STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
CPTAC clinical and proteomic data	Li et al.[37,160]	Proteomic Data Commons (PDC: https://pdc.cancer.gov/pdc/cptac-pancancer)
CPTAC genomic, transcriptomic data	Li et al.[37,160]	Proteomic Data Commons (PDC:https://pdc.cancer.gov/pdc/cptac-pancancer), and Cancer Data Service (CDS: https://dataservice.datacommons.cancer.gov/)
CPTAC precision proteogenomics data	This manuscript	Cancer Data Service (CDS: https://dataservice.datacommons.cancer.gov/)
Software and algorithms		
AlphaFold Protein Structure Database (AlphaFoldDB) v4	Varadi et al.[75] and Jumper et al.[76]	https://alphafold.ebi.ac.uk/
Ancestry prediction	Li Ding Lab	https://github.com/ding-lab/ancestry
bam-readcount v0.7.4 and v0.8	McDonnell Genome Institute	https://github.com/genome/bam-readcount
BWA v0.7.17-r1188	Li[161]	http://bio-bwa.sourceforge.net/
CharGer v0.5.4	Scott et al.[33]	https://github.com/ding-lab/CharGer/tree/v0.5.4
clusterProfiler v4.4.2	Wu et al.[162]	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
DNAScope	Freed et al.[163]	https://doi.org/10.1101/115717
Ensembl Variant Effect Predictor (VEP) v100	McLaren et al.[164]	https://github.com/Ensembl/ensembl-vep
FastQC v0.11.8	Andrews[165]	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
GATK DepthOfCoverage v3.8-0	McKenna et al.[166]	https://github.com/broadinstitute/gatk
GATK HaplotypeCaller v4.0.0.0	McKenna et al.[166]	https://github.com/broadinstitute/gatk
GATK VariantEval v3.8-0	McKenna et al.[166]	https://github.com/broadinstitute/gatk
GermlineWrapper pipeline v1.1	Li Ding Lab	https://github.com/ding-lab/germlinewrapper
GLIMPSE v2.0.0	Rubinacci et al.[27]	https://github.com/odelaneau/GLIMPSE
HotSpot3D/HotPho v1.8.2 for PDB	Li Ding Lab; Niu et al.[72]	https://github.com/ding-lab/hotspot3d/tree/ding_lab_internal
HotSpot3D/HotPho for AlphaFoldDB	Li Ding Lab; Niu et al.[72]	https://github.com/ding-lab/hotspot3d/tree/alphaFold_implementation
Integrative Genomics Viewer (IGV) v2.8.2	Robinson et al.[167]	https://software.broadinstitute.org/software/igv/
ImmuneRegulation	Kalayci et al.[168]	https://immuneregulation.mssm.edu/
Matrix eQTL	Andrey A. Shabalin	https://CRAN.R-project.org/package=MatrixEQTL
Mosdepth v0.2.4	Pedersen and Quinlan[169]	https://github.comim/brentp/mosdepth
Mutect v1.7.7	Cibulskis et al.[170]	https://github.com/broadinstitute/mutect
Picard Toolkit v2.22.4-0	Broad Institute of MIT and Harvard	https://github.com/broadinstitute/picard
Pindel v0.2.5	Ye et al.[171]	https://github.com/genome/pindel
Pymol v2.5.4	The PyMOL Molecular Graphics System, Version 2.5.4, Schrödinger, LLC.	https://pymol.org/2/
Python v2.7 and v3.7	Python Software Foundation	https://www.python.org/
QUILTS v3	Ruggles et al.[29]	https://quilts.fenyolab.org

(*Continued on next page*)

Please cite this article in press as: Martins Rodrigues et al., Precision proteogenomics reveals pan-cancer impact of germline variants, Cell (2025), https://doi.org/10.1016/j.cell.2025.03.026

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
R v4.0.3	R Development Core Team	https://www.R-project.org
Spectrum Mill (SM) v7.08	Broad Institute of MIT and Harvard	https://proteomics.broadinstitute.org
Strelka v2.9.2	Kim et al.[172]	https://github.com/Illumina/strelka
STRINGdb v11.5	Szklarczyk et al.[173]	https://www.string-db.org
survminer R package v0.4.9	Kassambara and Kosinski[174]	https://github.com/kassambara/survminer
RCSB Protein Data Bank (RCSB PDB) as of June 24th, 2021	Berman et al.[77]; Berman et al.[78]	https://www.rcsb.org/
TransVar v2.5.10.20211024	Zhou et al.[175]	https://github.com/zwdzwd/transvar
Uniprot Knowledge Base v2023_01	The Uniprot Consortium[79]	https://www.uniprot.org/
VarScan v2.3.8	Koboldt et al.[176]	https://dkoboldt.github.io/varscan/
vcf2maf	Kandoth et al.[177]	https://doi.org/10.5281/zenodo.593251

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Human subjects

This study includes samples from a total of 1,064 participants for which biospecimens were prospectively collected (tumor, germline blood, and adjacent normal samples when possible) from more than 30 tissue source sites both domestically and internationally. All samples were processed by a central biospecimen core resource following a tumor type specific protocol and standard operating procedures (SOPs). Pathology for all samples was verified by a general pathologist and reviewed by a disease-specific expert pathologist using histopathology and immunohistochemistry assays. Full details appear in our Pan-Cancer Data and Resource manuscript[160] and Pan-Cancer Driver manuscript.[37]

Clinical data annotation

Clinical data including sex at birth, age, and self-reported ancestry, race, and ethnicity information can be obtained from the CPTAC Data Portal and https://pdc.cancer.gov/pdc/cptac-pancancer. Full details appear in our Pan-Cancer Data and Resource manuscript[160] and Pan-Cancer Driver manuscript.[37]

METHOD DETAILS

Harmonized genome alignment

WGS, WES, and RNA-Seq sequence data were harmonized by NCI Genomic Data Commons (GDC) (https://gdc.cancer.gov/about-data/gdc-data-harmonization) based on GDC's HRCh38 human reference genome (GRCh38.d1.vd1), as described in the Pan-Cancer Data and Resource and Pan-Cancer Driver manuscripts.[37,160]

Germline variant calling and filtering from WES

WES data from 1,093 normal samples from all 10 cancer types were initially collected for this project. After pathology and clinical review, 1,064 cases were selected and assessed for quality using FastQC (version 0.11.8 with default parameters).[165] Coverage within target regions was calculated using Mosdepth[169] (version 0.2.4 with default parameters, except where -Q 20). Coverage ranged from 105X - 357X (Figure S1A). All 1,064 samples passed quality control criteria and had >20X average coverage (mapping quality $\geq$ 20) across target regions.

As described in our Pan-Cancer Data and Resource and Pan-Cancer Driver manuscripts,[37,160] germline variants for samples passing quality control criteria were identified using the GermlineWrapper pipeline (v1.1; https://github.com/ding-lab/germlinewrapper), which integrates multiple tools for the identification of germline SNVs and indels. SNVs were detected with VarScan[176] (version 2.3.8 with default parameters, except where –min-var-freq 0.08,–p value 0.10,–min-coverage 3,–strand-filter 1, -min-avg-qual 15, -min-reads2 2, -min-freq-for-hom 0.75) operating on a mpileup stream from SAMtools (version 1.2 with default parameters, except where -q 1 -Q 13) and GATK[166] (version 4.0.0.0, using its Haplotype Caller in single-sample mode excluding duplicate and unmapped reads and retaining calls with a minimum quality of 10). Germline indels were identified using VarScan (version and parameters as above), GATK (version and parameters as above) in single-sample mode, and Pindel[171] (version 0.2.5b9 with default parameters, except where -m 6, -w 1, and excluded centromere regions (genome.ucsc.edu)). We used the GRCh38 reference genome and specified an insertion size of 500 whenever this information was not provided in the BAM header. Single nucleotide variants (SNVs) were based on the union of raw GATK and VarScan calls. We required that indels were called by Pindel or at least two out of the three callers (GATK, VarScan, Pindel). Cutoffs of minimal 10X coverage and 20% VAF were used in the final step to report the high-quality germline variants.

