




Open-source curation of a pancreatic ductal adenocarcinoma gene expression analysis platform (pdacR) supports a two-subtype model

Luke A. Torre-Healy ^{1,11}, Ryan R. Kawalerski^{1,2,11}, Ki Oh ¹, Lucie Chrastecka³, Xianlu L. Peng^{4,5}, Andrew J. Aguirre⁶, Naim U. Rashid^{5,7}, Jen Jen Yeh^{4,5,8} & Richard A. Moffitt ^{1,9,10}✉

Pancreatic ductal adenocarcinoma (PDAC) is an aggressive disease for which potent therapies have limited efficacy. Several studies have described the transcriptomic landscape of PDAC tumors to provide insight into potentially actionable gene expression signatures to improve patient outcomes. Despite centralization efforts from multiple organizations and increased transparency requirements from funding agencies and publishers, analysis of public PDAC data remains difficult. Bioinformatic pitfalls litter public transcriptomic data, such as subtle inclusion of low-purity and non-adenocarcinoma cases. These pitfalls can introduce non-specificity to gene signatures without appropriate data curation, which can negatively impact findings. To reduce barriers to analysis, we have created pdacR (<http://pdacR.bmi.stonybrook.edu>, github.com/rmoffitt/pdacR), an open-source software package and web-tool with annotated datasets from landmark studies and an interface for user-friendly analysis in clustering, differential expression, survival, and dimensionality reduction. Using this tool, we present a multi-dataset analysis of PDAC transcriptomics that confirms the basal-like/classical model over alternatives.

¹Department of Biomedical Informatics, Stony Brook Medicine, Stony Brook, NY, USA. ²Department of Pathology, Stony Brook Medicine, Stony Brook, NY, USA. ³Department of Pharmacological Sciences, Stony Brook Medicine, Stony Brook, NY, USA. ⁴Department of Pharmacology, University of North Carolina, Chapel Hill, NC, USA. ⁵Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, USA. ⁶Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. ⁷Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. ⁸Department of Surgery, University of North Carolina, Chapel Hill, NC, USA. ⁹Department of Biomedical Informatics, Emory University, Atlanta, GA, USA. ¹⁰Department of Hematology and Medical Oncology, Emory University, Atlanta, GA, USA. ¹¹These authors contributed equally: Luke A. Torre-Healy, Ryan R. Kawalerski. ✉email: Richard.austin.moffitt@emory.edu

Over the past decade, several genomics studies of pancreatic ductal adenocarcinoma (PDAC) have contributed a wealth of molecular-level data to guide investigation of the disease. A subset of these studies have sought to define transcriptomic subtypes. Although the datasets generated from these studies are accessible as part of the public domain, bioinformatic hurdles ranging from decentralized data storage to insufficiently annotated samples complicates multi-dataset investigation, resulting in inappropriate and inconsistent application of datasets. Thus, there is an unmet need for a PDAC transcriptomic compendium that provides an easily accessible interface for analysis by the PDAC research community, while retaining the capacity for diverse analytical methods in dataset- and application-specific situations.

Multiple PDAC gene expression subtype models have been proposed toward the goal of using individualized therapeutic approaches to treat patients with this disease^{1–7}. Collisson et al. first defined three prognostically-relevant PDAC-specific subtypes (classical, quasi-mesenchymal (QM), and exocrine-like) using non-negative matrix factorization consensus clustering (NMF-CC) on a merged microarray dataset comprised of laser capture microdissected PDAC tumors ($n = 66$)². We later decoupled NMF and consensus clustering to create a two-subtype model (classical, with a similar gene set to Collisson's classical type, and basal-like, similar to basal breast and bladder cancers), using microarrays from bulk primary and metastatic PDAC tumors, normal pancreas and distant site normal tissue, and PDAC cell lines ($n = 357$)³. Similar to Collisson et al., Bailey et al. subsequently used NMF-CC to analyze RNA sequencing (RNA-seq) data from a combination of PDAC and other uncommon pancreatic cancer histological types, including acinar cell carcinomas, finding four tumor subtypes (pancreatic progenitor, squamous, aberrantly differentiated endocrine exocrine (ADEX), and immunogenic, $n = 96$)⁴. Then, in 2018, Puleo et al. proposed a five-subtype model using a consensus clustering approach on 309 formalin fixed paraffin embedded whole PDAC tumor samples (pure classical, immune classical, desmoplastic, stroma activated, and pure basal-like). However, they confirmed two-subtypes (classical and basal-like) when using only samples with purity (defined by the predicted mean variant allele frequency, or VAF) in the upper quartile of their dataset ($n = 78$)⁵.

In 2020, researchers have sought to further refine these definitions. Chan-Seng-Yue et al. utilized NMF to identify four gene signatures, subdividing basal-like and classical subtypes into “basal-like A”, “basal-like B”, “classical A”, and “classical B”⁶. Separate work by Nicolle et al. then utilized Independent Component Analysis (ICA) to generate a molecular gradient that correlates with histological differentiation and seeks to assign a continuous value to tumor samples using weighted valuations of between 4000 and 20,434 genes⁷.

Independent studies have comparatively evaluated the clinical utility of these proposed schemas. The Comprehensive Molecular Characterization of Advanced Pancreatic Ductal Adenocarcinoma for Better Treatment Selection (COMPASS) trial's earliest results showed that the Moffitt et al. classification system provided accurate prediction of the tumors which would respond or not respond to modified FOLFIRINOX or gemcitabine/nab-paclitaxel⁸. Furthermore, Tiriach et al. showed that the basal-like/classical system could be applied to transcriptomic data from patient-derived PDAC organoids to infer patient tumors which would respond to either oxaliplatin or gemcitabine⁹, and, shortly afterward, Aguirre et al. demonstrated utility using the Moffitt subtypes to characterize metastatic PDAC biopsies for treatment guidance in a future clinical care setting¹⁰. In 2019, Camolotto et al. showed that knockout of the transcription factor *Hnf4a* in mice resulted in a loss of classical expression, a decrease in tissue

differentiation, and worse prognosis¹¹. Even more recent work by Hayashi et al. was able to show concordance between the Moffitt subtypes, morphological presentation, and clinical course¹². In addition, O'Kane et al. used *GATA6* as a surrogate for the Moffitt basal-like/classical identification and recapitulated robust response to tumor and overall survival in the COMPASS cohort¹³. To enhance the clinical applicability of the basal-like/classical model, we previously generated a single-sample classifier termed purity independent subtyping of tumors (PurIST)^{3, 14}. This classifier showed a robust ability to predict a patient's molecular subtype and response to therapy across multiple technologies and without the need for a cohort⁸.

