A central research portal for mining pancreatic clinical and molecular datasets and accessing biobanked samples.

J. Oscanoa[1,a], H Ross-Adams[1,a], Abu Z M Dayem Ullah[1,2], TS Kolvekar[1,2], L Sivapalan[1], E Gadaleta[1], GJ Thorn[1], M Abdollahyan[1], A Imrali[2], A Saad[2, 4], R Roberts[2], C Hughes[2], PCRFTB, HM Kocher[2, 3, 4], C Chelala[1,2, *]

[1]Centre for Cancer Biomarkers and Biotherapeutics, Barts Cancer Institute, Queen Mary University London, UK, EC1M 6BQ

[2] Pancreatic Cancer Research Fund Tissue Bank, Centre for Tumour Biology, Barts Cancer Institute, Queen Mary University London, UK, EC1M 6BQ

[3] Centre for Tumour Biology, Barts Cancer Institute, Queen Mary University London, UK, EC1M 6BQ

[4] Barts and the London HPB Centre, The Royal London Hospital, Barts Health NHS Trust, Whitechapel, London E1 1BB

*Corresponding Author: Prof C Chelala, c.chelala@qmul.ac.uk

[a]These authors contributed equally.

PCRFTB: Mo Abu Hilal (Southampton), Bilal Al Sarireh (Swansea), Somaiah Aroori (Plymouth), Ali Arshad (Southampton), Satyajit Bhattacharya (The London Clinic, London), Brian Davidson (Royal Free, London), John Isherwood (Leicester), Deep Malde (Leicester), Stuart Robinson (Newcastle), Zahir Soonawalla (Oxford).

Abstract

The Pancreatic Expression Database (PED) is a powerful resource dedicated to the mining and analysis of pancreatic -omics datasets. Here, we demonstrate the biological interpretations that are possible because of vital updates that have transformed PED into a dynamic analytics hub accommodating an extensive range of publicly available datasets. PED now hosts clinical and molecular datasets from four primary sources (Cancer Genome Atlas, International Cancer Genome Consortium, Cancer Cell Line Encyclopaedia and Genomics Evidence Neoplasia Information Exchange) that together form the foundation of omics profiling of pancreatic malignancies and related lesions (n=7,760 specimens). Several user-friendly analytical tools to explore and integrate the molecular data derived from these primary specimens and cell lines are now available. Crucially, PED is integrated as the data access point for Pancreatic Cancer Research Fund Tissue Bank – the only national pancreatic cancer biobank in the UK. This will pioneer a new era of biobanking to promote collaborative studies and effective sharing of multi-modal molecular, histopathology and imaging data from biobank samples (>60,000 specimens from >3,400 cases and controls; 2,037 H&E images from 349 donors) and accelerate validation of *in silico* findings in patient-derived material. These updates place PED at the analytical forefront of pancreatic biomarker-based research, providing the user community with a distinct resource to facilitate hypothesis-testing on public data, validate novel research findings, and access curated, high-quality patient tissues for translational research. To demonstrate the practical utility of PED, we investigate somatic variants associated with established transcriptomic subtypes and disease prognosis: several patient-specific variants are clinically actionable and may be leveraged for precision medicine.

Introduction

Pancreatic ductal adenocarcinoma (PDAC) is predicted to become the second leading cause of cancer-related mortality worldwide before 2040[1]. It has dismal 5-year survival rates of 3-15%[1,2], largely due to late disease detection and few effective treatment options. Alarmingly, the incidence of early-onset pancreatic cancer is also increasing, in contrast to most other solid tumours[3]. Better tools for patient stratification and treatment response are thus

essential to improve survival outcomes. However, the pancreatic cancer research community is relatively small – investigators tend to develop bespoke collections of samples that may be unusable beyond the breadth and scope of their ethical approval and storage conditions. Sample collection protocols also vary widely, further limiting the translation of derived results to clinical benefits.

Most existing biomarkers for monitoring treatment or assessing prognosis are not based on the molecular attributes of PDAC tumours and have shown limited sensitivity and/or specificity in prospective settings[4]. Numerous studies into the genomic and transcriptomic determinants of tumour development and progression are available, but their findings are dispersed across multiple resources, and can be difficult to access and translate into meaningful survival or treatment benefits for patients by those without computational expertise. This highlights the pressing need for simplified, integrated data mining and analysis tools to improve accessibility to clinical and molecular information from disparate sources and enable laboratory and clinical researchers to easily and effectively cross-query large multi-omics datasets, to fuel new discoveries in pancreatic diseases.

Here, we present the latest release of the Pancreatic Expression Database (PED), an intuitive, online portal that links numerous multi-modal datasets to an active biobank where users can both validate their findings, and/or apply for samples to confirm *in silico* findings. To remain abreast of the evolving nature of integrative multi-omics workflows and in response to user feedback, we have made vital new updates to PED's Analytics Hub, broadened the range of datasets available and placed this essential resource at the centre of the established framework for PED (Table 1). We have also integrated PED as the major bioinformatics platform of the UK's national Pancreatic Cancer Research Fund Tissue Bank (PCRFTB), to facilitate investigative biomarker-based research and data sharing between clinicians and scientists. This is powered by a customised version of SNPNexus, a versatile platform for the functional annotation of known and novel sequence variation[5], designed to reduce the analytical burden associated with large-scale genomic datasets and facilitate the straightforward identification of biologically and clinically relevant genetic variants in patients.

The PCRFTB is the world's first national pancreas tissue bank and has been collecting blood, urine, saliva and solid tissue samples from patients recruited at nine participating centres across the UK NHS since 2015, making it a valuable resource for translational research. Tissues are available from patients with pancreatic and hepatobiliary diseases, including resectable and unresectable cancer. Blood, urine and saliva samples from patients, their first-degree relatives and other healthy volunteers are also available, as well as cancer organoids and cancer-associated fibroblasts. (Figure 1). These are accompanied by extensive, verified clinical, histological and imaging data that is continually updated – median 300 data points per visit, with some donors providing longitudinal samples at multiple visits throughout their treatment journey. Best practice and ongoing technical research ensures available biological materials are of high quality, to support reliable and reproducible results[6,7] In addition to samples, digitised radiological images are available for 171 patients with malignant, pre-malignant and benign pancreatic diagnoses, and 2,037 H&E images from 349 donors are also currently available.

The PCRFTB is further supported by a Data Return policy to maximise the use of available samples by linking each patient/donor with an enriched 'digital fingerprint' encompassing molecular, transcriptomic, proteomic, imaging and longitudinal clinical data. All donors provide written, informed consent, and all samples are collected, processed and stored at each of the participating centres (Barts, Leicester, Swansea, Oxford, Royal Free (London), Southampton, Newcastle, Plymouth, The London Clinic) under one Research Ethics Committee reference (13/SC/0593, renewed 18/SC/0629, renewed 23/SC/0282) and using standardised protocols, quality assurance and quality control policies ensuring consistency across the collection.

Via PED, researchers can directly query and apply to PCRFTB for samples, specimens and/or imaging data that match user-determined criteria. A link to the PCRFTB Tissue Request System allows users to submit Expressions of Interest directly to the Tissue Bank using an online application form.

By providing a dynamic hub for the analysis of publicly available pancreatic datasets and ongoing research data generated from biobanked samples, PED allows researchers to access a broad range of pancreas-specific molecular information freely and quickly. The flexibility of this hub allows molecular alterations with biological and clinical relevance to be identified and prioritised for downstream validation.

The Analytics Hub

We have updated the web-based Analytics Hub to include a broad range of pancreas-related, publicly available -omics datasets, together with expanded analytical features and visualisation options (Table 1), based on feedback from our diverse international user community.

