ORIGINAL ARTICLE

Experimental microdissection enables functional harmonisation of pancreatic cancer subtypes

Hans Carlo Maurer,[1,2,3] Sam R Holmstrom,[1,2,3] Jing He,[1,4,5] Pasquale Laise,[1,4,5] Tao Su,[1,3] Aqeel Ahmed,[1,3] Hanina Hibshoosh,[1,5] John A Chabot,[1,6] Paul E Oberstein,[7] Antonia R Sepulveda,[1,5] Jeanine M Genkinger,[1,8] Jiapeng Zhang,[9] Alina C Iuga,[1,5] Mukesh Bansal,[10] Andrea Califano,[1,4,5] Kenneth P Olive[1,2,3]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/gutjnl-2018-317706).

For numbered affiliations see end of article.

Correspondence to
Professor Andrea Califano, Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York, New York, USA; ac2248@cumc.columbia.edu and
Dr Kenneth P Olive, Columbia University Medical Center, New York, NY 10032, USA; kenolive@columbia.edu

HCM, SRH and JH contributed equally.

Received 8 October 2018
Revised 3 December 2018
Accepted 8 December 2018

To cite: Maurer HC, Holmstrom SR, He J, *et al*. *Gut* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/gutjnl-2018-317706

ABSTRACT

Objective Pancreatic ductal adenocarcinoma (PDA) has among the highest stromal fractions of any cancer and this has *complicated* attempts at expression-based molecular classification. The goal of this work is to profile purified samples of human PDA epithelium and stroma and examine their respective contributions to gene expression in bulk PDA samples.

Design We used laser capture microdissection (LCM) and RNA sequencing to profile the expression of 60 matched pairs of human PDA malignant epithelium and stroma samples. We then used these data to train a computational model that allowed us to infer tissue composition and generate virtual compartment-specific expression profiles from bulk gene expression cohorts.

Results Our analysis found significant variation in the tissue composition of pancreatic tumours from different public cohorts. Computational removal of stromal gene expression resulted in the reclassification of some tumours, reconciling functional differences between different cohorts. Furthermore, we established a novel classification signature from a total of 110 purified human PDA stroma samples, finding two groups that differ in the extracellular matrix-associated and immune-associated processes. Lastly, a systematic evaluation of cross-compartment subtypes spanning four patient cohorts indicated partial dependence between epithelial and stromal molecular subtypes.

Conclusion Our findings add clarity to the nature and number of molecular subtypes in PDA, expand our understanding of global transcriptional programmes in the stroma and harmonise the results of molecular subtyping efforts across independent cohorts.

Significance of this study

What is already known on this subject?

- Pancreatic ductal adenocarcinoma (PDA) is one of the most aggressive malignancies with currently no targetable genetic alterations. At the pathological level, it is a complex mixture of tumour cells, normal pancreatic tissues and stromal cell types, thus impeding the straightforward molecular characterisation of transcriptional profiles.
- Previous approaches to molecular subtyping have relied on bulk PDA samples leading to the proposal of anywhere between two to four distinct tumour classes. One study used indirect inference to identify two stromal subtypes associated with the activation state of pancreatic stellate cells.
- While a systematic evaluation of cross-compartment subtypes is lacking for PDA, current evidence suggests that epithelial and stromal programmes evolve independently.

What are the new findings?

- We used laser capture microdissection and RNA sequencing to directly sample pathologically verified PDA epithelia and their adjacent stroma for >60 patients.
- Tumour epithelia naturally separate into 'classical' and 'basal-like' subtypes while additional subtypes such as 'exocrine' or 'ADEX' are not supported.
- Unsupervised class detection among 110 stromal laser capture microdissection–RNA sequencing profiles detects two groups reflecting immune signalling and matricellular fibrosis, respectively.
- Systematic analysis of epithelial and stromal subtypes on nearly 400 PDA specimens found functional consistency across multiple cohorts.
- Across these same tumours, epithelial and stromal subtypes were partially linked, indicating potential dependence in the evolution of tissue compartments in PDA.

INTRODUCTION

All carcinomas harbour both transformed malignant cells and non-transformed stromal cells, in varying proportions.[1] Pancreatic ductal adenocarcinoma (PDA) is among the most stroma-rich cancers, with a complex inflammatory microenvironment that typically dominates the tumour parenchyma. Expected to be responsible for over 43 000 deaths per year in the USA, it is a common, aggressive malignancy that responds poorly to therapeutic intervention.[2 3] Within the stromal compartment of PDA, diverse fibroblast, myeloid, lymphoid, endothelial and other cell lineages contribute to both pro-tumour and anti-tumour processes, including angiogenesis and epithelial differentiation,[4] tissue stiffness,[5 6] drug delivery[7] and local immunosuppression.[8] These functions are orchestrated through

Significance of this study

How might it impact on clinical practice in the foreseeable future?

- The ability to robustly assess both epithelial and stromal subtypes for patients will facilitate the discovery of theranostic relationships between molecular composition and treatment studies, and could form the basis of future precision medicine approaches for PDA.

a host of paracrine signals that pass between and within the epithelial and stromal compartments communication that is quickly altered on tissue disruption. Thus, efforts to parse transcriptional programmes of PDA should take into account active processes in both compartments, ideally in an in situ context.

Despite extensive genomic characterisation,[9–13] individual DNA mutations have to date provided limited prognostic or theranostic information for PDA. Indeed, only a small fraction of pancreatic tumours is predicted to harbour 'druggable' genetic alterations.[11 13] As an alternative to genetic biomarkers, transcriptional classifiers for PDA have been explored using bulk tumour samples.[13–16] While these studies differ in the number of subtypes described, a shared message is that ductal pancreatic tumours include at least two groups distinguished by markers of epithelial differentiation state, with the more poorly differentiated subtype (ie, 'basal-like', 'squamous' or 'quasi-mesenchymal') exhibiting reduced overall survival relative to well-differentiated subtypes (ie, 'classical' or 'progenitor'). However, the contributions of stromal cells are handled differently in each instance, leading to some debate as to the merits of different proposed subtypes. To clarify this issue, we endeavoured to directly profile gene expression from purified neoplastic epithelium and associated stroma isolated from frozen human PDA samples.