Please cite this article in press as: Martins Rodrigues et al., Precision proteogenomics reveals pan-cancer impact of germline variants, Cell (2025), https://doi.org/10.1016/j.cell.2025.03.026

Variants called by GermlineWrapper were required to have an Allelic Depth (AD) ≥ 5 for the alternative allele. Additionally, we filtered out any indels longer than 100bp. A total of 185,724,997 variants passed these filters (Figure 2A). Variants were also filtered based on coding regions of full-length transcripts obtained from Ensembl release 100 (Gencode v34) plus the additional two base pairs flanking each exon that cover splice donor/acceptor sites, resulting in a total of 27,104,152 germline exonic variants across 1,064 samples, or 563,036 unique variants (Figure 2A).

Finally, variants passing filters were assessed for quality by calculating concordance with dbSnP (release 151) and average transition-transversion (TiTv) ratio using GATK's[166] VariantEval tool (v3.8-0 with default parameters). We achieved 97.43% concordance with dbSnP, and our germline exomes displayed high quality, with an average TiTv ratio of 2.74. All VCF files were converted to MAF format using vcf2maf[177] with VEP Ensembl v100 annotation.

It is important to clarify that a total of 27,838,075 germline exonic variants (570,645 unique variants) were originally called for the initial number of 1,093 patients, which were used as inputs for the generation of the precision peptidomics dataset before the cohort was reduced to 1,064 patients (see STAR Methods: proteomics LC-MS/MS data interpretation section for more details). That is the only section of the manuscript where the larger cohort was used as input. Results reported everywhere in the manuscript, however, only focus on events detected in the final cohort of 1,064 patients.

Somatic mutation and copy number variant calling from WES

Full details appear in the Pan-Cancer Data and Resources and Pan-Cancer Driver manuscripts.[37,160]

Germline variant calling and filtering from WGS

We performed germline variant calling on WGS of blood derived samples from CCRCC, GBM, HNSCC, LSCC, LUAD, PDAC and UCEC patients using DNAScope.[163] Briefly, we implemented a pipeline based on the GATK best-practices and functional equivalence recommendations. We first aligned the raw paired-end WGS FASTQ files to the latest human genome build GRCh38 (GDC GRCh38.d1.vd1 version) using bwa-mem,[161] and then performed duplicate marking. Next, we called variants producing one gVCF file per-sample, using the DNAScope[163] Haplotyper with '–emit_mode gvcf' using default setting. Next, we genotyped the samples at a set of high quality variants from 2,504 unrelated samples from Phase 3 of the 1000 Genomes Project, which were resequenced to a depth of 30X by the New York Genome Center (NYGC).[28] Finally, to account for the low WGS depth of the CPTAC samples, we used GLIMPSE[27] with default settings to perform genotype imputation and phasing with the same NYGC 1000 Genomes Project genome as the reference panel.

Comparison of WES and WGS variant calls

As a part of quality control, we compared the WES and WGS germline variant calls for seven cancer types (CCRCC, GBM, HNSCC, LSCC, LUAD, PDAC and UCEC) for which WGS data was available. Towards this end, we studied the variants in the common regions between WES and WGS in chromosomes 1-22. In these regions, we identified the number of WES variants and WGS variants and finally estimated the number of common WES and WGS variants sharing the same genotype (Figure S1E; Table S1D).

Ancestry prediction

We identified likely ancestry for each individual in the CPTAC dataset based on WES data using an in-house random forest classifier for genetic ancestry (https://github.com/ding-lab/ancestry). By using a reference panel of genotypes and clustering based on principal components, we selected a set of 107,765 coding SNPs with minor allele frequency (MAF) > 0.02 from the 1000 Genomes Project[178] (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/) and measured their depth and allele counts in each sample using bam-readcount (version 0.8 with default parameters, https://github.com/genome/bam-readcount). Following, we genotyped each sample as follows: 0/0 if reference allele count ≥ 8 and alternative allele count < 4; 0/1 if reference allele count ≥ 4 and alternative allele count ≥ 4; 1/1 if reference allele count < 4 and alternative allele count ≥ 8; and./. (missing) otherwise. Further, we filtered out markers with missingness > 5%, after which 70,049 markers remained for analysis. We performed principal component analysis (PCA) for each group of markers on the 1000 Genomes Project data to identify the top 20 principal components and projected our cohorts onto the 20-dimensional space representing the 1000 Genomes data. We then trained a random forest classifier with the 1000 Genomes dataset using the 20 principal components we identified, splitting the 1000 Genomes datasets 80/20 for training and validation, respectively. Our classifier achieved 99.6% accuracy on the validation dataset using models trained with the elected markers. The fitted classifiers were then used to classify samples into African (AFR), Ad Mixed American (AMR), East Asian (EAS), European (EUR), or South Asian (SAS). Due to the absence of individuals of Slavic origin in the training dataset, our model misclassified 9 individuals in the GBM, HNSCC, LSCC, PDAC, and UCEC cohorts into the AMR ancestry. This was confirmed using the available WGS data for these samples, for which we performed PCA using the EIGENSOFT software with the 1000 Genomes reference dataset[28] to estimate ancestry, as described in the Pan-Cancer Data and Resource and Pan-Cancer Driver manuscripts.[37,160] Briefly, for this analysis, we used common variants with a call rate of at least 0.99 and inferred the ancestry of each participant by visualizing the PCA plot and selecting cutoffs on principal components 1 through 10 corresponding to the five major populations. We then successfully classified the 9 individuals into the EUR group (Figure S1F). For the purposes of downstream analyses including principal component values of ancestry, we have excluded those individuals in order to consistently use the principal components based on WES data.

Gene list curation for pathogenic variant classification

We extended the list of 152 cancer predisposition genes (CPGs) previously compiled by Huang et al.[16] to a total of 160 CPGs, by adding 8 genes that contribute to cancer susceptibility based on literature review. This extended gene list of 160 genes was used as input for our tool CharGer[33] (described below) using the *–inheritanceGeneList* parameter. The source and reference for each curated predisposition gene are provided in Table S1C. This list of 160 CPGs is used throughout the study for multiple analyses.

Inference of the ancestral state of germline variants

To avoid potential confusion due to unclear major and minor allele status, that at many variants may vary across human cohorts of different ancestries, we have derived the ancestral status information for the 27,104,152 exonic germline variant calls from WES data in order to polarize according to conservation and assign their effects referring to the novel allele by default. To infer the ancestral state of each germline variant in our call set, we took advantage of the *AncestralAllele.pm* plugin provided within the Ensembl's Variant Effect Predictor (VEP) tool[164] (release 100), which retrieves ancestral allele sequences from a FASTA file for each base position in the input VCFs. These sequences are based on the Ensembl Compara ancestral sequences for *Homo sapiens* (GRCh38) corresponding to Ensembl release 100 and are created using the Enredo-Pecan-Ortheus (EPO) multiple sequence alignment method for inference of ancestor alignments based on sequences from multiple primates: human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), bonobo (*Pan paniscus*), gorilla (*Gorilla gorilla gorilla*), orangutan (*Pongo abelli*), gibbon (*Nomascus leucogenys*), vervet-AGM (*Chlorocebus sabaeus*), crab-eating macaque (*Macaca fascicularis*), macaque (*Macaca mulatta*), mouse lemur (*Microcebus murinus*).[179,180]

We then parsed and assigned the ancestral state of our germline variants, consisting of three different cases. First, for those variants where an ancestral state status was not available, or those insertions, deletions, and oligonucleotide variants in which the inferred ancestral allele did not match any of the two alleles called in our study, the polarization state was left undetermined. Second, for variants in which the ancestral sequence matched the reference allele called from the *Homo sapiens* reference genome, the assigned ancestral state was *ancestral* (i.e. the *Homo sapiens* reference allele is the same as the ancestral allele and the *Homo sapiens* alternative allele is the derived allele). Third, for variants in which the ancestral sequence matched the alternative allele called from the *Homo sapiens* reference genome, the assigned ancestral status was *derived* (i.e. the *Homo sapiens* alternative allele is the same as the ancestral allele and the *Homo sapiens* reference allele is now the derived allele). Although the reference allele generally corresponds to both the ancestral allele and the major allele for most variants in the human genome, by polarizing our analyses to describe the effects of the derived (novel) allele for all variants instead of the major/minor status particular to our cohort, our procedure ensures the evolutionary interpretation and eases future transferability across cohorts of different ancestries. In summary, across our analyses we refer to ancestral (ANC) and derived (DER) alleles instead of major and minor alleles, respectively.

