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

Seong Kyoon Park,[a,1] Jeongha Lee,[b,1] Seon Ju Park,[a,c] Ye Na Kim,[a,c] Gi Hyun Shin,[a,c] Kisoon Dan,[d,e] Hee-Jung Choi,[f] Dohyun Han,[d,e] Byung Joon Hwang,[a,c,2] and Murim Choi[b,2]

[a]Department of Molecular Bioscience, College of Biomedical Science, Kangwon National University, Chuncheon 24341, Republic of Korea; [b]Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul 03080, Republic of Korea; [c]Nabigene Inc., Kangwon Daehakgil, Chuncheon 24341, Republic of Korea; [d]Proteomics and Metabolomics Core Facility, Biomedical Research Institute, Seoul National University Hospital, Seoul 08032, Republic of Korea; [e]Department of Transdisciplinary Medicine, Seoul National University Hospital, Seoul 08032, Republic of Korea; [f]Department of Biological Sciences, Seoul National University, Seoul 08826, Republic of Korea

[1]S.K.P. and J.L. equally to this work.

[2]To whom correspondence may be addressed. Email: bjhwang@kangwon.ac.kr or 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. Finally, our DMS model enables a clear distinction of likely-loss-of-function (LoF) variants that are enriched in prostate cancers 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.

Significance statement

Genetic mutations often play different roles in cancer, driving or suppressing tumor growth depending on their molecular context. The protein SPOP is a striking example that acts as either a tumor-suppressor or oncogene depending on the specific mutation and tissue. However, understanding how thousands of possible mutations alter its function has been a major challenge. Here, we applied deep mutational scanning to experimentally measure the functional effects of nearly every possible amino acid change in *SPOP*. This work provides the first comprehensive functional map of *SPOP* variants, not only advancing basic knowledge of cancer biology but also establishing a framework for interpreting patient mutations in precision medicine.

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) and machine learning models (3-6)–are frequently employed to predict the potential impact of VUS. However, these *in silico* predictions often lack biological validation and can suffer from inconsistent results. While human population frequency data serves as a robust line of evidence (7,8), its utility is often limited for rare variants and can be less informative for individuals of non-European ancestries due to data underrepresentation. 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 (9-13). Indeed, a number of genes with disease association have been analyzed through this method to provide insights into the function of pathogenic variants (14-16) and sensitivity to drugs (17,18). 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 (19-21). It is especially crucial to dissect the

molecular mechanisms by which distinct mutations in these bivalent genes contribute to cancer. Indeed, recent deep mutational scans of *CARD11*, *PTPN11*, and *DDX3X* effectively stratified variants into LoF and GoF classes, linking them to differential pathologies or caner types (10,22,23).

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 any of these properties are known to contribute to tumorigenesis by activating malignant transcriptional mechanisms (27,28), including genomic instability (29,30), or modulating *SPOP* function through aberrant assembly (31). 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 (32). 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 (33). These divergent effects have been examined in the context of differential mutation hotspots observed between prostate and endometrial cancers (25,34), 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 fitness 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 (35), activated by ß-estradiol to drive I-Sce I expression through the LexA promoter (top construct in SI Appendix, Fig. S1A); pADH1(cr)-Zif268-PR-MSN2, activated by progesterone to drive DMS library expression through the pZ promoter (36); and the landing pad construct, integration site for the DMS library (the second construct in SI Appendix, 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 synonymous 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 (37). All designed sequences were synthesized as oligonucleotides (Dataset S1) by Twist Bioscience (South San Francisco, CA). Oligonucleotides from different tiles were separated and amplified using tile specific PCR primers (Dataset 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 (SI Appendix, Fig. S1B). Completed library constructs were cloned into a vector backbone compatible with the landing pad (SI Appendix, 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 SI Appendix, Fig. S1A).

Yeast cells were transformed through the LiAC/ss carrier DNA/polyethylene glycol method (38). 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 (SI Appendix, Fig. S1B) (39-41). 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 (42).

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 (43). Variants were called using AnalyzeSaturationMutagenesis in GATK (44), and the resulting *variantCounts* files used as per-variant count matrices. These matrices served as input for DESeq2 (45), and the enrichment ratio between *SPOP*-expressed and non-expressed pools was calculated by adapting a pipeline from a previous study (46). 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. Prior to DESeq2 differential enrichment analysis and z-score calculation, we removed variants with inconsistent detection across biological replicates.

Specifically, variants were retained only if they had non-zero counts in both replicates in the untreated samples. This pre-filtering step reduces stochastic artifacts driven by replicate dropouts. 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 (47). Error-corrected long-reads were subsequently filtered by length using *Seqkit* (48), and regions outside the ORF were trimmed using *cutadapt* (49). Filtered reads were aligned to *SPOP* cDNA using *minimap2* with the -ax map-ont option (50), 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 average 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 SI Appendix, 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 (51,52). Preparation and analysis of peptides were described previously (53). 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 (54), with criteria of absolute log2 fold-change >0.7 and adjusted *P*-value <0.05. Gene Ontology enrichment analysis utilized the clusterProfiler package (55).

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 and adjusted *P*-value > 0.1, indicating no depletion in the *SPOP*-expressed pool, were classified as ‘Likely 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 Log2FC > -1 and adjusted *P*-value > 0.1, which did not significantly impact cell number or produced accelerated growth, were collectively classified as ‘Likely LoF’ due to failing

to induce growth inhibition. And other remaining variants displaying depletion were classified as 'Tolerated'.

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 (56) 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. ClinVar variants annotated as *Pathogenic* or *Likely Pathogenic* were used as the pathogenic reference set and were stratified into somatic and germline subsets based on ClinVar classification type. Each 'ClinVar-Somatic' and 'ClinVar-Germline' set were composed with 4 and 10 variants. To increase power for somatic benchmarking, we additionally compiled a COSMIC somatic pathogenic reference set consisting of recurrent *SPOP* variants (occurrence count > 1), focusing on prostate cancer (n = 32). A curated germline LoF subset associated with

Nabais Sa-de Vries syndrome was assembled from Nabais Sa *et al.* (57) (*n* = 4). As a benign reference set, we used missense variants observed in gnomAD (v4.1.0) (*n* = 60)**.**

For each reference set comparison, we evaluated the binary classification to distinguish the pathogenic set from the benign reference set**.** Positive/negative likelihood ratio (LR+/-) was computed from sensitivity and specificity at the 'Likely LoF cutoff' (LR+ = sensitivity / (1 − specificity), LR- = (1-sensitivity)/specificity). AUC was calculated from the corresponding ROC curve using the continuous score, leveraging pROC R package.

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 (58), 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 (4), EVE (5), and ESM1b (6). Variants annotation was performed using public databases, including gnomAD (v4.1.0; accessed on 12/02/2025), ClinVar (accessed on 12/01/2025), and COSMIC (accessed on 12/02/2025) (59). Structural and positional visualization utilized PyMOL with the 8DWV (for Fig. 2E, Fig. 4C-E, and SI Appendix Fig. S6, Fig. S8) and 3HQI (for Fig. 3F) PDB structures as templates (60).