Several techniques may be employed to isolate cellular subsets from bulk tissue including magnetic separation, fluorescence-assisted cell sorting (FACS) and laser capture microdissection (LCM). The first two techniques rely on population-specific antibodies to isolate specific cell types, but require disruption of the tumour using prolonged enzymatic digestion, during which time transcriptional profiles are invariably altered. Moreover, PDA diffusely infiltrates the surrounding pancreatic parenchyma[17] so that even tumour samples enriched by FACS for epithelial cell markers can include contributions from normal, atrophic, preneoplastic or metaplastic epithelial cells. LCM provides a powerful solution, allowing the isolation of pathologically verified compartment-specific tissue samples based on morphological features, without disrupting the delicate interplay of intercellular communication.

We present here expression profiles of laser capture microdissected malignant epithelium and matched reactive stroma for 60 human PDAs, providing both the opportunity to study each compartment in isolation and to examine their interplay across samples. Furthermore, we provide a novel stromal classification signature derived from the direct analysis of a total of 110 experimentally purified stromal profiles, yielding two prominent subtypes. In contrast with a prior signature derived indirectly using blind-source separation techniques,[15] this direct signature highlights the contribution of immune signalling pathways in one subtype (immune-rich) versus extracellular matrix-associated pathways (ECM-rich) in the other.

Using the compartment-specific profiles, we used a machine learning technique called ADVOCATE[18] to model the compartment specificity of every gene expressed in PDA. Applying this information to new bulk PDA expression profiles, we can then infer the epithelial and stromal fractions of that tumour and, critically, generate a pair of virtual compartment-specific gene expression profiles for each bulk tumour, which may then be used by a variety of downstream analytical pipelines. Using this approach, we examined the composition of multiple public PDA expression datasets and inferred both epithelial and stromal molecular subtypes from over 350 human pancreatic tumours. Critically, we found that consideration of compartment-specific molecular subtypes led to harmonisation of results across datasets and the validation of functionally similar subtypes that span human pancreatic cancer.

RESULTS

Transcriptional profiling of isolated pancreatic cancer epithelium and stroma

To study the separate transcriptional programmes of intact pancreatic tumour epithelium and stroma, we optimised a robust protocol for maintaining RNA integrity during laser capture microdissection of frozen tumour tissues, yielding total RNA suitable for library preparation and RNA sequencing. We first applied this LCM–RNA-Seq technique to 60 primary PDA specimens that were harvested and frozen intraoperatively by the Columbia University Tumor Bank in collaboration with the Pancreas Center at Columbia University/New York Presbyterian Hospital (see the online supplementary tables S1 and S2 for patient characteristics). For each tumour, we generated paired gene expression profiles from the malignant epithelium and nearby reactive stroma, as distinguished by cell morphology (figure 1A). Extensive quality control metrics confirmed the high quality of resulting RNA libraries (figure 1B,C and online supplementary figure S1A–D).[19 20] Critically, samples from the two compartments separated spontaneously along the first component of a Principal Component Analysis (PCA) with virtually no overlap (figure 1D) and were distinguished by expression of established marker genes for epithelial cells (KRT19, EPCAM, CDH1) versus markers of various stromal cell types, including leucocytes (PTPRC, CD4, CD163), endothelial cells (VWF, ENG, CDH5) and cancer-associated fibroblasts (CAFs) (ACTA2, DCN, FAP) (figure 1E). We observed that technical variance was substantially lower than biological variance (online supplementary figure S1E,F) and found that different malignant areas captured from a single tumour clustered closely, suggesting that the intratumoural transcriptional heterogeneity of that tumour was less than the intertumoural heterogeneity of PDA (online supplementary figure S1G,H).

We next validated the paired LCM–RNA-Seq profiles by assessing the immunohistochemical staining pattern of proteins that were predicted to be highly compartment-specific at the RNA level (online supplementary table S3), making use of data from The Human Protein Atlas (HPA) pathology database.[21] We restricted our analysis to proteins for which the highest-quality antibodies were available (n=321), based on established HPA criteria (online supplementary table S4). Of these, we evaluated the immunostaining patterns for the 50 genes whose LCM–RNA-Seq expression was most differentially expressed for each compartment (online supplementary table S5), examining a minimum of six PDA samples per tested protein. This analysis yielded confirmatory staining patterns for 47 of 50 epithelial proteins and 36 of the 50 stromal proteins (figure 1F, G). For example, figure 1H, I shows two members of the galectin protein family, LGALS4 and LGALS1, with inverse staining patterns in

