
10. Howard, M.K., Hoppe, N., Huang, X.P., Mitrovic, D., Billesbolle, C.B., Macdonald, C.B., Mehrotra, E., Rockefeller Grimes, P., Trinidad, D.D., Delemotte, L. *et al.* (2025) Molecular basis of proton sensing by G protein-coupled receptors. *Cell*, **188**, 671-687 e620.
11. Matreyek, K.A., Starita, L.M., Stephany, J.J., Martin, B., Chiasson, M.A., Gray, V.E., Kircher, M., Khechaduri, A., Dines, J.N., Hause, R.J. *et al.* (2018) Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat Genet*, **50**, 874-882.
12. Starr, T.N., Greaney, A.J., Hilton, S.K., Ellis, D., Crawford, K.H.D., Dingens, A.S., Navarro, M.J., Bowen, J.E., Tortorici, M.A., Walls, A.C. *et al.* (2020) Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*, **182**, 1295-1310 e1220.
13. Majithia, A.R., Tsuda, B., Agostini, M., Gnanapradeepan, K., Rice, R., Peloso, G., Patel, K.A., Zhang, X., Broekema, M.F., Patterson, N. *et al.* (2016) Prospective

functional classification of all possible missense variants in PPARG. *Nat Genet*, **48**, 1570-1575.
14. van Loggerenberg, W., Sowlati-Hashjin, S., Weile, J., Hamilton, R., Chawla, A., Sheykhkarimli, D., Gebbia, M., Kishore, N., Fresard, L., Mustajoki, S. *et al.* (2023) Systematically testing human HMBS missense variants to reveal mechanism and pathogenic variation. *Am J Hum Genet*, **110**, 1769-1786.
15. Jia, X., Burugula, B.B., Chen, V., Lemons, R.M., Jayakody, S., Maksutova, M. and Kitzman, J.O. (2021) Massively parallel functional testing of MSH2 missense variants conferring Lynch syndrome risk. *Am J Hum Genet*, **108**, 163-175.
16. Howard, C.J., Abell, N.S., Osuna, B.A., Jones, E.M., Chan, L.Y., Chan, H., Artis, D.R., Asfaha, J.B., Bloom, J.S., Cooper, A.R. *et al.* (2025) High-resolution deep mutational scanning of the melanocortin-4 receptor enables target characterization for drug discovery. *Elife*, **13**.
17. Brenan, L., Andreev, A., Cohen, O., Pantel, S., Kamburov, A., Cacchiarelli, D., Persky, N.S., Zhu, C., Bagul, M., Goetz, E.M. *et al.* (2016) Phenotypic Characterization of a Comprehensive Set of MAPK1/ERK2 Missense Mutants. *Cell Rep*, **17**, 1171-1183.
18. Shi, Q., Xue, C., Zeng, Y., Yuan, X., Chu, Q., Jiang, S., Wang, J., Zhang, Y., Zhu, D. and Li, L. (2024) Notch signaling pathway in cancer: from mechanistic insights to targeted therapies. *Signal Transduct Target Ther*, **9**, 128.
19. Jamroze, A., Chatta, G. and Tang, D.G. (2021) Androgen receptor (AR) heterogeneity in prostate cancer and therapy resistance. *Cancer Lett*, **518**, 1-9.
20. Chen, X., Zhang, T., Su, W., Dou, Z., Zhao, D., Jin, X., Lei, H., Wang, J., Xie, X., Cheng, B. *et al.* (2022) Mutant p53 in cancer: from molecular mechanism to therapeutic modulation. *Cell Death Dis*, **13**, 974.
21. Boettcher, S., Miller, P.G., Sharma, R., McConkey, M., Leventhal, M., Krivtsov, A.V., Giacomelli, A.O., Wong, W., Kim, J., Chao, S. *et al.* (2019) A dominant-negative effect drives selection of TP53 missense mutations in myeloid malignancies. *Science*, **365**, 599-604.
22. Mighell, T.L., Thacker, S., Fombonne, E., Eng, C. and O'Roak, B.J. (2020) An Integrated Deep-Mutational-Scanning Approach Provides Clinical Insights on PTEN Genotype-Phenotype Relationships. *Am J Hum Genet*, **106**, 818-829.