Calculating extent of disease clustering (EDC)

EDC for the residues of interest was determined described in a previous study (61) with Chain E of the 8DWV PDB structure as reference. Residues were classified as disease residues if the fraction of variants at that position that were either 'Likely LoF' or 'Tolerated' with extreme lower-end z-scores was ≥ 0.5. Variants in the extreme low tail of the 'Tolerated' distribution were defined as those with a z-score below the 5th percentile of the synonymous variants distribution,

using synonymous variants as an empirical reference. Randomly selected residues (n = 26, matching the size of disease residues) served as the negative control. Statistical significance was assessed over 1,000 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 (SI Appendix, 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 (SI Appendix, 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, Dataset S3 and Fig. S4). 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. D, SI Appendix, Fig. S4) reduced by the addition of *SPOP*. Notably, FBA1 encodes fructose-bisphosphate aldolase and catalyzes a reversible reaction for both glycolysis and glyconeogenesis (62); its impacts on cell growth and proliferation can be attributed to its role in glycolysis (63,64) and its regulation of the polymerase III RNA complex (65). 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 (66-68). Notably, additional six out of 26 downregulated proteins are essential components of the glycolysis pathway (Fig.1D). Since the overexpression of CDC19 and FBA1 alone failed to rescue the lethal trait conferred by *SPOP* expression (SI Appendix, Fig. S6), it is possible that overall metabolic burden may have initiated the cell growth in our *SPOP*-expression system.

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

Many recent DMS studies have leveraged barcoding strategies combined with long-read sequencing to directly link mutations with their functional outcomes (69-71), while others have relied on short-read approaches that enable cost-efficient, high-depth quantification (11,15,72). However, library construction and sequencing can introduce additional unintended mutations that may not be fully captured by barcode-based designs alone. To take advantage of the complementary strengths of both platforms, we analyze our mutant libraries using short-read and long-read sequencing (Fig. 2A, SI Appendix, Table S1). The enrichment ratio of variants in the *SPOP*-expressed pool relative to the non-expressed pool was calculated separately for each platform (see Methods, Dataset S4). Two biological replicates were generated, and exhibited high correlation across treatment conditions and sequencing platforms (SI Appendix, Fig. S7).

To fully leverage the results from both sequencing platforms, we generated combined scores using inverse-variance weighted average values (Dataset S4). This led to the list of variants, including synonymous 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, SI Appendix, Table S2).

Next, we stratified variants into two groups–‘Likely Loss-of-Function (LoF)’ and ‘Tolerated’–based on log2 fold changes and Benjamini-Hochberg adjusted *P*-values. Since the variant assessment assay is based on the yeast survival and functional *SPOP* variants are expected to reduce cell viability, we defined variants showing insignificant changes in viability as ‘Likely LoF’ (Log2FC > -1 and FDR > 0.1) and all remaining variants as ’Tolerated’. Then we assigned a single cutoff from the intersection of the z-score distributions of the ‘Likely LoF’ versus ‘Tolerated’ variant groups (Fig, 2C). This classification strategy demonstrated good concordance with predicted variant impact (Fig. 2C, SI Appendix, Fig. S8). For example, 88.7% of stop-gain and 85.7% of frameshift mutations were enriched in the ‘Likely LoF’ classification. Missense-causing variants and in-frame deletions exhibited a bimodal distribution with subsets enriched either across the ‘Likely LoF’ and ‘Tolerated’ classifications. Lastly, 90.7% of synonymous mutations were enriched in ‘Tolerated’ classification. 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, SI Appendix, Fig. S9). Indeed, only 4.5% of the ‘Likely LoF’ variants (4 out of 89) statistically downregulated *SPOP* expression.

To better characterize variants classified as ‘Likely LoF’, we annotated each variant with EVE and ESM1b scores and observed that the ‘Likely LoF’ set was shifted toward more deleterious predictions compared to other variants (Fig. 2D). This finding suggests that variants in the ‘Likely LoF’ classification are likely to confer functional effects distinct from those of ‘Tolerated’ variants.

In addition, when we overlaid the evolutionary conservation scores of each amino acid residue from ConSurfDB with the median of missense-causing variant scores (73), we found significant

association between median missense z-scores and conservation grades (*P*-value = 6.2×10^{-8} from linear regression, Fig. 2E).

When we compared scores derived from each sequencing platform, short- and long-read data showed strong concordance (Fig. 2F, Pearson's correlation $R = 0.85$). Although a subset of variants were classified as 'Likely LoF' in a platform-specific manner (grid a and b in Fig. 2F; SI Appendix, Fig. S10), this discrepancy was not explained by differences in predicted deleteriousness: among missense variants, platform-specific calls showed no significant differences in matched ESM1b and AlphaMissense scores (Fig. 2G; Two-tailed t-test $P = 0.22$ and 0.11, respectively). Overall, while a subset of variants showed platform-specific 'Likely LoF' calls, we did not observe systematic biological differences. All told, our scoring model classified 713 out of 6,261 missense variants as 'Likely 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 missense variants listed in gnomAD; of these, 96.7% (58 out of 60) classified as 'Tolerated' (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 $P = 4.6 \times 10^{-4}$, Cohen's $d = 0.68$) (Fig. 3C and SI Appendix, Fig. S11). This distinction was also observed when plotting the COSMIC variants from each tumor by frequency (SI Appendix, Fig. S12). To further compare prostate- and endometrial cancer-associated variants, we visualized recurrently mutated residues (count > 5) on the 3D structure of the

MATH domain (Fig. 3D). Five of the six prostate-enriched residues localized near the SPOP-binding consensus (SBC) motif, and four of these five residues showed a high fraction (≥ 0.5) of 'Likely LoF' variants at that position. In contrast, recurrent endometrial cancer-associated residues showed lower 'Likely LoF' fractions (< 0.5) and were predominantly located at the MATH-domain interaction interface. These patterns are concordant with our 'Likely LoF' classification and with prior mechanistic models in which prostate cancer-associated *SPOP* mutations are predominantly LoF, whereas endometrial cancer-associated mutations exhibit more GoF-like properties, and they are consistent with the observed structural localization.

To define which classes of pathogenic variants are captured by our DMS readout, we examined which structural implications 'Likely LoF' variants have and where previously reported and experimentally validated variants fall in our score. We first assessed the structural distribution of 'Likely LoF' calls by collapsing variants into residues. Residues in structurally buried regions harbored a significantly higher fraction of 'Likely LoF' variants than solvent exposed residues (t-test $P = 2.0 \times 10^{-9}$ and Cohen's $d = 0.71$; Fig. 3E and SI Appendix, Fig. S13), consistent with sensitivity to variants that compromise folding/stability. Next, we performed a comprehensive analysis integrating databases with clinical relevance (ClinVar and COSMIC) and experimentally validated variants (Fig. 3F). ClinVar-annotated somatic pathogenic variants and recurrent prostate cancer-associated variants from COSMIC were shifted toward the 'Likely LoF' range relative to endometrial cancer-associated variants (grey boxes in Fig. 3F). Consistent with this, variants known to interact with oncogenic proteins or proteins responsible for genomic stability (*e.g.*, BET (33), Caprin1 (74), SRC-3 (27), Geminin (75), and 53BP1 (30)) generally scored as 'Likely LoF', whereas validated endometrial cancer variants with GoF-like properties (enhanced polyubiquitination (33) and altered oligomer assembly (25)/phase separation (31)) were predominantly classified as 'Tolerated' (green boxes in Fig. 3F). In contrast, germline pathogenic

variants showed weaker concordance with the ‘Likely LoF’ classification, including ClinVar variants and *de novo* variants associated with Nabais Sa-de Vries syndrome (NSDVS).