Publicly available data sources

Building on the 2018 release[8], the newly formed analytics hub now hosts publicly available clinical and molecular datasets from four core sources: The Cancer Genome Atlas (TCGA)[9] whole exome sequencing (WES) data, updated to filter specifically for adenocarcinoma samples vs others; the Cancer Cell Line Encyclopaedia (CCLE)[10], updated to include somatic mutation and mRNA expression data for the complete set of 60 pancreas cell lines from primary and metastatic tumours; the Genomics Evidence Neoplasia Information Exchange (GENIE)[11] v13.0, updated to include simple somatic mutations and clinical data from the *complete* set of 6,633 patients with pancreatic cancer of any type; and the now archived complete International Cancer Genome Consortium (ICGC)[12,13] dataset, including whole genome and RNA sequencing data from both adenocarcinoma (PACA-AU; PACA-CA) and neuro/endocrine tissues (PAEN-IT; PAEN-AU). These sources host data generated by both national and international consortia efforts to sequence and analyse cancer genomes and biology, including pancreatic malignancies. Analysed and quality-controlled data files were downloaded from the respective sources and used without further processing. PED 2024 uses the most recent data releases, including linked clinical data when available.

Advanced filtering of public datasets

Public datasets may be queried according to the clinical characteristics of each study cohort (Figure 2A). Filtering options have been selected based on relevance to disease development and pathogenesis, and the depth of annotation provided in available clinical data for each cohort. Implemented filters are accompanied by dedicated visualizations of clinical summaries for each respective study cohort (Figure 2B). These include filters based on patient-related factors (cancer type, sex, age, diabetes, family history, ethnicity, survival; Figure 2C), and tumour characteristics (stage, grade, *KRAS* somatic mutation status) (Figure 2D), which allow for trends in data to be clearly observed: e.g. survival beyond 3 years is very low for PDAC compared to neuroendocrine tumours (Figure 2B). Crucially, it is possible to filter each dataset by diagnosis, allowing researchers to focus on different pancreatic lesions individually (e.g. IPMN, ductal adenocarcinoma, neuroendocrine, adenosquamous, mucinous) that have different molecular alterations and clinical prospects, since using unstratified sample sets has been shown to yield unreliable results.[14,15]

Characterising the genomic characteristics of established PDAC molecular subtypes

PED facilitates the stratification and analysis of TCGA and ICGC cohorts according to their molecular subtype classifications, as determined by hallmark transcriptomic[16–18] and genomic (ICGC only)[19] studies, and recent histopathology-based artificial intelligence (AI) predictions in matched TCGA samples[20](Figure 2E). We demonstrate the implementation of this clinically useful feature below.

Collisson et al. (2011) originally identified 3 subclasses of PDAC tumours with different clinical outcomes and treatment responses, termed *quasi-mesenchymal* (QM) (worst prognosis), *classical* (best prognosis) and *exocrine-like*, using hybridisation array-based mRNA expression data from primary untreated resected PDAC[16]. Next, Moffitt et al (2015) analysed bulk tumour tissues from treatment-naïve primary resected PDAC tumours using virtual microdissection to exclude transcripts native to the normal pancreas and the tumour microenvironment, and reported 2 distinct tumour subtypes (*basal* and *classical*) as well as 2

classifications based on peritumoural stromal tissues (*activated* and *normal*)[17]. *Basal* subtype tumours were associated with a poorer overall patient survival compared to *classical* tumours, which overlapped significantly with the Collisson *classical* subtype. Subsequently, Bailey et al. performed RNA-sequencing of bulk primary untreated resected tumour tissues from 328 PDAC tumours and resolved four stable tumour classes (*squamous*, *pancreatic progenitor*, *immunogenic* and *aberrantly differentiated endocrine exocrine* (ADEX)), each governed through the differential expression of transcription factors and their targets involved in lineage specification during pancreatic development[18]. *Squamous* subtype tumours overlapped with previous *basal* (Moffitt) and *QM* (Collisson) classifications and were associated with the poorest overall prognosis in patients. Recent investigations of these proposed classifications have corroborated the presence of 2 overarching transcriptomic subtypes of PDAC tumours comprising *basal-like/squamous* and *classical/progenitor* that have shown relevance for defining survival outcomes in patients, with remaining subtypes (*exocrine-like, ADEX*) shown to have confounding associations with poor tumour cellularity[21,22]. Most recently, Saillard et al (2023) used an artificial intelligence model trained and validated on 5 independent surgical and biopsy cohorts with RNAseq and histology data (n=598), including n=126 TCGA samples to further refine these tumour subtypes[20]. This approach recapitulated the known *basal/classical*[17] tumour subtypes at the whole H&E slide level but detected variable proportions of basal cells in samples previously categorised as classical subtype when slides were analysed at 112 μm tile size level. This changed survival outcomes in 39% of cases classed as *classical* subtype by bulk RNAseq analysis, with the impact apparently proportional to the percentage basal cell content.[20]

At the DNA level, whole genome sequencing (WGS) and copy number variant (CNV) analysis performed on 100 treatment-naïve, macro-dissected PDAC tissues (ICGC Australia) identified four disease subtypes with distinct patterns of structural variation (*scattered, locally rearranged, stable, unstable)* and clinical utility, with *unstable* subtype characterised by a very high degree of genomic instability throughout the genome and encompassing defects in DNA damage repair (DDR) pathways that confer susceptibility to PARP inhibition (PARPi) and/or platinum chemotherapies[19].

Application of subtype-specific filtering criteria in study cohorts

Subtype-specific characteristics can be explored using the TCGA (n=185) and newly added ICGC Australia (PACA-AU, n=461) and Canada (PACA-CA, n=317) pancreatic cancer cohorts. Molecular subtype classifications according to Collisson[16], Moffitt[17], Bailey[18] and/or Saillard[20] are available for n=134 of the 156 confirmed PDAC patients included in the TCGA cohort[21]. Alternatively, ICGC-AU cohorts can be analysed according to the subtype classifications proposed by Bailey et al (2016)[18] (n=95 patients total; 81 PDAC) and/or Waddell et al (2015)(n=86 patients total; 85 PDAC).[19] Given the prognostic relevance of these transcriptomic subtypes, here we explore the associated genomic features of TCGA PDAC tumours classed unanimously as either classical/progenitor (n=27) or QM/basal-like/squamous (n=16) by all three subtyping systems (Collisson/Moffit/Bailey), as an example of PED Analytics Hub.

Subtype-specific somatic variations

Comparisons between the genomic characteristics of each TCGA subtype showed different gene sets mutated in *classical* and *basal* subtypes (Figure 3A, B). This was also true for PACA-AU prognostic subtypes (progenitor vs squamous) (Supplementary Figure 1A, B), but with little consensus between the two cohorts (Supplementary Figure 1C). To more reliably identify subtype-specific genetic variations robust to inevitable inter-study variability (e.g. tissue heterogeneity; WES vs WGS[23]), we considered the *union* of the top 25 most frequently mutated genes between similar prognostic groups for the two largest datasets (TCGA+ICGC). This revealed a handful of common genes detected at >10% prevalence (*KRAS, TP53, CDKN2A, MUC16, LRP1B, AFF2, FAT4*), but with most genes/variants being subtype-specific (Figure 3C), consistent with (1) the early acquisition of these common alterations during PDAC tumour development and (2) the molecular heterogeneity of PDAC tumours[21].

Identifying treatment biomarkers

In silico functional analysis of *all* patient-specific somatic variants identified, using the most recent, freely available embedded Cancer Genome Interpreter (CGI)[24] analytic tool, identified numerous biomarkers of response/resistance to existing clinical treatments and/or pharmacological inhibitors (Supplementary Table S2) in various cancer contexts, including multiple variants in *KRAS*, *TP53*, *CDKN2A* and *LRP1B*. Although *KRAS* is extensively mutated across both prognostic groups, *KRAS* p.G12C driver variants were unique to the best prognosis subtype tumours (Supplementary Table S2). This variant, rare in PDAC (<1% patients)[25], has been shown to preferentially drive the RAF/RAL pathway, while the more common $KRAS^{G12D}$ mutation (~30% PDAC patients) favours the PI3K/AKT pathway[25,26]. Targeted $KRAS^{G12C}$ inhibitor sotorasib has recently received FDA approval for treatment of mNSCLC[27], and has been shown to be safe and effective in treatment of advanced mPDAC, in a Phase I/II trial in n=38 patients[28]. Additionally, certain variants in *CDKN2A* (L104V, E120*, R58*, R80*) have been linked with treatment resistance to PD1 inhibitors, and treatment response to CDK4/6 inhibitors in cutaneous melanoma (CM; Supplementary Table S2), highlighting the usefulness of PED in identifying potentially clinically relevant patient- and cancer-specific therapeutic targets.