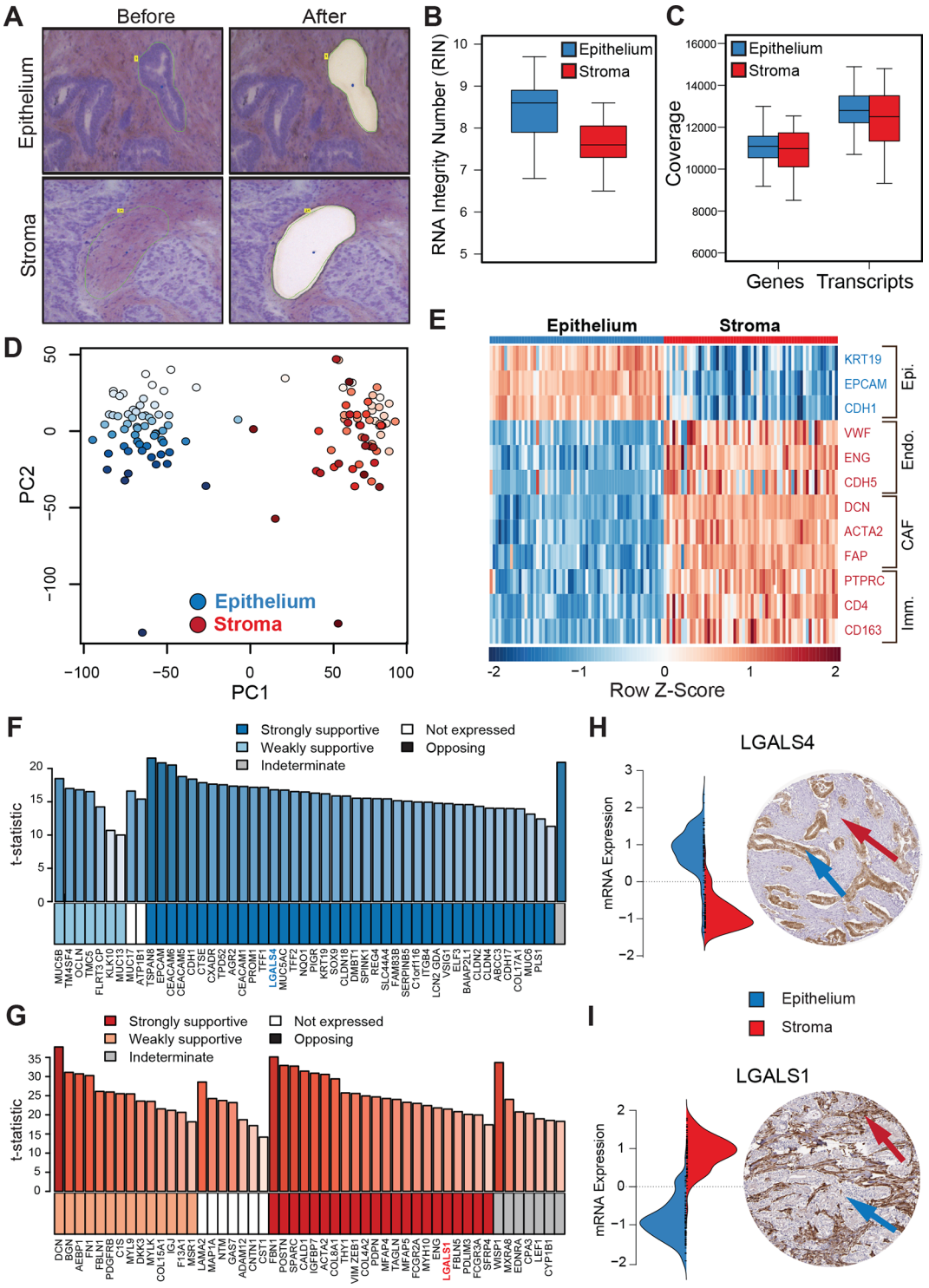

Figure 1 Compartment-specific gene expression profiling of pancreatic tumours. (A) Images of Cresyl Violet stained human PDA frozen sections before and after laser capture microdissection of malignant epithelial and adjacent stromal cells. (B) RIN values for RNA samples derived from the indicated compartment (n=60 each). (C) Number of genes and transcripts detected at >1 FPKM in the samples from (B). (D) Principal component analysis of the 60 paired epithelial and stromal LCM expression profiles from (C). Colour graduation shows pairing of samples from the same tumour. Three samples discussed later are labelled. (E) Heatmap showing the expression of marker genes for Epi. cells, Endo. cells, CAF and Imm. (F, G) Protein validation of genes predicted as epithelium-specific (F) or stroma-specific (G) based on mRNA expression. Bar height and colour shading reflect the certainty (t-statistic) of differential expression. The box colour below each bar summarises results of IHC on PDA sections from the HPA. IHC staining pattern was categorised as strongly or weakly supportive of the predicted compartment (blue/red), indeterminate (grey), absent (white) or opposite the predicted pattern (black). (H) An example epithelium-specific gene, LGALS4, showed a protein staining pattern that was strongly consistent with its mRNA expression (at left). Blue and red arrows indicate PDA epithelium and stroma, respectively. (I) LGALS1 exhibited a highly stroma-specific expression pattern. CAF, cancer-associated fibroblast; Epi, epithelial cells; Endo, endothelial cells; HPA, Human Protein Atlas; IHC, immunohistochemistry; Imm, immunocytes; LCM, laser capture microdissection; PDA, pancreatic ductal adenocarcinoma.

the two compartments, consistent with our predictions. Critically, none of the proteins was found to be expressed in a pattern opposite that predicted; rather, genes lacking supportive IHC staining were simply not detected, perhaps due to post-translational regulation. Thus, through the use of LCM–RNA-Seq, we compiled a comprehensive repertoire of compartment-specific genes serving as a novel, tumour-specific resource for the pancreatic cancer field.

Compartment fraction analysis reveals distinct compositions of public PDA datasets

Multiple large-scale gene expression datasets for PDA have been reported,[13–16] each providing important contributions to our understanding of the disease. However, cross-comparative analysis of these datasets has been challenging, due to differences in expression profiling platforms, inclusion criteria, sample preparation and other technical details. As a result, a consistent interpretation of the gene expression profile clusters emerging from these studies is still elusive, especially as it relates to stromal subtypes.

We reasoned that there are three potential sources of relevant heterogeneity in these data: (1) differences in the epithelium/stroma ratio in areas of frank carcinoma; (2) variation in the representation of uncharacterised tissue (eg, normal pancreas ductal epithelium, pancreatitis, lymph nodes and so on) in the bulk sample; and (3) technical differences (eg, expression platform and library preparation method). To manage these issues, we made use of a machine learning algorithm called ADVOCATE[18] to model the epithelial and stromal expression of every gene based on the 60 matched epithelial and stromal PDA LCM–RNA-Seq profiles above. After training, ADVOCATE can perform two functions on new bulk PDA expression profiles: (1) infer the fractions of epithelial and stromal tissues that make up the bulk sample; and (2) generate a pair of complete virtual compartment-specific expression profiles for each bulk tumour. Extensive validation of these functions using in silico analyses, mixing approaches, paired LCM/bulk profiles and histopathological evaluations are presented in a preprint on Bioarxiv.[22] We used ADVOCATE to perform a systematic analysis of over 350 published PDA expression profiles.