Furthermore, we evaluated the performance of our combined score model in distinguishing pathogenic variants from benign variants. Given the absence of a ClinGen Variant Curation Expert Panel (VCEP) for *SPOP*, we used ClinVar as the primary source of previously reported variants. ClinVar variants annotated as *Pathogenic/Likely Pathogenic* were used to define a reference pathogenic set and were stratified into ClinVar-Somatic and ClinVar-Germline**,** with gnomAD missense variants used as a benign reference (SI Appendix, Table S3). Our scoring system achieved the highest selection for the ‘ClinVar-Somatic’ set (LR+ = 22.5). Given the limited size of the ‘ClinVar-Somatic’ set (*n* = 4), we expanded somatic pathogenic set with recurrent (count >1) prostate cancer-associated COSMIC variants (COSMIC-PC), which retained robust performance (LR+ = 16.94). In contrast, ‘ClinVar-Germline’ showed poor concordance with our LoF-based assay readout (LR+ = 0), whereas the curated germline subset associated with Nabais Sa-de Vries syndrome and proposed LoF mechanisms showed slightly improved performance (LR+ = 7.5). This suggests that our scoring system, which captures the SPOP-mediated degradation of essential proteins of yeast growth, specifically identifies variants leading to LoF via interaction defects and structural destabilization. While this robustly predicts the pathogenicity of somatic variants driving prostate cancer, it shows limited sensitivity towards germline mutations.

Next, we experimentally tested variants found in high frequencies among the prostate cancer patients through assessing yeast survival (Fig. 3G). As expected, variants classified as ‘Likely LoF’ (*i.e*., p.Y87N, p.F102C, and p.W131G) failed to arrest yeast growth, whereas ‘Tolerated’ variants that were annotated as benign, found in endometrial cancers, and other variants from prostate cancers (*e.g.*, Y123C and F133L) maintained the growth inhibition phenotype (Fig. 3G and SI Appendix, Fig. S14).

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 'Likely LoF' classification for in-frame deletion (Fisher's exact test $P = 3.3 \times 10^{-3}$) (Fig. 4A). Meanwhile, most residues recurrently observed in endometrium cancer (4/6) were categorized into the 'Tolerated' classification 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 $P = 2.4 \times 10^{-3}$, Cohen's $d = 0.35$). 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 (25,33). This structural distinction has been exploited to explain the divergent pathways by which *SPOP* mutations promote tumorigenesis in those two cancer types (25,31). In our 3D protein structure annotated with the z-score of in-frame deletions as color of spheres (Fig. 4C), residues with 'Likely LoF' z-scores are buried inside by those classified as 'Tolerated'. 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 (SI Appendix, Fig. S15).

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 (SI Appendix, Fig. S16A). 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 (SI Appendix, Fig. S16B).

Disease-causing variants that operate by different molecular mechanisms are known to exhibit distinct tendencies with regard to clustering on 3D protein structure (61). To assess whether our DMS model reflects this phenomenon, we extracted residues that are enriched with ‘Likely LoF’ variants and variants at the other end of the z-score distribution. This assignment is based on our assumption that GoF variants would be enriched in the low end of the z-score distribution (SI Appendix, Fig. S17). Extent of disease clustering (EDC) values of residues enriched with ‘Tolerated’ variants with low z-scores showed an EDC of 1.69, whereas both those associated with ‘Likely LoF’ variants and randomly selected residues exhibited EDC values close to 1 (Fig. 4D; see Methods). Upon 1,000 permutations of randomly selected residues, the EDC observed for residues associated with low z-scores was significantly higher (permutation-based $P = 1.3 \times 10^{-23}$, Fig. 4E). This observation indicates that ‘Tolerated’ variants with extreme z-scores might have additional functionality compared to other ‘Tolerated’ variants.

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 binary classification: ‘Likely LoF’ and ‘Tolerated’. Most stop-gain and frameshift variants fell into the ‘Likely LoF’, and synonymous variants into the ‘Tolerated’; missense variants and in-frame deletions showed wider classification distributions. Notably, our DMS scores classified the majority of gnomAD variants into the ‘Tolerated’ classification and differential score distributions were evident between variants recurrently observed in prostate and in endometrium cancer.

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 comparisons against known variants define the scope of our scoring system for SPOP. We observed that curated somatic driver variant sets were enriched in the ‘Likely LoF’ class, and literature-curated variants with experimentally validated LoF mechanisms similarly clustered within the ‘Likely LoF’ range. In contrast, endometrial cancer-associated variants were preferentially assigned outside the ‘Likely LoF’ range, consistent with the previous reports that many of these mutations exhibit GoF-like properties. Accordingly, our score showed robust performance when evaluated using the ‘ClinVar-Somatic’ and ‘COSMIC-PC’ reference set (LR+ = 22.5 and 16.94, respectively). Together, these results support that our assay reliably assigns variants as ‘Likely LoF’ when they impair SPOP function in a manner consistent with loss of substrate recognition and/or degradation activity.

Multiple DMS studies have been performed in the cancer context to categorize tumorigenesis variants and establish scoring models based on cellular fitness readouts (76-79). The assay strategy often includes cell abundance (12), homology-directed repair reporter gene activity (80), and physical interaction with a partner protein (81). Our DMS library was synthesized through tile-based cloning, whereas several studies introduced mutations via CRISPR-based editing (82,83). 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 (77) and impacts on splicing or untranslated regions (78).

Our study has several limitations. The assay system is yeast-based and may not cover all substrate interactions that occur in human cells. Yeast lacks an endogenous *SPOP* ortholog and does not encode recognizable homologs of major mammalian SPOP substrates (such as ERG (84,85), TRIM24 (86), c-MYC (87), NANOG (88), PD-L1 (89), IRF1 (90), SRC-3 (27), DAXX (91), DEK (92), GLI2 (93), and AR (94)). However, yeast encodes a diverse repertoire of SCF/F-box

proteins, which operate as substrate adaptors in an analogous manner to mammalian substrate-recognition modules. Moreover, the core elements of the ubiquitin-proteasome system required for SPOP-mediated ubiquitination are evolutionarily conserved. Yeast expresses orthologues of Cullin3 (Cdc53), Rbx1 (Hrt1), and multiple E2 ubiquitin-conjugating enzymes (including Ubc4/Ubc5), which together form a functional CUL3-type Cullin-RING ubiquitin ligase (CRL) scaffold (SI Appendix, Fig. S3). Collectively, these conserved components provide a mechanistically compatible cellular environment that allows heterologous human SPOP to engage the ubiquitination machinery despite the absence of its native substrates.

Secondly, it would be also possible to use the library to assay against a specific substrate, oligomerization status, or formation of phase separation of SPOP (31,95). However, our assay also showed limitations in distinguishing neutral variants from GoF variants, as endometrial-associated variants enriched in ‘Tolerated’ classification. This likely reflects that the ‘Tolerated’ class contains a mixture of truly neutral variants and pathogenic variants acting through non-LoF mechanisms, which in turn reduces the interpretability of negative calls and contributes to weaker negative likelihood ratios (SI Appendix, Table 3). Therefore, ‘Tolerated’ calls should be interpreted with caution and may not be used as a filter for excluding potentially relevant variants.