While a limited number of studies have begun comparisons of subtype performance in clinical care, diversity of opinions in the field and difficulty in obtaining data has led to confusion. Indeed, even recent studies show mixed strategies among investigators choosing which datasets or gene signatures to use. In particular, we highlight three common issues. First, validation is frequently only performed on one publicly available dataset. Second, there are multiple examples of inappropriate use of The Cancer Genome Atlas (TCGA) Program PDAC dataset^{15–30}. The mRNA dataset includes 182 samples, but only 150 have been confirmed histologically as PDAC. The remaining 32 samples are classified as either other types of pancreatic cancer, no evidence of cancer, or adjacent normal samples. Including such samples in derivation or validation of expression signatures results in misleading results. Finally, a growing consensus surrounding the two-subtype model has led researchers to collapse other 3+ subtype models into two in an attempt to harmonize with the basal-like/classical model^{31–33}. However, this approach does not address confounding factors related to purity and tissue specificity inherent in the approaches used to derive the models. Given that the two-subtype and collapsed 3+ subtype signatures are not completely redundant, attempting to interchangeably label a subset of the data as basal-like, squamous, or QM is not equivalent.

To facilitate standardized investigation of PDAC data, we have created an R package, termed *pdacR*, for public use. As part of this package, we have gathered and carefully annotated thirteen human PDAC datasets from the past decade for a thorough comparison of the performance of subtyping schemas in identifying only the tumor cells within PDAC tumors (Table 1). In keeping with the Findability, Accessibility, Interoperability, and Reusability (FAIR) practices promoted by the NIH, we also created a graphical user interface (GUI) for easy-to-use interrogation of these datasets in a point-and-click manner that aims to enable a wide audience to succinctly pursue hypotheses in existing PDAC transcriptomic data^{34–36}. This GUI is available as web-hosted software for ease of access and use. The flexibility of *pdacR*, combined with its inclusion of a wide range of PDAC datasets that are focused on the most reliable, large, and commonly referenced studies, expands upon on previous work by Marzec et al. and Tan et al. to develop the Pancreatic Expression Database (PED) and Human Pancreatic Cancer Database (HPCDB), respectively, for pancreatic cancer data centralization and user interface development for genomics analysis^{37–40}.

Results

Features and organization of the *pdacR* package. To facilitate access and investigation of the datasets used for this paper, we compiled thirteen of the most referenced transcriptomic human PDAC datasets in an open-source R package we call *pdacR*. Beyond gene expression and sample information for these datasets, the package includes curated gene sets defined in earlier PDAC and immunological studies, the R code used to parse and

Table 1 Datasets in the pdacR package and those used in this analysis.

Dataset	Type	Sample description	Samples	Features	Used
Chen, 2015 ⁵²	MA	63 fresh frozen microdissected PDAC tumor samples	63	11	no
ICGC PACA-AU, 2016 ⁴	MA	131 primary pancreatic tumors, 101 PDAC, 30 other	131	32	yes
ICGC PACA-AU, 2016 ⁴	RNAseq	82 primary pancreatic tumors, eight cell lines, and two metastatic tumors	92	32	yes
ICGC PACA-CA, 2016	RNAseq	40 cell lines, 11 xenografts, 16 metastatic tumors, and 195 primary pancreatic tumors	49	11	no
Moffitt, 2015 ³	MA	145 primary PDAC, 61 metastatic PDAC, 17 cell lines, 46 normal pancreas, 88 distant site normal tissue	357	25	yes
Moffitt, 2015 ³	RNAseq	15 primary PDAC, 37 patient-derived xenografts, three PDAC cell lines, and six cancer-associated fibroblast lines	61	7	yes
Olive, 2019 ⁴²	RNAseq	15 triplicate-matched PDAC bulk tumor and laser capture microdissected (LCM) epithelium and stroma, 51 additional LCM epithelium-stroma pairs, and 57 additional unmatched LCM stroma	204	19	yes
Puleo, 2018 ⁵	MA	309 primary PDAC FFPE block tumor samples	309	19	yes
Seino, 2018 ⁵³	MA	2 normal, 39 PDAC, 6 normal-like PDAC, and 10 genetically engineered organoids	57	11	no
TCGA-PAAD, 2017 ⁴¹	RNAseq	181 macrodissected primary tumors, with 150 whitelisted by histologic identification of neoplastic cellularity and cancer type	181	55	yes
Nones, 2013 ⁵⁴	MA	103 primary pancreatic ductal adenocarcinoma samples	103	18	no
ScAtlas, 2022	scRNA	43 pseudobulked primary PDAC, 43 pseudobulked Normal pancreas	86	5	no
CPTAC3-PDA ⁵⁵⁻⁵⁷	RNAseq	140 Primary PDAC and 39 adjacent normal pancreas	179	78	no

List of datasets included in pdacR. Dataset indicates publication where data was initially published. Type indicates method of transcriptomic measurement (MA = Microarray, scRNA = Pseudobulked scRNA). Samples indicate number of samples in the dataset. Features indicate number of associated sample metadata (e.g., Tumor grade, histology, purity, etc.). The 'Used' column indicates if the dataset appears in this manuscript (yes) or if it is solely provided as part of the pdacR tool (no).

visualize the data shown in this work, and a *shiny*-based graphical user interface (UI) for point-and-click analysis of public and user-generated data (Fig. 1).

UI structure and data analysis. The UI is composed of four primary functional components that enable a range of investigative methods: heatmapping, dimension reduction, survival analysis, and differential expression. Users begin by choosing their dataset of choice from the list of those already curated or, on a local instance, may choose to load a user-generated dataset. User-generated datasets may be imported as R-formatted data which has been converted to a *SummarizedExperiment* container format. While the application will automatically check and convert data format from *SummarizedExperiment* to the appropriate list-style upon loading, specifics of this format and the conversion are provided on the pdacR GitHub page's README file. Once a dataset and a gene list has been selected, many common analyses can be performed. The heatmapping component enables users to generate ordered heatmaps using user-defined gene sets and customizable clustering or sorting methods. The gene sets included are pulled directly from a variety of seminal publications in the field of PDAC transcriptomics, allowing for comparison of their performance and robustness across the available datasets. The dimension reduction component generates projections of data by either principal component analysis or t-SNE methods, with data point coloring by sample metadata or up to three gene set expression scores in an RGB color space. The survival analysis component can generate either a Kaplan Meier or Cox-based survival curve on continuous (e.g., gene expression) or categorical (e.g., tumor grade) data, and enables the user to define continuous data as categorical by toggling a quantile cutoff. The differential expression component allows users to generate volcano plots with full user freedom to define head-to-head comparisons and select which genes or gene sets to highlight.

