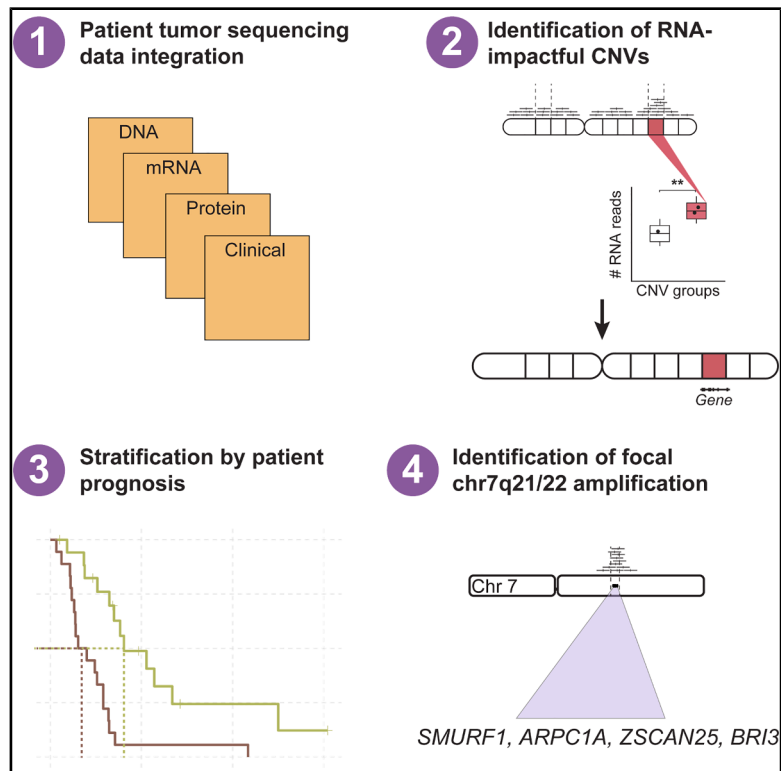


Multi-omics whole-genome characterization of the copy number landscape of metastatic pancreatic ductal adenocarcinoma

Graphical abstract



Authors

James T. Topham, Joanna M. Karasinska, Andrew Metcalfe, ..., Steven J.M. Jones, Daniel J. Renouf, David F. Schaeffer

Correspondence

david.schaeffer@vch.ca

In brief

Cancer; Genomics; Proteomics; Transcriptomics

Highlights

- Integrated sequencing analysis identified recurrent CNVs in mPDAC
- CNVs were stratified by their impact on mRNA and prognosis
- Focal amplification of chr7q21/22 is a poor prognostic somatic event
- Pan-cancer analysis further highlights *SMURF1* over-expression and amplification in PDAC



Article

Multi-omics whole-genome characterization of the copy number landscape of metastatic pancreatic ductal adenocarcinoma

James T. Topham,¹ Joanna M. Karasinska,¹ Andrew Metcalfe,¹ Hassan A. Ali,¹ Steve E. Kalloger,^{1,2} Maya Kevorkova,¹ Emma Titmuss,³ Gian Luca Negri,⁴ Sandra E. Spencer,⁴ Gun Ho Jang,⁵ Grainne M. O'Kane,⁵ Richard A. Moore,⁴ Andrew J. Mungall,⁴ Jonathan M. Loree,³ Faiyaz Notta,⁵ Julie M. Wilson,⁵ Oliver F. Bathe,⁶ Patricia A. Tang,⁶ Rachel Goodwin,⁷ Gregg B. Morin,^{4,8} Jennifer J. Knox,⁹ Steven Gallinger,^{5,9} Janessa Laskin,^{3,4} Marco A. Marra,^{4,10} Steven J.M. Jones,^{4,10} Daniel J. Renouf,^{1,3,11,13} and David F. Schaeffer^{1,2,12,13,14,*}

¹Pancreas Centre BC, Vancouver, BC, Canada

²Department of Pathology and Laboratory Medicine, UBC, Vancouver, BC, Canada

³Division of Medical Oncology, BC Cancer, Vancouver, BC, Canada

⁴Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC, Canada

⁵Ontario Institute for Cancer Research, Toronto, ON, Canada

⁶Departments of Surgery and Oncology, Cummings School of Medicine, University of Calgary, Calgary, AB, Canada

⁷The Ottawa Hospital Cancer Centre, Ottawa Hospital Research Institute, Ottawa, ON, Canada

⁸Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

⁹University Health Network, University of Toronto, Toronto, ON, Canada

¹⁰Department of Medical Genetics and Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada

¹¹Department of Medicine, University of British Columbia, Vancouver, BC, Canada

¹²Division of Anatomic Pathology, Vancouver General Hospital, Vancouver, BC, Canada

¹³These authors contributed equally

¹⁴Lead contact

*Correspondence: david.schaeffer@vch.ca

<https://doi.org/10.1016/j.isci.2025.113176>

SUMMARY

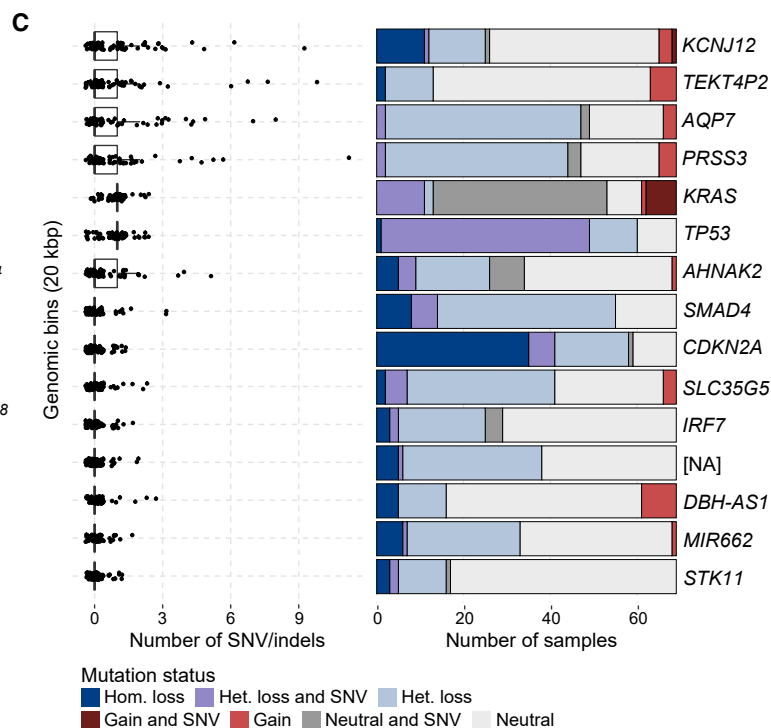
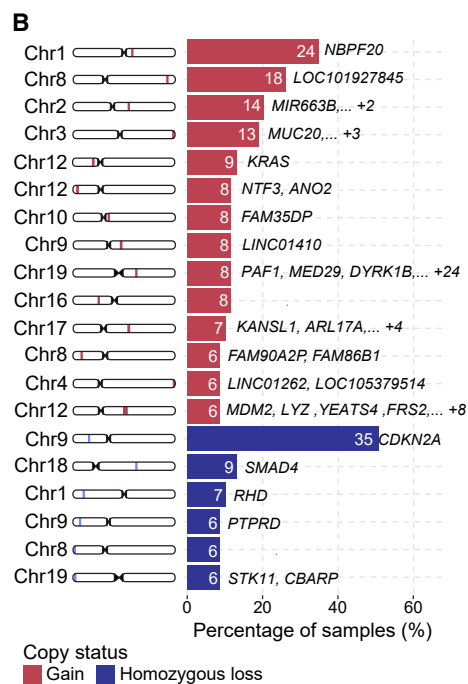
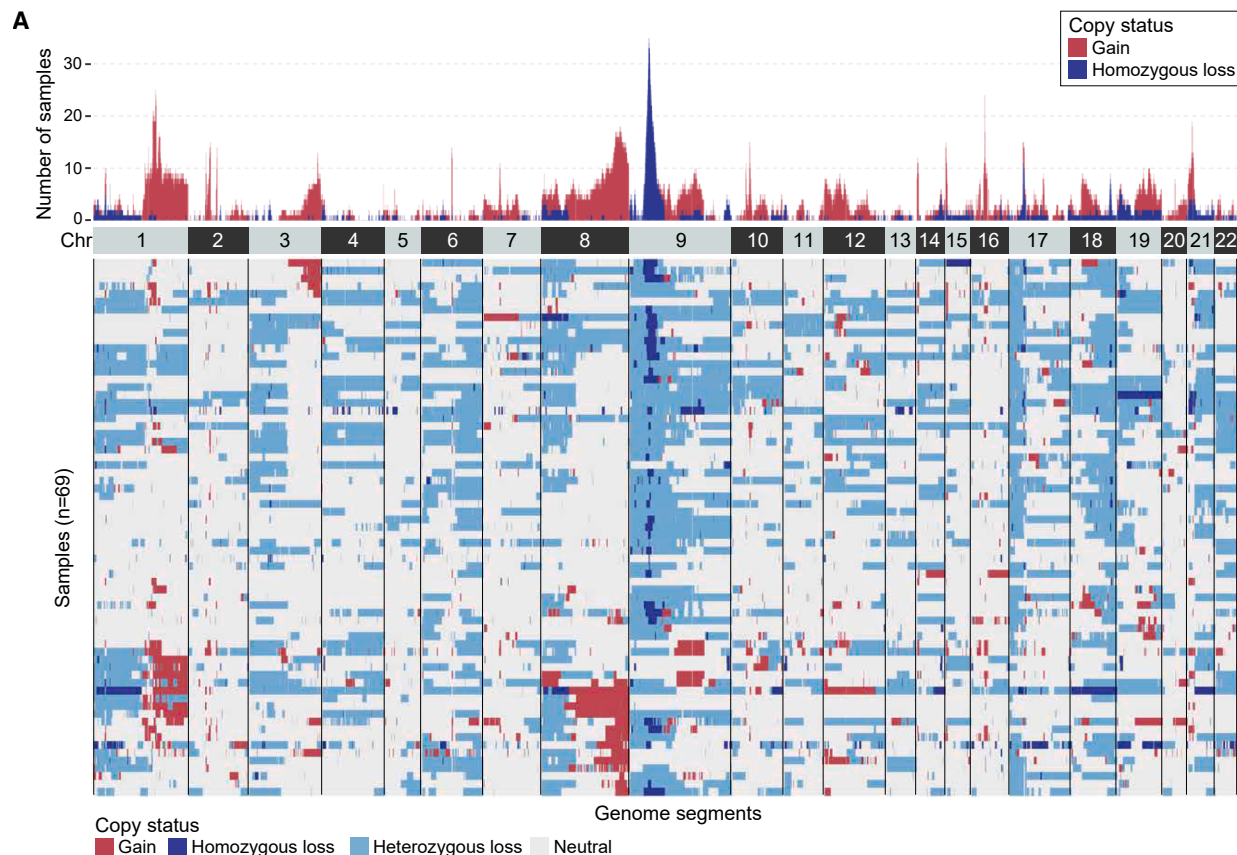
Pancreatic ductal adenocarcinoma (PDAC) is a highly aggressive disease with limited therapeutic options. While genome profiling has benefitted clinical decision-making in many cancer types, PDAC remains among the most lethal solid cancers and has relatively few clinically actionable molecular targets. Using a multi-omics approach, we integrate whole-genome and transcriptome datasets along with proteomics and clinical metadata to identify copy-altered regions with significant impact on expression at the whole-genome scale. This approach identified *SMURF1*, *ARPC1A*, *ZSCAN25*, and *BRI3* as focally amplified and over-expressed genes located on chr7q21/22 with expression negatively correlated with patient survival. We extrapolate these results into a pan-cancer analysis to demonstrate that the association between *SMURF1* amplification versus expression is strongest in PDAC among 23 other cancer types. Taken together, these data provide a detailed overview of the copy number landscape in PDAC while highlighting chr7q21/22 amplification as a recurrent somatic event impacting both gene expression and patient survival.