Lastly, 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. This discrepancy may reflect distinct underlying molecular pathway in somatic versus germline contexts, which are associated with different disease symptoms (*i.e.*, somatic mutations cause tumors, while germline mutations cause neurodevelopmental disorder; OMIM #618828 and #618829).

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, individual variant scores are available at MaveDB (ID: urn:mavedb:00001258-a-1), and Mass spectrum based based proteomics data were deposited at PRIDE (ID: PXD067557, will be accessible upon publication). All codes for calculating scores in DMS models are available on Zenodo at https://doi.org/10.5281/zenodo.16787854.

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. 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.
4. 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.
5. Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J.K., Brock, K., Gal, Y. and Marks, D.S. (2021) Disease variant prediction with deep generative models of evolutionary data. *Nature*, **599**, 91-95.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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**.
18. 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.
19. 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.
20. Jamroze, A., Chatta, G. and Tang, D.G. (2021) Androgen receptor (AR) heterogeneity in prostate cancer and therapy resistance. *Cancer Lett*, **518**, 1-9.
21. 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.
22. Jiang, Z., van Vlimmeren, A.E., Karandur, D., Semmelman, A. and Shah, N.H. (2025) Deep mutational scanning of the multi-domain phosphatase SHP2 reveals mechanisms of regulation and pathogenicity. *Nat Commun*, **16**, 5464.
23. Radford, E.J., Tan, H.K., Andersson, M.H.L., Stephenson, J.D., Gardner, E.J., Ironfield, H., Waters, A.J., Gitterman, D., Lindsay, S., Abascal, F. *et al.* (2023) Saturation genome editing of DDX3X clarifies pathogenicity of germline and somatic variation. *Nat Commun*, **14**, 7702.
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. Geng, C., He, B., Xu, L., Barbieri, C.E., Eedunuri, V.K., Chew, S.A., Zimmermann, M., Bond, R., Shou, J., Li, C. *et al.* (2013) Prostate cancer-associated mutations in speckle-type POZ protein (SPOP) regulate steroid receptor coactivator 3 protein turnover. *Proc Natl Acad Sci U S A*, **110**, 6997-7002.
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. 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**.
30. Wang, D., Ma, J., Botuyan, M.V., Cui, G., Yan, Y., Ding, D., Zhou, Y., Krueger, E.W., Pei, J., Wu, X. *et al.* (2021) ATM-phosphorylated SPOP contributes to 53BP1 exclusion from chromatin during DNA replication. *Sci Adv*, **7**.
31. 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.
32. 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.
33. 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.
34. 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.
35. 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.
36. 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.
37. Lee, J., Park, S.K., Hwang, B.J. and Choi, M. (2025) CodonMutator: A Python-based Automated Oligonucleotide Design Framework for Deep Mutational Scanning Library Construction. *BMB Rep*.
38. 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.

39. 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.
40. 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.
41. 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.
42. Elsaesser, R. and Paysan, J. (2004) Liquid gel amplification of complex plasmid libraries. *Biotechniques*, **37**, 200, 202.
43. Bushnell, B. (2015) BBMap - sourceforge.net/projects/bbmap/.
44. 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.
45. 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.
46. 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.
47. 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.
48. 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.
49. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011*, **17**, 3.
50. Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094-3100.
51. Wisniewski, J.R. and Gaugaz, F.Z. (2015) Fast and sensitive total protein and Peptide assays for proteomic analysis. *Anal Chem*, **87**, 4110-4116.
52. von der Haar, T. (2007) Optimized protein extraction for quantitative proteomics of yeasts. *PLoS One*, **2**, e1078.
53. 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.
54. 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.
55. 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.

56. 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.
57. Nabais Sa, M.J., El Tekle, G., de Brouwer, A.P.M., Sawyer, S.L., Del Gaudio, D., Parker, M.J., Kanani, F., van den Boogaard, M.H., van Gassen, K., Van Allen, M.I. *et al.* (2020) De Novo Variants in SPOP Cause Two Clinically Distinct Neurodevelopmental Disorders. *Am J Hum Genet*, **106**, 405-411.
58. 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.
59. 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.
60. Schrodinger, LLC. (2015) The PyMOL Molecular Graphics System, Version 1.8.
61. 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.
62. Lobo, Z. (1984) Saccharomyces cerevisiae aldolase mutants. *J Bacteriol*, **160**, 222-226.
63. 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.
64. 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.
65. 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.
66. 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.
67. 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.
68. 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.
69. Gronbaek-Thygesen, M., Voutsinos, V., Johansson, K.E., Schulze, T.K., Cagiada, M., Pedersen, L., Clausen, L., Nariya, S., Powell, R.L., Stein, A. *et al.* (2024) Deep mutational scanning reveals a correlation between degradation and toxicity of thousands of aspartoacylase variants. *Nat Commun*, **15**, 4026.
70. Carr, C.R., Crawford, K.H.D., Murphy, M., Galloway, J.G., Haddox, H.K., Matsen, F.A.t., Andersen, K.G., King, N.P. and Bloom, J.D. (2024) Deep mutational

scanning reveals functional constraints and antibody-escape potential of Lassa virus glycoprotein complex. *Immunity*, **57**, 2061-2076 e2011.
71. Frank, F., Keen, M.M., Rao, A., Bassit, L., Liu, X., Bowers, H.B., Patel, A.B., Cato, M.L., Sullivan, J.A., Greenleaf, M. *et al.* (2022) Deep mutational scanning identifies SARS-CoV-2 Nucleocapsid escape mutations of currently available rapid antigen tests. *Cell*, **185**, 3603-3616 e3613.
72. Beltran, A., Jiang, X., Shen, Y. and Lehner, B. (2025) Site-saturation mutagenesis of 500 human protein domains. *Nature*, **637**, 885-894.
73. Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T. and Ben-Tal, N. (2016) ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res*, **44**, W344-350.
74. Shi, Q., Zhu, Y., Ma, J., Chang, K., Ding, D., Bai, Y., Gao, K., Zhang, P., Mo, R., Feng, K. *et al.* (2019) Prostate Cancer-associated SPOP mutations enhance cancer cell survival and docetaxel resistance by upregulating Caprin1-dependent stress granule assembly. *Mol Cancer*, **18**, 170.
75. Ma, J., Shi, Q., Cui, G., Sheng, H., Botuyan, M.V., Zhou, Y., Yan, Y., He, Y., Wang, L., Wang, Y. *et al.* (2021) SPOP mutation induces replication over-firing by impairing Geminin ubiquitination and triggers replication catastrophe upon ATR inhibition. *Nat Commun*, **12**, 5779.
76. 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.
77. 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.
78. 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.
79. 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.
80. 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.
81. 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**.
82. Hsu, P.D., Lander, E.S. and Zhang, F. (2014) Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, **157**, 1262-1278.