23. Sahu, S., Galloux, M., Southon, E., Caylor, D., Sullivan, T., Arnaudi, M., Zanti, M., Geh, J., Chari, R., Michailidou, K. *et al.* (2025) Saturation genome editing-based clinical classification of BRCA2 variants. *Nature*, **638**, 538-545.
24. Zhuang, M., Calabrese, M.F., Liu, J., Waddell, M.B., Nourse, A., Hammel, M., Miller, D.J., Walden, H., Duda, D.M., Seyedin, S.N. *et al.* (2009) Structures of SPOP-substrate complexes: insights into molecular architectures of BTB-Cul3 ubiquitin ligases. *Mol Cell*, **36**, 39-50.
25. Cuneo, M.J., O'Flynn, B.G., Lo, Y.H., Sabri, N. and Mittag, T. (2023) Higher-order SPOP assembly reveals a basis for cancer mutant dysregulation. *Mol Cell*, **83**, 731-745 e734.
26. Marzahn, M.R., Marada, S., Lee, J., Nourse, A., Kenrick, S., Zhao, H., Ben-Nissan, G., Kolaitis, R.M., Peters, J.L., Pounds, S. *et al.* (2016) Higher-order oligomerization promotes localization of SPOP to liquid nuclear speckles. *EMBO J*, **35**, 1254-1275.

27. Boysen, G., Barbieri, C.E., Prandi, D., Blattner, M., Chae, S.S., Dahija, A., Nataraj, S., Huang, D., Marotz, C., Xu, L. *et al.* (2015) SPOP mutation leads to genomic instability in prostate cancer. *Elife*, **4**.
28. Blattner, M., Liu, D., Robinson, B.D., Huang, D., Poliakov, A., Gao, D., Nataraj, S., Deonarine, L.D., Augello, M.A., Sailer, V. *et al.* (2017) SPOP Mutation Drives Prostate Tumorigenesis In Vivo through Coordinate Regulation of PI3K/mTOR and AR Signaling. *Cancer Cell*, **31**, 436-451.
29. Sondka, Z., Dhir, N.B., Carvalho-Silva, D., Jupe, S., Madhumita, McLaren, K., Starkey, M., Ward, S., Wilding, J., Ahmed, M. *et al.* (2024) COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Res*, **52**, D1210-D1217.
30. Janouskova, H., El Tekle, G., Bellini, E., Udeshi, N.D., Rinaldi, A., Ulbricht, A., Bernasocchi, T., Civenni, G., Losa, M., Svinkina, T. *et al.* (2017) Opposing effects of cancer-type-specific SPOP mutants on BET protein degradation and sensitivity to BET inhibitors. *Nat Med*, **23**, 1046-1054.
31. Cavalcante, L., Deshmukh, S.K., Ribeiro, J.R., Carneiro, B.A., Dizon, D.S., Angara, K., Mattox, T., Wu, S., Xiu, J., Walker, P. *et al.* (2023) Opposing Roles of SPOP Mutations in Human Prostate and Endometrial Cancers. *JCO Precis Oncol*, **7**, e2300088.
32. Dossani, Z.Y., Reider Apel, A., Szmidt-Middleton, H., Hillson, N.J., Deutsch, S., Keasling, J.D. and Mukhopadhyay, A. (2018) A combinatorial approach to synthetic transcription factor-promoter combinations for yeast strain engineering. *Yeast*, **35**, 273-280.
33. Aranda-Diaz, A., Mace, K., Zuleta, I., Harrigan, P. and El-Samad, H. (2017) Robust Synthetic Circuits for Two-Dimensional Control of Gene Expression in Yeast. *ACS Synth Biol*, **6**, 545-554.
34. Gietz, R.D. and Schiestl, R.H. (2007) High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat Protoc*, **2**, 31-34.
35. Rolloos, M., Hooykaas, P.J. and van der Zaal, B.J. (2015) Enhanced targeted integration mediated by translocated I-SceI during the Agrobacterium mediated transformation of yeast. *Sci Rep*, **5**, 8345.
36. Katz, S.S., Gimble, F.S. and Storici, F. (2014) To nick or not to nick: comparison of I-SceI single- and double-strand break-induced recombination in yeast and human cells. *PLoS One*, **9**, e88840.