QUANTIFICATION AND STATISTICAL ANALYSIS

Pathogenicity assessment of rare germline variants

Germline variants called with GermlineWrapper were annotated with the Ensembl Variant Effect Predictor (VEP)[164] (version 100 with default parameters, except where –everything) and their pathogenicity was determined (as described in the Pan-Cancer Data and Resource and Pan-Cancer Driver manuscripts.[37,160]) with our automatic pipeline CharGer[33] (version 0.5.4 with default CharGer scores, https://github.com/ding-lab/CharGer/tree/v0.5.4), which prioritizes variants based on the guidelines by the American College of Medical Genetics and Genomics - Association for Molecular Pathology (ACMG-AMP).[181] CharGer retrieves information from the ClinVar (release as of 08/15/2019 parsed using codes from MacArthur lab ClinVar, https://github.com/macarthur-lab/clinvar) and gnomAD[182] (r2.1.1) databases, as well as computational tools, including SIFT[183] (v5.2.2) and Polyphen[184] (v2.2.2), to inform the implementation of 12 pathogenic and 4 benign evidence levels for the classification of germline variants. The detailed implementation and score of each evidence level, as well as parameters used are as previously described.[16]

We further selected rare variants with $\leq$ 0.05% allele frequency (AF) in gnomAD (r2.1.1) or 1000 Genomes.[178] We also performed read count analysis using bam-readcount (https://github.com/genome/bam-readcount; version 0.8 with parameters -q 10, -b 15) to evaluate the number of reference and alternative alleles for each variant. We required variants to have at least 5 counts of the alternative allele and a variant allele frequency (VAF) of at least 20% in both tumor and normal samples. Variants remaining after these filters were manually reviewed with the Integrative Genomics Viewer (IGV) software[167] (v2.8.2). We considered variants to be pathogenic (P) if they were known pathogenic variants in ClinVar; likely pathogenic (LP) if CharGer score > 8; and prioritized variant of uncertain significance (PVUS) if CharGer score > 4. A list of all variants passing manual review and their information are displayed in Table S2.

Burden testing analyses of rare P/LP germline variants

We performed burden testing of rare P/LP variants using the Total Frequency Test (TFT),[185] which applies a one-sided Fisher test to detect genes enriched for P/LP variants in the combined set of samples from TCGA and CPTAC cohorts versus controls. For this, we collapsed P/LP germline variants detected in the same gene and applied the total allele counts of P/LP variants identified in the gnomAD (r2.1.1) non-cancer cohort (n=118,479) using the CharGer pipeline described above as controls. We also tested burden for each cancer type and each gene using all other cancer types as controls, subtracting out the cohorts with suggestive enrichment for the specific gene in the gnomAD analyses. We used the standard Benjamini-Hochberg procedure to adjust the resulting p-values to FDR. We defined significant events if FDR $\leq$ 0.05, and suggestive events if FDR $\leq$ 0.15.

LOH analysis of rare P/LP germline variants

Analysis of loss-of-heterozygosity (LOH) events can help identify germline variants that are positively selected in the tumor by comparing the VAF in the tumor to that in the normal. We first estimated read counts for each variant in both normal and tumor samples for our CPTAC cases using bam-readcount (https://github.com/genome/bam-readcount) (v0.8 with parameters -q 10, -b 15). Then, LOH events were identified using a one-tailed Fisher's exact test between tumor and matched normal samples to detect germline variants for which VAF in the tumor was significantly higher than the VAF in the matched normal. The resulting p-values were adjusted to FDR using the Benjamini-Hochberg procedure. We considered LOH to be significant if FDR $\leq$ 0.05 and suggestive if FDR $\leq$ 0.15.

Proteomics LC-MS/MS data interpretation

MS/MS spectra from all omes were interpreted from using Spectrum Mill (SM) v7.08 (proteomics.broadinstitute.org) to provide identification and relative quantitation at the protein, peptide, and post-translational modification (PTM) site (phospho and acetyl) site levels.

Precision sequence databases

For searching with LC-MS/MS datasets from all omes we generated a cohort-level precision protein sequence database for each tumor type starting with a base human reference proteome to which we appended non-redundant somatic mutations and germline variants and indels for each of the ~100 participants/cohort. The base proteome consisted of the human reference proteome Gencode v34 (ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_34/) with 47,429 non-redundant protein coding transcript biotypes mapped to the human reference genome GRCh38, 602 common laboratory contaminants, 2,043 curated smORFs (lncRNA and uORFs), 237,427 novel unannotated ORFs (nuORFs) supported by ribosomal profiling nuORF DB v1.0[186] for a total of 287,501 entries. The personalized protein sequence entries were prepared by processing the individual participant's somatic and germline variant calls from whole exome sequencing data, described above, using QUILTS v3[29] with no further variant quality filtering using an Ensembl v100 reference proteome and reference genome for sequence identifiers consistent with the variant calling. Gencode v34 is a contemporaneous subset of Ensembl v100 (March 2020). From the unique germline variants originally called for the initial cohort of 1,093 participants (570,645 unique variants) in the 10 cohorts (see above STAR Methods: germline variant calling and filtering from WES), 342,311 unique coding, non-synonymous germline SAAVs and indels from the standard chromosomes (1-22, X, and Y) were mapped into proteins in the Gencode v34 reference proteome to use for peptide searches (337,469 unique variants pertain to the final cohort of 1,064 patients). A total of 232,228 unique somatic variants and indels were similarly mapped into Gencode v34. It is important to note that numbers of unique germline variants reported elsewhere in the manuscript are slightly different than what is reported in this section of STAR Methods because the number of patients included in the study was reduced to 1,064 after clinical and pathology review. The rest of the manuscript only reports results focusing on this final cohort. The range of germline variant counts across tissues (75K UCEC to 108K BRCA) was considerably narrower than for somatics (5K PDAC to 57K UCEC). Germline variants were also much more frequently shared amongst multiple participants in a cohort (4.6% UCEC to 51% BRCA) than somatics (0.02% PDAC to 0.37% COAD). Using the SM Protein Database utilities the base reference proteome and individual patient proteomes were combined and redundancy removed to produce a cohort-level protein sequence database and a variant summary table to enable subsequent mapping of sequence variants identified in Tandem Mass Tag (TMT) multiplexed LC-MS/MS datasets back to individual patients. After accounting for unique tryptic peptides of length 8-40 with 0 missed cleavages the LUAD search space is 58% Gencode v34 reference proteome, 39% nuORFs, and 9% germline/somatic variants. Other cohorts in the study had greater or lesser germline/somatic content in proportion to the sample size.

Each somatic and germline variant is included in the database by way of a full length copy of its reference protein sequence with a single AA change to retain the positional location within the full-length protein. Driven by the germline variants, the average redundancy of each tryptic peptide rises from 1.9 fold for the reference proteome alone to a range of 7 to 10 fold for these 10 cohort-level precision databases.

Our Spectrum Mill workflow incorporates 3 features that mitigate a search bogging down due to the wild type peptide redundancy. 1) A single copy of an SAAV containing protein is included in the sequence database when it occurs in multiple patients in the cohort. 2) Matches to wild-type peptides in the personalized protein entries are not reported, otherwise the results would be swamped with all the protein identifiers from the germline entries. 3) While SM's search engine digests all protein entries when processing a database during a search, peptide spectrum matching with an MS/MS spectrum is done on only one copy of each peptide by constructing and consulting a hash of all tryptic peptides and the protein identifiers in which they occur.

Spectrum quality filtering

For all omes, similar MS/MS spectra with the same precursor m/z acquired in the same chromatographic peak were merged. The precursor MH+ inclusion range was 800-6,000, and the spectral quality filter was a sequence tag length > 0 (i.e., minimum of two peaks separated by the in-chain mass of an amino acid).