83. 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.
84. Gan, W., Dai, X., Lunardi, A., Li, Z., Inuzuka, H., Liu, P., Varmeh, S., Zhang, J., Cheng, L., Sun, Y. *et al.* (2015) SPOP Promotes Ubiquitination and Degradation of the ERG Oncoprotein to Suppress Prostate Cancer Progression. *Mol Cell*, **59**, 917-930.
85. An, J., Ren, S., Murphy, S.J., Dalangood, S., Chang, C., Pang, X., Cui, Y., Wang, L., Pan, Y., Zhang, X. *et al.* (2015) Truncated ERG Oncoproteins from TMPRSS2-ERG Fusions Are Resistant to SPOP-Mediated Proteasome Degradation. *Mol Cell*, **59**, 904-916.
86. Groner, A.C., Cato, L., de Tribolet-Hardy, J., Bernasocchi, T., Janouskova, H., Melchers, D., Houtman, R., Cato, A.C.B., Tschopp, P., Gu, L. *et al.* (2016) TRIM24 Is an Oncogenic Transcriptional Activator in Prostate Cancer. *Cancer Cell*, **29**, 846-858.
87. Geng, C., Kaochar, S., Li, M., Rajapakshe, K., Fiskus, W., Dong, J., Foley, C., Dong, B., Zhang, L., Kwon, O.J. *et al.* (2017) SPOP regulates prostate epithelial cell proliferation and promotes ubiquitination and turnover of c-MYC oncoprotein. *Oncogene*, **36**, 4767-4777.
88. Wang, X., Jin, J., Wan, F., Zhao, L., Chu, H., Chen, C., Liao, G., Liu, J., Yu, Y., Teng, H. *et al.* (2019) AMPK Promotes SPOP-Mediated NANOG Degradation to Regulate Prostate Cancer Cell Stemness. *Dev Cell*, **48**, 345-360 e347.
89. Zhang, J., Bu, X., Wang, H., Zhu, Y., Geng, Y., Nihira, N.T., Tan, Y., Ci, Y., Wu, F., Dai, X. *et al.* (2018) Cyclin D-CDK4 kinase destabilizes PD-L1 via cullin 3-SPOP to control cancer immune surveillance. *Nature*, **553**, 91-95.
90. Schwartz, I., Vunjak, M., Budroni, V., Cantoran Garcia, A., Mastrovito, M., Soderholm, A., Hinterndorfer, M., de Almeida, M., Hacker, K., Wang, J. *et al.* (2023) SPOP targets the immune transcription factor IRF1 for proteasomal degradation. *Elife*, **12**.
91. Kwon, J.E., La, M., Oh, K.H., Oh, Y.M., Kim, G.R., Seol, J.H., Baek, S.H., Chiba, T., Tanaka, K., Bang, O.S. *et al.* (2006) BTB domain-containing speckle-type POZ protein (SPOP) serves as an adaptor of Daxx for ubiquitination by Cul3-based ubiquitin ligase. *J Biol Chem*, **281**, 12664-12672.
92. Theurillat, J.P., Udeshi, N.D., Errington, W.J., Svinkina, T., Baca, S.C., Pop, M., Wild, P.J., Blattner, M., Groner, A.C., Rubin, M.A. *et al.* (2014) Prostate cancer. Ubiquitylome analysis identifies dysregulation of effector substrates in SPOP-mutant prostate cancer. *Science*, **346**, 85-89.
93. Chen, M.H., Wilson, C.W., Li, Y.J., Law, K.K., Lu, C.S., Gacayan, R., Zhang, X., Hui, C.C. and Chuang, P.T. (2009) Cilium-independent regulation of Gli protein function by Sufu in Hedgehog signaling is evolutionarily conserved. *Genes Dev*, **23**, 1910-1928.
94. Geng, C., Rajapakshe, K., Shah, S.S., Shou, J., Eedunuri, V.K., Foley, C., Fiskus, W., Rajendran, M., Chew, S.A., Zimmermann, M. *et al.* (2014) Androgen receptor is the key transcriptional mediator of the tumor suppressor SPOP in prostate cancer. *Cancer Res*, **74**, 5631-5643.

95. 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.

Figure Legends

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 fitness score of each variant was based on the enrichment ratio in the *SPOP*-expressed pool compared to the non-expressed pool. (B) Effect of galactose-inducible *SPOP* expression on growth of the yeast strain FY23. Cells transformed with CEN plasmids were spotted onto plates containing galactose to induce overexpression. (C) Volcano plot of proteins differentially expressed upon overexpression of wild-type *SPOP* . (D) Position of highlighted proteins in the simplified glycolysis pathway. Significantly down-regulated enzymes were highlighted with blue. 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.

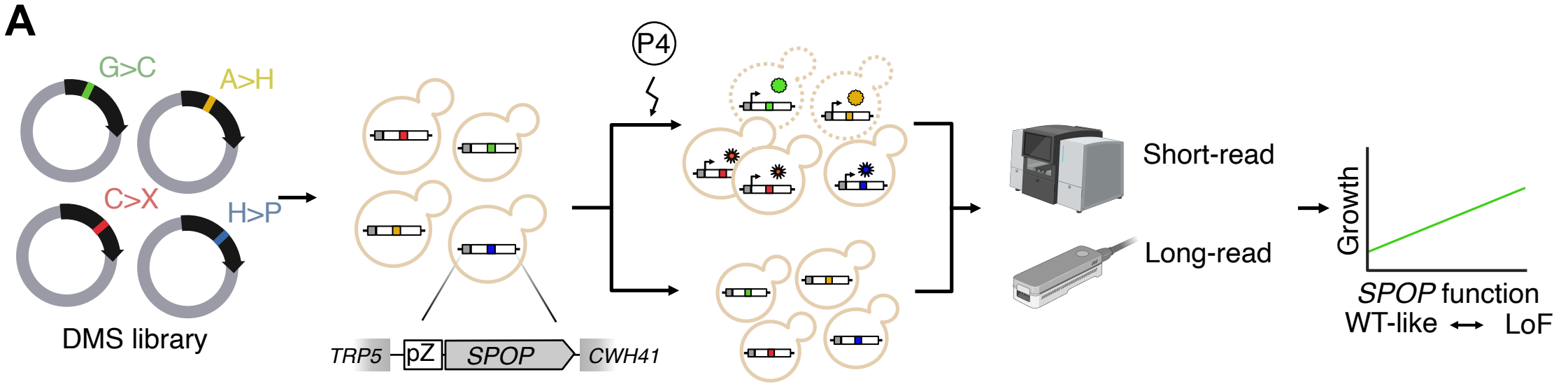
Figure 2. Generation of combined z-scores for the variant functional assessment. (A) Read coverage distributions of the reference template for short-read (left) and long-read (right) sequencing, and schematic illustration of read coverage patterns from each platform (center). Short-read sequencing showed a smooth, bell-shaped coverage profile across the ~1.5 kb amplicon, with reduced depth toward the 5′ and 3′ ends consistent with fragmentation during library preparation. In contrast, long-read sequencing exhibited a relatively uniform coverage profile across the full amplicon, indicating even amplicon representation. (B) Heatmap showing