INTRODUCTION

Genomic profiling has been a catalyst for the development of novel therapeutics in cancer care. *BCR-ABL* fusions and HER2 overexpression are early examples of targetable genomic events that evolved the personalized medicine paradigm for chronic myelogenous leukemia and breast cancer, respectively.^{1,2} In colorectal cancer, *RAS* mutation, *BRAF* mutation and microsatellite instability (MSI) are genomic markers that guide clinical decision-making.³ While advancements in precision medicine have contributed to increases in survival for many cancer types, pancreatic ductal adenocarcinoma (PDAC) still faces significant challenges, with 5-year survival of only 12% - the lowest of all cancer types included in the 2023 US cancer statistics.⁴

Patients diagnosed with PDAC often experience a delayed onset of symptoms and rapid disease progression, and this limits opportunities for clinical trial enrollment and tumor profiling. *NRG1*, *NTRK*, and *ALK* fusions, germline *BRCA/PALB2* mutations, and, more recently, *KRAS* G12C mutation, are among the few clinically actionable genomic events that have been identified in PDAC, and only account for approximately 20% of patients.⁵ While mRNA and chromosome structural variation subtypes have been associated with patient survival, strategies targeting such subtypes remain in early stages of development.^{6–8} As the mutational landscape of PDAC was elucidated with the emergence of large sequencing-based trials, studies often reported on the most frequently mutated genes in PDAC, converging on driver events affecting *KRAS*, *TP53*, *CDKN2A*,





(legend on next page)

and *SMAD4*.^{8–11} Meanwhile, some of the most successful targets in PDAC, such as *NRG1* fusion, have been found to occur in less than 5% of patient tumors.¹² To date, a genome-wide assessment of copy number variants (CNVs) remains a critical opportunity to uncover additional molecular targets in PDAC. Here, we perform an integrated bioinformatics analysis using whole-genome DNA, RNA, proteomics, and clinical metadata from a cohort of tumor samples collected from patients with treatment-naïve metastatic PDAC.

RESULTS

Assessing the copy number landscape of pancreatic ductal adenocarcinoma reveals recurrently copy-altered genes

To comprehensively investigate the CNV landscape in metastatic PDAC, we utilized whole-genome sequencing data from a cohort of 69 patient tumor samples. The majority of samples were biopsied from metastatic sites on the liver ($n = 59$; 85.51%), while other samples were biopsied from the pancreas ($n = 3$; 4.35%) and other metastatic sites including lymph nodes ($n = 3$; 4.35%), peritoneum ($n = 1$; 1.45%), lung ($n = 1$; 1.45%), omentum ($n = 1$; 1.45%), and abdominal wall ($n = 1$; 1.45%). Median age at diagnosis was 60 years (min = 38, max = 78) and 60.87% ($n = 42$) of patients were male. The majority of patients received either FOLFIRINOX ($n = 37$; 53.6%) or gemcitabine plus nab paclitaxel ($n = 20$; 29.0%) as first line therapy for their metastatic disease, and all treatments were initiated after the biopsy for sequencing was performed.

Using tumor sequencing data, copy number status was first calculated for each of 144,061 non-overlapping, 20 kb-sized bins across all chromosomes for each sample and identified copy-neutral, amplification, heterozygous loss, and homozygous loss bins. For visualization purposes, adjacent bins with identical copy status in all samples were merged, resulting in 5,271 unique genome segments (Figure 1A). To aid in the visualization of recurrent CNV events, the number of samples with a CNV was counted for each genome segment. To investigate recurrent events in more detail, we identified segments that showed amplification in three (4.35%) or more samples (56/5,271; 1.06%) or homozygous loss in three or more samples ($n = 31/5,271$; 0.59%) and mapped each segment to any overlapping genes (Table S1). Genes overlapping the most recurrently amplified segments included neuroblastoma breakpoint family gene *NBPF20* ($n = 24$ samples; 34.78%), microRNA *MIR663B* ($n = 14$; 20.29%) and mucin family member *MUC20* ($n = 13$; 18.84%). Recurrent homozygous deletion segments

included genes such as blood group antigen gene *RHD* ($n = 7$; 10.14%), protein tyrosine phosphatase member *PTPRD* ($n = 6$; 8.70%) as well as a focal segment on chr8 (hg19 coordinates 2,240,001–2,260,001) that did not overlap with any coding elements ($n = 6$; 8.70%; Figure 1B). Known PDAC driver genes, including *KRAS* (copy amplification in nine samples), *CDK2NA* (deletion in 35 samples) and *SMAD4* (deletion in nine samples), were also among the top recurrent targets of copy alteration.

Amplification of alleles harboring oncogenic *KRAS* single nucleotide variants (SNVs) has been well-documented in a subset of patients with metastatic PDAC.¹³ Motivated by this observation, and to investigate associations between CNV events and overlapping point mutations at a genome-wide scale, we first calculated average mutation density (mean number of SNVs/indels) for each bin across all samples to compile a list of highly mutated bins (Table S2). We then assessed, for each bin, the number of samples with “two hits,” defined as having heterozygous deletion of the entire bin and at least one SNV/indel observed in the same bin, or having homozygous deletion of the entire bin (Figure 1C). Bins with highest SNV/indel density, regardless of copy status, were found to overlap potassium channel subfamily member *KCNJ12* (mean = 1.0 SNV/indels per sample; $n = 12$ samples with two hits), pseudogene *TEKT4P2* (mean = 0.96 SNV/indels; $n = 2$ samples with two hits), aquaporin *AQP7* (mean = 0.94 SNV/indels; $n = 2$ samples with two hits), serine protease *PRSS3* (mean = 0.94 SNV/indels; $n = 2$ samples with two hits) and *KRAS* (mean = 0.91 SNV/indels; $n = 11$ samples with two hits). Bins with the highest number of samples having two hits overlapped genes including tumor suppressors *TP53* ($n = 49$ samples with two hits; mean = 0.87 SNV/indels), *CDKN2A* ($n = 41$; mean = 0.17 SNV/indels) and *SMAD4* ($n = 14$; mean = 0.19 SNV/indels) as well as nucleoprotein *AHNAK2* ($n = 9$; mean = 0.46 SNV/indels), solute carrier family member *SLC35G5* ($n = 7$; mean = 0.16 SNV/indels) and serine/threonine kinase *STK11* ($n = 7$; mean = 0.13 SNV/indels). We also observed a region on chr4 (hg19 coordinates 49,480,001–49,500,001) that had two hits in five samples and mean = 0.14 SNV/indels but did not overlap any coding element. Taken together, these data offer a comprehensive, genome-wide summary of the copy number mutation landscape in metastatic PDAC and offer a compendium of genes of interest for future study.