MS/MS search conditions

Using the SM MS/MS search module for all omes parameters included: ‘‘trypsin allow P’’ enzyme specificity with up to 4 missed cleavages; precursor and product mass tolerance of ± 20 ppm; 30% minimum matched peak intensityScoring parameters were ESI-QEXACTIVE- HCD-v2, for whole proteome datasets, and ESI-QEXACTIVE-HCD-v3, for phosphoproteome and acetylome. Allowed fixed modifications included carbamidomethylation of cysteine and selenocysteine. TMT labeling was required at lysine, but peptide N-termini were allowed to be either labeled or unlabeled. Allowed variable modifications for whole proteome datasets were acetylation of protein N-termini, oxidized methionine, deamidation of asparagine, hydroxylation of proline in PG motifs, pyro-glutamic acid at peptide N-terminal glutamine, and pyro-carbamidomethylation at peptide N-terminal cysteine with a precursor MH+ shift range of -18 to 97 Da. For all PTM-omes variable modifications were revised to omit hydroxylation of proline and allow deamidation only in NG motifs. The phosphoproteome was revised to allow phosphorylation of serine, threonine, and tyrosine with a precursor MH+ shift range of -18 to 272 Da. The acetylome was revised to allow acetylation of lysine with a precursor MH+ shift range of -400 to 70 Da.

We used the *in silico* enzyme specificity, trypsin allow P, which omits the followed by P exception, to accommodate our two protease lysis/digestion protocol (Lys-C/Trypsin). While trypsin cleaves after lysine (K) and arginine (R) except when followed by proline (P), Lys-C cleaves after K with a reduced capacity for cleaving at KP. Extending to 4 missed cleavages (rather than a more conventional 2 for a trypsin only protocol) accommodates RP sites being counted as missed cleavages where they otherwise would not be with an ordinary trypsin specificity.

The allowed modifications were included because they are quite common in studies of this type, and failure to allow for them will otherwise lead to loss of these identifications with some of the spectra for those modified peptides becoming lower scoring false-positive identifications to an unmodified peptide. Modifications with positional constraints contribute tiny increases in search space size: <1.05 fold for pyro-Q/C at peptide N-termini, < 1.01 - fold for protein N-terminal acetylation, and <1.05 fold for hydroxyproline at PG sites, typically detected only in collagen domains of proteins. With Spectrum Mill the N-terminal modifications could not be considered without also allowing peptide N-termini to lack a TMT label (2-fold increase in search space size). The TMT labeling in our studies is >90% complete, with incomplete labeling mostly at N-termini rather than lysine due to the difference in reactivity of the primary amines at those sites. Search space size increases >2-fold due to each phosphorylation at serine(S), threonine(T), and tyrosine(Y), acetylation of lysine, oxidation at methionine(M), and deamidation at asparagine(N). In phosphoproteome and acetylome searches the deamidation contribution is diminished to ∼1.05 fold increase with the chemically favored positional constraint of an NG motif.

PTM site localization

Using the SM Autovalidation and Protein/Peptide Summary modules for the PTM-ome datasets results were filtered and reported at the phospho and acetyl site levels. When calculating scores at the variable modification (VM) site level and reporting the identified VM sites, redundancy was addressed in SM as follows: a VM-site table was assembled with columns for individual TMT-plex experiments and rows for individual VM-sites. PSMs were combined into a single row for all non-conflicting observations of a particular VM-site (e.g., different missed cleavage forms, different precursor charges, confident and ambiguous localizations, and different sample-handling modifications). For related peptides, neither observations with a different number of VM-sites nor different confident localizations were allowed to be combined. Selecting the representative peptide for a VM-site from the combined observations was done such that once confident VM-site localization was established, higher identification scores and longer peptide lengths were preferred. While an SM PSM identification score was based on the number of matching peaks, their ion type assignment, and the relative height of unmatched peaks, the VM site localization score was the difference in identification score between the top two localizations. The score threshold for confident localization, > 1.1, essentially corresponded to at least 1 b or y ion located between two candidate sites that has a peak height > 10% of the tallest fragment ion (neutral losses of phosphate from the precursor and related ions as well as immonium and TMT reporter ions were excluded from the relative height calculation). The ion type scores for b-H_3PO_4, y-H_3PO_4, b-H_2O, and y-H_2O ion types were all set to 0.5. This prevented inappropriate confident localization assignment when a spectrum lacked primary b or y ions between two possible sites but contained ions that could be assigned as either phosphate-loss ions for one localization or water loss ions for another localization.

Protein grouping of PSMs, peptides and PTM sites

Using the SM Autovalidation and Protein/Peptide summary modules results were filtered and reported at the protein level. Identified proteins were combined into the same protein group if they shared a peptide with sequence length greater than 8. A protein group could be expanded into subgroups (isoforms or family members) when distinct peptides were present which uniquely represent a subset of the proteins in a group. For the proteome dataset the protein grouping method ‘‘expand subgroups, top uses shared’’ (SGT) was employed which allocates peptides shared by protein subgroups only to the highest scoring subgroup containing the peptide. For the PTM-ome datasets the protein grouping method ‘‘unexpand subgroups’’ was employed which reports a VM-site only once per protein group allocated to the highest scoring subgroup containing the representative peptide. The SM protein score is the sum of the scores of distinct peptides. A distinct peptide is the single highest scoring instance of a peptide detected through an MS/MS spectrum. MS/MS spectra for a particular peptide may have been recorded multiple times (e.g., as different precursor charge states, in adjacent bRP fractions, modified by deamidation at Asn or oxidation of Met, or with different phosphosite localization), but are still counted as a single distinct peptide.

Please cite this article in press as: Martins Rodrigues et al., Precision proteogenomics reveals pan-cancer impact of germline variants, Cell (2025), https://doi.org/10.1016/j.cell.2025.03.026

Peptide spectrum match (PSM) filtering and false discovery rates (FDR)

Using the SM Autovalidation module peptide spectrum matches (PSMs) for individual spectra were confidently assigned by applying target-decoy based FDR estimation to achieve <1.0% FDR at the PSM, peptide, VM site and protein levels. For the whole proteome dataset thresholding was done in 3 steps: at the PSM level, the protein level for each TMT-plex, and the protein level for the cohort of 2 TMT-plexes. For the PTM omes (phosphoproteome and acetylome datasets), thresholding was done in two steps: at the PSM level for each TMT-plex and at the VM site level for the cohort of 2 TMT-plexes. In step 1 for all datasets, PSM level autovalidation was done first and separately for each TMT-plex experiment using an auto-thresholds strategy with a minimum sequence length of 7; automatic variable range precursor mass filtering; with score and delta Rank1 - Rank2 score thresholds optimized to yield a PSM level FDR estimate for precursor charges 2 through 4 of < 0.8% for each precursor charge state in each LC-MS/MS run. To achieve reasonable statistics for precursor charges 5-6, thresholds were optimized to yield a PSM-level FDR estimate of < 0.4% across all runs per TMT-plex experiment (instead of per each run), since many fewer spectra are generated for the higher charge states.

In step 2 for the PTM omes: phosphoproteome and acetylome datasets VM site polishing autovalidation was applied across both TMT plexes to retain all VM site identifications with either a minimum id score of 8.0 or observation in n TMT plexes (n=4, 3, or 2 if > 20, 7, or 1 plexes/cohort, respectively). The intention of the VM site polishing step is to control FDR by eliminating unreliable VM site level identifications, particularly low scoring VM-sites that are only detected as low scoring peptides that are also infrequently detected across TMT plexes in the study. Using the SM Protein/Peptide Summary module to make VM-site reports the ubiquitylome and acetylome datasets are further filtered to remove peptides ending with the regular expression [K][K]k since trypsin and Lys-C cannot cleave at a acetylated lysine. The [K] means retain if unmodified Lys present in one of the last two positions to allow for a missed cleavage with ambiguous PTM-site localization. C-terminally acetylated lysines are present in the acetylome dataset, but have been shown to arise from artifactual modification during TMT-labeling after trypsin digestion.

In step 2 for the whole proteome dataset, protein polishing autovalidation was applied separately to each TMT-plex experiment to further filter the PSMs using a target protein level FDR threshold of zero. The primary goal of this step was to eliminate peptides identified with low scoring PSMs that represent proteins identified by a single peptide, so-called ‘‘one-hit wonders.’’ After assembling protein groups from the autovalidated PSMs, protein polishing determined the maximum protein level score of a protein group that consisted entirely of distinct peptides estimated to be false-positive identifications (PSMs with negative delta forward-reverse scores). PSMs were removed from the set obtained in the initial peptide level autovalidation step if they contributed to protein groups that had protein scores below the maximum false-positive protein score. Step 3 was then applied, consisting of protein polishing autovalidation across all TMT plexes in a cohort together using the protein grouping method ‘‘expand subgroups, top uses shared’’ to retain protein subgroups with either a minimum protein score of 25 or observation in TMT plexes (n=4, 3, or 2 if > 20, 7, or 1 plexes/cohort, respectively). The primary goal of this step was to eliminate low scoring proteins that were infrequently detected in a cohort. As a consequence of these two protein- polishing steps, each identified protein reported in the study comprised multiple peptides, unless a single excellent scoring peptide was the sole match and that peptide was observed in multiple TMT-plexes.