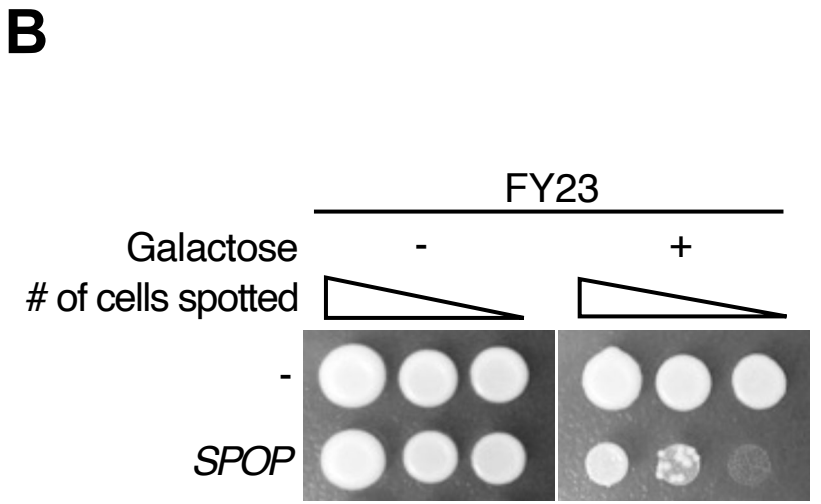
combined z-scores for the covered *SPOP* variants. The x-axis represents the residue number and the y-axis the alternative amino acid. SPOP functional domains are indicated along the top. Amino acids are color-coded based on their biochemical properties. (C) Violin plot displaying z-score distribution by variant impact. The cutoff of 'Likely LoF' is represented as red dashed line. (D) Scatter plot of EVE and ESM1b scores for variants. 'Likely LoF' variants are depicted in the red dotted rectangles. 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 for comparison with conservation score in the residue level. *P*-value denotes the significance of the trend in a linear regression (F) Scatter plot showing correlation of z-scores from short-read and long- and short-read data. (G) ESM1b and AlphaMissense scores of variants classified as 'Likely LoF' by a single sequencing platform. 'a' and 'b' represent variants exclusively assigned as 'Likely LoF' from short- and long-read platforms, respectively. Two-tailed t-test *P*-values are shown.

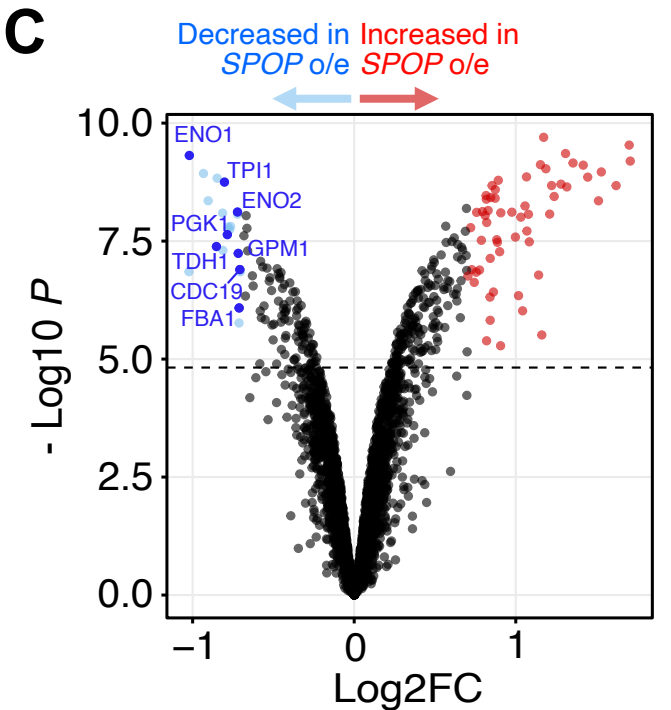
Figure 3. Validation of SPOP fitness scores. (A) Density plot of observed DMS variants and gnomAD variants (left). Scatter plot of z-scores with allele frequency in gnomAD (right). Variants on the far left part of the right panel are singletons. Variant impacts are distinguished by dot color; the linear 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 assigned as 'Likely LoF'; and the bottom panel the distribution of z-scores per tumor site. H&L, Hematopoietic and lymphoid. CNS, Central nervous system. (D) Structural placement diagrams of recurrent variant residues (count > 5) found in prostate and endometrial cancers. Residues of interest were shown as color-coded spheres by fraction of 'Likely LoF' varaint per residue (PDB ID: 3HQI). Only the MATH domain and the SBC

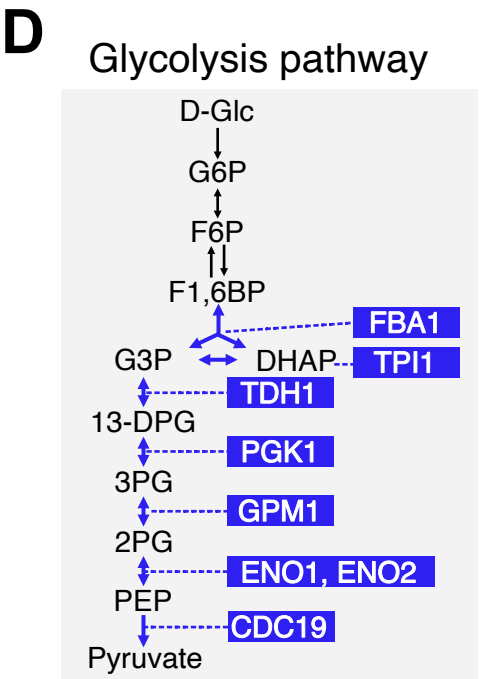
motf are displayed, represented as cartoons and yellow sticks, respectively. (E) Comparison of the fraction of 'Likely LoF' variants per residue value between buried or exposed regions. (F) Boxplots displaying the z-scores of variants reported in public databases or validated in previous studies. The plots are color-coded by source: grey for cohort-based data and green for variants from previous studies. (G) Yeast growth assay of variants associated with prostate cancer. Abbreviations: Deg., degradation; Polyubi., polyubiquitylation; Dissoc., dissociation.