Comparison to other resources. Other graphical user interfaces have been developed to facilitate transcriptomics investigation in pancreatic cancer, though there are clear advantages offered in pdacR (Supplementary Table 1). Similar to the pancreatic expression database (PED) and the human pancreatic cancer database (HPCDB), pdacR combines data from several independent studies to enable hypothesis testing in different sample types from a range of datasets. PED, like pdacR, features a graphical user interface to make analyses available to users without computer programming abilities and allows users to filter their datasets based on some clinical parameters, though it is limited in other areas critical for accurate and reproducible investigation. Unlike other tools, pdacR features datasets that were parsed in a manner consistent with the recommendations of the original authors. Also, pdacR allows for extensive and flexible sample filtration prior to analysis. This is a unique feature critical to ensuring that analysis is performed on the appropriate sub-populations of datasets, while also allowing for the generation of forward-looking analyses. Importantly, the structure of pdacR as a functional R package provides investigators the ability to move their hypothesis testing from the graphical user interface to more intensive code-based analyses in an ad-hoc and specialized manner when necessary, a functionality not offered in other pancreatic cancer investigation software. Finally, pdacR uniquely combines the most frequently cited datasets from the PDAC literature with the ability for researchers to locally load their own data for ease of access to computational tools. While this package is currently limited in its ability to analyze large-scale mutation data and lacks pathways-based investigation modes, it is also, importantly, open-source to enable continuous interface

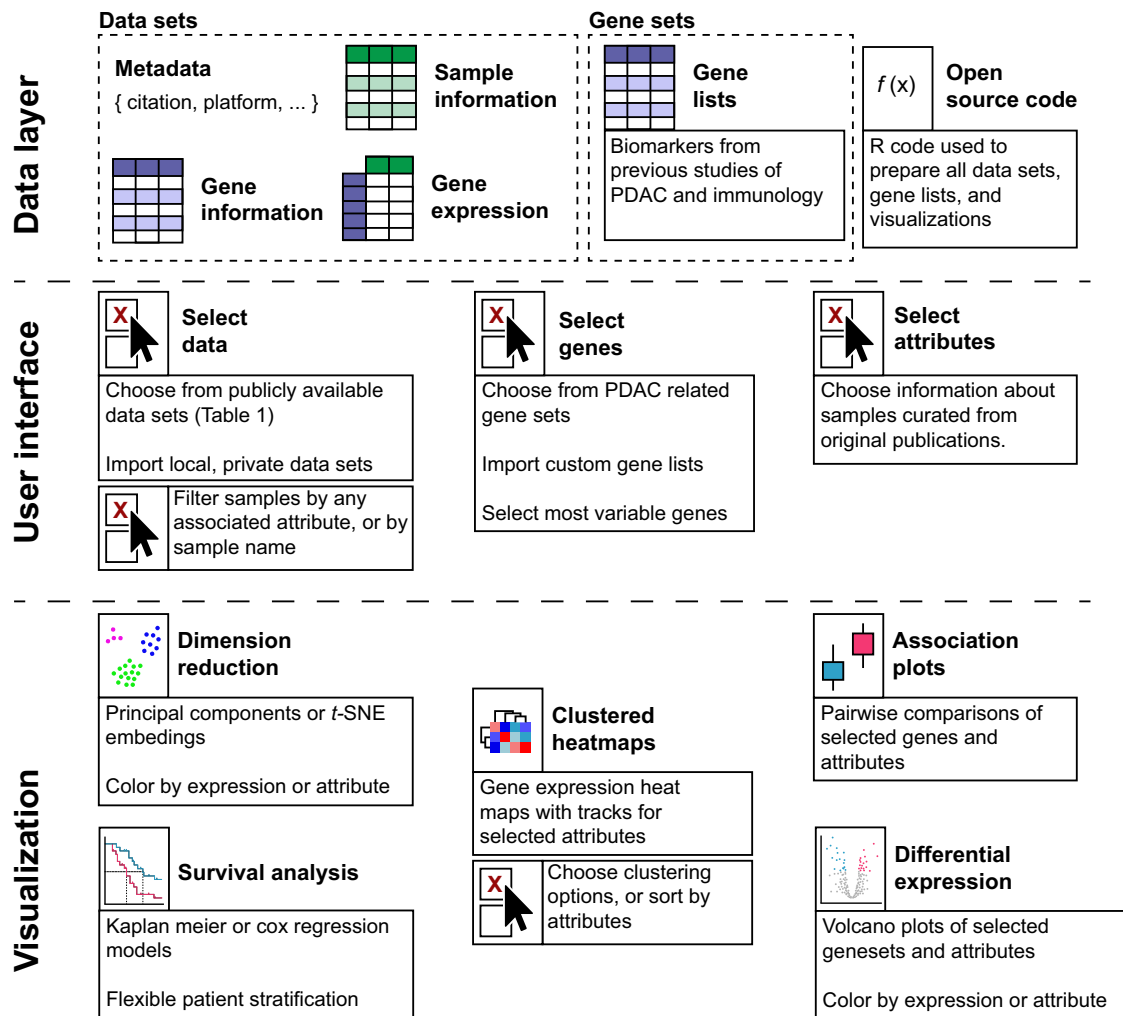


Fig. 1 Features and organization of the pdacR package. The pdacR package contains hand-curated and annotated datasets from many of the landmark PDAC studies conducted over the past decade and others. *Data layer:* Dataset sample- and gene-based data are organized by the primary author of the dataset and the year of publication. Gene lists defined in these and other studies are compiled for ease of analysis, named by author and cell/tissue type defined by the list (e.g., “CIBERSORT Monocytes”, a gene list that specifically identifies cells of the monocyte lineage). *User interface:* A GUI that enables point-and-click is hosted online and also may be called locally by the user through the R Shiny interface. Users may select from the compiled datasets, filter samples, select genes to use in their analysis (either user-defined or built into the package), and select factors from the samples with which to visualize the data. *Visualization:* The package includes functions to facilitate data analysis and visualization, either generated for this package or from others in the literature. In the GUI, users may save images from their analysis as PNGs or PDFs for further illustration and publication.

development and dataset inclusion, submitted using github pull requests.

ADEX, Exocrine, and Immunogenic subtypes are not tumor intrinsic. Recent work from Rashid et al. compared the prognostic values of the Collisson, Moffitt, and Bailey subtyping schemas across clinical studies¹⁴. Here, we used transcriptomic data from several pancreatic cancer studies to evaluate the tumor cell specificity of these schemas. Using data from the TCGA study, we determined that high expression of genes specific to the Bailey ADEX (“ADEX”, Fig. 2a) gene signature in PDAC tumors predicts low tumor cellularity as estimated by ABSOLUTE (Spearman $\rho = -0.260$, $p = 0.0014$) as does the Collisson exocrine gene set (Supplementary Fig. 1A). Low tumor cellularity, however, was not predictive of high ADEX gene signature expression. Using data from Puleo et al., with predicted sample average variant allele frequency (VAF) as a proxy for sample tumor cellularity, we then confirmed that the Collisson

Exocrine signature is similarly predictive of low sample purity (Spearman $\rho = -0.221$, $p = 0.0003$, Fig. 2b, association with Bailey ADEX gene set in Supplementary Fig. 2B). Importantly, these trends were not found in other gene signatures of the Collisson or Bailey subtyping schemas, suggesting that ADEX and Exocrine subtypes may be driven by expression from non-neoplastic cells.