We began by examining the compartment fractions from the gene expression profiles of three independent cohorts: (a) University of North Carolina Chapel Hill (UNC, n=125), (b) the International Cancer Genome Consortium (ICGC, PACA-AU RNA-Seq dataset, n=93), and (c) The Cancer Genome Atlas (TCGA, PAAD dataset, n=137) (see Methods for inclusion criteria). The compartment fractions of these cohorts had not previously been directly compared using a single, common analysis method, perhaps due to differences in expression platforms (array vs RNA sequencing), or their available metadata. Using ADVOCATE, we found that the epithelial and stromal fractions varied significantly between the cohorts with 46%, 67% and 55% epithelium for the ICGC, UNC and TCGA cohorts, respectively ($p<0.001$, one-way analysis of variance) (figure 2A). These results align with 'tumour purity' analyses performed on the TCGA and ICGC cohorts using DNA-based techniques.[13 16] Our findings highlight critical differences in tissue composition between tumour collections curated with different inclusion criteria or enrichment practices.

Prior classification efforts built from individual cohorts have yielded divergent gene signatures that stratify pancreatic cancer into various subgroups, leading to ongoing debate as to their relative merits. We realised that our compartment-specific expression data could provide some context as to the nature of the genes that comprise each classification signature. Therefore, we extracted the list of signature genes overexpressed in each of the 11 proposed subtypes in the Collisson, Moffitt and Bailey classification schemes (online supplementary table S6–10). We then examined the overall expression level and compartment specificity of these genes in our LCM–RNA-Seq dataset (figure 2B).[13–15] We noted that the Bailey classifier was developed using Ensembl gene annotation and that the Bailey-immunogenic subtype includes numerous recombined immunoglobulin genes that are not designated in the NCBI annotation used for the CUMC dataset (online supplementary table S10). Therefore, to assess the compartment specificity of Bailey classifier genes, we used a version of the CUMC dataset that was remapped using the Ensembl GRCh37 gene annotation.

Examining each of the proposed subtype groups in turn, we noted that the genes used to define the Collisson-classical, Moffitt-classical, Moffitt-basal-like and Bailey-progenitor subtypes were all heavily weighted towards epithelial expression, suggesting that regardless of the amount of stromal infiltration, these genes are predominantly providing information about the malignant compartment. Conversely, those used to define the Moffitt-activated, Moffitt-normal and Bailey-immunogenic subtypes were weighted towards stromal expression, suggesting that these subtypes report on information that is largely independent of the malignant compartment. The Collisson-quasi-mesenchymal and Bailey-squamous gene sets were both well-expressed and represented a mixture of epithelial and stromal identity, consistent with a more poorly differentiated state. Finally, the majority of genes that define the Collisson-exocrine and Bailey-ADEX subtypes exhibited very low expression in the LCM–RNA-Seq datasets, suggesting that their expression in bulk tissue is derived from cell types that are largely absent from our microdissected samples. Together, these data provide insight into the cellular compartments that contribute to previous molecular gene signatures built from bulk tumour tissue samples.

Transcriptional deconvolution improves functional classification across cohorts

An important feature of a robust classification system is its capacity to identify sample subsets that are functionally similar across independent datasets. Given the uncertainty in the current literature regarding the actual number of these subsets in PDA, we first carried out an unsupervised analysis of CUMC epithelial LCM profiles using multiple independent approaches, all of which favoured a two-cluster solution (online supplementary figures S2A–D). Functional annotation of these groups was in agreement with that of the basal-like and classical groups (online supplementary figure S2E, supplementary tables S11 and S12) described by Moffitt *et al* and the respective UNC classifier genes were significantly enriched towards their counterpart among CUMC samples (online supplementary figure S2F). This functional alignment with our LCM data together with a superior compartment specificity (see figure 2B) led us to prioritise the UNC tumour classifier lists for molecular subtyping of the epithelium across PDA cohorts.

We next examined the relationship between compartment fraction and inferred epithelial subtype in each cohort (online supplementary table S13). In the ICGC and UNC cohorts, basal-like and classical tumours were inferred to have similar epithelial fractions. By contrast, in the TCGA cohort, basal-like tumours were inferred to have a significantly higher epithelial fraction (online supplementary figure S2G). This suggested the possibility that subtype calls were confounded by tumour composition. We