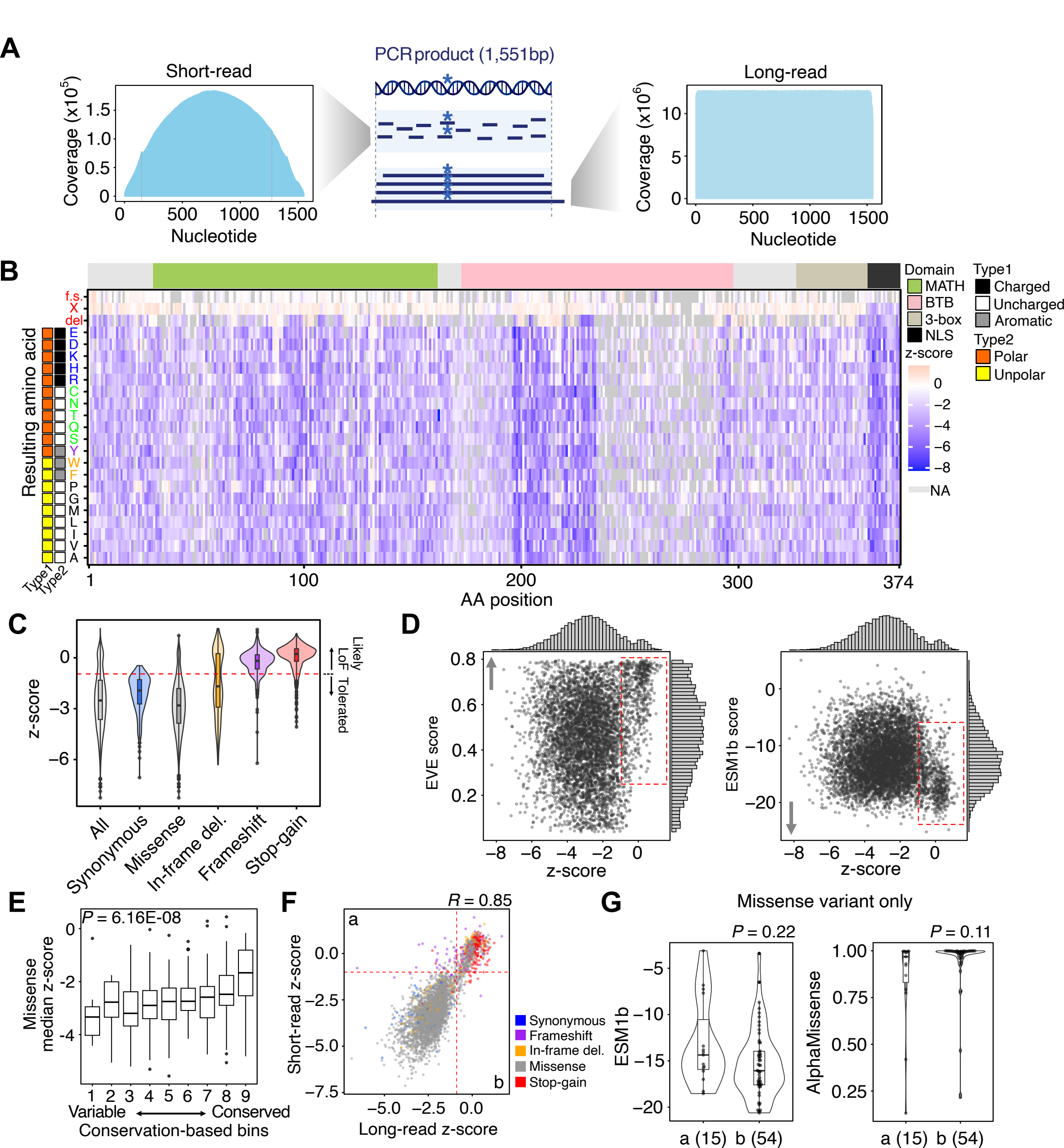
Figure 4. Interpretation of variant impact based on amino acid biochemical properties and structural context. (A) Dissection of COSMIC-overlapping residues for their z-score of the corresponding in-frame deletion on that position. (B) Stratification of in-frame deletion z-scores by residue solvent accessibility. (C) Residues with in-frame deletions are displayed as color-coded spheres on the oligomeric structure of SPOP. Colors indicate z-scores of in-frame deletions and "Chain" refers to SPOP monomer. (D) Residues exclusively enriched by variants assigned as 'Likely LoF' (left), 'Tolerated' variants with low scores (Tol.^{Low}; 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 1,000 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).

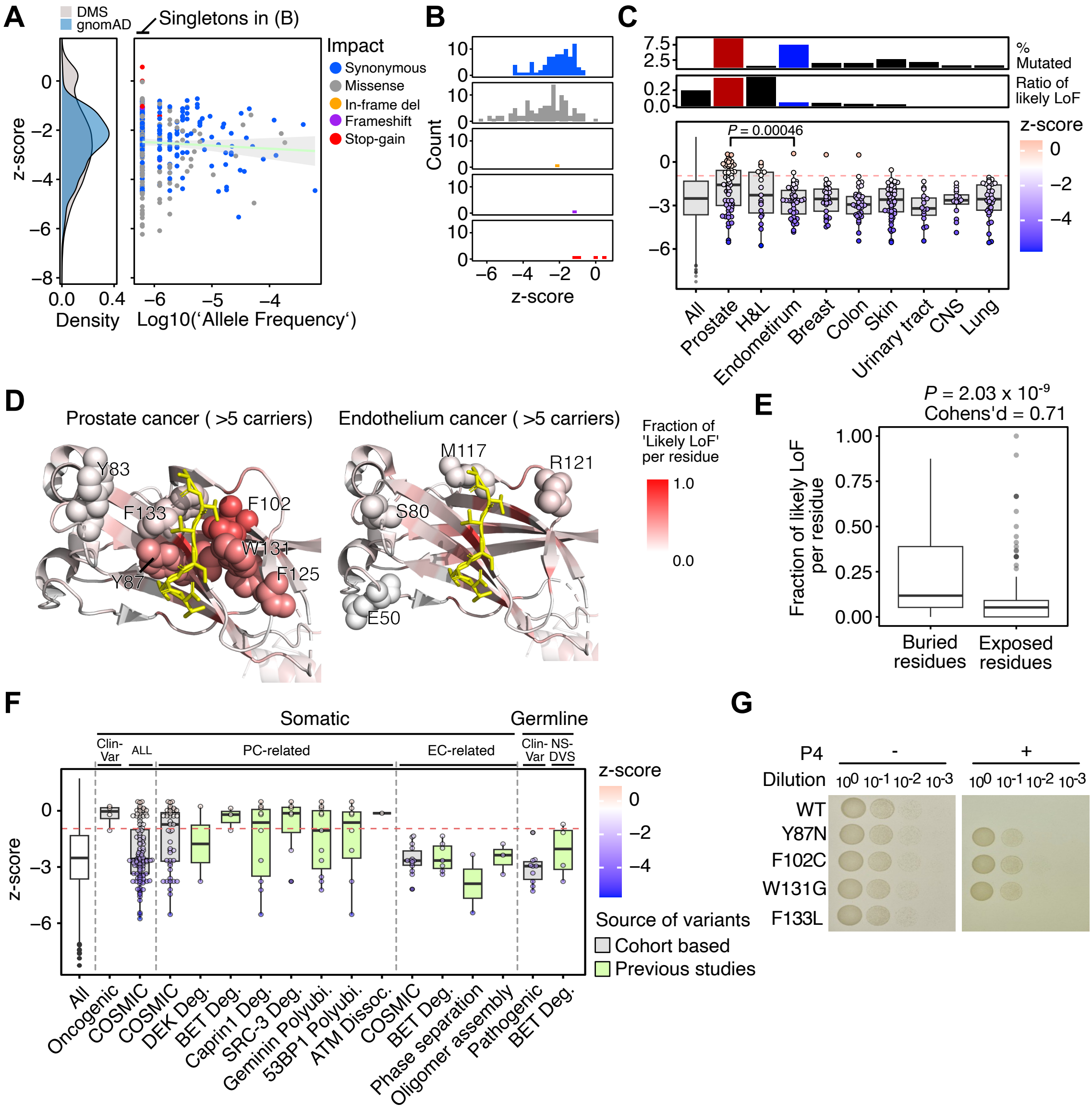

A
G>C
A>H
C>X
H>P
DMS library
TRP5
pZ
SPOP
CWH41
P4
Short-read
Long-read
Growth
SPOP function
WT-like
LoF