37. Kuijpers, N.G., Chroumpi, S., Vos, T., Solis-Escalante, D., Bosman, L., Pronk, J.T., Daran, J.M. and Daran-Lapujade, P. (2013) One-step assembly and targeted integration of multigene constructs assisted by the I-SceI meganuclease in Saccharomyces cerevisiae. *FEMS Yeast Res*, **13**, 769-781.
38. Elsaesser, R. and Paysan, J. (2004) Liquid gel amplification of complex plasmid libraries. *Biotechniques*, **37**, 200, 202.
39. Bushnell, B. (2015) BBMap - sourceforge.net/projects/bbmap/.
40. Auwera, G.A.V.d. and O'Connor, B.D. (2020) *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media, Incorporated.
41. Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, **15**, 550.

42. Olvera-Leon, R., Zhang, F., Offord, V., Zhao, Y., Tan, H.K., Gupta, P., Pal, T., Robles-Espinoza, C.D., Arriaga-Gonzalez, F.G., Matsuyama, L. *et al.* (2024) High-resolution functional mapping of RAD51C by saturation genome editing. *Cell*, **187**, 5719-5734 e5719.
43. Holley, G., Beyter, D., Ingimundardottir, H., Moller, P.L., Kristmundsdottir, S., Eggertsson, H.P. and Halldorsson, B.V. (2021) Ratatosk: hybrid error correction of long reads enables accurate variant calling and assembly. *Genome Biol*, **22**, 28.
44. Shen, W., Le, S., Li, Y. and Hu, F. (2016) SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One*, **11**, e0163962.
45. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011*, **17**, 3.
46. Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094-3100.
47. Wisniewski, J.R. and Gaugaz, F.Z. (2015) Fast and sensitive total protein and Peptide assays for proteomic analysis. *Anal Chem*, **87**, 4110-4116.
48. von der Haar, T. (2007) Optimized protein extraction for quantitative proteomics of yeasts. *PLoS One*, **2**, e1078.
49. Kim, H., Woo, J., Dan, K., Lee, K.M., Jin, M.S., Park, I.A., Ryu, H.S. and Han, D. (2021) Quantitative Proteomics Reveals Knockdown of CD44 Promotes Proliferation and Migration in Claudin-Low MDA-MB-231 and Hs 578T Breast Cancer Cell Lines. *J Proteome Res*, **20**, 3720-3733.
50. Zhang, X., Smits, A.H., van Tilburg, G.B., Ovaa, H., Huber, W. and Vermeulen, M. (2018) Proteome-wide identification of ubiquitin interactions using UbIA-MS. *Nat Protoc*, **13**, 530-550.
51. Yu, G., Wang, L.G., Han, Y. and He, Q.Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, **16**, 284-287.
52. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15-21.
53. Pejaver, V., Byrne, A.B., Feng, B.J., Pagel, K.A., Mooney, S.D., Karchin, R., O'Donnell-Luria, A., Harrison, S.M., Tavtigian, S.V., Greenblatt, M.S. *et al.* (2022) Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *Am J Hum Genet*, **109**, 2163-2177.
54. Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E. and Ben-Tal, N. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163-164.
55. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E. *et al.* (2019) COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*, **47**, D941-D947.
56. Schrodinger, LLC. (2015) The PyMOL Molecular Graphics System, Version 1.8.
57. Delgado, J., Radusky, L.G., Cianferoni, D. and Serrano, L. (2019) FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics*, **35**, 4168-4169.

58. Gerasimavicius, L., Livesey, B.J. and Marsh, J.A. (2022) Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *Nat Commun*, **13**, 3895.
59. Lobo, Z. (1984) Saccharomyces cerevisiae aldolase mutants. *J Bacteriol*, **160**, 222-226.
60. Rodaki, A., Young, T. and Brown, A.J. (2006) Effects of depleting the essential central metabolic enzyme fructose-1,6-bisphosphate aldolase on the growth and viability of Candida albicans: implications for antifungal drug target discovery. *Eukaryot Cell*, **5**, 1371-1377.
61. Ritterson Lew, C. and Tolan, D.R. (2012) Targeting of several glycolytic enzymes using RNA interference reveals aldolase affects cancer cell proliferation through a non-glycolytic mechanism. *J Biol Chem*, **287**, 42554-42563.
62. Ciesla, M., Mierzejewska, J., Adamczyk, M., Farrants, A.K. and Boguta, M. (2014) Fructose bisphosphate aldolase is involved in the control of RNA polymerase III-directed transcription. *Biochim Biophys Acta*, **1843**, 1103-1110.