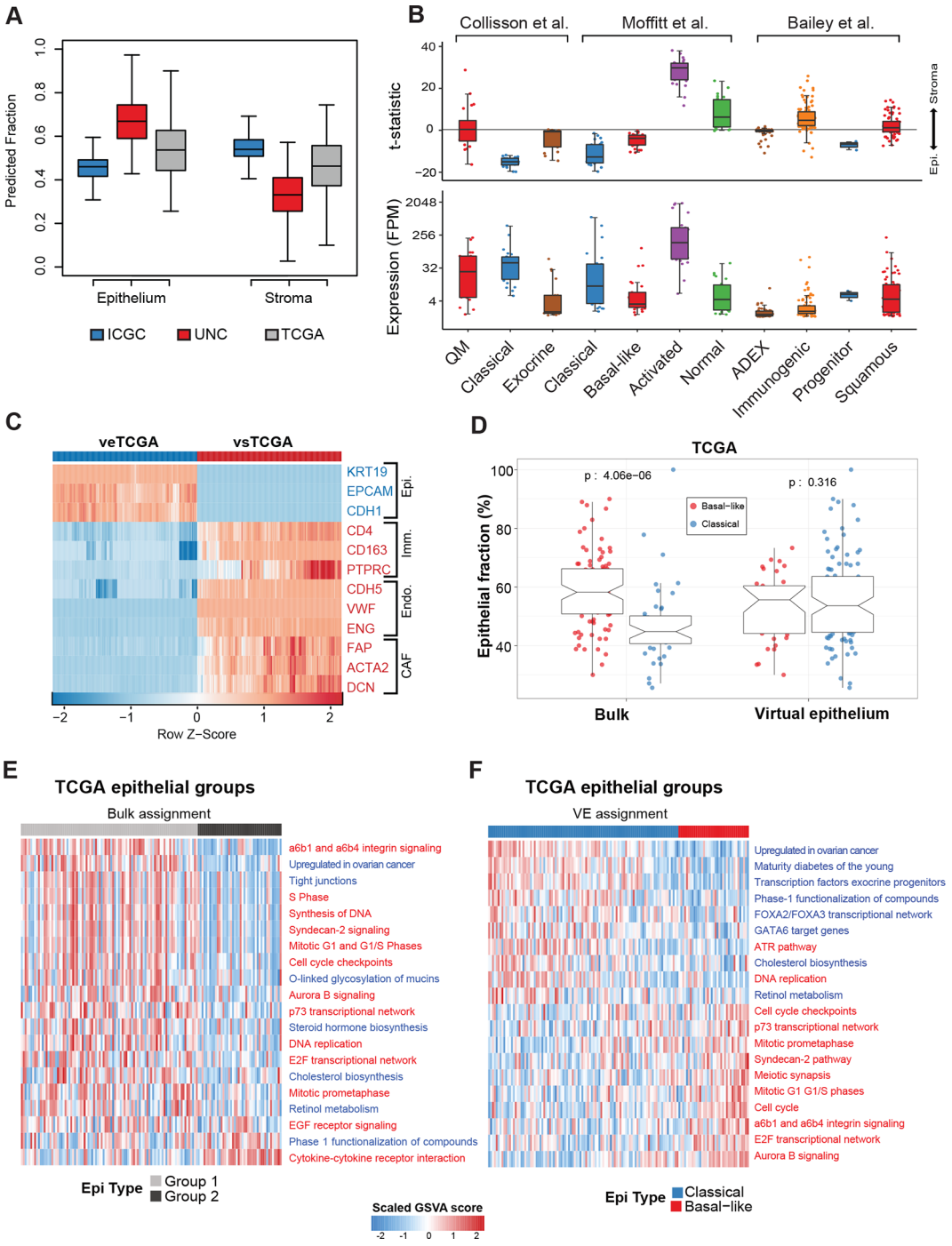

Figure 2 Analysis and classification of pancreatic tumour cohorts and classifiers. (A) Tumour and stroma content analysis of pancreatic tumours from the ICGC (blue), UNC (red) and TCGA (grey) cohorts. (B) Analysis of gene expression across 60 pairs of PDA epithelium and stroma LCM–RNA-Seq profiles, highlighting the genes used to determine each subtype from the Collisson, Moffitt and Bailey classifiers. Top panel displays the compartment specificity of the signature genes for each subtype based on the t-statistic of their differential expression between PDA epithelium and stroma samples; positive values indicate stromal enrichment. Lower panel depicts the average expression of each signature genes across all LCM–RNA-Seq samples, in FPM mapped fragments. (C) Heatmap depicting the differential expression of indicated marker genes in deconvolved veTCGA and vsTCGA profiles from the TCGA cohort. (D) Epithelial fraction of TCGA pancreatic tumours allocated to the basal-like (red) and classical (blue) subtypes based on analysis of either bulk or virtual epithelial expression profiles. (E, F) Analysis of gene sets associated with the Moffitt basal-like (red) and classical (blue) subtypes based in bulk expression profiles (E) from TCGA versus virtual epithelial profiles (F) of the same tumours. Heatmap depicts GSVA scores per sample for indicated gene sets. Stratification of bulk TCGA profiles using the Moffitt classifier results in groups that are not differentially enriched in the gene sets classically associated with basal-like versus classical subtypes. However, following deconvolution, the virtual epithelial profiles stratify into two groups that reflect the functional biology of the basal-like and classical subtypes. FPM, fragments per million; GSVA, gene set variance analysis; LCM, laser capture microdissection; PDA, pancreatic ductal adenocarcinoma; TCGA, The Cancer Genome Atlas cohort; UNC, University of North Carolina cohort; ve, virtual epithelial; vs, virtual stromal.

reasoned that removing the non-epithelial signals from bulk expression profiles might lead to more consistent molecular classification across cohorts with varied tissue composition. We, therefore, used ADVOCATE to generate virtual epithelial and stromal expression profiles from the bulk samples of each PDA cohort (producing new datasets: vUNC, vTCGA and vICGC). In each case, virtual profiles displayed a clear expression of established cell-specific marker genes (figure 2C and online supplementary figure S3A,B). Notably, bulk samples were distributed between the corresponding virtual epithelium and virtual stroma samples by hierarchical clustering (online supplementary figure S3C-E). Strikingly, subtype calls made from deconvolved TCGA expression data yielded two groups whose distributions of epithelial fractions were now balanced (figure 2D). Moreover, we found the impact of deconvolution to be most apparent by functional analysis (online supplementary figure S3F,G; online supplementary tables S14–S16). Prior to deconvolution, analysis of the TCGA bulk samples classified by the UNC epithelial signature show substantial mixing of gene sets that are otherwise associated with basal-like (red) or classical (blue) tumours (figure 2E), whereas after deconvolution, these groups closely aligned (figure 2F). The variable stromal composition of the bulk TCGA dataset was thus interfering with the ability of the UNC epithelial signature to identify functionally meaningful groups of tumours.

We also noted that application of the Moffitt-E classifier to the veICGC dataset revealed excellent alignment with the pancreatic progenitor and squamous subtypes described by Bailey *et al*.[13] (SMC=0.91) (online supplementary table S13). Together, these data indicate that removal of stromal expression data from bulk tumour datasets results in the reclassification of many bulk tumour samples, particularly from the TCGA cohort, and this can improve the functional similarity of groups identified by different classification systems.

Identification of immune-rich and ECM-rich subtypes of PDA stroma

Prior work on the classification of pancreatic tumour stroma used indirect inference from bulk tissue profiles and focused primarily on the biology of quiescent or activated fibroblasts. In order to capture the contributions of all of the dozens of distinct cell types present in PDA stroma, we expanded the stromal LCM–RNA-Seq cohort described above to include samples from a total of 110 unique patients. Non-negative matrix factorisation with consensus clustering identified two prominent molecular subtypes among these samples (see the online supplementary figure S4 for additional details). Clear functional identities were established for these subtypes using gene set variance analysis (GSVA), leading to their designations as an 'immune-rich' group characterised by numerous immune and interleukin signals; and an 'ECM-rich' group, characterised by numerous extracellular matrix-associated pathways (figure 3A, online supplementary tables S17 and S18). We next extracted a gene signature distinguishing these two stromal subtypes, making use of the compartment specificity analysis described above to filter for stroma-specific genes (online supplementary tables S19–S21, see the online supplementary methods). Application of this signature to the virtual stroma profiles yielded two prominent clusters, as reflected across the UNC, ICGC and TCGA cohorts (figure 3B–D, online supplementary tables S22–S24). Critically, in each cohort, the two clusters were again characterised by their enrichment for gene sets associated with ECM deposition or immune processes, indicating a robust and consistent performance of this new, stroma-specific 'CUMC-S' signature.