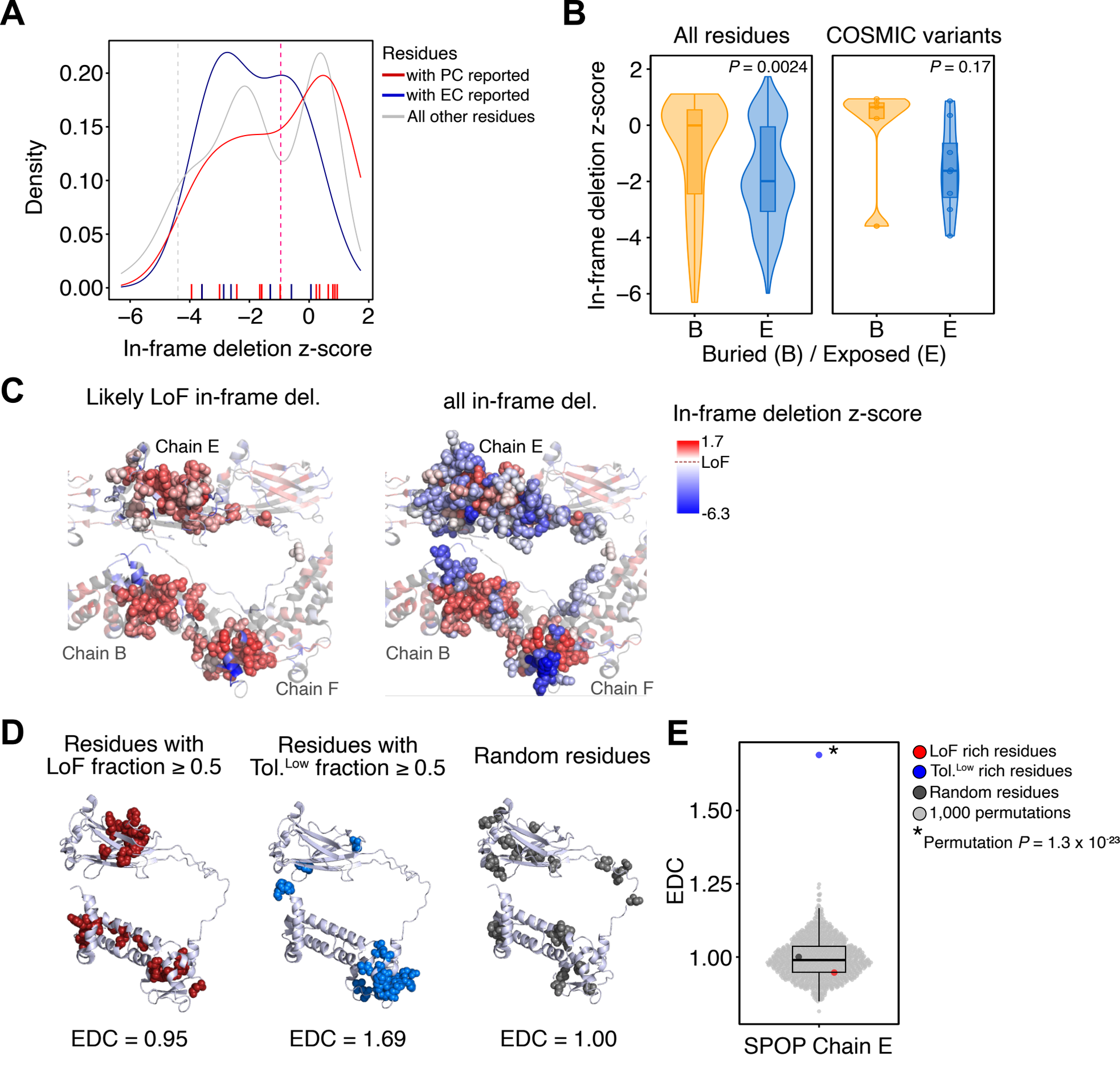

B
FY23
Galactose
-
+
# of cells spotted
-
SPOP


C
Decreased in SPOP o/e
Increased in SPOP o/e
ENO1
TPI1
ENO2
PGK1
GPM1
TDH1
CDC19
FBA1
10.0
7.5
5.0
2.5
0.0
- Log10 P
−1
0
1
Log2FC


D
Glycolysis pathway
D-Glc
G6P
F6P
F1,6BP
FBA1
G3P
DHAP
TPI1
TDH1
13-DPG
PGK1
3PG
GPM1
2PG
ENO1, ENO2
PEP
CDC19
Pyruvate


A
Short-read
Coverage (x10^5)
Nucleotide
PCR product (1,551bp)
Long-read
Coverage (x10^6)
Nucleotide
B
Domain
MATH
BTB
3-box
NLS
Type1
Charged
Uncharged
Aromatic
Type2
Polar
Unpolar
z-score
NA
Resulting amino acid
f.s.
X
del
Type1
Type2
AA position
C
z-score
Likely LoF
Tolerated
All
Synonymous
Missense
In-frame del.
Frameshift
Stop-gain
D
EVE score
z-score
ESM1b score
z-score
E
P = 6.16E-08
Missense median z-score
Variable
Conserved
Conservation-based bins
F
R = 0.85
Short-read z-score
Long-read z-score
Synonymous
Frameshift
In-frame del.
Missense
Stop-gain
G
Missense variant only
P = 0.22
ESM1b
a (15)
b (54)
P = 0.11
AlphaMissense
a (15)
b (54)


A
DMS
gnomAD
Singletons in (B)
Impact
Synonymous
Missense
In-frame del
Frameshift
Stop-gain
z-score
Density
Log10('Allele Frequency')
B
Count
z-score
C
% Mutated
Ratio of likely LoF
P = 0.00046
z-score
All
Prostate
H&L
Endometirum
Breast
Colon
Skin
Urinary tract
CNS
Lung
D
Prostate cancer (>5 carriers)
Endothelium cancer (>5 carriers)
Y83
F133
F102
W131
Y87
F125
M117
R121
S80
E50
Fraction of 'Likely LoF' per residue
1.0
0.0
E
P = 2.03 x 10-9
Cohens'd = 0.71
Fraction of likely LoF per residue
Buried residues
Exposed residues
F
Somatic
Germline
Clin-Var
ALL
PC-related
EC-related
Clin-Var
NS-DVS
z-score
All
Oncogenic
COSMIC
COSMIC
DEK Deg.
BET Deg.
Caprin1 Deg.
SRC-3 Deg.
Geminin Polyubi.
53BP1 Polyubi.
ATM Dissoc.
COSMIC
BET Deg.
Phase separation
Oligomer assembly
Pathogenic
BET Deg.
Source of variants
Cohort based
Previous studies
G
P4
Dilution
WT
Y87N
F102C
W131G
F133L


A
Density
0.20
0.15
0.10
0.05
0.00
−6
−4
−2
0
2
In-frame deletion z-score
Residues
with PC reported
with EC reported
All other residues
B
All residues
COSMIC variants
P = 0.0024
P = 0.17
In-frame deletion z-score
2
0
−2
−4
−6
B
E
B
E
Buried (B) / Exposed (E)
C
Likely LoF in-frame del.
all in-frame del.
In-frame deletion z-score
1.7
LoF
-6.3
Chain E
Chain B
Chain F
Chain E
Chain B
Chain F
D
Residues with LoF fraction ≥ 0.5
Residues with Tol.Low fraction ≥ 0.5
Random residues
EDC = 0.95
EDC = 1.69
EDC = 1.00
E
EDC
1.50
1.25
1.00
SPOP Chain E
LoF rich residues
Tol.Low rich residues
Random residues
1,000 permutations
* Permutation P = 1.3 x 10-23