Transcriptome data integration identifies RNA-impactful copy number variant events

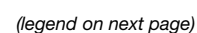
Given the large list of recurrently copy-altered bins, we sought to stratify our results to identify a subset of bins that are most likely to contribute to tumorigenesis. Intuitively, the impact that a CNV

Figure 1. Copy number landscape of PDAC reveals recurrently copy-altered genes

(A) Heatmap depicts copy number status across all samples ($n = 69$) for all segments of the genome. Chromosome mapping provided by the top overlay. Upper bars indicate the number of samples in which each genome segment was found to be copy number amplified (red) or deleted (blue).

(B) Bar plot summarizes the most recurrently amplified (red) or deleted (blue) genome segments across the sample cohort. Numbers near the edge of each bar indicate the number of samples in which the CNV was observed. Accompanying text lists genes that overlap each genome segment. Karyogram-like images (left) depict the location of each genome segment on the respective chromosome.

(C) Boxplot shows the mutation (SNV/indel) density for genome segments that had the highest average (mean) mutation density and/or the highest number of two hits. “NA” signifies that no gene was found to overlap the segment. Accompanying bar plots (right) show the distribution of copy number statuses for each genome segment. Boxplots indicate median (central line) and 25–75% IQR (bounds of box), and whiskers extend from box bounds to the largest value no further than 1.5 times the IQR.



has on the cell can be correlated to the magnitude of effect it has on the expression of the affected gene(s), as has been demonstrated for a subset of oncogenes and tumor suppressors,¹⁴ including *KRAS* at an allelic-specific level in PDAC.¹³ Identification of amplification and homozygous deletion events that result in significantly higher or lower (respectively) expression changes at the affected region therefore presents a strategy to filter for events that are more likely to be biologically relevant. To achieve this, we identified bins that were (a) recurrent targets of amplification or homozygous deletion and (b) had significantly increased or decreased (respectively) mRNA levels using a three-step bespoke bioinformatics pipeline (Figure 2A). This identified 2,444 and 269 RNA-impactful amplification and deletion bins, respectively, hereafter referred to as “RNA-impactful bins” (Figure 2B). RNA-impactful bins were further filtered using a threshold absolute difference (copy-altered vs. neutral) in RNA coverage of 0.25 log₁₀ RPKM ($n = 274$, Table S3). Gene set enrichment analysis (GSEA) using genes overlapped by RNA-impactful bins ($n = 164$) revealed significant enrichment of pathways including: genes residing on chr7q21/22 that are amplified in breast cancer ($p = 6.77\text{e-}27$) and genes previously found to be amplified in pancreatic ($p = 3.10\text{e-}7$) and glioma ($p = 6.90\text{e-}4$) tumor types (Table S4).

Finally, amplification events were filtered to only include amplification bins part of focal amplification events (size ≤ 3 Mb) in all affected samples ($n = 50$; Table S5) and bins were then mapped to affected genes ($n = 38$ genes; 24 and 14 affected by amplification or deletion, respectively). Co-expression and co-occurrence analysis were performed to assess relationships between genes affected by RNA-impactful bins, and a cluster of genes co-amplified along chr7q21/22 was identified (Figure 2C). Mass spectrometry protein quantification data were available for a subset of genes affected by filtered RNA-impactful bins ($n = 21$), and enabled side-by-side comparison of the effect of copy status on protein and mRNA levels (Figures 2D and S1). Copy amplification was associated with significantly higher protein and mRNA levels for *PDAP1* (copy amplified versus neutral; protein $p = 5.35\text{e-}4$; mRNA $p = 0.0021$), *BAIAP2L1* (protein $p = 5.81\text{e-}4$; mRNA $p = 0.0026$), *ASNS* (protein $p = 0.0029$; mRNA $p = 0.0018$), *RIOK3* (protein $p = 0.027$; mRNA $p = 1.43\text{e-}5$), *NPC1* (protein $p = 0.020$; mRNA $p = 3.59\text{e-}5$), and *TECPR1* (protein $p = 0.020$; mRNA $p = 0.0011$). Taken together, these data highlight a group of genes affected by recurrent copy amplifications that are strongly tied to upregulation at both the mRNA and protein levels.

A subset of the filtered RNA-impactful bins (25/274; 9.12%) did not overlap any gene coding region and were assessed for overlap with enhancer-related epigenetic marks H3K4me1, H3K27ac and H3K27me3 indicative of poised (H3K4me1 and H3K27me3), primed (H3K4me1) and active (H3K4me1 and H3K27ac) enhancers, using healthy pancreatic tissue data from the International Human Epigenome Consortium (IHEC) database.¹⁵ All 25 bins were amplification events. In the two available IHEC samples, 11/25 (44.0%) of bins were found to overlap epigenetic marks indicative of at least one active enhancer region (Figure S2). Of these 11 bins, four were located on chr1, six on chr8, and one on chr18. One bin located on chr8 was found to overlap a primed enhancer region.

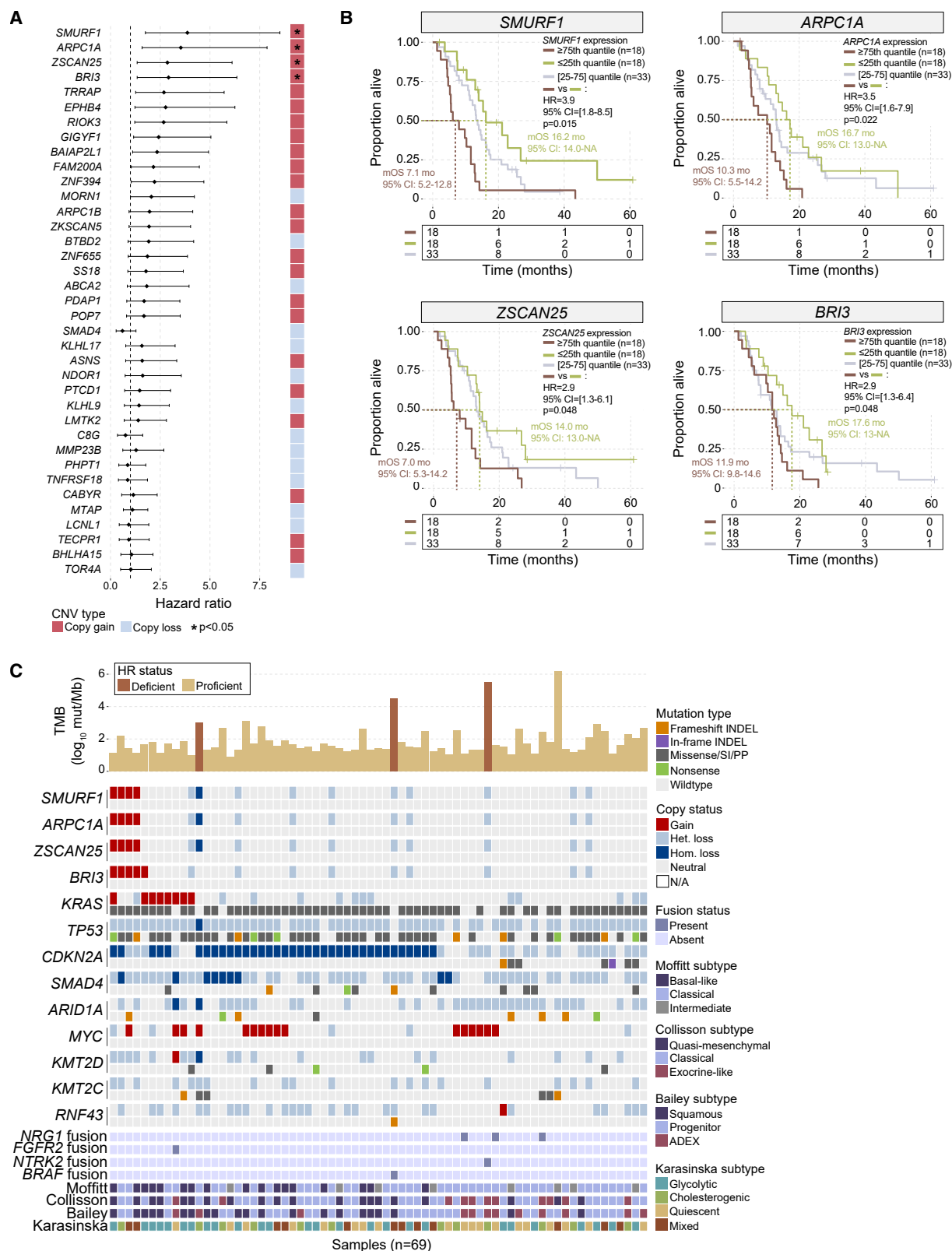
Survival analysis further stratifies genes affected by RNA-impactful copy number variants

To further stratify the list of genes affected by RNA-impactful CNVs, expression levels of each gene were assessed for their association with overall survival (OS). Genes with expression values significantly associated with OS were *SMURF1* (hazard ratio (HR) = 3.9; 95% confidence interval (CI) = 1.8–8.5; $p = 0.015$), *ARPC1A* (HR = 3.5; 95% CI = 1.6–7.9; $p = 0.022$), *ZSCAN25* (HR = 2.9; 95% CI = 1.3–6.1; $p = 0.048$) and *BRI3* (HR = 2.9; 95% CI = 1.3–6.4; $p = 0.048$; Figure 3A). Median OS levels between high ($>75^{\text{th}}$ quantile) and low ($<25^{\text{th}}$ quantile) expression groups included 7.1 versus 16.2 months (respectively) for *SMURF1*, 10.3 vs. 16.7 months for *ARPC1A*, 7.0 vs. 14.0 months for *ZSCAN25*, and 11.9 vs. 17.6 months for *BRI3* (Figure 3B).