Subset-specific FDR filtering for germline variant containing peptides in the proteome

While peptides in the proteome dataset matched to reference proteome sequences are subject to multi-step, protein-level and cohort level FDR filtering as described above, FDR for subsets of rarely observed (<5% of total) classes of peptides required more stringent score thresholding to reach a suitable subset-specific FDR < 1.0%. To this end, we devised and applied subset-specific filtering approaches.

The subset of peptides containing single amino acid variants (SAAVs) and indels observed in the proteome was extracted after step 1 of PSM filtering described above using the SM Protein/Peptide Summary module to create a proteogenomics (PG) site report, with quantitation normalized to nullify the effect of differential protein loading using the aggregate protein-level normalization factors from the fully filtered proteome dataset. Germline variants containing peptides were split up into 4 subsets (SAAVs and indels, with each further split by multiple or single representation in a cohort) and each subset was filtered to <1% FDR.

Subsets were thresholded independently in each subset using a 2-step approach. First, PSM scoring metric thresholds were tightened in a fixed manner so that distributions for each metric improved to meet or exceed the aggregate distributions. The fixed thresholds were: minimum score: 7; minimum percent scored peak intensity: 50%; normalized precursor mass error: +/-5 ppm. Second, individual subsets with FDR estimates remaining above 1% were further subject to a grid search to determine the lowest values of backbone cleavage score (sequence coverage metric) and score (fragment ion assignment metric) that improved FDR to < 1% for each subset.

Quantitation using TMT ratios

Using the SM Protein/Peptide Summary module, a protein comparison report was generated for the proteome dataset using the protein grouping method ‘‘expand subgroups, top uses shared’’ (SGT). For the PTM omes (phosphoproteome and acetylome datasets) Variable Modification site comparison reports limited to either phospho, or acetyl sites, respectively, was generated using the protein grouping method ‘‘unexpand subgroups.’’ Relative abundances of proteins and VM-sites were determined in SM using TMT reporter ion $\log_2$ intensity ratios from each PSM. TMT reporter ion intensities were corrected for isotopic impurities in the SM Protein/Peptide Summary module using the afRICA correction method, which implements determinant calculations according to Cramer’s Rule and correction factors obtained from the reagent manufacturer’s certificate of analysis for each cohort. Each protein-level or PTM site-level TMT ratio was calculated as the median of all PSM-level ratios contributing to a protein subgroup or PTM site. PSMs were excluded from the calculation if they lacked a TMT label, had a precursor ion purity < 50% (MS/MS has significant precursor isolation

Please cite this article in press as: Martins Rodrigues et al., Precision proteogenomics reveals pan-cancer impact of germline variants, Cell (2025), https://doi.org/10.1016/j.cell.2025.03.026

contamination from co-eluting peptides), or had a negative delta forward-reverse identification score (half of all false-positive identifications). Using the SM Process Report module non-quantifiable proteins and PTM sites (ex: unlabeled peptides containing an acetylated protein N-terminus and ending in arginine rather than lysine) were removed, and median/MAD normalization was performed on each TMT channel in each ome to center and scale the aggregate distribution of protein-level or PTM site-level log-ratios around zero in order to nullify the effect of differential protein loading and/or systematic MS variation. When subsets of an ome (nuORF or SAAVs, etc) the TMT ratios were normalized using the normalization factors for the aggregate distribution of the corresponding ome.

It is worth noting that current precision database methods separately quantify different forms of a peptide (reference sequence, variant–containing, phosphorylated, unphosphorylated, etc.) having distinct peptide masses and retention times in TMT labeled ratio-based LC-MS/MS experiments. A TMT labeled experiment is purpose-built to measure ratios of an individual peptide form across samples, which are combined so that each sample in a TMT-plex produces a reporter ion of distinct m/z in each MS/MS spectrum. The TMT reporter ion intensities of the reference sequence and variant-containing forms of a peptide cannot be directly combined to form a single value representing the overall peptide abundance since the MS/MS spectra will have been briefly sampled at different points in their corresponding chromatographic peaks.[187] Protein- or gene-level quantification will mitigate this effect by relying on multiple other wild-type (WT)-only peptides. In contrast, PTM measurements may be more affected since they are usually measured as single peptides.

Germline Variants Co-localizing with or Around PTM sites

Input data

From a total of 27,104,152 germline variants called from WES data, we selected 11,962,341 missense germline variants across our 1,064 samples over 10 cancer types to find germline variants directly co-localizing or nearby a PTM site.

As per PTM data, we obtained a total of 141,330 unique phosphorylation sites detected in at least one of the samples in our CPTAC cohort (134,244 on reference peptides and 7,086 on variant peptides affected by germline SAAVs) and 23,756 unique acetylation sites (23,190 on reference peptides and 566 on variant peptides affected by germline SAAVs). Sites detected on the same peptide sequence were considered as separate individual sites yielding a total of 168,423 and 9,018 phosphorylation sites on reference and variant peptides, respectively, and 24,109 and 639 acetylation sites on reference and variant peptides, respectively.

Calculation of linear distances

Missense variants co-localizing with PTM sites involving serine (S), threonine (T), tyrosine (Y), or lysine (K) codons were cross-referenced in the PTM data for cognate positions. PTM associated germline variants were grouped according to the three types of consequences at the PTM level: (1) an amino acid change caused loss of the PTM site; (2) a variant caused gain of a PTM site not encoded by the reference allele; or, (3) one phosphorylated residue changed to another (such as from a serine to a tyrosine, with phosphorylation detected in both). The ancestral and derived alleles were compiled for all the co-localizing variants. In three specific cases: *AHNAK* S4516N, *FAM83B* S729T, and *FLG* S3174C the reference-associated phosphorylated serine detected in the PTM data was derived from the ancestral annotation (T4516N, P729T, and G3174C, respectively). Therefore, these variants were excluded from the analysis.

We also detected variants around a PTM site by calculating the linear distance of missense germline variants relative to PTM sites based on amino acid position as extracted from reference peptides, classifying events using 2 categories: missense variants affecting an amino acid within 5 amino acids of the PTM site were categorized as *proximal* events; variants affecting amino acids beyond 5 amino acids of the PTM site were categorized as *distal*. We further confirmed if the amino acid changes predicted from germline variant information matched what was detected in the variant peptide information, when existent. In terms of variants proximal or distal to a site, because most variants distal to a PTM site and a portion of proximal variants fell outside the peptide capture of the PTM site in question, we would not expect to detect a variant-derived peptide for such cases. These direct, proximal, and distal events were used for downstream analyses.

Analyses of Direct, Proximal, and Distal Impact of Germline Variants on Protein and PTM Levels

We assessed the potential influence of a germline variant direct, proximal, or distal to a PTM site on the overall protein abundance levels using a general linear model approach. We also tested the effects of germline variants on phosphorylation and acetylation levels of reference peptides using the same approach, but only those variants for which the position fell outside the peptide capture of the PTM site in question in order to limit the possibility of bias in the mass spec measures (See limitations of the study and quantitation using TMT ratios STAR Methods section). Therefore, for variants directly overlapping a PTM site, we only tested their impact on the overall protein abundance, not on PTM levels. Common germline variants (gnomAD AF $\geq$1%) were tested individually. In the case of low frequency and rare germline variants (gnomAD AF <1%), to increase statistical power, we collapsed all individuals harboring a low frequency/rare variant proximal (within 5 amino acids), or all individuals with a low frequency/rare variant distal (> 5 amino acids) to a PTM site into a single variable, at the gene level. In order to test the pan-cancer differences in protein, phosphorylation, or acetylation levels between carriers and non-carriers of a certain germline variant, we ran the following model to learn the β coefficients:

$$Y = \beta_0 + \beta_1 M_v + \beta_2 P_1 + \beta_3 P_2 + \beta_4 P_3 + \beta_5 C + \epsilon$$

where Y is a (n x 1) vector representing the protein, phosphorylation, or acetylation abundance of the protein of interest for the site of interest; M is a binary vector indicating the germline variant status for the site of interest (v) for each sample; $P_{1\text{-}3}$ denote the first three PCs for patient genetic ancestry determination (WES-based); and C is the one-hot encoded cancer type for the samples. The error (ϵ) is assumed to be normally distributed with a constant variance. Tumor samples and matching NAT samples were tested separately. Cancer-type specific analyses were also performed. All resulting p-values were adjusted to FDR using the standard Benjamini-Hochberg procedure. The results from these tests are provided in Table S3.