Of the somatic variants identified, CGI oncogenic classifications (bioactivity) (Supplementary Table S1), revealed several to be TIER 1 predicted driver variants, i.e. the gene activity is confirmed relevant to cancer, with mutations identified effecting oncogenic transformation. Only one was identified in the poorest outcome patient group – a splice donor variant in central cell-cycle regulator gene *ATM*, and linked with response to cisplatin chemotherapy, PARPi by olaparib, and PD1/PD-L1 inhibition in other solid tumours (Supplementary Table S2). However, this variant was uncommon in the sample set (1/16). Conversely, several TIER 1 oncogenic driver variants were identified in the best prognosis patient group, with some associated with response/resistance to specific drugs in other solid tumours, suggesting possible utility in pancreatic cancer: *ARID1A* (p.E1542*, p.Q1277*; responsive to EZH2, PD1, ATR & PARP inhibitors), *RNF43* (W159*, responsive to porcupine inhibitor) and

TP53 (S166*, Y205S, D259V, M246R, S241F, L194H, N131I; resistance to CDK4/6 inhibitor abemaciclib, cisplatin, MDM2 inhibitor; responsive to ATR inhibitor AZD6738, doxorubicin, decitabine, gemcitabine, mitomycin C).

Results of the above detailed output are also summarised in alluvial plots, showing clinically-actionable targets present in selectable proportions of the filtered dataset (5%-25%) and their responsiveness/resistance to available drugs (Supplementary Figure 2A, B), as well as a visual summary of the number of druggable gene categories represented in the selected dataset (Supplementary Figure 2C, D), that shows different clinically actionable genome targets between the two prognostic groups. While the most commonly mutated genes are common between prognostic subtypes (≥25% of patients from both groups contain similar *KRAS* and *TP53* variants), distinct biomarker/drug combinations are apparent when less common variations are considered: only one other gene in the poor prognosis group was identified as harbouring variants linked to (among others) responsiveness to small molecule AURKA-VEGF inhibitor ilorasertib (CDKN2A R58*), whereas the better prognosis subgroup is associated with 6 additional gene/biomarker candidates (*RNF43*, *PIK3CA*, *ERCC4*, *CTNNB1*, *CDKN2A*, *ARID1A*), that have shown promise in treating other solid tumours. Additionally, by exploring the available CCLE database of 60 pancreatic cell lines, *in vitro* models with/without *KRAS* and/or *TP53* variants may be identified to support downstream functional studies.

These results demonstrate the value of PED in contextualising individual patient genetic profiles in suggesting possible treatment options, or refining research areas to pursue for more effective, stratified approaches in pancreatic cancer.

Characterising patterns of gene expression in best- and worst-outcome tumours

Inspecting the top 250 differentially expressed genes in both TCGA and ICGC filtered patient subgroups confirms scant overlap between genes differentially expressed in best and worst-outcome PDAC tumours. However, *classical/progenitor* tumours appear to have higher *TP53* expression compared to *QM/basal/squamous* tumours, consistent with its role as

tumour suppressor (Figure 4A) and mirrored in ICGC *progenitor* and *squamous* patient subgroups[18] (Supplementary Figure 3). Subtype-specific differences in expression were also observed for *MUC16* (encoding CA125 membrane glycoprotein) which was over-expressed in *QM/basal/squamous* subtype tumours, compared to *classical/progenitor* cases (Figure 4B). Considering all n=402 confirmed PDAC in PACA-AU (the largest single transcriptomic dataset), PED shows that *MUC16* over-expression is associated with significantly reduced patient survival (logrank p=0.011; HR=2.23) (Figure 4C), where no association was found for neuroendocrine tumours (Figure 4D). Elevated CA125 has also recently been shown to be an independent prognostic marker of significantly shorter survival in n=207 resectable PDAC patients, both before and after treatment[29]. Functionally, CA125 over-expression has been shown to promote tumourigenesis *in vitro* and *in vivo*[30,31], and monoclonal antibody mAb AR9.6 has recently shown potential as a specific and effective inhibitor of CA125 and its oncogenic effects in pancreatic and ovarian cancers[31,32].

Identifying clinically actionable genomic alterations in KRAS wild-type PDAC tumours

Several studies have explored the genetic landscape of *KRAS* wild-type tumours, delineating several alterations that occur frequently in the absence of mutant *KRAS* [21,33–35]. In addition, a significant enrichment for somatic aberrations that target the RAS-MAPK pathway, either upstream or downstream of *KRAS*, has been observed in up to one-third of *KRAS* wild-type tumours[21,35],where *BRAF* alterations were prevalent and mutually exclusive with *KRAS* mutations[35]. However, alterations within genes that are not typically associated with *RAS* signalling have also been widely identified across *KRAS* wild-type PDAC tumours, and require further investigation to determine functional relevance[18,21].

Alternative oncogenic drivers amongst KRAS wild-type PDAC tumours

We used GENIE[11] as the largest available resource to identify n=756 *KRAS* wild-type PDAC samples. To identify other likely molecular drivers in these tumours, predictions from CGI were analysed to evaluate the distribution of altered genes and their associated pathways. Mutations were detected across several genes previously reported to be altered in *KRAS*

wild-type PDAC cases, including *TP53* (mutated in >40% of the samples), *GNAS* and *BRAF*[36] (Figure 5A). *In silico* biomarker predictions also showed that *ARID1A*, *BRAF, CDKN2A, GNAS, PIK3CA and TP53* variants demonstrated therapeutic potential in response to PARP, tyrosine kinase and VEGF inhibitors, immunotherapies and several chemotherapies (Figure 5B).

The analysis of altered signalling pathways amongst mutated genes across *KRAS* wild-type samples revealed frequent alterations in pathways associated with MAPK signalling, P53 signalling, neurotrophin, cell cycle, wnt and apoptosis signalling, consistent with previous characterisations of core biological pathways involved in PDAC development and progression[18,19,21,30,36]. (Figure 5C). These findings highlight the utility of PED for the prioritisation of functionally and biologically relevant variants amongst subgroups of PDAC tumours, with important implications for the characterisation of distinct molecular pathologies and the identification of novel therapeutic opportunities.

Using the PED Analytics Hub, we demonstrate how our integrated high-performance visualisation and analysis tool can be used to investigate the link between genomic and transcriptomic features and phenotypes of pancreatic cancer, providing an important step in defining potential subtype-specific therapeutic vulnerabilities.

The PCRFTB Data Return Module

The vision of precision medicine has driven unprecedented interest into biomarker-based studies (genomics, transcriptomics, proteomics) for pancreatic cancer, which are being adopted across research and development from early discovery through to clinical research and trials[37]. Fundamental to biomarker research is access to quality biospecimens and samples that have been well annotated with clinical and molecular data[6]. Whilst many biobanks have invested heavily in the IT infrastructure of sample management, most platforms are facing challenges in the effective sharing of returned data to drive investigative research across the pancreatic research community. A major challenge is the rise of so-called ‘big data’ from e.g. NGS and images that need to be integrated with large quantities of

primary/secondary care information and other real-world healthcare data. Traditional biobanks are not usually set up to leverage these innovations, which have the potential to improve patient outcomes and accelerate the development and delivery of new therapies. PED is now the primary bioinformatics platform of the PCRFTB, providing a unique integrated resource of biological materials and associated clinical, molecular and radiological/imaging data.

In addition to providing a direct link to sample requests from the PCRFTB, the newly incorporated data return module in PED hosts clinical and molecular data returned to the bank from studies undertaken using PCRFTB specimens, where published findings have been made available for researchers to review and analyse prior to submitting a tissue request. Studies are categorised according to the type of -omics data generated for each project (i.e. genomic, transcriptomic or proteomic), with alternative data types (e.g. summaries of staining or imaging results generated from experimental investigations) classified “other”, to simplify use. These classifications are presented in a summary table, which also provides a description of each project and details the different sample types (i.e. blood, tissues, cell lines) and cohort sizes used for each project. Users also have the option view this information as clinical summary plots for each individual cohort, prior to exploring available molecular data from each study.