We next investigated the relationship between copy amplifications affecting *SMURF1*, *ARPC1A*, *ZSCAN25* and *BRI3* and mutation events that are common to PDAC, including CNV and SNV/indel status of *KRAS*, *TP53*, *CDKN2A*, *SMAD4*, *ARID1A*, *MYC*, *KMT2D/C* and *RNF43* (Figure 3C). At least one of *SMURF1*, *ARPC1A*, *ZSCAN25*, and *BRI3* was amplified in 5/69 (7.2%) of samples, while all four genes were co-amplified in 4/69 (5.8%) of samples. The frequency of CNV and SNV/indel events in genes commonly mutated in PDAC was not significantly different between patients with co-amplification of *SMURF1/ARPC1A/ZSCAN25/BRI3* vs. other patients. Between these two groups, there were no significant differences in tumor mutational burden (TMB; $p = 0.27$) as well as the distribution of Moffitt ($p = 0.58$), Collisson ($p = 0.63$), Bailey ($p = 0.64$), and Karasinska ($p = 0.19$) molecular subtypes. Patients with the

Figure 2. Transcriptome data integration identifies RNA-impactful amplifications

(A) Analysis overview for the identification of RNA-impactful bins, spanning RNA-seq read mapping to bins, bin-wise comparison of RNA-seq coverage between copy number altered and neutral status, and mapping of bins to their affected gene.
(B) Scatterplot depicts the results of RNA-impactful bin identification analysis. Vertical line indicates p value = 0.05. Horizontal lines indicate threshold expression difference (copy-altered vs. neutral) at which RNA-impactful amplifications and deletions were defined (0.25 and -0.25 , respectively).
(C) Hybrid co-expression (bottom left) and co-occurrence (top right) heatmap for genes affected by RNA-impactful CNVs ($n = 38$). Upper heatmap indicates the availability of protein quantification for each gene and strength of mRNA expression difference between copy-altered vs. neutral samples. Right side heatmap indicates the chromosome location of each gene. A cluster of co-expressed, co-amplified genes is highlighted and projected onto their location on chromosome 7 (right side).
(D) Boxplots show differences in normalized protein (upper) and mRNA (lower) expression levels between samples with copy-altered versus neutral copy status. Genes affected by RNA-impactful CNVs that had protein quantification available are shown. Boxes indicate median (central line) and 25–75% IQR (bounds of box), and whiskers extend from box bounds to the largest value no further than 1.5 times the IQR. Wilcoxon mean rank-sum p values are shown.



(legend on next page)

Table 1. *SMURF1*, *ARPC1A*, *ZSCAN25*, and *BRI3* copy number amplification frequencies across TCGA cancer type groupings

Cancer group	<i>SMURF1</i> (%)	<i>ARPC1A</i> (%)	<i>ZSCAN25</i> (%)	<i>BRI3</i> (%)
Brain	16.82	16.98	16.82	16.82
Other	8.88	8.88	8.67	8.88
Gastrointestinal	8.57	8.15	7.66	8.32
Urological	6.26	6.26	6.09	6.2
Respiratory	6.1	5.99	5.89	5.58
Gynecological	5.88	5.88	5.75	5.35
Head/neck	4.13	4.13	3.96	3.88
Pancreas	2.34	2.34	2.34	2.34
Breast	2.26	2.17	2.26	1.88

co-amplification of *SMURF1*, *ARPC1A*, *ZSCAN25*, and *BRI3* did not show significant enrichment of any specific clinical features.

We next sought to determine whether the frequency of amplifications in *SMURF1*, *ARPC1A*, *ZSCAN25* and *BRI3*, and their effect on mRNA levels, could be generalized to other cohorts of PDAC. To this end, we leveraged CNV and expression datasets from TCGA ($n = 129$; resectable PDAC), ICGC ($n = 149$; resectable PDAC), COMPASS ($n = 195$; metastatic PDAC) and Hartwig ($n = 46$; metastatic PDAC) studies. *SMURF1/ARPC1A/ZSCAN25/BRI3* amplifications were present in 4.7/5.4/5.4/3.9% of TCGA samples, 6.0/6.0/6.0/5.4% of ICGC samples, 7.7/7.7/8.2/6.2% of COMPASS samples, and 15.2/15.2/15.2/15.2 of Hartwig samples, respectively. *SMURF1* amplification was associated with increased expression in TCGA ($p = 0.0017$), COMPASS ($p = 9.7\text{e-}6$) and Hartwig ($p = 0.019$) samples (Figure S3). *ARPC1A* amplification was associated with increased expression in TCGA ($p = 1.3\text{e-}4$), ICGC ($p = 4.2\text{e-}4$), COMPASS ($p = 3.0\text{e-}6$) and Hartwig ($p = 0.019$) samples. *ZSCAN25* amplification was associated with increased expression in TCGA ($p = 0.0013$), ICGC ($p = 0.0012$), COMPASS ($p = 1.6\text{e-}4$) and Hartwig ($p = 2.0\text{e-}4$) samples. Finally, *BRI3* amplification was associated with increased expression in COMPASS ($p = 1.6\text{e-}4$) and Hartwig ($p = 0.019$) samples. Survival analysis of external datasets showed significantly lower OS for patients with high expression of *SMURF1* (HR = 2.9; 95% CI = 1.0–8.0; $p = 0.049$), *ARPC1A* (HR = 5.1; 95% CI = 1.5–17.3; $p = 0.018$) and *BRI3* (HR = 3.3; 95% CI = 1.1–10.0; $p = 0.049$) in the Hartwig study. Meanwhile, high expression of *SMURF1* was associated with improved OS in the COMPASS study (HR = 0.57; 95% CI = 0.37–0.89; $p = 0.048$; Figure S4). To broaden the cross-cohort comparison of expression-based survival analysis, a second analysis was performed that included all

genes tested for survival differences in the PanGen discovery cohort (regardless of their prognostic significance in the PanGen cohort) that were assayed in the validation cohorts, though no individual genes in this analysis were statistically significant after multiple test correction (Figure S5).

Pan-cancer analysis of prognostic genes affected by RNA-impactful copy number variants

To examine the presence and strength of the RNA-impactful copy alterations in prognostic genes found in PDAC in other cancer types, we extrapolated our results using a pan-cancer analysis of TCGA CNV and RNA-seq datasets from 8,163 samples across 24 cancer types. Amplification rates of *SMURF1*, *ARPC1A*, *ZSCAN25*, and *BRI3* varied across TCGA cancer type groupings, with the highest frequencies seen in cancers located in the brain (genes amplified in 16.82, 16.98, 16.82, and 16.82% of samples, respectively; Table 1; Figure 4A). Across all cancer groupings, the similarity in frequency of amplification for each gene was reflective of the tendency for these genes to be co-amplified. Strength of correlation between copy status and expression of each gene was assessed across individual cancer types, and *SMURF1* copy status was significantly correlated with increased expression in 21/24 (87.50%) of cancer types, with pancreatic cancer (PAAD) showing the strongest correlation (defined by Spearman rho; rho = 0.72, $p = 1.22\text{e-}21$; Figure 4B; Table S6). Meanwhile, copy status of *ARPC1A*, *ZSCAN25*, and *BRI3* was positively correlated with expression in 24/24 (100.00%), 22/24 (91.67%), and 23/24 (95.83%) of cancer types, respectively. Unlike *SMURF1*, copy status versus expression correlations for *ARPC1A*, *ZSCAN25* and *BRI3* were strongest in SKCM (rho = 0.75, $p < 2.2\text{e-}16$), KIRP (rho = 0.78, $1.35\text{e-}56$) and COAD (rho = 0.54, $3.14\text{e-}21$), respectively. Across individual cancer types, PAAD showed the highest differences in median expression (copy amplification versus rest) for *SMURF1* (amplification frequency = 2.34%; expression difference = 0.66 log₁₀ RSEM, $p = 0.0039$), *ARPC1A* (2.34%; 0.69 log₁₀ RSEM, $p = 0.0034$), *ZSCAN25* (2.34%; 0.35 log₁₀ RSEM, $p = 0.0039$) and *BRI3* (2.34%; 0.47 log₁₀ RSEM, $p = 0.0066$; Figure 4C). Overall, these data demonstrate the occurrence of *SMURF1*, *ARPC1A*, *ZSCAN25*, and *BRI3* amplifications across several disease types. Importantly, both correlation and magnitude of copy status versus expression of *SMURF1* were strongest in pancreatic cancer versus all other cancer types.

DISCUSSION

We present a comprehensive and integrative analysis of the CNV landscape in metastatic PDAC. In addition to highlighting the

Figure 3. Survival analysis identifies prognostic impact of the expression of genes affected by RNA-impactful CNVs

(A) Forest plot shows hazard ratio (x axis), based on overall survival, of gene expression groups ($\geq 75^{\text{th}}$ percentile versus $\leq 25^{\text{th}}$ percentile) for each gene affected by RNA-impactful CNVs (y axis). Dots indicate hazard ratios, with bars representing 95% confidence intervals. Side heatmap indicates the type of CNV affecting each gene. Asterisks indicate significance based on a log rank test.