In an analysis of paired samples (primary tumor-primary tumor, primary tumor-metastatic tumor, or primary tumor-cell line) from the PACA-AU RNAseq and PACA-AU microarray datasets, we determined that expression of the Bailey ADEX, Bailey Immunogenic, and Collisson exocrine gene signatures are lost in cell lines derived from primary PDAC tumors with a range of expression of these same signatures (Fig. 2c). Contrastingly, we found that tumor cell lines retain expression of the Collisson classical and QM, Moffitt classical and basal-like, and Bailey pancreatic progenitor and squamous subtypes (Supplementary Fig. 1). The retention of certain signatures describing both basal-like and classical populations suggests that a non-tumor cell

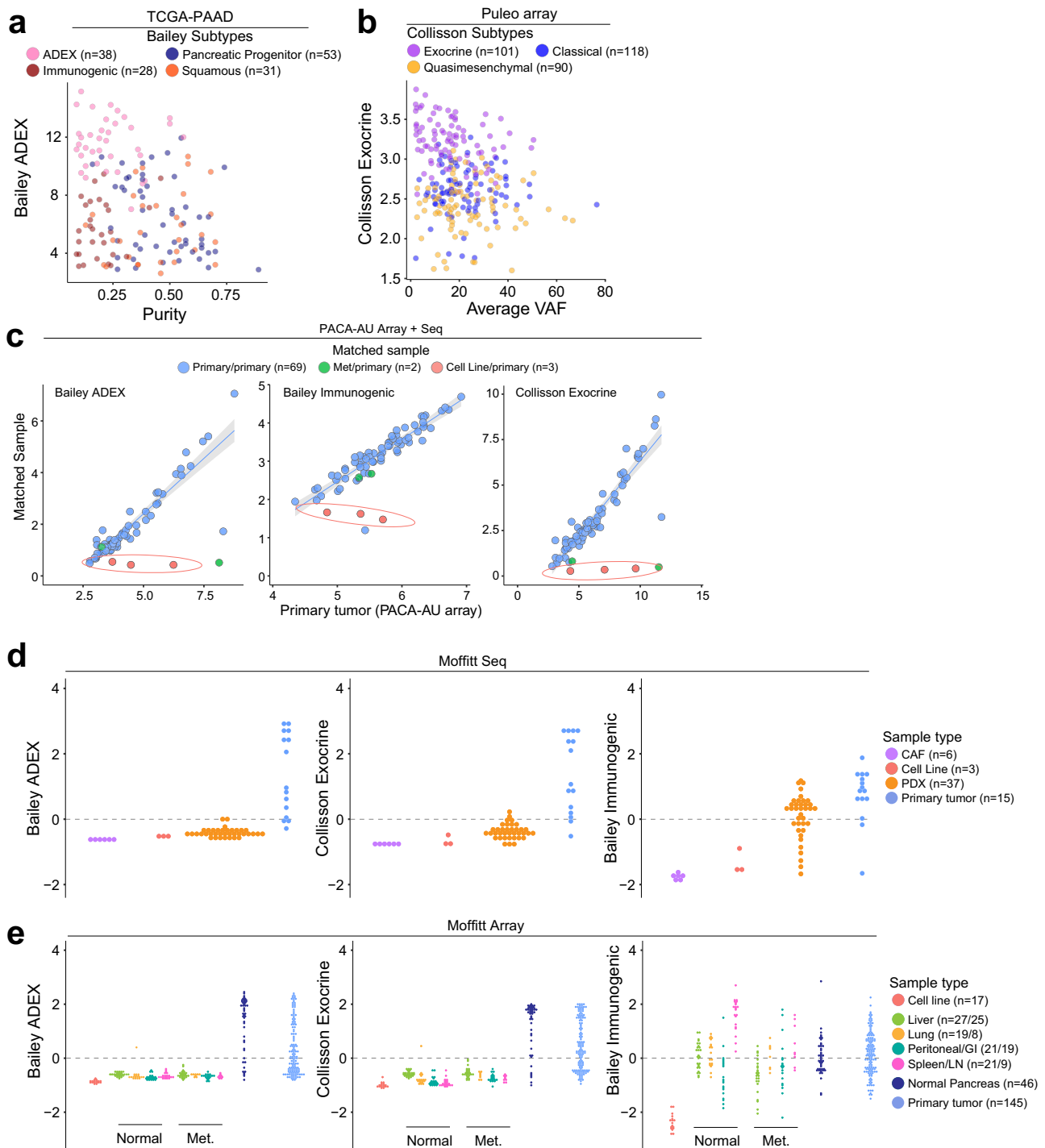


Fig. 2 ADEX, Exocrine, and Immunogenic PDAC subtypes are not tumor-intrinsic. **a** Expression of genes in the Bailey ADEX geneset is inversely correlated with tumor sample purity, assessed by ABSOLUTE, in the TCGA dataset (Spearman $\rho = -0.260$, $p = 0.0014$, $n = 150$). Low tumor sample purity is predictive of high ADEX geneset expression. **b** The Collisson exocrine geneset exhibits the pattern seen in the TCGA dataset (**a**), extended to the Puleo microarray (Spearman $\rho = -0.221$, $p = 0.0003$, $n = 309$). **c** Matched samples between the PACA-AU RNAseq and PACA-AU array datasets show that the ADEX, Collisson Exocrine, and Bailey Immunogenic gene signatures are lost in cell lines and metastatic tumors derived from primary PDAC tumors expressing either high or low levels of the same gene signatures (gene expression plotted as the average of the log2 transformation of the genes which define the subtype). **d** The ADEX and Exocrine genesets are highly expressed in primary tumor samples compared to cancer-associated fibroblasts (CAFs), cell lines, and patient-derived xenografts (PDXs), though the Bailey Immunogenic geneset demonstrates appreciable expression in PDXs. **e** ADEX and Exocrine gene signatures are highly expressed in normal pancreas or primary PDAC tumors, and are comparatively absent from metastatic tumors, cell lines, and normal non-pancreas samples, whereas the Immunogenic gene signature is expressed across normal non-pancreas, metastatic, normal pancreas, and primary PDAC tumors but is lost in cell lines. Y-axes from (**d**) and (**e**) represent the mean-centered, scaled average expression for the indicated genesets.

component of the whole tumor PDACs used in this study is contributing to the expression of genes defining the ADEX, Immunogenic, and Exocrine subtypes.