Like other tissue repositories[38,39], PCRFTB has implemented a data return policy, where anonymised data derived from banked samples is returned to the tissue bank on completion of the study and made freely available to the research community, regardless of whether the study is ultimately published[6]. As associated sample datasets develop, more in-depth integrative analyses will be possible. Under *PCRFTB Data*, PED lists the details of any returned data and associated sample characteristics available for analysis, while *PCRFTB Research Projects* links directly to the relevant published report. So far, this has further enriched the data available for banked tissue sample and/or patients, and currently includes data on stromal[40], urinary miRNA[41], metallomic[42] , and volatile organic compound[43] biomarkers, proteomics (ELISA)[44], circulating tumour cells (CTCs) and xenotransplantation

models[45,46], germline and somatic mutations[47], a phase I clinical trial[48], and risk[49] and recurrence[50] predictions incorporating electronic health record data. To date, PCRFTB has processed 38 EoI, supported 33 research projects and 19 peer-reviewed publications. Ultimately, each study contributes to the development of a 'digital fingerprint' for each patient, linking multi-modal data with longitudinal clinical information.

PED Analytics Hub is the web-based portal through which this enriched dataset can be accessed and compared with large scale pancreatic -omics data, with the unique benefit of also providing access to additional patient samples (*via* PCRFTB) for subsequent validation of molecular alterations with clinical potential. Here, we demonstrate the added value of tissue banking to precision cancer medicine, to translate research findings into prognostic and therapeutic tools using well-annotated curated tissues and associated clinical data.

Discussion

A rapid expansion in high-throughput genomic and transcriptomic profiling of pancreatic diseases necessitates the development of sophisticated yet user-friendly analytics hubs to host the growing compendium of molecular and clinical datasets and enable integrated mining and analysis of available results. The updated PED now supports a range of data modalities to enable users across the diverse international pancreatic research community to identify and investigate trends in molecular data across disparate cohorts of patients, samples and cell lines easily and effectively. PED now provides an unprecedented opportunity to characterise distinct variations in both tumour mutation landscapes and gene expression profiles that are associated with prognostic molecular subtypes. Candidate tumour drivers and biomarkers predictive of response to existing and novel clinical treatments can be identified and visualized, allowing suitable targets for downstream validation and pharmacological testing to be prioritised without the need for laborious data retrieval or processing tasks. Furthermore, PED is the gateway to a national tissue bank repository of >60,000 samples from >3,400 patients and a growing repository of digitised radiological and H&E images, that researchers can access independently and apply for donor samples that match with their research question. The selection of samples is

significantly improved by the availability of high-quality clinical data, curated and maintained by PCRFTB.

These major updates to the PED infrastructure are further underpinned by its recent adoption as the bioinformatics platform of the PCRFTB, pioneering a new generation in biobanking to support effective data sharing and promote collaborative studies, democratizing access to complex cancer genomics. By harmonising PCRFTB samples with clinical and molecular information from datasets returned to the biobank, PED provides an essential platform to support translational pancreatic research and fuel discoveries that can manifest clinically meaningful benefits for patients. PCRFTB recently launched internationally (https://www.pcrf.org.uk/news/tissue-bank-launches-internationally/), improving opportunities for high-quality research into earlier diagnosis and treatment of pancreatic cancer. As -omics driven research continues to drive efforts to advance the characterisation of pancreatic diseases, PED's design supports the ongoing data analysis, integration and visualisation needs of the growing research community.

446 **Table 1. Summary of the 2024 updates to PED**

Features	2018 release[8]	2024 release
The PED Analytics Hub		
Publicly available data sources (pancreas-specific)		
PubMed[a]	X	
TCGA[b]	X	X
ICGC		X
GENIE[c]	X	X
CCLE[d]	X	X
Analytical features		
Principal components analysis	X	X
Gene expression profiles	X	X
Correlation analyses	X	X
Gene networks	X	X
Survival analyses[e]	X	X
Variant identification	X	X
Somatic gene interactions	X	X
Reactome & oncogenic pathway analyses		X
Improved clinical annotations to visualize & query publicly available datasets		X
MAFtools genomic analyses and summary visualisations		X
Tumour mutational burden		X
Clinically actionable genes/proteins & associated drugs		X
Cohort comparison by clinical/molecular feature		X
Gene intersections between filtered datasets		X
The PCRFTB Data Module		
Data return module to host both -omics and experimental datasets		X
Integrated primary/secondary care clinical data		X
Apply for samples		X

447

448

449

[a] Less-used literature mining module
[b] Updated to include essential filters based on cancer subtype
[c] Expanded from 445 adenocarcinoma or neuroendocrine tumours to full set of 6,633 pancreatic cancers of all types, with additional clinical and molecular information.
[d] Somatic variant dataset; now expanded to include the full set of 60 primary and metastatic tumour derived cell lines.
[e] Expanded to include analyses based on mutational status and mRNA level

References

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2020. *CA Cancer J Clin* **70**, 7–30 (2020).
2. Pereira, S. P. *et al.* Early detection of pancreatic cancer. *Lancet Gastroenterol Hepatol* **5**, 698–710 (2020).
3. Abboud, Y. *et al.* Increasing Pancreatic Cancer Incidence in Young Women in the. *Gastroenterology* 1–12 (2023) doi:10.1053/j.gastro.2023.01.022.
4. Brezgyte, G., Shah, V., Jach, D. & Crnogorac-jurcevic, T. Non-invasive biomarkers for earlier detection of pancreatic cancer—a comprehensive review. *Cancers (Basel)* **13**, 1–25 (2021).
5. Oscanoa, J. *et al.* SNPnexus: A web server for functional annotation of human genome sequence variation (2020 update). *Nucleic Acids Res* **48**, W185–W192 (2020).
6. Balarajah, V. *et al.* Pancreatic cancer tissue banks: where are we heading? *Future Oncology* **12**, 2661–2663 (2016).
7. Imrali, A. *et al.* Validation of a Novel, Flash-Freezing Method: Aluminum Platform. *Current Protocols in Essential Laboratory Techniques* **21**, 1–17 (2020).
8. Marzec, J. *et al.* The Pancreatic Expression Database: 2018 update. *Nucleic Acids Res* **46**, D1107–D1110 (2018).
9. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113–1120 (2013).
10. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
11. Sweeney, S. M. *et al.* AACR project genie: Powering precision medicine through an international consortium. *Cancer Discov* **7**, 818–831 (2017).
12. Zhang, J. *et al.* International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data. *Database (Oxford)* **2011**, bar026 (2011).
13. Zhang, J. *et al.* The International Cancer Genome Consortium Data Portal. *Nat Biotechnol* **37**, 367–369 (2019).
14. Peran, I., Madhavan, S., Byers, S. W. & McCoy, M. D. Curation of the pancreatic ductal adenocarcinoma subset of the cancer genome atlas is essential for accurate conclusions about survival-related molecular mechanisms. *Clinical Cancer Research* **24**, 3813–3819 (2018).
15. Nicolle, R. *et al.* Prognostic biomarkers in pancreatic cancer: Avoiding errata when using the TCGA dataset. *Cancers (Basel)* **11**, 1–10 (2019).
16. Collisson, E. A. *et al.* Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat Med* **17**, 500–503 (2011).
17. Moffitt, R. A. *et al.* Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat Genet* **47**, 1168–1178 (2015).
18. Bailey, P. *et al.* Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* **531**, 47–52 (2016).
19. Waddell, N. N. N. *et al.* Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495–501 (2015).
20. Saillard, C. *et al.* Pacpaint: a histology-based deep learning model uncovers the extensive intratumor molecular heterogeneity of pancreatic adenocarcinoma. *Nat Commun* **14**, (2023).
21. Raphael, B. J. *et al.* Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell* **32**, 185-203.e13 (2017).