(B) Kaplan-Meier curves show overall survival differences between percentile expression groups for *SMURF1*, *ARPC1A*, *ZSCAN25*, and *BRI3*. Hazard ratio (HR) and 95% confidence intervals (CIs) based on Cox models are shown along with log rank p values ($\geq 75^{\text{th}}$ percentile versus $\leq 25^{\text{th}}$ percentile). Dashed lines indicate median overall survival times.

(C) OncoPrint shows the mutation status of *SMURF1*, *ARPC1A*, *ZSCAN25*, and *BRI3* alongside TMB, HRD status, other mutations commonly observed in PDAC, fusion events, and molecular subtypes.

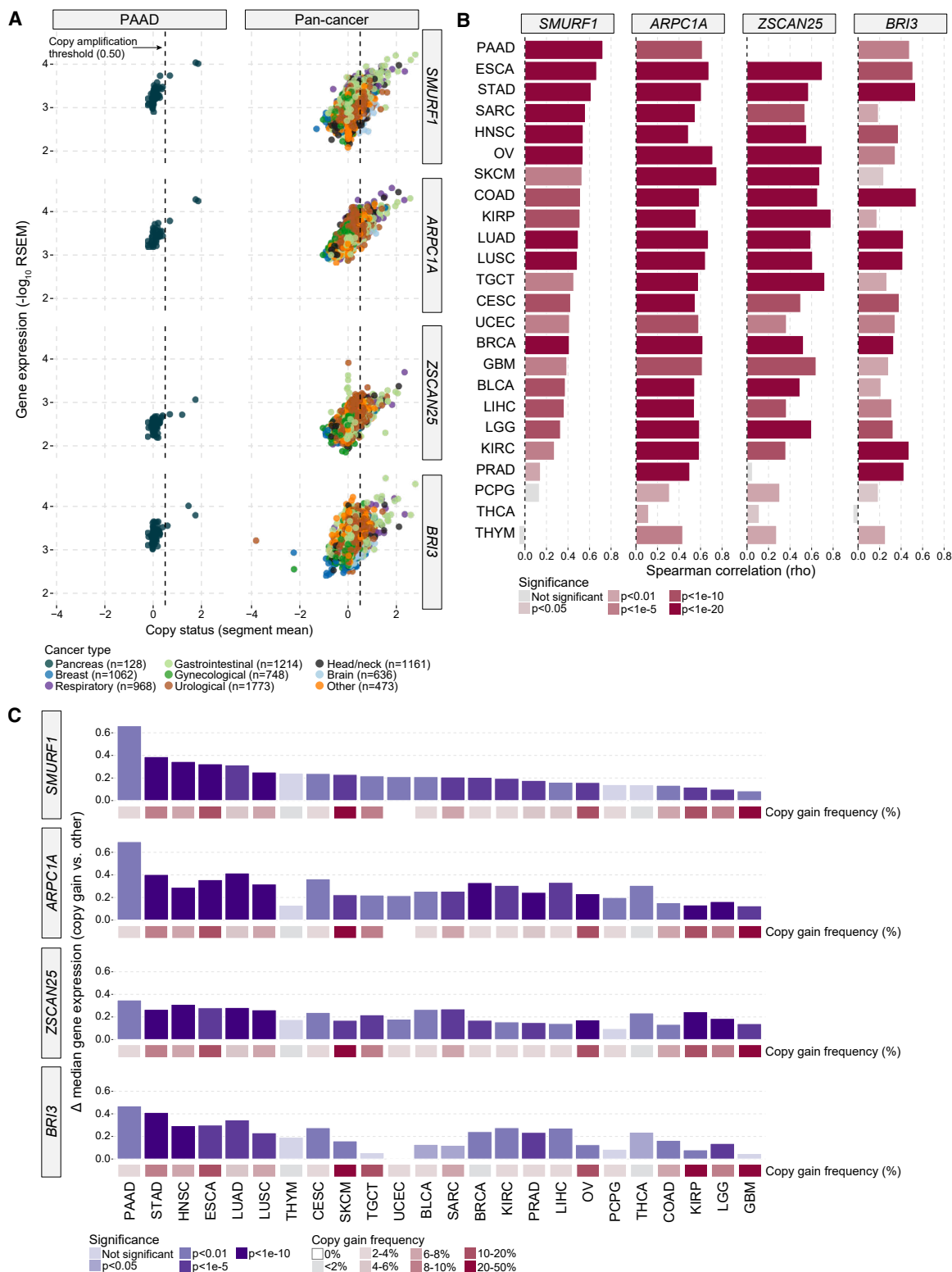


Figure 4. Extrapolation of copy number versus expression analysis to 24 cancer types

(A) Scatterplot shows gene expression (y axis) versus copy status (segment mean; x axis) for *SMURF1*, *ARPC1A*, *ZSCAN25*, and *BRI3* (top to bottom) in pancreatic cancer samples (PAAD; left) and all other cancer types (right). Cancer type acronyms are based on TCGA barcodes. Vertical dashed line indicates threshold segment mean to be considered as copy amplification (segment mean = 0.50).

(legend continued on next page)

most frequently occurring events, our integration of matched RNA-seq data enabled the stratification of CNVs by their transcriptional impact. This strategy revealed copy amplification coupled with overexpression across multiple genes, which were further stratified by the relationship between the expression of affected genes and patient survival.

SMURF1, *ARPC1A*, *ZSCAN25*, and *BRI3* were identified as genes that exhibited recurrent focal amplifications with concomitant mRNA overexpression, with each having mRNA expression levels associated with poor prognosis in the PanGen cohort. These four genes are located on a focal region of chromosome 7 (chr7q21/22) and were co-amplified in the majority of affected samples. Somatic copy amplification of chr7q21/22 has been previously reported in both breast and glioma cancers.^{16,17} While protein levels of *ARPC1A* were not significantly different between amplified versus copy-neutral samples, protein data were not available for *SMURF1*, *ZSCAN25*, and *BRI3*, leaving the impact of these copy alterations on protein expression an open question that remains outside the scope of this study. *ARPC1A* has been found to play a role in cytoskeletal formation, cell migration, and is associated with poor prognosis in prostate cancer.¹⁸ *ZSCAN25*, also known as ZNF498, is a member of the KZFP family of transcriptional regulators and has been found to directly inhibit p53 activity, and is associated with poor prognosis, in hepatocellular carcinoma.¹⁹ While less studied, *BRI3* has been linked to tumor necrosis factor (TNF)-related cell death in L929 cell lines.²⁰ Despite the prognostic co-amplification of these four genes residing in close proximity on chr7q21/22, the functional driver within this genomic region remains unclear and is a topic for future study.

Our analysis highlighted focal *SMURF1* amplifications among the CNVs with the highest transcriptional impact in PDAC. *SMURF1* functions as an E3 ubiquitin ligase, tagging target substrates for proteasome degradation. Specifically, *SMURF1* is involved in TGF-beta and bone morphogenic protein (BMP) pathways through interaction with SMAD family proteins, and has been found to promote tumorigenesis in multiple cancer types.²¹ The oncogenic role of *SMURF1* has been demonstrated in pancreatic cancer cell lines, where *SMURF1* amplification and overexpression lead to increased cell invasiveness.²² Here, we demonstrate that *SMURF1* amplification and overexpression is among the most significant correlations seen, is generalizable across multiple independent studies and is strongest in PDAC compared to 23 other cancer types. While *SMURF1* expression was found to be associated with poor prognosis in the PanGen discovery cohort and the Hartwig external cohort (both metastatic PDAC studies), an opposite effect was observed in the COMPASS external cohort (metastatic PDAC), where high *SMURF1* expression was associated with improved survival outcome. In addition to stark fundamental differences across the validation PDAC cohorts, such as disease stage (ICGC and TCGA primary disease; COMPASS and Hartwig metastatic disease) and laser capture microdissection

(present in ICGC and COMPASS), previous studies have indicated differences in the molecular profiles of these cohorts, such as *NRG1* fusion rates.¹² Considering this limitation, the reliability for *SMURF1* expression to be prognostic across metastatic PDAC cohorts remains an open question that may depend on factors such as biopsy location and sample purity. As successful *SMURF1* targeting has been reported using proteasome inhibitors such as bortezomib,²³ our findings provide rationale for future study of this gene in PDAC.

Taken together, these data compile a detailed overview of copy alterations and their impact on expression in PDAC. By integrating genome, transcriptome, proteomics and clinical metadata, our study highlights the co-amplification of chr7q21/22 genes *SMURF1*, *ARPC1A*, *ZSCAN25* and *BRI3* as an important molecular event for future study that may offer an opportunity toward generating personalized treatment strategies for PDAC.

Limitations of the study

Due to the retrospective nature of this study, our results are observational and hypothesis-generating. By studying biopsy samples obtained from patients diagnosed with PDAC, our study is inherently limited by a small sample size. Late onset of symptoms, rapid disease progression and urgency to initiate first line chemotherapy, as well as technical challenges in obtaining high tumor content biopsies, present challenges in profiling tumor tissue from patients. While the incorporation of external datasets helps to alleviate this issue, fundamental differences between datasets, such as sequencing methodology and chemotherapies used at the time the data were generated, prevent the data from being directly aggregated into a single large-scale analysis. It is well-established that many cancers, including PDAC, have genomic and transcriptomic profiles that are spatially heterogeneous across the bulk of the tumor sample. As our study utilizes core needle-based biopsies, only a small section of the tumor is sampled, and the degree to which the location sampled is representative of the rest of the tumor remains unknown.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Dr. David Schaeffer (David.Schaeffer@vch.ca).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Genomic data generated within the PanGen and COMPASS studies are available in the European Genome-phenome Archive (EGA) under accession numbers #EGAS00001001159 and #EGAS00001002543, respectively. Raw protein data are available in the Proteomics Identifications Database (PRIDE) under accession number [PXD036632](https://www.ebi.ac.uk/pride/archive/study/PSX036632). Hartwig data were accessed through

(B) Bar plot shows the strength of correlation (x axis) between gene expression versus copy number for *SMURF1*, *ARPC1A*, *ZSCAN25*, and *BRI3* (left to right) across 24 cancer types (y axis). Bars are colored by Spearman correlation p value.