Epithelial and stromal subtypes are partially linked and associated with survival differences

Having determined the epithelial and stromal subtypes of all CUMC, UNC, ICGC and TCGA samples, a comprehensive analysis revealed substantial variation in subtype composition across the four datasets. Within the epithelium, the basal-like group comprised 29%, 41% and 27% of cases in the veUNC, veICGC and veTCGA cohorts, respectively (figure 3E–G) and 36% of our epithelial LCM–RNA-Seq profiles (online supplementary figure S5A). Within the stroma, the ECM-rich subtype comprised 62%, 52% and 31% of cases in the vsUNC, vsICGC and vsTCGA cohorts, respectively (figure 3E–G), and 47% of our stromal LCM–RNA-Seq samples (online supplementary Figure S5A). These observations serve to further highlight the significant heterogeneity between independent collections of pancreatic tumour specimens.

We next assessed the associations of epithelial and stromal subtypes with survival outcomes. Examining the epithelial subtypes, we found that removing stromal gene expression with ADVOCATE increased the survival association between classical and basal-like tumours in all three bulk datasets, with a particularly strong effect on TCGA outcomes (figure 4A–C) where 45% of the samples were re-classified after deconvolution. For the stromal subtypes, we observed at least a trend towards reduced survival among ECM-rich tumours in all three datasets (a finding made more apparent by deconvolution); however, this only reached significance in the ICGC cohort (figure 4D–F). Together, these data indicate that (i) variations in tumour composition between different large-scale gene expression datasets can affect the predictive power of established classifier signatures for PDA, and (ii) transcriptional deconvolution can help overcome this hurdle, improving the reproducibility of outcome prediction.

The existence of numerous paracrine signalling pathways whose activity is affected by oncogenic mutations implies that stromal transcriptional programmes should be influenced by epithelial identity.[23] We examined this corollary by ascertaining the association of epithelial and stromal subtypes in our experimental LCM dataset as well as in those from the virtual UNC, ICGC and TCGA datasets. We found that in the ICGC and TCGA cohorts, the ECM-rich stroma subtype was preferentially associated with the basal-like epithelial subtype; the UNC and CUMC cohorts trended in this direction but did not reach significance. However, a meta-analysis of the 393 samples from all four datasets yielded an OR of 2.7 for the association of basal-like epithelium and ECM-rich stroma (online supplementary figure S5B, random effects model: OR 2.7 (1.33–5.53), $p<0.001$), indicating a partial association between epithelial and stromal compartments.

The imperfect alignment of the epithelial and stromal subtypes offered the possibility that combination subtypes might vary in their survival associations as compared with either compartment alone (figure 4G–I, online supplementary figure S5C–H). Indeed, consideration of combined epithelial and stromal combination subtypes affected the outcome prediction, particularly in the case of the UNC cohort where combination subtyping of deconvolved samples found a particularly poor outcome for basal-like/ECM-rich tumours relative to classical/immune-rich tumours (HR=3.76 for combined subtyping vs 2.11 for epithelial subtyping alone, figure 4I vs. 4C). Together, these data highlight the relationship between basal-like epithelium with