22. Sinkala, M., Mulder, N. & Martin, D. Machine Learning and Network Analyses Reveal Disease Subtypes of Pancreatic Cancer and their Molecular Characteristics. *Sci Rep* **10**, 1212 (2020).
23. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* **6**, 271-281.e7 (2018).
24. Tamborero, D. *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med* **10**, 25 (2018).
25. Kwan, A. K., Piazza, G. A., Keeton, A. B. & Leite, C. A. The path to the clinic: a comprehensive review on direct KRASG12C inhibitors. *Journal of Experimental and Clinical Cancer Research* **41**, (2022).
26. Ihle, N. T. *et al.* Effect of KRAS oncogene substitutions on protein behavior: Implications for signaling and clinical outcome. *J Natl Cancer Inst* **104**, 228–239 (2012).
27. Nakajima, E. C. *et al.* FDA Approval Summary: Sotorasib for KRAS G12C-Mutated Metastatic NSCLC. *Clinical Cancer Research* **28**, 1482–1486 (2022).
28. Strickler, J. H. *et al.* Sotorasib in KRAS p.G12C–Mutated Advanced Pancreatic Cancer . *New England Journal of Medicine* **388**, 33–43 (2023).
29. Napoli, N. *et al.* Ca 125 is an independent prognostic marker in resected pancreatic cancer of the head of the pancreas. *Updates Surg* **75**, 1481–1496 (2023).
30. Qi, Z. H. *et al.* RIPK4/PEBP1 axis promotes pancreatic cancer cell migration and invasion by activating RAF1/MEK/ERK signaling. *Int J Oncol* **52**, 1105–1116 (2018).
31. Thomas, D. *et al.* Isoforms of MUC16 activate oncogenic signaling through EGF receptors to enhance the progression of pancreatic cancer. *Molecular Therapy* **29**, 1557–1571 (2021).
32. Sharma, S. K. *et al.* ImmunoPET of Ovarian and Pancreatic Cancer with AR9.6, a Novel MUC16-Targeted Therapeutic Antibody. *Clinical Cancer Research* **28**, 948–959 (2022).
33. Heining, C. *et al.* NRG1 Fusions in KRAS Wild-Type Pancreatic Cancer. *Cancer Discov* **8**, 1087–1095 (2018).
34. Luchini, C. *et al.* KRAS wild-type pancreatic ductal adenocarcinoma: molecular pathology and therapeutic opportunities. *Journal of Experimental & Clinical Cancer Research* **39**, 227 (2020).
35. Singhi, A. D. *et al.* Real-Time Targeted Genome Profile Analysis of Pancreatic Ductal Adenocarcinomas Identifies Genetic Alterations That Might Be Targeted With Existing Drugs or Used as Biomarkers. *Gastroenterology* **156**, 2242-2253.e4 (2019).
36. Philip, P. A. *et al.* Molecular Characterization of KRAS Wild-type Tumors in Patients with Pancreatic Adenocarcinoma. *Clinical Cancer Research* **28**, 2704–2714 (2022).
37. Herbst, B. & Zheng, L. Precision medicine in pancreatic cancer: treating every patient as an exception. *Lancet Gastroenterol Hepatol* **4**, 805–810 (2019).
38. Gadaleta, E., Pirrò, S., Dayem Ullah, A. Z., Marzec, J. & Chelala, C. BCNTB bioinformatics: The next evolutionary step in the bioinformatics of breast cancer tissue banking. *Nucleic Acids Res* **46**, D1055–D1061 (2018).
39. Speirs, V. Quality Considerations When Using Tissue Samples for Biomarker Studies in Cancer Research. *Biomark Insights* **16**, (2021).
40. Goulart, M. R. *et al.* Pentraxin 3 is a stromally-derived biomarker for detection of pancreatic ductal adenocarcinoma. *NPJ Precis Oncol* **5**, (2021).
41. Debernardi, S. *et al.* Noninvasive urinary miRNA biomarkers for early detection of pancreatic adenocarcinoma. *Am J Cancer Res* **5**, 3455–3466 (2015).

42. Schilling, K. *et al.* Urine metallomics signature as an indicator of pancreatic cancer. *Metallomics* **12**, 752–757 (2020).
43. Daulton, E. *et al.* Volatile organic compounds (VOCs) for the non-invasive detection of pancreatic cancer from urine. *Talanta* **221**, (2021).
44. Debernardi, S. *et al.* A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case-control study. *PLoS Med* **17**, 1–23 (2020).
45. Raj, D. *et al.* Switchable CAR-T cells mediate remission in metastatic pancreatic ductal adenocarcinoma. *Gut* **68**, 1052–1064 (2019).
46. Raj, D. *et al.* CEACAM7 is an effective target for CAR T-cell therapy of pancreatic ductal adenocarcinoma. *Clinical Cancer Research* **27**, 1538–1552 (2021).
47. Sivapalan, L. *et al.* Longitudinal profiling of circulating tumour DNA for tracking tumour dynamics in pancreatic cancer. *BMC Cancer* **22**, 1–17 (2022).
48. Kocher, H. M. *et al.* Phase I clinical trial repurposing all-trans retinoic acid as a stromal targeting agent for pancreatic cancer. *Nat Commun* **11**, 4841 (2020).
49. Zardab, M. *et al.* Differentiating Ductal Adenocarcinoma of the Pancreas from Benign Conditions Using Routine Health Records: A Prospective Case-Control Study. *Cancers (Basel)* **15**, (2023).
50. Ang, A., Michaelides, A., Chelala, C., Ullah, D. & Kocher, H. M. Prognostication for recurrence patterns after curative resection for pancreatic ductal adenocarcinoma. *Ann Hepatobiliary Pancreat Surg* **28**, 248–261 (2024).

Funding

The PCRFTB is funded by PCRF. Organoid generation is supported by Barts Charity. HMK and CC acknowledge the support of NIHR Barts BRC. This work was supported by Barts Charity (grant code MGU0504) and Barts NIHR BRC (grant code BTXH1A1R), part of the Precision Medicine programme.

Competing interests

H.M.K. received a research grant for conducting this trial (Celgene: institutional) and educational grant support for attending or organizing conferences (Celgene, Baxalta, Mylan, Medtronic, Oncosil: institutional) which are unrelated to this work. No other authors have interests to declare.

Ethics approvals

All tissue/sample donors provide written, informed consent, and all samples were collected, processed and stored at each of the participating centres (Barts, Leicester, Swansea, Oxford, Royal Free (London), Southampton, Newcastle, Plymouth, The London Clinic) under one Research Ethics Committee reference (13/SC/0593, renewed 18/SC/0629, renewed 23/SC/0282).

Author contributions

Conceptualization: CC; Methodology: JO, AZMDU, TSK, EG, GJT, MA; Software: JO; Validation: LS, HRA; Investigation: LS, HRA; Resources: PCRFTB, AI, AS, RR, CH; Data Curation: MA, TSK, AZMDU, EG, JO, RR, CH; Writing – original draft: LS; Writing – Review & Editing: HRA, HMK, CC; Supervision: CC; Funding acquisition: HMK, CC. All authors read and approved the final manuscript.