(C) Bar plot indicates the magnitude difference in gene expression in copy-amplified samples compared to non-copy-amplified samples across cancer types (x axis) for *SMURF1*, *ARPC1A*, *ZSCAN25*, and *BRI3* (top to bottom). Bars are colored by the Wilcoxon mean rank-sum p value. Lower heatmaps indicate the frequency of copy number amplification in each cancer type.

the Hartwig Medical Foundation database (<https://www.hartwigmedicalfoundation.nl>). All original code has been deposited at https://github.com/jtopham/publications/tree/main/fxbin_prj and is publicly available as of the date of publication.

ACKNOWLEDGMENTS

We gratefully acknowledge the participation of patients and their families, and the PanGen, Genome Sciences Center and COMPASS teams. This research was supported through philanthropic donations received through the BC Cancer Foundation, as well as funding provided by the Terry Fox Research Institute (Project 1078), Ontario Institute for Cancer Research (PanCuRx Translational Research Initiative), Pancreatic Cancer Canada, Genome British Columbia (project B20POG) and VGH and UBC Hospital Foundation. The views expressed in this publication are the views of the authors and do not necessarily reflect those of The Terry Fox Research Institute or the Terry Fox Foundation. This publication and the underlying study have been made possible partly on the basis of the data that Hartwig Medical Foundation have made available to the study. The results published here are in part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI (<http://cancergenome.nih.gov>), as well as data generated by the International Cancer Genome Consortium (<https://icgc.org>).

AUTHOR CONTRIBUTIONS

J.T.T., J.M.K., D.J.R., and D.F.S. conceived and designed the study. J.T.T., J.M.K., M.K., G.L.N., S.E.S., G.H.J., G.M.O., R.A.M., A.J.M., J.M.L., F.N., J.M.W., O.F.B., P.A.T., R.G., G.B.M., J.J.K., S.G., J.L., M.A.M. S.J.M.J., D.J.R., and D.F.S. contributed to data acquisition. J.T.T. and J.M.K. contributed to data analysis. J.T.T., J.M.K., A.M., H.A., S.E.K., M.K., E.T., J.M.L., D.J.R., and D.F.S. contributed to data interpretations. J.T.T., J.M.K., E.T., J.M.L., M.A.M., D.J.R., and D.F.S. contributed to writing the article.

DECLARATION OF INTERESTS

We declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS**
 - PanGen study enrollment
- **METHOD DETAILS**
 - PanGen sample genome and transcriptome sequencing
 - Proteomics data generation
 - PanGen sample somatic mutation calling
 - Tumor mutational burden
 - PDAC subtyping
 - Genome-wide CNV assessment
 - RNA-impactful bin identification
 - Gene set enrichment analysis
 - Survival analysis
 - Validation PDAC datasets
 - Pan-cancer analysis
- **QUANTIFICATION AND STATISTICAL ANALYSIS**

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.isci.2025.113176>.

Received: May 30, 2024
Revised: April 8, 2025
Accepted: July 17, 2025
Published: July 22, 2025

REFERENCES

1. Shah, N.P., Kasap, C., Weier, C., Balbas, M., Nicoll, J.M., Bleickardt, E., Nicaise, C., and Sawyers, C.L. (2008). Transient potent BCR-ABL inhibition is sufficient to commit chronic myeloid leukemia cells irreversibly to apoptosis. *Cancer Cell* 14, 485–493.
2. Slamon, D.J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., Wolter, J., Pegram, M., et al. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N. Engl. J. Med.* 344, 783–792.
3. Koulis, C., Yap, R., Engel, R., Jardé, T., Wilkins, S., Solon, G., Shapiro, J. D., Abud, H., and McMurrick, P. (2020). Personalized Medicine-Current and Emerging Predictive and Prognostic Biomarkers in Colorectal Cancer. *Cancers* 12, 812.
4. Siegel, R.L., Miller, K.D., Wagle, N.S., and Jemal, A. (2023). Cancer statistics, 2023. *CA Cancer J. Clin.* 73, 17–48.
5. Topham, J.T., Renouf, D.J., and Schaeffer, D.F. (2023). Circulating tumor DNA: toward evolving the clinical paradigm of pancreatic ductal adenocarcinoma. *Ther. Adv. Med. Oncol.* 15, 17588359231157651.
6. Rashid, N.U., Peng, X.L., Jin, C., Moffitt, R.A., Volmar, K.E., Belt, B.A., Panni, R.Z., Nywening, T.M., Herrera, S.G., Moore, K.J., et al. (2020). Purity Independent Subtyping of Tumors (PuriST), A Clinically Robust, Single-sample Classifier for Tumor Subtyping in Pancreatic Cancer. *Clin. Cancer Res.* 26, 82–92.
7. Waddell, N., Pajic, M., Patch, A.M., Chang, D.K., Kassahn, K.S., Bailey, P., Johns, A.L., Miller, D., Nones, K., Quek, K., et al. (2015). Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* 518, 495–501.
8. Aung, K.L., Fischer, S.E., Denroche, R.E., Jang, G.H., Dodd, A., Creighton, S., Southwood, B., Liang, S.B., Chadwick, D., Zhang, A., et al. (2018). Genomics-Driven Precision Medicine for Advanced Pancreatic Cancer: Early Results from the COMPASS Trial. *Clin. Cancer Res.* 24, 1344–1354.
9. Bailey, P., Chang, D.K., Nones, K., Johns, A.L., Patch, A.M., Gingras, M. C., Miller, D.K., Christ, A.N., Bruxner, T.J.C., Quinn, M.C., et al. (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* 531, 47–52.
10. Wartenberg, M., Cibin, S., Zlobec, I., Vassella, E., Eppenberger-Castori, S., Terracciano, L., Eichmann, M.D., Worni, M., Gloor, B., Perren, A., and Karamitopoulou, E. (2018). Integrated Genomic and Immunophenotypic Classification of Pancreatic Cancer Reveals Three Distinct Subtypes with Prognostic/Predictive Significance. *Clin. Cancer Res.* 24, 4444–4454.
11. Raphael, B.J., Hruban, R.H., Aguirre, A.J., Moffitt, R.A., Yeh, J.J., Stewart, C., Robertson, A.G., Cherniack, A.D., Gupta, M., Getz, G., et al. (2017). Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell* 32, 185–203.e13.
12. Topham, J.T., Tsang, E.S., Karasinska, J.M., Metcalfe, A., Ali, H., Kalloger, S.E., Csizmek, V., Williamson, L.M., Titmuss, E., Nielsen, K., et al. (2022). Integrative analysis of KRAS wildtype metastatic pancreatic ductal adenocarcinoma reveals mutation and expression-based similarities to cholangiocarcinoma. *Nat. Commun.* 13, 5941.
13. Chan-Seng-Yue, M., Kim, J.C., Wilson, G.W., Ng, K., Figueroa, E.F., O’Kane, G.M., Connor, A.A., Denroche, R.E., Grant, R.C., McLeod, J., et al. (2020). Transcription phenotypes of pancreatic cancer are driven by genomic events during tumor evolution. *Nat. Genet.* 52, 231–240.
14. Shao, X., Lv, N., Liao, J., Long, J., Xue, R., Ai, N., Xu, D., and Fan, X. (2019). Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Med. Genet.* 20, 175.
15. Bujold, D., Morais, D.A.d.L., Gauthier, C., Côté, C., Caron, M., Kwan, T., Chen, K.C., Laperle, J., Markovits, A.N., Pastinen, T., et al. (2016). The International Human Epigenome Consortium Data Portal. *Cell Syst.* 3, 496–499.e2.