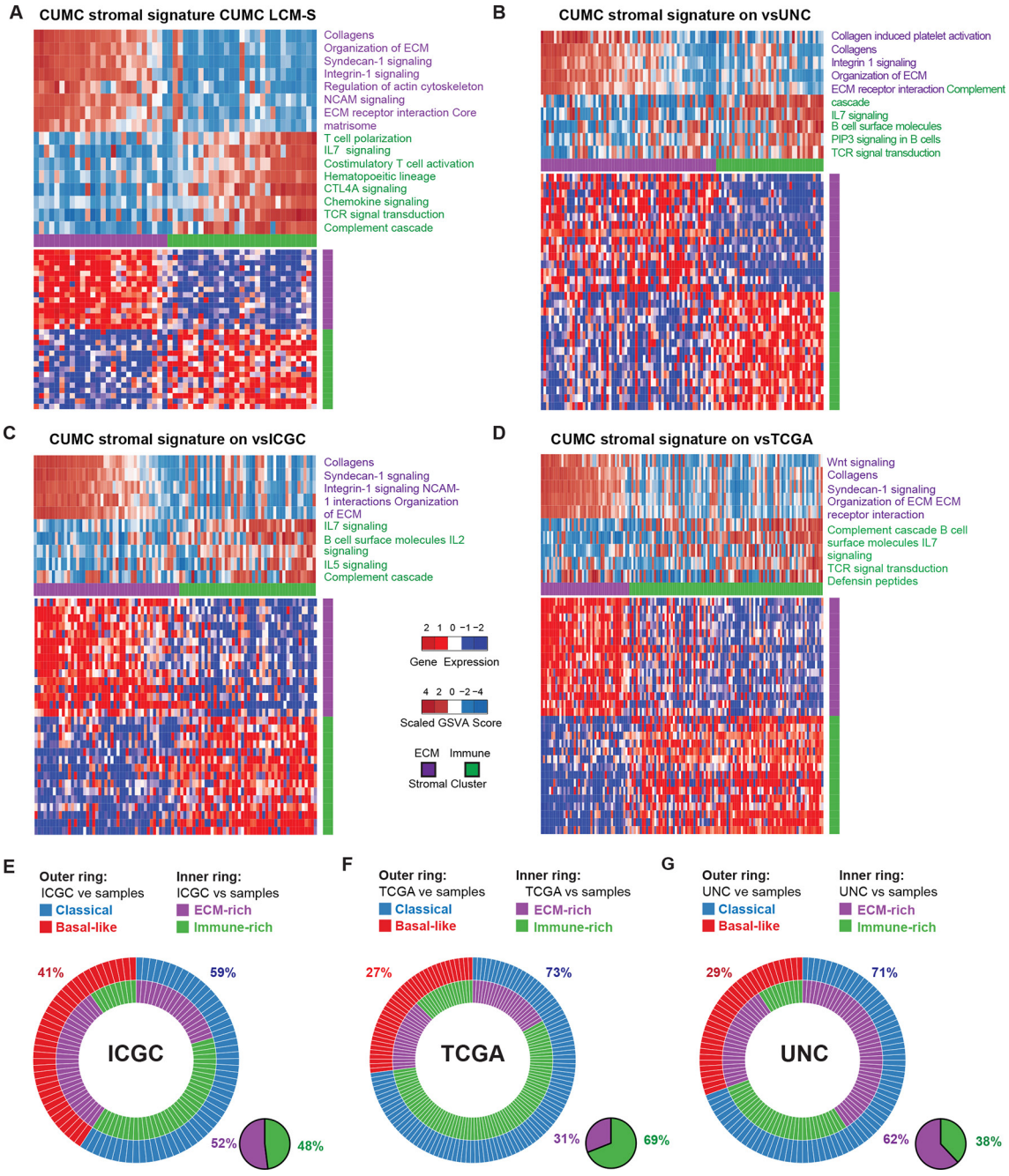

Figure 3 Systematic stromal subtyping of PDA. (A–D) Heatmaps of the top 30 DEG between groups obtained by clustering stromal LCM–RNA-Seq samples from CUMC tumours (A), and virtual stromal (vs) profiles from the UNC (B), ICGC (C) and TCGA (D) cohorts, respectively. Clustering was based on the expression of a signature derived from stromal LCM profiles from 110 individual patients (CUMC-S classifier, see the online supplementary methods). Top section of heat-map depicts GSVA scores per sample for indicated gene sets. In each virtual stroma dataset, two groups were identified, one with features indicating elevated extracellular matrix deposition and remodelling ('ECM-rich', purple) and another enriched in various immune and interleukin pathways ('immune-rich', green). (E–G) Multilayered donut plots showing (**i**) the alignment of epithelial with stromal subtypes for each tumour in each cohort and (ii) the proportion of each epithelial subtype. Separate pie charts summarise the proportion of stromal subtypes per cohort. DEG, differentially expressed genes; GSVA, gene set variance analysis; ICGC, International Cancer Genome Consortium cohort; PDA, pancreatic ductal adenocarcinoma; LCM, laser capture microdissection; TCGA, The Cancer Genome Atlas cohort; UNC, University of North Carolina cohort.

ECM-rich stroma in pancreatic cancer and the strong association of this combination with poor overall survival.

DISCUSSION

The traditional understanding of genetic mutations as drivers of tumour development has led to a focus on the malignant compartment that is exemplified by the term 'tumour purity', which regards the stroma as mere contamination. However, with the understanding that stromal cells play critical roles in both promoting and restraining pancreatic tumour progression,[24] the consensus view of the stromal compartment has shifted to that of a critical partner, or foil, to the malignant epithelium. Indeed,