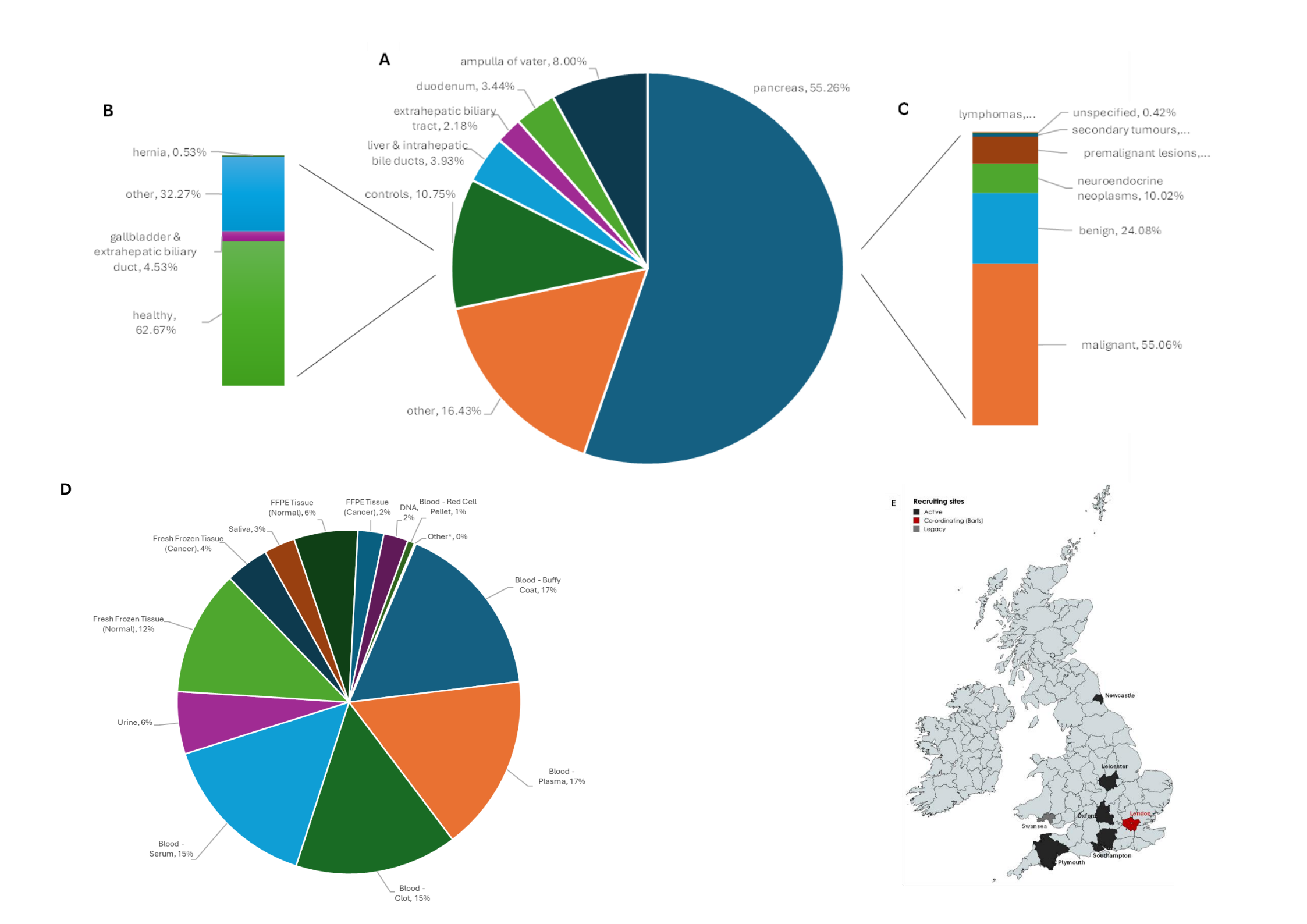
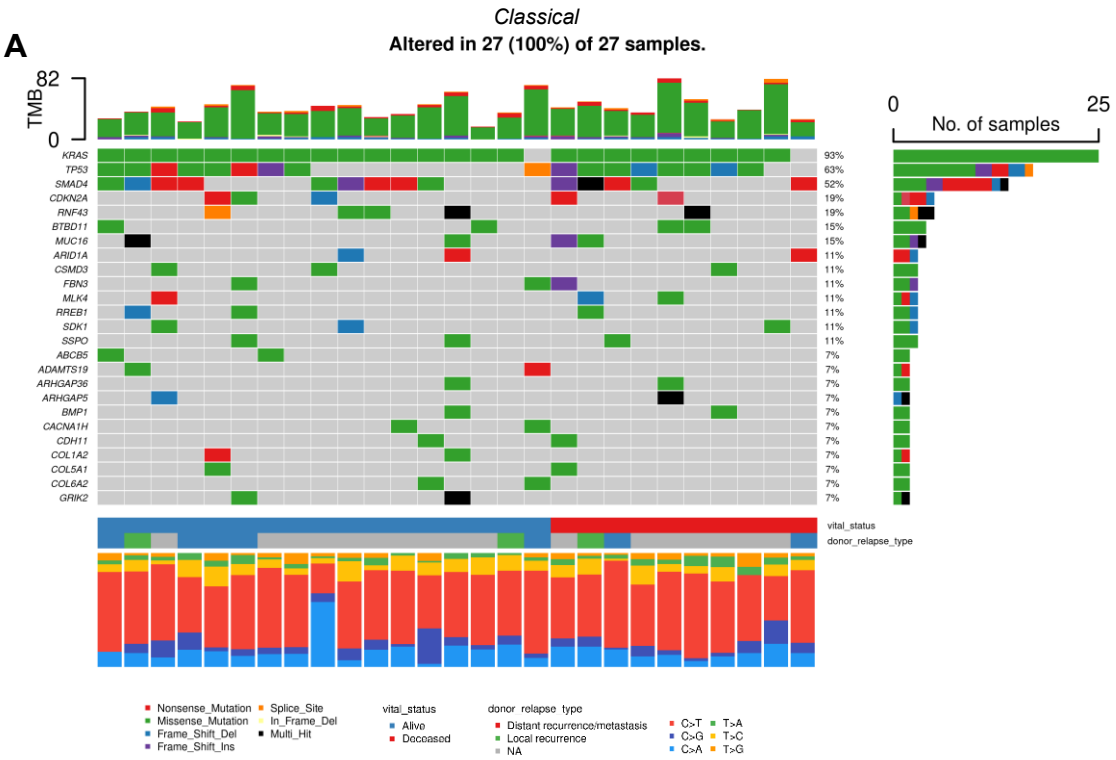

A
ampulla of vater, 8.00%
duodenum, 3.44%
extrahepatic biliary tract, 2.18%
liver & intrahepatic bile ducts, 3.93%
controls, 10.75%
other, 16.43%
pancreas, 55.26%
B
hernia, 0.53%
other, 32.27%
gallbladder & extrahepatic biliary duct, 4.53%
healthy, 62.67%
C
lymphomas,...
unspecified, 0.42%
secondary tumours,...
premalignant lesions,...
neuroendocrine neoplasms, 10.02%
benign, 24.08%
malignant, 55.06%
D
FFPE Tissue (Normal), 6%
FFPE Tissue (Cancer), 2%
DNA, 2%
Blood - Red Cell Pellet, 1%
Other*, 0%
Saliva, 3%
Fresh Frozen Tissue (Cancer), 4%
Blood - Buffy Coat, 17%
Fresh Frozen Tissue (Normal), 12%
Urine, 6%
Blood - Plasma, 17%
Blood - Serum, 15%
Blood - Clot, 15%
E
Recruiting sites
Active
Co-ordinating (Barts)
Legacy
Newcastle
Leicester
Oxford
London
Swansea
Southampton
Plymouth

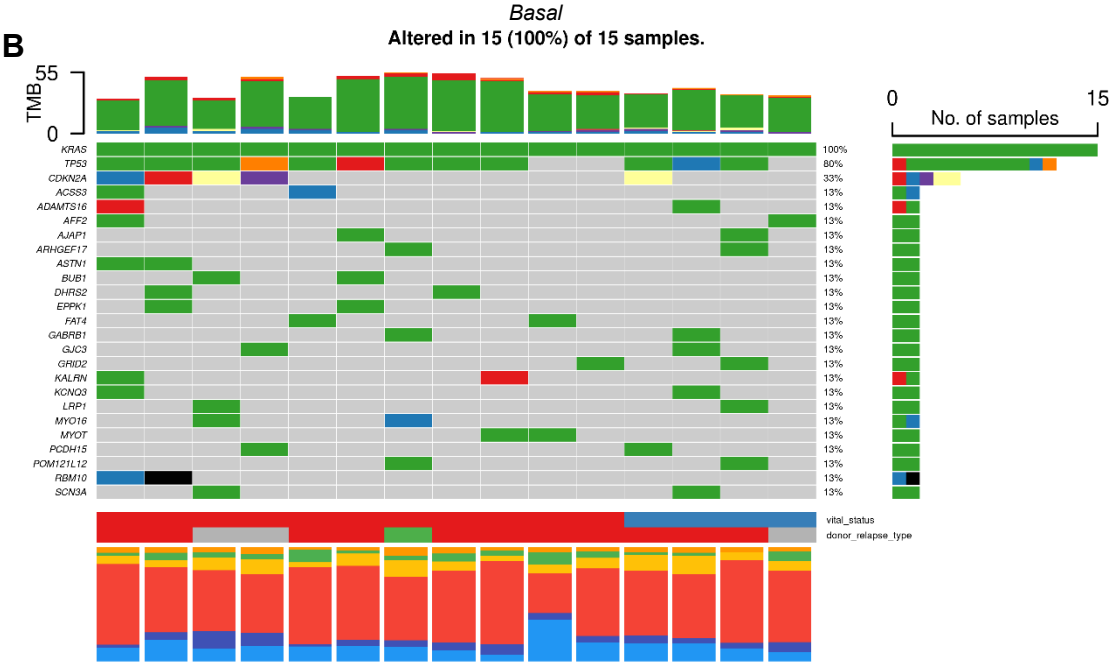
Figure 1. Summary of available tissue types.
(**A**) The proportion of >3 400 unique organ site tissues available for study in the UK national Pancreatic Cancer Research Fund Tissue Bank, with the breakdown of controls (**B**) and pancreas (**C**) highlighted. (**D**) Distribution of >60 000 PCRFTB specimens by type, across all patients. Details are updated weekly. Additionally, radiological imaging is available for 171 patients with malignant, pre-malignant and benign pancreatic diagnoses, and >2,000 H&E images from 349 donors. Samples can be applied for here. (**E**) Geographical locations of PCRFTB patient recruitment sites.