To further assess the specificity of existing subtype gene signatures, we used RNAseq and microarray data from Moffitt et al., which includes primary PDACs, patient-derived xenografts, PDAC cell lines, cancer-associated fibroblasts (CAFs), and normal and metastatic PDAC tissue from liver, lung, peritoneum/stomach/intestine, and spleen/lymph node. We found that the ADEX, Exocrine, and Immunogenic gene sets are highly expressed in primary PDACs compared to either PDAC cell lines or subcutaneously implanted patient-derived xenografts (PDXs). Notably, primary PDACs also demonstrate increased expression of these gene sets compared to CAF cell lines ($p < 0.0001$), suggesting that the non-tumor component driving ADEX, Exocrine, and Immunogenic gene expression is unlikely dominated by CAFs (Fig. 2d). Moffitt basal-like and classical subtypes were similarly resistant to confounding by CAF gene expression (Supplementary Fig. 1B). Similarly, this trend is maintained across the Collisson classical/QM subtypes and the Bailey progenitor/squamous subtypes, which are highly concordant (~60% overlap) with the Moffitt classical/basal-like subtypes, respectively (Supplementary Fig. 1b, c). Importantly, we found that the ADEX/Exocrine signatures are enriched in normal pancreata compared to PDACs ($p < 0.0001$), suggesting a pancreas cell intrinsic ADEX/Exocrine gene expression origin (Fig. 2e). The Bailey Immunogenic subtype, however, was enriched in normal pancreas, metastatic tumor, and primary PDACs compared to cell lines ($p < 0.0001$), suggesting that a non-tumor cell component drives expression of this gene set across several tissue types (Fig. 2e). Contrastingly, though Moffitt basal-like and classical subtypes demonstrate varied representation across normal pancreas, primary pancreatic tumor, and metastatic tumor, expression of these subtypes is retained in PDAC cell lines, with similar expression patterns to the analogous Collisson and Bailey subtypes (Supplementary Fig. 1C). It is important to note that there exist pancreatic tumors of high tumor cell purity that also show high expression of the ADEX and Exocrine gene sets. Histologic characterization of these tumors, however, reliably identifies these tumors as acinar cell carcinomas (ACCs), not PDAC (Fig. 3b, c, Supplementary Fig. 2c).

Subsets of the ADEX, Exocrine, and Immunogenic gene signatures mark non-adenocarcinoma samples. After identifying that the ADEX, Exocrine, and Immunogenic gene signatures are not tumor cell specific, we hypothesized that expression of these gene sets might be confounded by subsets of genes which are not PDAC-specific. To determine if this were the case, we performed gene-based consensus clustering of the PACA-AU RNAseq and Moffitt RNAseq datasets. We found that the Bailey ADEX gene set is comprised of primarily three subsets of genes: those that are generally nonspecific, those that are pancreas nonspecific, and those that are ACC-specific. This could explain why ADEX signatures are enriched in ACC samples, and provides justification for the non-tumor-specific nature of the signature (Fig. 3a).

A similar analysis of the Bailey Immunogenic gene set demonstrated 26% overlap with the pancreatic progenitor and 9% of the ADEX gene sets, suggesting a lack of specificity in the signature. Furthermore, we identified that 24 of the 180 immunogenic genes used in this analysis are highly expressed in primary PDACs but not PDXs, suggesting that these genes likely confer non-tumor specificity to the signature (Fig. 3d). In order to reliably investigate the role of tumor or stromal expression patterns on response to therapy and survival, the signals must be clearly separable by tissue compartment.

Large gene sets are confounded by expression in adjacent normal tissues. Gene sets become increasingly prone to noise and off-target effects as they increase in length and include more biologically multi-functional transcripts. For example, normal pancreas expresses even the neoplastic Collisson and Bailey gene sets (Supplementary Fig. 1D, E). The lists extracted from Chan-Seng-Yue et al. also show aberrant expression in a range of normal tissues as well as susceptibility to confounding by CAFs specifically in the Basal-like A list (Supplementary Fig. 3B, C). As an addition to categorical approaches, Nicolle et al. recently developed a Pancreatic Molecular Gradient (PAMG) to assign a continuous score that correlates with differentiation status and survival⁷⁷. The signature, derived from primary tumors and PDXs, incorporates up to 20,164 genes into scoring. As seen in other signatures, PAMG application results in a varying range of scores across purities and tissue types (Supplementary Figs. 3, 4). The basal-like/classical signatures were purposefully derived with this problem in mind (Moffitt 2015). As such, they are robust against both metastatic location (Supplementary Fig. 1C) and neoplastic purity (Supplementary Fig. 4)⁴¹. Thus, longer signatures result in more complex and potentially less useful tools than the basal-like/classical or PurIST models. The above data together suggest that in the context of unpurified samples, the basal-like/classical signatures or PurIST classifier should be used in favor of seemingly similar models.

Discussion

Several studies have proposed PDAC transcriptomic subtypes. Though efforts have been made to explain reasons for the differences between these proposals, widespread uncertainty is evident in the PDAC literature regarding which subtyping schemas are most relevant for prospective studies of tumor cells³¹. Here we used transcriptomic data from several landmark PDAC studies to understand which subtypes are the most tumor-cell specific. We show that the basal-like/classical and PurIST two-subtype schemas specifically are the most likely to identify tumor cell components of PDAC tumors across studies, reinforcing and expanding upon early suggestions by TCGA and Maurer et al.^{41, 42}.

One goal of describing molecular subtypes is to predict patient outcomes and response to therapy. We previously have shown that the logistic regression classifier PurIST remains robust across multiple datasets at predicting patient outcomes¹⁴. A benefit of our classifier is the lack of variation between datasets, cohorts, and data types. By utilizing pre-defined top-scoring pairs of genes, PurIST can be applied to a single patient and generate a prediction without the need for a cohort. This reduces uncertainty and promotes reliability regardless of available technology or dataset size. Conversely, continuous scores such as PAMG introduce uncertainty at the clinical level due to the requirement for normalization. This makes the score, and therefore the prediction, cohort dependent.

Beyond needing clarity regarding the clinical application and tumor-intrinsic quality of transcriptomic subtypes, the PDAC research field currently lacks data centralization. Furthermore, others have indicated that data that is readily accessible through online repositories has, in some cases, been either poorly formatted or poorly annotated for subsequent analysis^{43, 44}. In certain instances, this may lead to misguided or misinformed conclusions in well-meaning investigations^{15–30}. Clearly, these challenges pose a major hurdle in the development of novel hypotheses, particularly for those without first-hand familiarity with how the data were generated. To remedy this, we developed an R package termed pdacR that houses carefully organized and annotated public PDAC transcriptomics data for ease of access