16. Nikolsky, Y., Sviridov, E., Yao, J., Dosymbekov, D., Ustyansky, V., Kaznacheev, V., Dezso, Z., Mulvey, L., Macconail, L.E., Winckler, W., et al. (2008). Genome-wide functional synergy between amplified and mutated genes in human breast cancer. *Cancer Res.* 68, 9532–9540.
17. Tamayo, P., Cho, Y.-J., Tsherniak, A., Greulich, H., Ambrogio, L., Schouten-van Meeteren, N., Zhou, T., Buxton, A., Kool, M., Meyerson, M., et al. (2011). Predicting relapse in patients with medulloblastoma by integrating evidence from clinical and genomic features. *J. Clin. Oncol.* 29, 1415–1423.
18. Chen, Y.-H., Chen, H., Lin, T.T., Zhu, J.M., Chen, J.Y., Dong, R.N., Chen, S.H., Lin, F., Ke, Z.B., Huang, J.B., et al. (2023). ARPC1A correlates with poor prognosis in prostate cancer and is up-regulated by glutamine metabolism to promote tumor cell migration, invasion and cytoskeletal changes. *Cell Biosci.* 13, 38.
19. Zhang, X., Zheng, Q., Yue, X., Yuan, Z., Ling, J., Yuan, Y., Liang, Y., Sun, A., Liu, Y., Li, H., et al. (2022). ZNF498 promotes hepatocellular carcinogenesis by suppressing p53-mediated apoptosis and ferroptosis via the attenuation of p53 Ser46 phosphorylation. *J. Exp. Clin. Cancer Res.* 41, 79.
20. Wu, H., Liu, G., Li, C., and Zhao, S. (2003). bri3, a novel gene, participates in tumor necrosis factor- α -induced cell death. *Biochem. Biophys. Res. Commun.* 311, 518–524.
21. Xia, Q., Li, Y., Han, D., and Dong, L. (2021). SMURF1, a promoter of tumor cell progression? *Cancer Gene Ther.* 28, 551–565.
22. Kwei, K.A., Shain, A.H., Bair, R., Montgomery, K., Karikari, C.A., van de Rijn, M., Hidalgo, M., Maitra, A., Bashyam, M.D., and Pollack, J.R. (2011). SMURF1 Amplification Promotes Invasiveness in Pancreatic Cancer. *PLoS One* 6, e23924.
23. Wang, Z., Wang, J., Li, X., Xing, L., Ding, Y., Shi, P., Zhang, Y., Guo, S., Shu, X., and Shan, B. (2014). Bortezomib prevents oncogenesis and bone metastasis of prostate cancer by inhibiting WWP1, Smurf1 and Smurf2. *Int. J. Oncol.* 45, 1469–1478.
24. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291.
25. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* 25, 1754–1760.
26. Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., and Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. *Bioinforma. Oxf. Engl.* 31, 2032–2034.
27. 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. *Bioinforma. Oxf. Engl.* 29, 15–21.
28. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinforma. Oxf. Engl.* 30, 923–930.
29. Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., and Saunders, C.T. (2018). Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* 15, 591–594.
30. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinforma. Oxf. Engl.* 32, 1220–1222.
31. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6, 80–92.
32. Shen, R., and Seshan, V.E. (2016). FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* 44, e131.
33. Zhu, Y., Orre, L.M., Zhou Tran, Y., Mermelekas, G., Johansson, H.J., Malyutina, A., Anders, S., and Lehtiö, J. (2020). DEqMS: A Method for Accurate Variance Estimation in Differential Protein Expression Analysis. *Mol. Cell. Proteomics* 19, 1047–1057.
34. Uhrig, S., Ellermann, J., Walther, T., Burkhardt, P., Fröhlich, M., Hutter, B., Toprak, U.H., Neumann, O., Stenzinger, A., Scholl, C., et al. (2021). Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res.* 31, 448–460.
35. Merino, D.M., McShane, L.M., Fabrizio, D., Funari, V., Chen, S.J., White, J. R., Wenz, P., Baden, J., Barrett, J.C., Chaudhary, R., et al. (2020). Establishing guidelines to harmonize tumor mutational burden (TMB): in silico assessment of variation in TMB quantification across diagnostic platforms: phase I of the Friends of Cancer Research TMB Harmonization Project. *J. Immunother. Cancer* 8, e000147.
36. Ellrott, K., Bailey, M.H., Saksena, G., Covington, K.R., Kandath, C., Stewart, C., Hess, J., Ma, S., Chiotti, K.E., McLellan, M., et al. (2018). Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst.* 6, 271–281.e7.
37. Topham, J.T., Karasinska, J.M., Lee, M.K.C., Csizmek, V., Williamson, L. M., Jang, G.H., Denroche, R.E., Tsang, E.S., Kalloger, S.E., Wong, H.L., et al. (2021). Subtype-Discordant Pancreatic Ductal Adenocarcinoma Tumors Show Intermediate Clinical and Molecular Characteristics. *Clin. Cancer Res.* 27, 150–157.
38. Karasinska, J.M., Topham, J.T., Kalloger, S.E., Jang, G.H., Denroche, R. E., Culibrk, L., Williamson, L.M., Wong, H.L., Lee, M.K.C., O'Kane, G.M., et al. (2020). Altered Gene Expression along the Glycolysis-Cholesterol Synthesis Axis Is Associated with Outcome in Pancreatic Cancer. *Clin. Cancer Res.* 26, 135–146.
39. Collisson, E.A., Sadanandam, A., Olson, P., Gibb, W.J., Truitt, M., Gu, S., Cooc, J., Weinkle, J., Kim, G.E., Jakkula, L., et al. (2011). Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat. Med.* 17, 500–503.
40. Krijgsman, O., Carvalho, B., Meijer, G.A., Steenbergen, R.D.M., and Ylstra, B. (2014). Focal chromosomal copy number aberrations in cancer—Neddles in a genome haystack. *Biochim. Biophys. Acta* 1843, 2698–2704.
41. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* 1, 417–425.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Mass spectrometry protein quantification data	This paper	PRIDE: PXD036632
RNAseq data; PanGen cohort	This paper	EGA: EGAS00001001159
RNAseq data; COMPASS cohort	Aung et al. ⁸	EGA: EGAS00001002543
RNAseq data; Hartwig cohort	Hartwig Foundation	NA
RNAseq data; ICGC cohort	https://portal.gdc.cancer.gov/	NA
RNAseq data; TCGA cohort	https://portal.gdc.cancer.gov/	NA
Mutation data; PanGen cohort	This paper	EGA: EGAS00001001159
Mutation data; COMPASS cohort	Aung et al. ⁸	EGA: EGAS00001002543
Mutation data; Hartwig cohort	Hartwig Foundation	NA
Mutation data; ICGC cohort	https://portal.gdc.cancer.gov/	NA
Mutation data; TCGA cohort	https://portal.gdc.cancer.gov/	NA
Clinical data; PanGen cohort	This paper	NA
Clinical data; COMPASS cohort	Aung et al. ⁸	NA
Clinical data; Hartwig cohort	Hartwig Foundation	NA
Clinical data; ICGC cohort	https://portal.gdc.cancer.gov/	NA
Clinical data; TCGA cohort	https://portal.gdc.cancer.gov/	NA
mSigDB gene sets	https://www.gsea-msigdb.org/gsea/msigdb/	NA
IHEC epigenetics ChIP-seq dataset (ENCODE)	https://epigenomesportal.ca/ihec	ENCBS209WBW, ENCBS869MHP
Refseq gene annotations	https://www.ncbi.nlm.nih.gov/refseq/	NA
Consensus exon regions (GRCh37.p13)	https://www.ncbi.nlm.nih.gov	GCF_000001405.25
ExAC common variants (hg19)	Lek et al. ²⁴	ExAC_nonTCGA.r0.3.1.sites.vcf.gz
Software and algorithms		
BWA-mem	Li et al., 2009 ²⁵	v0.7.6a
sambamba	Tarasov et al. ²⁶	v0.5.5
STAR	Dobin et al. ²⁷	v2.7.3
PicardTools	https://broadinstitute.github.io/picard/	v2.17.3
Subread	Liao et al. ²⁸	v1.4.6
Strelka	Kim et al., 2018 ²⁹	v2.9.10
Manta	Chen et al., 2016 ³⁰	v1.5.0
SnEff	Cingolani et al. ³¹	V4.3
Facets	Shen et al. ³²	v0.6.0
bedtools	https://bedtools.readthedocs.io/en/latest/	v2.26.0
vcf2maf	https://github.com/mskcc/vcf2maf	v1.6.18
Human genome build hg19 (GRCh37)	https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.25/	NA
R	https://www.r-project.org	v3.6.3
Analysis code	https://github.com/jtopham/publications/tree/main/fxbin_prj	NA

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

PanGen study enrollment

All patients with tumor samples belonging to the discovery cohort ($n = 69$) had metastatic (unresectable) disease. Presence of a primary tumor located on the pancreas was noted for all patients, with subsequent histopathological confirmation of PDAC. Patients

were enrolled in the Prospectively Defining Metastatic Pancreatic Ductal Adenocarcinoma Subtypes by Comprehensive Genomic Analysis (PanGen; NCT02869802) clinical trial between October 2016 and May 2022. All patients received tumor biopsy and sequencing prior to initiating treatment for their metastatic disease. Metastatic tumor lesions were biopsied for sequencing analysis. Patient samples did not receive laser-capture microdissection prior to sequencing. The PanGen trial was approved by the University of British Columbia Research Ethics Committee (REB# H14-00291) and was conducted in accordance with international ethical guidelines. Written informed consent was obtained from each patient prior to molecular profiling. All sequencing data were housed in a secure computing environment.

METHOD DETAILS

PanGen sample genome and transcriptome sequencing

Tumor and matched normal (blood) samples received whole genome sequencing (WGS) with target depths of 80x and 40x, respectively. WGS libraries had reads trimmed to 75 base pairs (bp) and were aligned (hg19; GRCh37-lite) using BWA-mem v0.7.6a²⁵ (default parameters). WGS duplicate reads were marked using sambamba v0.5.5²⁶ (default parameters). RNA-sequencing (RNA-seq) was performed on tumor samples with a target depth of 200 million reads. RNA-seq reads were trimmed to 75bp and aligned (GRCh37-lite) using STAR v2.7.3,²⁷ with parameters: *-chimSegmentMin 20 -outSAMmultNmax 1 -outSAMstrandField intron-Motif -outFilterIntronMotifs RemoveNoncanonical*. RNA-seq duplicate reads were marked using PicardTools v2.17.3. Raw reads counts were assigned to Ensembl 75 genes using Subread v1.4.6,²⁸ normalized for library depth and gene size (RPKM) and subsequently log₁₀-transformed.