Using the same approach as above, we also tested for the effects of highlighted variants from the direct/proximal/distal analyses on their Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway partners' protein and phosphoprotein abundances. That is, we evaluated "mTOR signaling" for DEPTOR S389N (hsa04150), "ErbB signaling" for ERBB2 P1170A (hsa04012), "MAPK signaling" for MAP2K2 P298L (hsa04010), "Antigen processing and presentation" for HLA-B V69A (hsa04612), "Apoptosis" for CASP8 D344H (hsa04210), and "Cell cycle" for ATRX E929Q (hsa04110). Because MGMT is not a member of any KEGG pathways, it was not tested. We similarly did not test SBDS, which is only in the general "Ribosome biogenesis in eukaryotes" pathway. The analyses were done at both pan-cancer and cancer-specific levels, in which we required at least 5 observations each in variant carriers and non-carriers to test. Resulting p-values were FDR adjusted using the standard Benjamini-Hochberg procedure, and all hits from the general linear model with FDR $\leq$ 0.05 were prioritized for plotting by carrier status. Pairwise Wilcoxon tests between carrier groups were performed for plotting, and FDR adjusted p-values are provided within the boxplots.

To determine whether the genes harboring PTM-affecting germline variants exhibited any biological bias, we conducted an over-representation analysis of curated pathways from the MiSigDB Hallmark[188,189] set and Wiki Pathways.[190] For genes with variants directly overlapping PTM sites imposing phosphorylation loss and gain, the background gene set was defined as all genes detected in the phosphoproteome data. All acetylated proteins detected in the PTM data were similarly used for background adjustment for genes that experience acetylation site loss or gain. The R package clusterProfiler v4.4.2 was used to conduct these analyses for each PTM type and consequence group separately. Results were constricted to a cutoff of 0.05 FDR adjusted p-value and a q-value cutoff of 0.1. A similar analysis was performed for proximal and distal events. In this case, genes harboring variants proximal or distal to a PTM site were used as the test gene set, testing each group separately. The background gene sets and significance cut-offs were defined as above.

HotSpot3D / HotPho analyses

Input PTM data

Here, we collected information for every PTM site detected on both reference and variant peptides in at least one of the samples in our CPTAC cohort via our analyses of proteomics LC-MS/MS data (See proteomics LC-MS/MS data interpretation STAR Methods section for more details). In total, 8,046 PTM sites (7,353 phosphosites and 693 acetylation sites) were detected on variant peptides affected by germline SNVs or Indels in at least one of our samples. For the purposes of HotSpot3D/HotPho analyses, however, we have excluded PTM sites on variant peptides affected by germline Indels.

We obtained 141,330 unique phosphorylation sites detected in at least one of the samples in our CPTAC cohort, from which 134,244 are on reference peptides and 7,086 are on variant peptides affected by germline SAAVs. As per acetylation sites, we obtained 23,756 unique acetylation sites, from which 23,190 are on reference peptides and 566 are on variant peptides affected by germline SAAVs. Further, sites detected on the same peptide sequence were considered as separate individual sites for the purposes of using it as an input for HotSpot3D[72] due to the format required by the tool, yielding a total of 168,423 and 9,018 phosphorylation sites on reference and variant peptides, respectively, and 24,109 and 639 acetylation sites on reference and variant peptides, respectively. Of these, 123,676 phosphorylation sites and 23,646 acetylation sites are unique and were used as inputs for HotSpot3D/HotPho. To map amino acid residues on different protein isoforms between UniProt Knowledge Base (UniProtKB, version 2023_01)[79] and our dataset, we used Transvar,[175] which allowed us to map them to their unique genomic positions.

Input somatic mutation and germline variant data

Somatic mutations and germline variants detected from WES, as described above, were filtered for missense single nucleotide events. Therefore, from a total of 345,653 and 27,104,152 exonic somatic mutations and germline variants called from WES data, respectively, we selected 183,503 missense somatic mutations and 11,962,341 missense germline variants across our 1,064 samples over 10 cancer types as inputs for HotSpot3D/HotPho.

PDB and AlphaFoldDB structures

We used the GRCh38 assembly and Ensembl release 100 (Gencode v34) in order to preprocess residue pair data for all human proteins available in two databases: (1) the RCSB Protein Data Bank (RCSB PDB)[77,78] as of June 24th, 2021, which contains PDB structures for 7,780 proteins; and (2) the AlphaFold Protein Structure Database (AlphaFoldDB - AFDB)[75,76] v4, as of March 16th, 2023, which contains predicted protein structures from 19,966 proteins present in Uniprot. For PDB, we filtered out chains or structures due to artifacts, as previously described.[73] For AFDB, HotSpot3D's algorithm pulls information from the web page version of the database, which provides information for proteins up to 2700 amino acids long. For those proteins which are longer than 2700aa, AFDB provides 1400aa long overlapping fragments, for which only the first 1400aa are available in the webpage version used here.

Quality control

As described before,[73] HotSpot3D/HotPho takes as input a file containing all PTM sites of interest containing the following information for each: the HUGO gene symbol, the corresponding Ensembl transcript ID, the protein residue position, and a summarized description of the site (e.g. Phosphoserine, Acetyllysine, etc). This information is then passed through the software, together with the input germline variant and somatic mutation information, to find pairwise relationships between mutations and sites. For the purposes of these analyses, we use the word "mutations" to describe both somatic and germline events. For PDB, because the structures provided by uploaders in the database do not always directly map to the associated Uniprot entries, HotSpot3D/HotPho calculates offsets in residue numbers in PDB structures and transcripts. For AFDB, because we are dealing with computationally predicted structures, the residues at the same position between the database structure and the Uniprot entry may not always perfectly match. Therefore, we have filtered out any sites where the residue provided in the PDB or AFDB structure did not match the residue in the input phosphorylation or acetylation site data, resulting in the following results for each input database: (1) PDB: 41,748 mutation-mutation pairs, 13,072 mutation-site pairs (4,625 excluded), and 11,328 site-site pairs (5,414 excluded); (2) AFDB: 110,255 mutation-mutation pairs; 29,888 mutation-site pairs (3,282 excluded), and 32,946 site-site pairs (4,972 excluded).

Cluster discovery and filtering

We have implemented HotSpot3D[72] and HotPho[73] to allow for the co-clustering of both missense germline variants and somatic mutations with phosphorylation and acetylation sites on the 3D protein structures (Figure 4A), as previously described.[73] Briefly, we used HotSpot3D to calculate the 3D distances between mutations and PTM sites using structures from PDB, as well as predicted structures from AFDB. During this process, missense variants and PTM sites are considered as nodes and the 3D distances between them as edges on an undirected graph. The clusters are then calculated using the Floyd–Warshall shortest-paths algorithm and using recurrence as the vertex type and clustering distance of 10Å, as implemented in HotSpot3D.[72] These analyses yielded a total of 15,132 unfiltered clusters across 4,409 unique proteins using PDB structures (2,084 site-only, 9,558 mutation-only, 3,490 hybrid), and 96,719 unfiltered clusters in 15,655 unique proteins using AFDB structures (14,788 site-only, 62,437 mutation-only, 19,494 hybrid).

We further filtered clusters based on the cluster closeness score (Cc), for which a high score indicates a cluster enriched in mutations and PTM sites on the 3D protein structure. Here we use a threshold of top 5% to select high confidence intramolecular clusters for downstream analyses, as described in the previous HotSpot3D and HotPho studies.[72,73] This generated a final set of 210 hybrid, 509 mutation-only, and 111 site-only clusters from PDB and 978 hybrid, 3126 mutation-only, 731 site-only clusters from AFDB. These results are provided in Table S4.