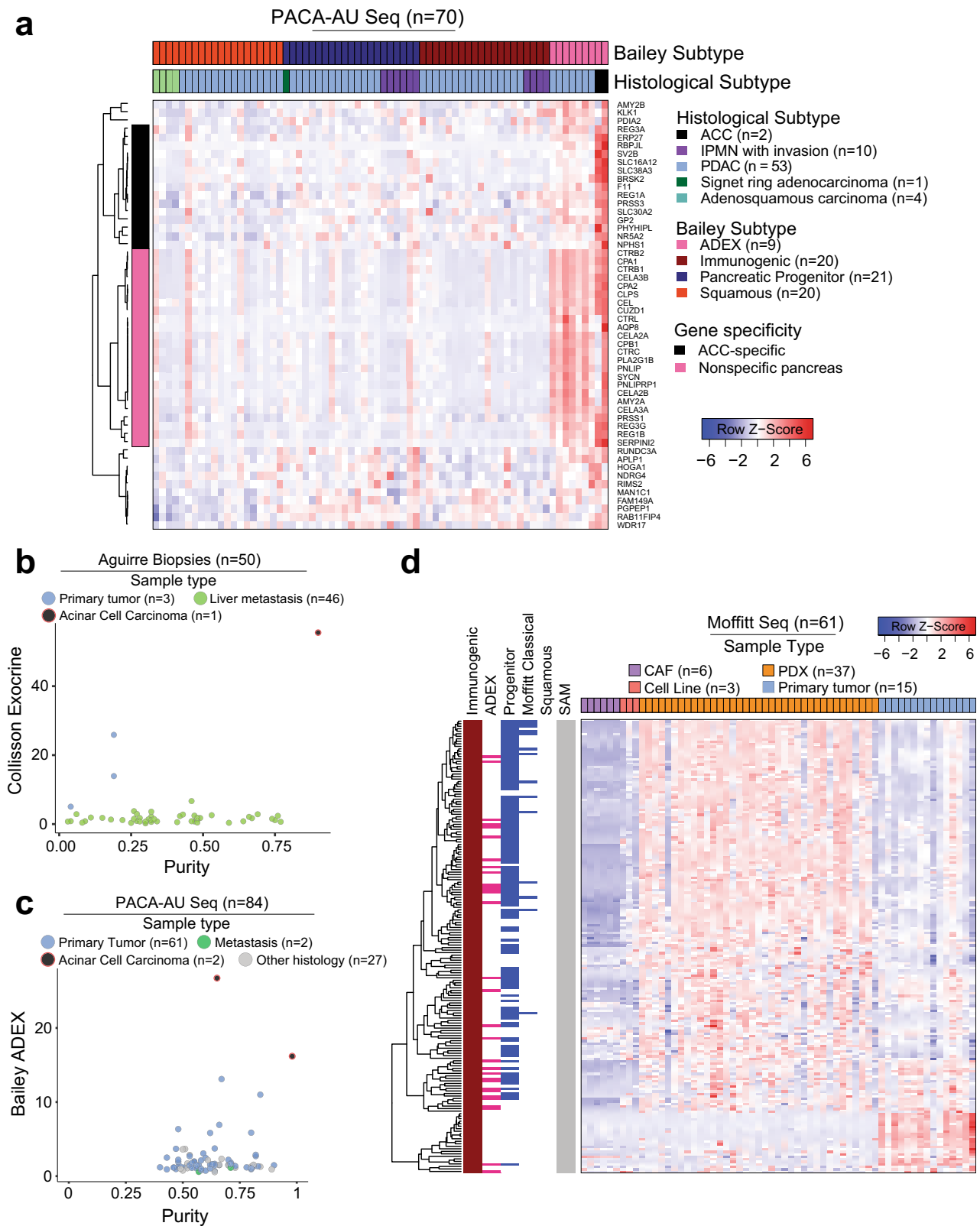


Fig. 3 Subsets of the ADEX, Exocrine, and Immunogenic gene signatures mark non-adenocarcinoma samples. **a** The Bailey ADEX geneset is comprised of genes that specifically identify Acinar Cell Carcinomas (black), genes that appear to be more highly expressed in the ADEX subtypes yet are conserved in non-ADEX samples (light pink), and genes that are nonspecific to a known sample type (uncolored). **b, c** High-purity samples from core needle biopsies or primary tumor samples marked by high expression of Exocrine or ADEX expression, respectively, were histologically identified as Acinar Cell Carcinomas. **d** The subset of genes from the Immunogenic geneset that is highly expressed in PDXs is also found in the Bailey ADEX (pink), Pancreatic Progenitor (blue, left), and Moffitt classical (blue, right) signatures, while genes exclusive to the Immunogenic subtype are not expressed in PDXs.

and heightened transparency. pdacR includes a point-and-click user interface to facilitate investigation of these datasets by researchers without formal training in computational statistics in an effort to democratize data access. We have shown here that the pdacR user interface provides the user with a wide variety of data visualization methods that are highly customizable yet user-friendly. Importantly, pdacR improves on existing pancreatic cancer user interfaces, namely by retaining its flexibility as a user interface and R package, with full customization by the experienced statistician. We expect that this tool will become widely adopted in the PDAC community, especially as RNA quantification methods become more widely used to study this disease.

We have presented data to suggest that the basal-like/classical subtyping schema is both the most clinically reliable model and the most appropriate in describing only the tumor cells within a non-purified PDAC tumor. This data and others are now compiled in a public R package, pdacR, for open use by the research community. While this work has helped further our understanding of PDAC subtypes, future studies on these subtypes in the context of patient outcomes and response to therapy are necessary to apply our understanding to clinically meaningful intervention strategies. We hope this tool (<http://pdacR.bmi.stonybrook.edu>) and the associated code and data (github.com/rmoffitt/pdacR) will be used by other researchers to facilitate robust external validation of future analyses.

Recent attempts to improve our understanding of PDAC have emphasized the utility of single-cell RNA sequencing (scRNA). Given that the datasets provided here are all bulk analyses, we are limited in our ability to fully separate cellular populations. We partially address this shortcoming using datasets derived from a multitude of pure cell type sources or those enhanced by LCM. While scRNA sequencing is expanding in its utilization, it remains prohibitively expensive for inclusion in the realm of clinical characterization and application. While incorporation of new scRNA datasets would improve this tool from a purely investigative standpoint, the computational burden and memory required are beyond the scope of a responsive GUI. We have, in parallel, made available a parsed single-cell atlas (<https://github.com/rmoffitt/scOh>). This atlas does not include a web interface, but covers many substantial hurdles in scRNA analysis, including parsing, integration, and normalization. Another limitation inherent to this sort of data collection is the inability to robustly merge datasets. Preliminary versions of this tool did have the option for data aggregation. However, in addition to being too computationally intensive for a point-and-click interface, we found that aggregation introduced more problems than it solved. If we cannot assume that datasets come from comparable distributions and cell type compositions (e.g., all samples in both datasets are primary tumors), then we necessarily confound our analysis. In addition, our datasets come from a variety of platforms, including bulk RNA sequencing and microarrays, adding another layer of complexity to aggregation and batch correction. Believing that providing the option of poorly-merged data would be more harmful than helpful, we ultimately decided was beyond the scope of this project. It is our belief that the rapidity with which a researcher could perform serialized analysis on different datasets overcomes much of the limitation of not being able to merge the data.

Methods

Data acquisition and processing. All datasets used in this study were obtained from public repositories, indicated in Table 1. All data were previously publicly available and were collected under the approval of the respective IRBs at the time of collection and publication of the respective study. Expression matrices were unmodified compared to the original publication, without re-alignment of RNAseq data, or re-normalization of array data.