Proteomics data generation

Mass spectrometry-based proteomics sequencing of clinical (PanGen) tumor samples was performed using the SP3-CTP pipeline as previously described.¹² Protein data were median-centre normalized using the DEqMS package in R.³³

PanGen sample somatic mutation calling

Somatic SNV/indels were called using paired tumor/normal WGS libraries using a combination of Strelka v2.9.10²⁹ and Manta v1.5.0³⁰ (default parameters; genome build GRCh37). Variant annotation was performed using SnpEff v4.3³¹ with parameters: *-v GRCh37.75 -canon -no-downstream -no-upstream -noLog -noStats -no-intergenic*. CNVs and tumor ploidy were calculated using Facets v0.6.0³² (default parameters). Copy number amplification was defined as a segment having a total copy number greater than or equal to twice the tumor ploidy. Copy number segments were mapped to Refseq genes using bedtools v2.26.0. Somatic fusion events were identified based on RNA-seq data using Arriba v1.2.0³⁴ (default parameters), with a separate STAR alignment (recommended fusion calling parameters: *-outFilterMismatchNmax 3 -chimSegmentMin 10 -chimOutType WithinBAM SoftClip -chimJunctionOverhangMin 10 -chimScoreMin 1 -chimScoreDropMax 30 -chimScoreJunctionNonGTAG 0 -chimScoreSeparation 1 -alignSJstitchMismatchNmax 5 -1 5 5 -chimSegmentReadGapMax 3*). Fusion events were filtered for in-frame events reported at a confidence level of “high” by Arriba.

Tumor mutational burden

For TMB calculation, VCF files were converted to MAF format using vcf2maf v1.6.18 with default parameters. Variants were filtered to exclude those located outside of exons (using consensus exon regions GRCh37.p13; GCF_000001405.25, downloaded June 19, 2020) and any common variants identified by ExAC²⁴ (ExAC_nonTCGA.r0.3.1.sites.vcf.gz). Variants were filtered based on established guidelines,^{35,36} including: VAF ≥ 0.05, tumor depth ≥ 25 and alternate allele count ≥ 3. Only missense, nonsense and in-frame/frameshift variants were included for TMB calculation. TMB was calculated using the sum of all filtered variants in a sample, and 32.102474 Mb was used as the denominator.³⁵

PDAC subtyping

Moffitt subtype classification was performed using the RNA-seq version of the Moffitt PuriST algorithm.⁶ PuriST scores (basal-like probability values) were used to stratify patients into basal-like (score > 0.75), classical (score < 0.25) and intermediate (score [0.25–0.75]) subtype groups.³⁷ Collisson and Bailey subtypes were determined using a consensus clustering approach, as previously described.^{38,39} Karasinska subtypes were determined based on the relative expression of glycolytic (*GAPDH*, *ALDOA*, *PKM*, *ENO1*, *TP11*, *PGK1*, *GPI*, *PGAM1*, *PFKP*, *PFKFB3*, *ENO2*, *PPP2R5D*, *PFKM*, *PFKFB4*) and cholesterologenic (*FDPS*, *FDFT1*, *DHCR24*, *EBP*, *IDI1*, *MVD*, *HMGCS1*, *SQLE*, *NSDHL*, *DHCR7*, *HMGCR*, *LSS*, *SC5D*, *MVK*, *HSD17B7*) genes, as previously described.³⁸

Genome-wide CNV assessment

Human genome build hg19 (GRCh37) was segmented into 144,061 non-overlapping 20 kb bins. To determine the copy status of each bin, for each sample, each bin was mapped to the highest-overlapping copy number segment for that sample using bedtools v2.26.0. Bins with coordinates overlapping a centromere were excluded from analysis. For visualization purposes only, adjacent bins with identical copy status in all samples were merged and referred to as ‘genome segments’. To calculate RNA-seq coverage, the number

of overlapping RNA-seq reads centers was determined for each bin, in each sample, using bedtools v2.26.0. Read coverage was normalized for sequencing depth and $\log_{10}$ -transformed. Mutation density was assessed for each bin, in each sample, by counting the number of overlapping somatic SNV and indels in the respective sample using bedtools v.2.26.0.

RNA-impactful bin identification

Bins with at least three samples (3/69; 4.35% of patients) with amplification ($n = 23,893$ bins) or at least three samples with homozygous deletions ($n = 1,622$) were chosen for analysis, and the number of RNA-seq read centers aligning to each bin was calculated (hereafter referred to as 'RNA coverage'). For each bin, RNA coverage was then compared between samples with the CNV event versus samples with neutral copy state (difference in median levels; Wilcoxon mean rank-sum test), and p values were subjected to Benjamini-Hochberg (BH) multiple test correction. Bins that had significantly different ($p < 0.05$) RNA coverage were labeled as 'RNA-impactful bins', and a threshold difference in median RNA coverage greater than 0.25 or less than $-0.25 \log_{10}$ RPKM were used to further stratify bins. For downstream analysis of genes affected by RNA-impactful bins, only amplifications that were part of a focal CNV segment in all affected patients were considered. Focal events were defined as CNV segments less than or equal to 3 Mb in size.⁴⁰ To assess overlap between bins and ChIP-seq marks indicative of enhancer regions, data from two healthy pancreatic tissue samples from IHEC were used (<https://epigenomesportal.ca/ihec/>; hg19 build 2010-10; ENCODE; downloaded March 28, 2024). Each bin was segmented into 20 non-overlapping windows of size 1 kb, and assessed for overlap with ChIP-seq peaks using bedtools. Enhancer types were defined for each 1 kb window, and active enhancers were windows that overlap both H3K4me1 and H3K27ac in the absence of H3K27me3. Poised enhancers were defined as overlapping H3K4me1 and H3K27me3 in the absence of H3K27ac, and primed enhancers as overlapping H3K4me1 only.

Gene set enrichment analysis

Genes overlapped by RNA-impactful bins were assessed for pathway enrichment using multiple hypergeometric tests, as previously described,¹² and included 32,284 gene sets accessed through the Molecular Signatures Database (mSigDB⁴¹).

Survival analysis

For genes affected by filtered RNA-impactful bins, high and low expression groups were defined as samples with higher than 75th percentile expression or lower than 25th percentile expression, respectively. Genes for which the Cox proportional hazards assumption was not met were excluded from survival analysis. HR and 95% CI values were calculated using Cox proportional hazard models, and log rank p values were generated and subjected to BH multiple test correction.

Validation PDAC datasets

Metastatic PDAC samples in the COMPASS validation cohort ($n = 195$) were derived from patients enrolled and sequenced as part of the Comprehensive Molecular Characterization of Advanced PDAC For Better Treatment Selection (COMPASS; NCT02750657) trial. COMPASS patient samples had RNA-seq ($n = 195$) and WGS ($n = 195$) data generated and processed as described previously.⁸ Overall survival data was available for all patients ($n = 195$). Metastatic PDAC samples in the Hartwig dataset were accessed through the Hartwig Medical Foundation database. Hartwig WGS data ($n = 113$) were accessed in the form of Purple and Linx tool outputs. Hartwig RNA-seq data ($n = 46$) were processed using the same pipeline as PanGen samples. For each PDAC validation dataset, normalized gene expression values (FPKM for COMPASS, RPKM for Hartwig) were $\log_{10}$ -transformed prior to analysis. All somatic mutation data were based on human genome build GRCh37 (hg19). Overall survival data were available for 84 patients in the Hartwig database. Mutation (SNV, indel, CNV) and RNA-seq data for resectable PDAC samples from the TCGA ($n = 94$) and ICGC ($n = 92$) database were downloaded and samples were filtered as described previously.³⁸ TCGA and ICGC RNA-seq data were available as TPM and FPKM, respectively. For validation analyses using the external datasets, allele-specific CNV data were used, and copy number amplifications at given loci were defined as a total copy status greater than twice the tumor ploidy.

Pan-cancer analysis

For all available tumor types, RNA-seq (RSEM) data was downloaded from the GDC TCGA data portal as previously described³⁸ and filtered for tumor samples based on TCGA barcodes. Samples with both RNA-seq and CNV (microarray, measured as the continuous variable 'segment mean') data belonging to tumor types with at least 100 samples were included for analysis ($n = 8,163$ samples from 24 cancer types). For visualization purposes only, the following TCGA cancer type codes were merged into more generalized groups: gastrointestinal ("COAD", "READ", "ESCA", "STAD", "LIHC"), urological ("BLCA", "KIRC", "KIRP", "PRAD", "TGCT"), gynecological ("CESC", "OV", "UCEC"), head/neck ("HNSC", "PCPG", "THCA"), respiratory ("LUAD", "LUSC") and brain ("GBM", "LGG"). Copy number segments with at least 10 supporting probes and a segment mean greater than 0.50 were defined as amplifications.

QUANTIFICATION AND STATISTICAL ANALYSIS

Fisher's exact tests were used to compare the distribution of discrete variables between two groups. Wilcoxon mean rank-sum tests were used for two-group comparison of continuous variables. Spearman correlations were performed to assess correlation between two continuous variables. HR and 95% CIs were calculated using Cox proportional hazards models. Log rank tests were used to calculate *p* values in OS analysis. All comparison tests were two-tailed. All *p* values were subjected to BH multiple test correction when applicable. All analyses were performed using R v3.6.3.