Impact on protein abundance analyses

We applied a linear model to evaluate the protein abundance level differences between carriers and non-carriers of co-clustered mutations and/or PTM sites within the same intramolecular cluster. We ran the model to learn the β coefficients as follows:

$$Y = \beta_0 + \beta_1 M_v + \beta_2 P_1 + \beta_3 P_2 + \beta_4 P_3 + \beta_5 C + \beta_5 N + \epsilon$$

where Y is a (n x 1) vector representing the protein abundance of the protein of interest for the cluster of interest; M is a binary vector indicating the co-clustered status (v) for each sample (i.e. if a sample had any event co-clustered in a particular cluster, it was grouped here); $P_{1\text{-}3}$ denote the first three PCs for patient genetic ancestry determination (WES-based); C is the one-hot encoded cancer type for the samples, and N is the CNV value for the gene being tested, as determined by GISTIC2. The error (ϵ) is assumed to be normally distributed with a constant variance.

Cancer-type specific analyses were also performed in the same way, where we evaluated the effect of germline and somatic variants involved in hybrid clusters on protein abundance levels between carriers and non-carriers to find genetic changes potentially associated with a certain cancer type.

Analyses of phosphorylation and acetylation levels were not performed in this case due to the limitations addressed in this manuscript (See limitations of the study).

Allele specific expression analysis using RNA-seq data

To identify allele specific expression (ASE) events based on RNA-seq, we used 1,057 tumor and 340 NAT samples with available RNA-seq data. For these analyses, we used only SNVs in cancer-related genes (624 cancer related genes[17]). First, germline variants were filtered to the ones that were detected in either of the three datasets: proteome, phosphoproteome, or acetylome. Next, we calculated read counts for each variant in each sample's RNA-seq BAM files using bam-readcount (v0.7.4 with parameters -q 10, -b 15, and -i so that reads overlapping with an insertion were not included in the per base counts). We retained only variants with at least 10 read counts covering reference and alternative alleles for this analysis. Then, to identify ASE events, we performed a two-sided binomial test with a null probability of success 0.5 in a Bernoulli experiment. The resulting p-values were adjusted using BH procedure, and ASE events were called significant if they reached FDR<0.05.

Indel variant analysis

Summary statistics of indel counts were measured according to the germline MAF files (above Methods) and restricted to a large set of cancer related genes as previously described.

Please cite this article in press as: Martins Rodrigues et al., Precision proteogenomics reveals pan-cancer impact of germline variants, Cell (2025), https://doi.org/10.1016/j.cell.2025.03.026

Indel positioning was performed by mapping variants to the exons and labeling them according to position (First, Middle, or Last exon). When only 1 or 2 exons made up the composition of a gene, then they were assigned first and last, and no did not receive a middle label. Relative position of the mutation within the gene model was calculated for each gene based on the size of the exon as cataloged by Ensembl v100.[191] The Penultimate region next to the last exon junction (<50bp from the last exon junctions [EJC]) was measured. This was performed for frameshift mutations and predicted inframe mutations as annotated in the germline MAF files (see above STAR Methods - germline variant calling and filtering from WES). Again, using the Ensembl gene annotations, relative positions to the last exon start-position was used to determine whether a mutation was assigned to the penultimate position. Kernel density information was estimated and plotted to identify gene position differences of inframe and frameshift mutations (Figure 6B).

We also developed two simple algorithms to discover the impact of these germline variants on protein abundances. The first method seeks to determine the impact of indels by looking at the upstream and downstream peptide-level abundances. Simply stated, we used a t-test as the crux of the first analysis. Second, we sought to find mutations that had an effect on protein abundance with respect to the RNA expression. Below we outline the implementation of a multi-omic LDA (moLDA) analysis to accomplish this objective.

We used the following criteria to discover variants that had variable upstream and downstream consequences of indels. First, we restricted our indel variants to those that had a predicted frameshift, splice-region, protein-altering designation according to VEP annotations (see above STAR Methods - germline variant calling and filtering from WES). Next, we restricted our search to variants that were observed in at least 20 samples. We ensured only variants with at least 6 measured peptides, up- and down-stream, were included. We then split the data based on whether there was a significant difference between upstream peptide abundances to downstream peptide abundances using a t-test. P-values and 95% confidence intervals for all indels and genes that met these criteria are provided in Table S6.

The second strategy we implemented to identify the role of indels on protein variability was to leverage an assumed relationship between RNA expression and protein abundance to find examples where mutations clearly associated with an expected relationship. To achieve this objective we implemented a multi-omic linear discriminant analysis (LDA) to classify indel status based on RNA and protein abundance. Briefly, LDA is a statistical method used for classifying or predicting the group membership of observations based on a set of predictor variables. It aims to find a linear combination of predictors that maximally separates two differentiating groups. Here the groups are defined as indel carriers and non-carriers and the predictors are protein abundance and RNA expression. First, we ensured that more than 30 samples had both RNA and protein abundance measurements for a given gene (in *cis*). Next, we excluded all mutations that didn't have at least 6 samples with the mutations and at least 6 samples without the mutations. Following a data integration step to merge RNA expression with protein abundance we used the 'lda' function as part of the MASS R library to find linear combinations of protein and RNA that segregated based on mutation status (Figure 6E). Genes and mutations were prioritized based on their singular value decomposition (SVD) scores which provide higher scores for improved separation between predictors Table S6.

Identification of expression and protein quantitative trait loci (eQTLs and pQTLs)

We performed quantitative trait loci (QTL) mapping to identify common germline genetic variants that affect gene expression (eQTL) and protein abundance (pQTL) in tumor and normal tissues utilizing the linear regression model in MatrixeQTL.[192] For this purpose, we used WGS germline SNPs with MAF $\geq$ 5% and included gender and ten principal components as covariates to adjust for population stratification. We analyzed the data on each cancer and tissue separately. Specifically, we conducted the eQTL and pQTL analyses for tumor and normal tissues of ccRCC, HNSCC, LSCC LUAD and PDAC for which both gene expression and protein abundance data are available (except for eQTL analysis on normal tissue of PDAC patients due to the limited number of samples with normal data). For the eQTL analysis, we utilized the FPKM normalized gene expression generated from the RNA-Seq data as discussed in the Pan-Cancer Data and Resource and Pan-Cancer Driver manuscripts,[37,160] and further performed TPM conversion, quantile normalization, and inverse normal transformation to remove technical noises and allow cross-sample comparisons. The eQTL analysis included individuals for whom genotype and gene expression data were available and genes with TPM > 0.1 in at least 20% of samples (Table S7A). To eliminate the hidden determinants in the expression data, we additionally selected 15 PEER factors as covariates using PEER software.[193] The pQTL analysis included individuals for whom genotype and protein abundance were available and proteins with data in at least 20% of samples (Table S7A). We deemed QTLs at FDR $\leq$ 1% as significant and considered variants within 1 Mb of a genes' transcription start site as cis-QTLs. The significant eQTLs can be viewed at https://immuneregulation.mssm.edu/.[168] Furthermore, we performed overall survival analysis based on the expression of interesting genes (*ERAP2*, *HLA-DQB1* and *PPIL3*) in the ccRCC, HNSCC, LSCC, and LUAD CPTAC cohorts using the best cutoff with Kaplan-Meier Plotter.[194] We also used Kaplan-Meier Plotter to assess the correlation between the expression and overall survival in the TCGA cohort.

Colocalization Analysis of eQTLs and pQTLs

We performed colocalization analysis to determine whether the leading variants among the cis- eQTLs and pQTLs are the same in certain genes of interest using 'coloc' R package's coloc.Abf function.[195] We applied the default values for the prior probabilities for a SNP being associated with gene expression only, protein abundance only and with both.

Polygenic Risk Scores and associations with protein abundance

Summary statistics, including risk allele, protective allele, odds ratio (OR) and annotated gene, were obtained for the largest genome-wide association study available for each cancer type. This included ccRCC, PDAC, UCEC, GBM, LUAD, and LSCC for a total of 133 risk variants (Table S7H). The same GWAS study was used for LUAD and LSCC as this discovery study comprised a balanced mixture of cases of both lung cancer subtypes. Polygenic risk scores (PRS) for each cancer type were calculated using the *score* routine available in PLINK 2.0, weighting the allele dosage at each variant by the effect size.