Gene expression analysis. Gene expression analyses were performed using log₂-scaled expression matrices. Heatmaps were generated using the “heatmap.3” function in R. Gene set scores were called as the mean expression value of the genes in gene set. For correlation between the PACA-AU RNAseq and Micro Array, these gene set scores were compared directly. For evaluation of expression across different tissue types in Moffitt Array and Moffitt Seq, these scores were mean-centered and scaled.

Clustering. For clustering and subtype calls (when lacking in original publication), datasets were trimmed down to include only PDAC samples and only the genes present in the relevant gene sets. We then applied the ConsensusClusterPlus function with kmeans clustering and euclidean distance to call the appropriate number of cluster ($k = \text{number of gene sets}$).

Statistics and reproducibility. To account for the potential disproportionate impact of outliers on correlations, correlation values were calculated using Spearman Rho. Difference in expression of gene sets by cell types was calculated using a non-parametric Wilcoxon Rank-Sum test. Reproducibility is one of our main motivators for this publication. To that end, we have provided a web application wherein all the analyses can be replicated with a series of clicks. We have also provided all our analysis code, along with parsed and ready-to-analyze datasets, in our [github](https://github.com/rmoffitt/pdacR)⁴⁵.

R package and UI design. The pdacR UI was designed using the *shiny* R package. Datasets were chosen for inclusion in the pdacR package based on public availability or investigator permission, as well as importance in the field as determined by the authors. Differential expression analysis on RNAseq or Array experiments is performed using the DESeq2⁴⁶ or limma⁴⁷ packages respectively. The logFC is calculated by dividing the mean of cohort A by the mean of cohort B for each gene. Survival analysis is performed using the survival and survminer packages in R. Censor and survival data are obtained from the initial publications, and are a combination of overall and disease-specific survivals. User-generated data may be converted from a *SummarizedExperiment* container style to a UI-compatible format using our ‘./R/Convert_GUI_data.R’ wrapper function.

Reporting summary. Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All datasets used in this study are compiled in our pdacR package, which is freely available to the public at github.com/rmoffitt/pdacR. All datasets have been pulled from publicly accessible databases, and their accessions are contained within the metadata of each R object. Accession numbers are also listed here: Chen - GSE57495; CPTAC - phs001287; Nones - GSE50827; Bailey (PACA-AU) - EGAS00001000154; PACA-CA - icgc.org; Moffitt Array - GSE71729; Moffitt Seq - PMID:26343385 (Supplement); Olive - GSE93326; Puleo - E-MTAB-6134; Seino - GSE107610; TCGA - phs000178. Within R, these data can be pulled from their original sources using packages such as TCGAbiolinks^{48–50} or GEOquery⁵¹, but we recommend using our pre-parsed formats.

Code availability

All parsing and analytical code used in this study are compiled in our pdacR package, which is available to the public at github.com/rmoffitt/pdacR. Figures were generated using the .rmd files in *inst/analysis*, while the entire app framework is located at *inst/shiny/app.R*. The Github version at time of manuscript preparation can be identified by SHA: 3a3f5683059b943cb3ef2d762ab3cddcd572bb0⁴⁵.

Received: 2 December 2021; Accepted: 11 January 2023;

Published online: 10 February 2023

References

1. Stratford, J. K. et al. A six-gene signature predicts survival of patients with localized pancreatic ductal adenocarcinoma. *PLoS Med.* **7**, e1000307 (2010).
2. Collisson, E. A. et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat. Med.* **17**, 500–503 (2011).
3. Moffitt, R. A. et al. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat. Genet.* **47**, 1168–1178 (2015).
4. Bailey, P. et al. Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* **531**, 47–52 (2016).