Map created with mapchart.net.
*=pancreatic juice, CTC, bile, organoids.

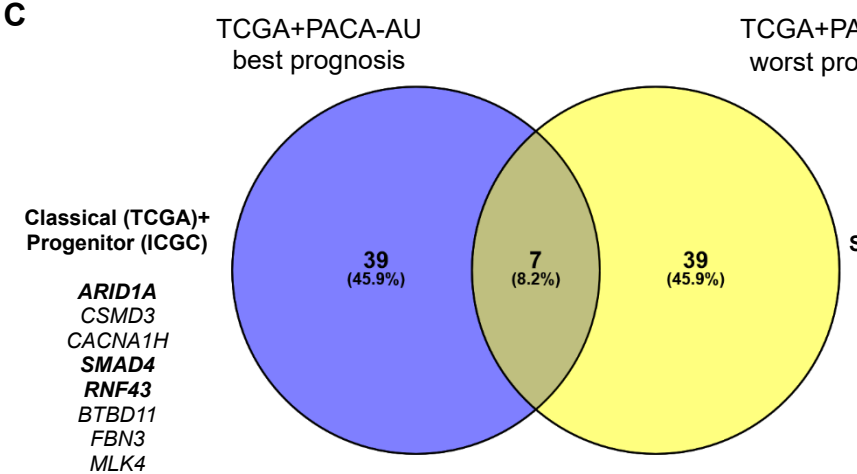

A
PED
Pancreatic Expression Database
The bioinformatics portal of the PCRFTB
Powered by SNPnexus
pancreatic cancer research fund
TISSUE BANK
Home
Publicly Available Data
PCRFTB Data
Cohort Comparison
Apply for Samples
The Cancer Genome Atlas
Clinical filters
Age (30-90):
Sex:
Female
Male
Race:
White
Black or African American
Asian
Diagnosis:
Ductal adenocarcinoma
Tumor stage:
I
II
III
IV
Survival:
Alive
Deceased
Survival Period:
<1 year
1-5 years
>5 years
History of Diabetes:
Yes
No
Family History of Cancer:
Yes
No
Molecular filters
Missense mutation - KRAS:
G12
G13
Q61
Absent
Missense mutation - TP53:
Present
Absent
Subtype - Moffitt:
Classical
Basal
Subtype - Bailey:
Progenitor
Squamous
Immunogenic
ADEX
Subtype - Collisson:
Classical
Exocrine
Quasi-mesenchymal
Subtype - PacPaint:
Classical
Basal
B
Survival Time
Age at Diagnosis
Well-differentiated
Moderately differentiated
Poorly differentiated
Undifferentiated
Undetermined
Survival Time (Years)
Number of Donors
Age at Diagnosis
Race
Tumor Stage
Tumor Grade
KRAS Missense Mutation
TP53 Missense Mutation
C
Tumor Stage
Tumor Grade
Missense Mutation - KRAS
Missense Mutation - TP53
D
Age at Diagnosis
Number of Cases
Sex
Male
Female
Survival Status
Deceased
Alive
Race
White
Black or African American
Asian
Not available
E
Molecular Subtype - Collisson
Classical
Exocrine-like
Quasi-mesenchymal
Unknown
Molecular Subtype - Moffitt
Classical
Basal
Unknown
Molecular Subtype - Bailey
Progenitor
Squamous
Immunogenic
ADEX
Unknown
Molecular Subtype - PacPaint
Classical
Basal
Unknown

Figure 2. Advanced filtering options and clinical summaries for publicly available PDAC datasets. (A) Available data can be filtered according to various patient-related factors and tumour characteristics, including the stratification and analysis of cohorts according to KRAS and TP53 mutational status and established transcriptomic (TCGA, ICGC), genomic (ICGC) or histologically-derived AI (PacPaint) subtypes. **(B)** Dynamic bar charts allow multiple covariates to be viewed in relation to each other: e.g. survival trends in PDAC (TCGA) and neuroendocrine (ICGC PAEN-AU). **(C, D, E)** Each filtered attribute can be visualized as clinical summaries for each study cohort**.** Alternatively, data can be downloaded as .csv or .xls files, for offline analysis.





C



Classical (TCGA)+ Progenitor (ICGC)

ARID1A
CSMD3
CACNA1H
SMAD4
RNF43
BTBD11
FBN3
MLK4
RREB1
SDK1
SSPO
ARHGAP5
GRIK2
BCB5
ADAMTS19
ARHGAP36
BMP1
CDH11
COL1A2
COL5A1
COL6A2
TGFBR2
ACVR1B
GATA6
THSD7B
OBSCN
ARR3
CDH4
DCLK1
FBN2
FCGBP
GGYF2
GRID2IP
HERC2
HCN2
ITPR2
LRP2
MAGEB1
PCDHB5

Common

KRAS
TP53
CDKN2A
MUC16
LRP1B
AFF2
FAT4

Basal (TCGA)+ Squamous (ICGC)

RBM10
ACSS3
ADAMTS16
AJAP1
ARHGEF17
ASTN1
BUB1
DHRS2
EPPK1
GABRB1
GJC3
GRID2
KALRN
KCNQ3
LRP1
MYO16
MYOT
PCDH15
POM121L12
SCN3A
HMCN1
MT-CYB
NRXN1
ACTL9
AKAP2
ATG2A
C6orf222
CACNA1B
CDCA2
CDH12
CDH13
COL5A3
CTNNA3
FGF6
FSTL5
GPR26
HGF
KDM6A
MASP1

Figure 3. Transcriptomic stratification in PDAC reveals subtype-specific somatic variants
Oncoplots* of the top 25 most frequently mutated genes for consensus **(A)** n=27 *classical*-type and **(B)** n=16 *basal*-type PDAC cases (TCGA). **(C)** Overlap^ between the somatically mutated genes associated with best/worst prognosis subtypes across TCGA and ICGC PACA-AU cohorts combined. Genes highlighted in **bold** contain Tier 1 predicted oncogenic driver variants that have associated pharmacological inhibitors or chemotherapies (see Supplementary Table S1).

*Mutated genes are ranked in order of the *total* number of *mutations* in each given gene (where genes may have >1 mutation present; black 'multi-hit'), while the percentage to the right of each bar reflects the proportion of *samples* altered in the cohort. ^Created in Venny 2.1.