In a first pass, we checked the discriminatory power of the PRS by integrating CPTAC and UKBB datasets. To control for population structure, for these specific analyses we selected individuals of European ancestry in both datasets. For each cancer type, we compared the PRSs of the corresponding subtype with: a) patients of other cancer types in CPTAC; b) individuals with cancer diagnosis in the UKBB; and c) individuals without cancer diagnosis in the UKBB. Three out of five tested cancers, namely PDAC, GBM, and LSCC, showed significantly higher PRSs in the corresponding CPTAC patients compared to controls (Figure S7B). We focused on these three cancers in subsequent analyses.

We identified tumor proteins that showed significant association with PRS using linear models. To avoid the effects of hidden variables inducing covariance in the protein abundance matrix, we first carried out a principal component analysis. Considering the relatively large number of potential covariates to the sample size in our study, we performed a supervised selection of covariates to be included in the linear models. We tested the correlation between the PRS as well as the first ten principal components (PCs) of protein abundance with relevant clinical, demographic and molecular variables, including: genetic ancestry (first 10 PCs), age at diagnosis, sex, tumor purity, and smoking in the case of lung cancers. In the linear models we only included as covariates those showing significant correlation with the corresponding PRS and/or proteomic PCs. Given our interest in germline variants (which are present in the different tumor compartments), tumor purity was not included in any of the models despite significant correlation. We excluded proteins with more than 20% of missing data across individuals. The effect of the PRS was estimated for each protein using lm() function in R with the following designs:

$$\begin{aligned}\mathrm{lm}\big(&\mathrm{protein} \sim \mathrm{PRS\_GMB} + \mathrm{AncestryPC1} + \mathrm{AncestryPC2} + \mathrm{AncestryPC3} + \mathrm{AncestryPC5} + \mathrm{AncestryPC7} + \mathrm{AncestryPC10}\\ &+ \mathrm{Age} + \mathrm{ProteinPC1} + \mathrm{ProteinPC2} + \mathrm{ProteinPC3} + \mathrm{ProteinPC4} + \mathrm{ProteinPC8}\big)\end{aligned}$$

$$\mathrm{lm}\big(\mathrm{protein} \sim \mathrm{PRS\_LSCC} + \mathrm{AncestryPC6} + \mathrm{ProteinPC1} + \mathrm{ProteinPC3} + \mathrm{ProteinPC7} + \mathrm{ProteinPC9}\big)$$

$$\begin{aligned}\mathrm{lm}\big(&\mathrm{protein} \sim \mathrm{PRS\_PDAC} + \mathrm{AncestryPC1} + \mathrm{AncestryPC2} + \mathrm{AncestryPC4} + \mathrm{AncestryPC5} + \mathrm{Age} + \mathrm{ProteinPC2}\\ &+ \mathrm{ProteinPC3} + \mathrm{ProteinPC4} + \mathrm{ProteinPC5} + \mathrm{ProteinPC6}\big)\end{aligned}$$

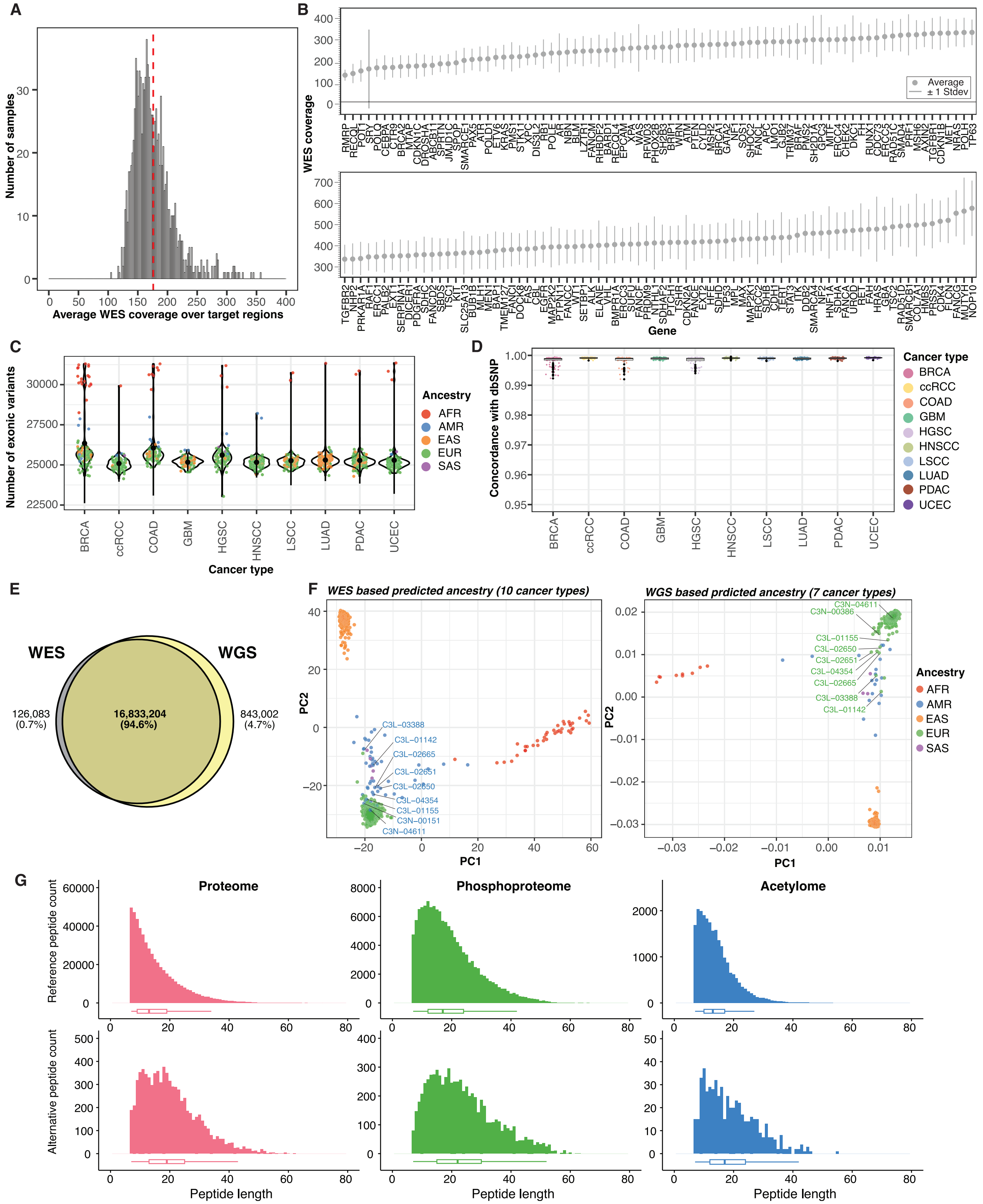
False discovery rates were estimated from the p-values using the fdrtool R package.[196] STRINGdb[173] R package (v11.5; https://www.string-db.org) was used to infer the protein-protein interaction networks and compute enrichments for the number of interactions among the top proteins associated with PRS. The database contains information for 19,566 proteins and over 2.9 million interactions. Over 93% of queried proteins were present in the STRING dataset. We performed Gene Set Enrichment Analyses (GSEA) analyses for Reactome pathways using the R package ReactomePA[197] with 10,000 permutations and significance threshold of 0.05 with BH FDR adjustment. Disease free survival and overall survival plots were generated using survminer (v0.4.9; https://github.com/kassambara/survminer) and survival (https://github.com/therneau/survival) R packages, stratifying the patients according to the median of the PRS scores.

ADDITIONAL RESOURCES

Comprehensive information about the CPTAC program, including program initiatives, investigators, and datasets, are available at the CPTAC program website: https://proteomics.cancer.gov/programs/cptac.

For the Pan-Cancer proteogenomics collection papers, along with links to the data and supplementary materials associated with these publications, please visit the Proteomic Data Commons (PDC) at https://pdc.cancer.gov/pdc/cptac-pancancer and the Cancer Research Data Commons at https://dataservice.datacommons.cancer.gov/#/data.

Supplemental figures



(legend on next page)