5. Puleo, F. et al. Stratification of pancreatic ductal adenocarcinomas based on tumor and microenvironment features. *Gastroenterology* **155**, 1999–2013.e1993 (2018).
6. Chan-Seng-Yue, M. et al. Transcription phenotypes of pancreatic cancer are driven by genomic events during tumor evolution. *Nat. Genet.* **52**, 231–240 (2020).
7. Nicolle, R. et al. Establishment of a pancreatic adenocarcinoma molecular gradient (PAMG) that predicts the clinical outcome of pancreatic cancer. *EBioMedicine* **57**, 102858 (2020).
8. Aung, K. L. et al. In *Clinical Cancer Research* Vol. 24, 1344–1354 (American Association for Cancer Research Inc., 2018).
9. Tiriach, H. et al. Organoid profiling identifies common responders to chemotherapy in pancreatic cancer. *Cancer Disco.* **8**, 1112–1129 (2018).
10. Aguirre, A. J. et al. Real-time genomic characterization of advanced pancreatic cancer to enable precision medicine. *Cancer Disco.* **8**, 1096–1111 (2018).
11. Camolotto, S. A. et al. Reciprocal regulation of pancreatic ductal adenocarcinoma growth and molecular subtype by HNF4a and SIX1/4. *bioRxiv* <https://doi.org/10.1101/814525> (2019).
12. Hayashi, A. et al. A unifying paradigm for transcriptional heterogeneity and squamous features in pancreatic ductal adenocarcinoma. *Nat. Cancer* **1**, 59–74 (2020).
13. O’Kane, G. M. et al. GATA6 expression distinguishes classical and basal-like subtypes in advanced pancreatic cancer. *Clin. Cancer Res.* **26**, 4901–4910 (2020).
14. Rashid, N. U. et al. Purity independent subtyping of tumors (PurIST), a clinically robust, single-sample classifier for tumor subtyping in pancreatic cancer. *Clin. Cancer Res.* **26**, 82–92 (2020).
15. Mishra, N. K. & Guda, C. Genome-wide DNA methylation analysis reveals molecular subtypes of pancreatic cancer. *Oncotarget* **8**, 28990–29012 (2017).
16. Li, H. et al. Integrated expression profiles analysis reveals novel predictive biomarker in pancreatic ductal adenocarcinoma. *Oncotarget* **8**, 52571–52583 (2017).
17. Hu, H. et al. Elevated COX-2 expression promotes angiogenesis through EGFR/p38-MAPK/Sp1-dependent signalling in pancreatic cancer. *Sci. Rep.* **7**, 470 (2017).
18. Moncada, R. et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat. Biotechnol.* **38**, 333–342 (2020).
19. Wei, R. et al. IMUP and GPRC5A: two newly identified risk score indicators in pancreatic ductal adenocarcinoma. *Cancer Cell Int* **21**, 620 (2021).
20. Dong, C. et al. Intron-retention neoantigen load predicts favorable prognosis in pancreatic cancer. *JCO Clin. Cancer Inf.* **6**, e2100124 (2022).
21. Hu, Y. et al. A reciprocal feedback between N6-methyladenosine reader YTHDF3 and lncRNA DICER1-AS1 promotes glycolysis of pancreatic cancer through inhibiting maturation of miR-5586-5p. *J. Exp. Clin. Cancer Res.* **41**, 69 (2022).
22. Loeffler, C. M. L. et al. Predicting mutational status of driver and suppressor genes directly from histopathology with deep learning: a systematic study across 23 solid tumor types. *Front. Genet.* **12**, 806386 (2021).
23. Wang, X. et al. Reveal the heterogeneity in the tumor microenvironment of pancreatic cancer and analyze the differences in prognosis and immunotherapy responses of distinct immune subtypes. *Front. Oncol.* **12**, 832715 (2022).
24. Zhu, T. et al. A pan-tissue DNA methylation atlas enables in silico decomposition of human tissue methylomes at cell-type resolution. *Nat. Methods* **19**, 296–306 (2022).
25. Wei, R. et al. Type 1 T helper cell-based molecular subtypes and signature are associated with clinical outcome in pancreatic ductal adenocarcinoma. *Front. Cell Dev. Biol.* **10**, 839893 (2022).
26. Zhan, Q. et al. Identification of copy number variation-driven molecular subtypes informative for prognosis and treatment in pancreatic adenocarcinoma of a Chinese cohort. *EBioMedicine* **74**, 103716 (2021).
27. Frey, P. et al. SMAD4 mutations do not preclude epithelial-mesenchymal transition in colorectal cancer. *Oncogene* **41**, 824–837 (2022).
28. Martínez-Bosch, N. et al. Soluble AXL is a novel blood marker for early detection of pancreatic ductal adenocarcinoma and differential diagnosis from chronic pancreatitis. *EBioMedicine* **75**, 103797 (2022).
29. Chen, G. et al. Identification and validation of constructing the prognostic model with four DNA methylation-driven genes in pancreatic cancer. *Front. Cell Dev. Biol.* **9**, 709669 (2021).
30. Liu, X. et al. Multi-omics analysis of intra-tumoural and inter-tumoural heterogeneity in pancreatic ductal adenocarcinoma. *Clin. Transl. Med.* **12**, e670 (2022).
31. Collisson, E. A., Bailey, P., Chang, D. K. & Biankin, A. V. Molecular subtypes of pancreatic cancer. *Nat. Rev. Gastroenterol. Hepatol.* **16**, 207–220 (2019).
32. Brunton, H. et al. HNF4A and GATA6 loss reveals therapeutically actionable subtypes in pancreatic cancer. *Cell Rep.* **31**, 107625 (2020).
33. Raghavan, S. et al. Microenvironment drives cell state, plasticity, and drug response in pancreatic cancer. *Cell* **184**, 6119–6137.e6126 (2021).
34. Boeckhout, M., Zielhuis, G. A. & Bredenoord, A. L. The FAIR guiding principles for data stewardship: fair enough? *Eur. J. Hum. Genet.* **26**, 931–936 (2018).
35. Corpas, M., Kovalevskaya, N. V., McMurray, A. & Nielsen, F. G. G. A FAIR guide for data providers to maximise sharing of human genomic data. *PLOS Computat. Biol.* **14**, e1005873 (2018).
36. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
37. Chelala, C. et al. Pancreatic Expression database: a generic model for the organization, integration and mining of complex cancer datasets. *BMC Genomics* **8**, 439 (2007).
38. Dayem Ullah, A. Z. et al. The pancreatic expression database: recent extensions and updates. *Nucleic Acids Res.* **42**, D944–D949 (2014).
39. Marzec, J. et al. The pancreatic expression database: 2018 update. *Nucleic Acids Res.* **46**, D1107–d1110 (2018).
40. Tan, Y. et al. HPCDB: an integrated database of pancreatic cancer. Preprint at *bioRxiv* <https://doi.org/10.1101/169771> (2017).
41. Raphael, B. J. et al. in *Cancer Cell* Vol. 32, 185–203.e113 (Cell Press, 2017).
42. Maurer, C. et al. Experimental microdissection enables functional harmonisation of pancreatic cancer subtypes. *Gut* **68**, 1034–1043 (2019).
43. Nicolle, R. et al. Prognostic biomarkers in pancreatic cancer: avoiding errata when using the TCGA dataset. *Cancers* **11** <https://doi.org/10.3390/cancers11010126> (2019).
44. 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. *Clin. Cancer Res.* **24**, 3813–3819 (2018).
45. Torre-Healy, L., Kawalerski, R. & Moffitt, R. rmoffitt/pdacR: ready for publication. <https://doi.org/10.5281/zenodo.7383375> (2022).
46. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
47. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
48. Colaprico, A. et al. TCGAAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**, e71–e71 (2015).
49. Silva, T. C. et al. TCGA Workflow: analyze cancer genomics and epigenomics data using Bioconductor packages. *F1000Res* **5**, 1542 (2016).
50. Mounir, M. et al. New functionalities in the TCGAAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLOS Computat. Biol.* **15**, e1006701 (2019).
51. Davis, S. & Meltzer, P. S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **23**, 1846–1847 (2007).
52. Chen, D. T. et al. Prognostic fifteen-gene signature for early stage pancreatic ductal adenocarcinoma. *PLoS ONE* **10**, e0133562 (2015).
53. Seino, T. et al. Human pancreatic tumor organoids reveal loss of stem cell niche factor dependence during disease progression. *Cell Stem Cell* **22**, 454–467.e456 (2018).
54. Grimont, A. et al. SOX9 regulates ERBB signalling in pancreatic cancer development. *Gut* **64**, 1790–1799 (2015).
55. Cao, L. et al. Proteogenomic characterization of pancreatic ductal adenocarcinoma. *Cell* **184**, 5031–5052.e5026 (2021).
56. Clark, K. et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J. Digital Imaging* **26**, 1045–1057 (2013).
57. National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium pancreatic ductal adenocarcinoma collection (CPTAC-PDA) (version 11) [data set]. *Cancer Imaging Arch.* <https://doi.org/10.7937/K9/TCIA.2018.SC20FO18> (2018).

Acknowledgements

We would like to thank all included teams for supporting use of their data in pdacR via depositing to a non-controlled repository. This work was supported by a Research Scholar Grant, RSG-21-026-01 - CCG, from the American Cancer Society, as well as an R01, R01CA237404, from the National Cancer Institute.

Author contributions

Data acquisition: R.R.K., K.O., L.C., N.U.R., X.L.P., R.A.M., and A.J.A. Data analysis: L.T.H., R.R.K., and R.A.M. User interface development: L.T.H., R.R.K., and R.A.M. Writing of manuscript: L.T.H., R.R.K., and R.A.M. Critical evaluation of manuscript: L.T.H., R.R.K., K.O., L.C., N.U.R., X.L.P., A.J.A., J.J.Y., and R.A.M.

Competing interests

R.A.M., N.U.R., and J.J.Y. hold financial interests in PurIST (patents US11053550, WO2016061252A1), which has been licensed to GeneCentric Technologies. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-023-04461-6>.

Correspondence and requests for materials should be addressed to Richard A. Moffitt.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Ana Teresa Maia and Eve Rogers.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023