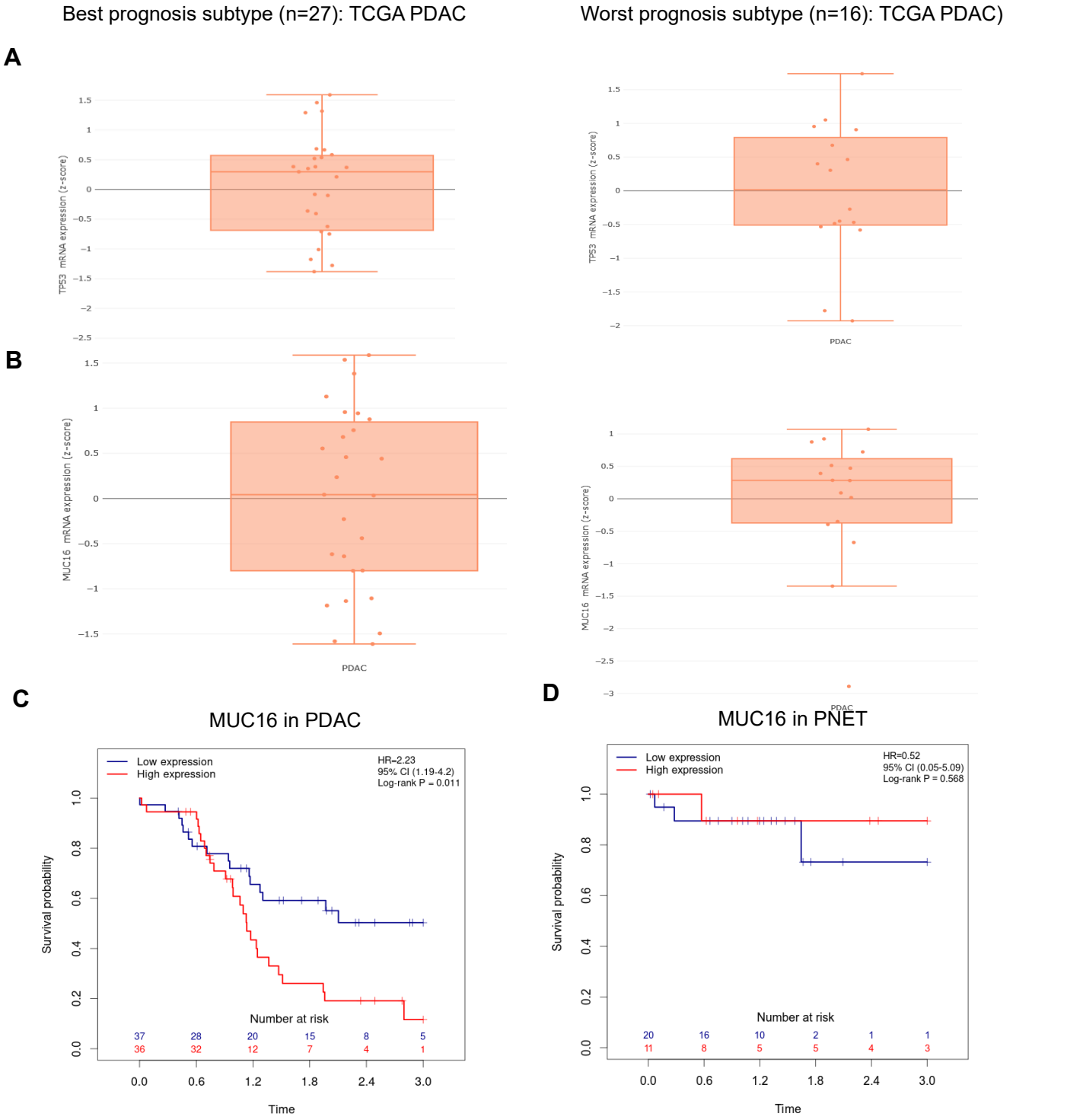


Figure 4. Differentially expressed genes between classical/progenitor and basal-like/QM/squamous TCGA PDAC tumours.

Box plots showing the trends of **(A)** TP53 and **(B)** MUC16 mRNA expression levels across all patients in each filtered TCGA PDAC group; best prognosis (n=27; classical/progenitor; left) and worst prognosis (n=16; basal/squamous/QM; right). **C**. Kaplan-Meier curve showing elevated MUC16 expression significantly associated with lower patient survival over 3 years, from n=402 PACA-AU PDAC patients with expression and outcome data (logrank p=0.011; hazard ratio (HR)=2.23). **D.** No association between MUC16 mRNA expression levels and outcome were observed in n=65 neuroendocrine carcinomas (PAEN-AU).

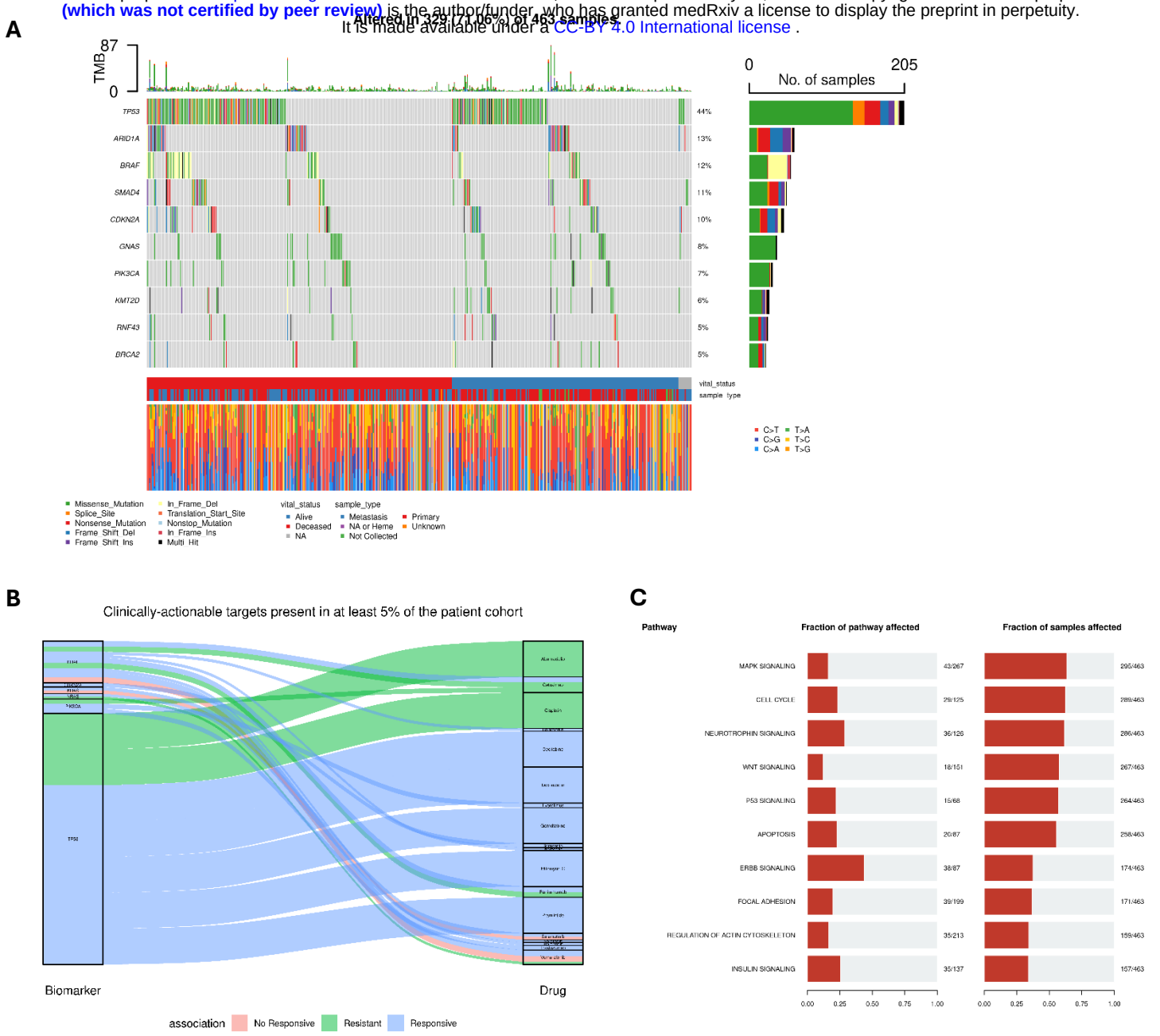


Figure 5. Frequently altered genes and biological pathways amongst n=756 *KRAS* wild-type PDAC tumours from GENIE

(A) Oncoplot showing the top 10 most frequently mutated genes in *KRAS* wild-type PDAC tumours (confirmed somatic missense mutations filtered out; insertions or duplications may still be present). **(B)** Alluvial plot showing gene targets harbouring any variants with therapeutic biomarker potential in ≥5% of patients, as identified by the Cancer Genome Interpreter and based on data from OncoKB, CIVic (Clinical Interpretation of Variants in Cancer) and the Cancer Biomarkers database. **(C)** Altered biological pathways amongst *KRAS* wild-type PDAC tumours include MAPK and p53 signalling, as derived from the KEGG pathway database. The proportion of genes mutated in each pathway (left) and the proportion of all KRAS wild-type patients affected (x-axis) are given.