63. Kuivanen, J., Kannisto, M., Mojzita, D., Rischer, H., Toivari, M. and Jantti, J. (2021) Engineering of Saccharomyces cerevisiae for anthranilate and methyl anthranilate production. *Microb Cell Fact*, **20**, 34.
64. Boles, E., de Jong-Gubbels, P. and Pronk, J.T. (1998) Identification and characterization of MAE1, the Saccharomyces cerevisiae structural gene encoding mitochondrial malic enzyme. *J Bacteriol*, **180**, 2875-2882.
65. Cereghetti, G., Wilson-Zbinden, C., Kissling, V.M., Diether, M., Arm, A., Yoo, H., Piazza, I., Saad, S., Picotti, P., Drummond, D.A. *et al.* (2021) Reversible amyloids of pyruvate kinase couple cell metabolism and stress granule disassembly. *Nat Cell Biol*, **23**, 1085-1094.
66. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E. *et al.* (2015) 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**, 405-424.
67. Tavtigian, S.V., Greenblatt, M.S., Harrison, S.M., Nussbaum, R.L., Prabhu, S.A., Boucher, K.M., Biesecker, L.G. and ClinGen Sequence Variant Interpretation Working, G. (2018) Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet Med*, **20**, 1054-1060.
68. Sabri, N., Cuneo, M.J., Marzahn, M.R., Lee, J., Bouchard, J.J., Gullulu, O., Vaithiyalingam, S., Borgia, M.B., Schmit, J. and Mittag, T. (2023) Reduction of oligomer size modulates the competition between cluster formation and phase separation of the tumor suppressor SPOP. *J Biol Chem*, **299**, 105427.
69. Mighell, T.L., Evans-Dutson, S. and O'Roak, B.J. (2018) A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. *Am J Hum Genet*, **102**, 943-955.
70. Giacomelli, A.O., Yang, X., Lintner, R.E., McFarland, J.M., Duby, M., Kim, J., Howard, T.P., Takeda, D.Y., Ly, S.H., Kim, E. *et al.* (2018) Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat Genet*, **50**, 1381-1387.

71. Findlay, G.M., Daza, R.M., Martin, B., Zhang, M.D., Leith, A.P., Gasperini, M., Janizek, J.D., Huang, X., Starita, L.M. and Shendure, J. (2018) Accurate classification of BRCA1 variants with saturation genome editing. *Nature*, **562**, 217-222.
72. Waters, A.J., Brendler-Spaeth, T., Smith, D., Offord, V., Tan, H.K., Zhao, Y., Obolenski, S., Nielsen, M., van Doorn, R., Murphy, J.E. *et al.* (2024) Saturation genome editing of BAP1 functionally classifies somatic and germline variants. *Nat Genet*, **56**, 1434-1445.
73. Starita, L.M., Islam, M.M., Banerjee, T., Adamovich, A.I., Gullingsrud, J., Fields, S., Shendure, J. and Parvin, J.D. (2018) A Multiplex Homology-Directed DNA Repair Assay Reveals the Impact of More Than 1,000 BRCA1 Missense Substitution Variants on Protein Function. *Am J Hum Genet*, **103**, 498-508.
74. Hidalgo, F., Nocka, L.M., Shah, N.H., Gorday, K., Latorraca, N.R., Bandaru, P., Templeton, S., Lee, D., Karandur, D., Pelton, J.G. *et al.* (2022) A saturation-mutagenesis analysis of the interplay between stability and activation in Ras. *Elife*, **11**.
75. Hsu, P.D., Lander, E.S. and Zhang, F. (2014) Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, **157**, 1262-1278.
76. Findlay, G.M., Boyle, E.A., Hause, R.J., Klein, J.C. and Shendure, J. (2014) Saturation editing of genomic regions by multiplex homology-directed repair. *Nature*, **513**, 120-123.
77. Bouchard, J.J., Otero, J.H., Scott, D.C., Szulc, E., Martin, E.W., Sabri, N., Granata, D., Marzahn, M.R., Lindorff-Larsen, K., Salvatella, X. *et al.* (2018) Cancer Mutations of the Tumor Suppressor SPOP Disrupt the Formation of Active, Phase-Separated Compartments. *Mol Cell*, **72**, 19-36 e18.