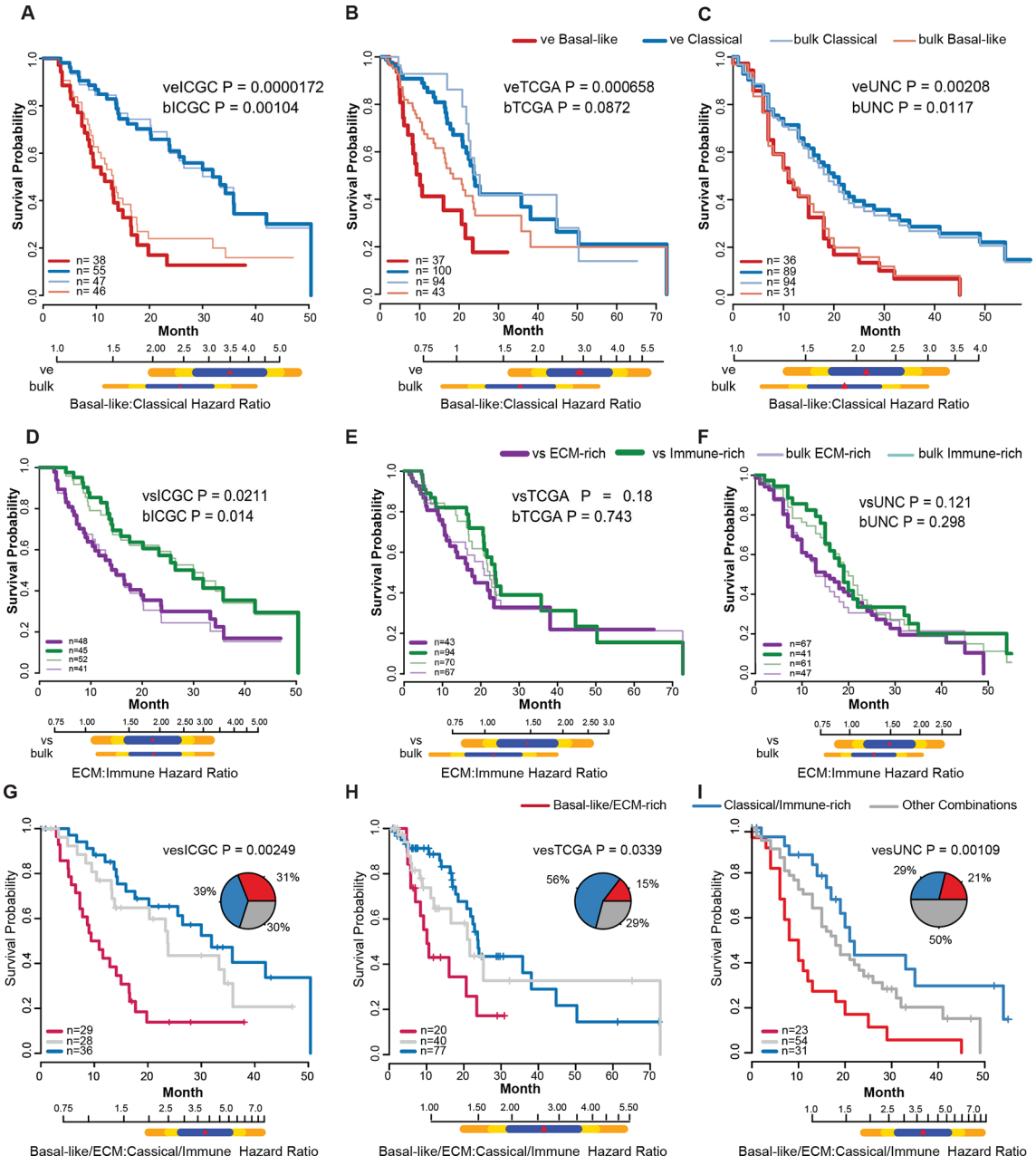

Figure 4 Combined epithelial and stromal subtypes associate with overall survival. KM survival analysis of patients with resected PDA from the ICGC (n=93), TCGA (n=137) or UNC (n=125) cohorts, stratified by the indicated signatures applied to either bulk expression profiles (thin lines) or transcriptionally deconvolved versions of the same (thick lines). Below each KM plot, horizontal bars indicate the HRs from a CPHM, along with their 80% (blue), 90% (yellow) and 95% (orange) confidence intervals. (A–C) KM plot of patients from the indicated cohorts using the Moffitt-E signature to stratify basal-like (red) versus classical (blue) tumours, showing that the detection of a differential prognosis among the epithelial subtypes is generally enhanced by transcriptional deconvolution. Bars indicate HR for basal-like tumours in virtual epithelial and bulk profiles. (D–F) KM plot of patients from the indicated cohorts using the CUMC-S signature to stratify ECM-rich (purple) versus immune-rich (green) tumour. KM survival analysis depicts overall survival relative to stromal subtype. Stromal subtypes are statistically associated with outcome in the ICGC cohort with ECM-rich tumours having a worse prognosis. Bars indicate HR for ECM-rich tumours in virtual stromal and bulk profiles. (G–I) KM plot of patients from the indicated cohorts using a combination of the Moffitt-E and CUMC-S signatures. Red lines indicate basal-like tumours with an ECM-rich stroma while blue lines indicate classical tumours with an immune-rich stroma; all other tumours are represented as a grey line. Bars indicate HR for basal-like::ECM-rich tumours in bulk and virtual epithelial/stroma (ves) profiles. CPHM, Cox proportional hazards model; ECM, extracellular matrix; KM, Kaplan-Meier; ICGC, International Cancer Genome Consortium cohort; PDA, pancreatic ductal adenocarcinoma; TCGA, The Cancer Genome Atlas cohort; UNC, University of North Carolina cohort. .

in some contexts, the stroma can even play a dominant role, as epitomised by the success of stroma-targeted immunotherapy in treating aggressive cancers such as metastatic melanoma and non-small cell lung cancer. In this light, we sought to study the interplay of PDA epithelium and stroma in their native state, separated by LCM from otherwise intact samples but matched by patient so that the reciprocal signals active in each compartment might be examined.

A key outcome of this work is to unify our understanding of molecular subtypes in pancreatic ductal adenocarcinoma. To do this, we first examined the properties of subtypes resulting from existing classification schemes. We noted that among >60 individual epithelial tumour profiles, there was little evidence for the existence of the Collisson-exocrine or Bailey-ADEX subtype, as evidenced by the general lack of expression of marker genes associated with these subtypes. Conversely, signature genes for the Bailey-immunogenic subtype are generally well expressed, but predominantly in stromal samples, suggesting that this subtype, which was presented as being mutually exclusive with the epithelial squamous and progenitor subtypes, in fact, arises from the stromal compartment.

Given the fact that none of the classification signatures was perfectly epithelium-specific, we suspected that varying levels of stromal tissue content might impact the assignment of tumours to different molecular subtypes. Indeed, removal of stromal expression signals from bulk expression data resulted in the reclassification of nearly half the TCGA samples using the Moffitt-E signature and improved detection of the functional processes associated with the classical and basal-like subtypes in each cohort. Classification efforts may thus benefit from virtual purification of gene expression prior to supervised clustering.

In our effort to establish a novel classification system for PDA stroma, we placed the greatest emphasis on the reproducibility of molecular phenotypes across multiple cohorts. Following this process, we observed with great interest the emergence of two prominent molecular subtypes in the stroma with pronounced enrichment for two different aspects of stromal biology: ECM deposition and remodelling versus immune-related processes. This concept refines the idea of 'activated' and 'normal' stromal subtypes, which was derived largely from the biology of pancreatic stellate cells[15] and thus did not take into account the substantial contributions of immune cells to the PDA microenvironment. We also note that although the Bailey-immunogenic tumours in the ICGC cohort are generally identified as immune-rich by our analysis, there are important distinctions between these classification schemes. Specifically, the Bailey-immunogenic subtype is one of four mutually exclusive classes and picks up classical/progenitor tumours with a high abundance of immune infiltration. This structure precludes both the possibility of tumours having a low abundance of stroma but for which the stroma has an immunogenic quality and tumours with high stromal abundance lacking an immunogenic character.

By examining all four tumour cohorts, we found a strong association between an ECM-rich stroma and basal-like epithelium while immune-rich stroma occurred more often in association with classical epithelia. The latter finding corroborates the concept that epithelial traits promoting dedifferentiation in PDA, such as the loss of SMAD4 expression, may, in fact, shape a more matricellular stromal phenotype.[23] Interestingly, a recent study[25] in patient-derived xenografts showed that basal-like and classical tumour cells, respectively, implanted subcutaneously into mice almost unequivocally induced microenvironments dominated by fibrosis (ie, ECM-rich) and immune infiltration (ie, immune-rich), respectively. We also found that cross-compartment subtypes are associated with differences in outcome, with basal-like/ECM-rich tumours having a substantially worse overall survival when compared with classical/immune-rich tumours (overall HR=3.76, 3.81, and 2.63 for UNC, ICGC and TCGA, respectively). Although a direct comparison is not possible, this effect size is in the same general range as other known single variables in pancreatic cancer biology, including lymph node status (HR=1.5), postoperative CA19-9 level (HR=3.6) or the number of high penetrance driver genes (HR=1.4).[26 27] Unfortunately, differences in the clinicopathological data reported for each cohort precluded a more sophisticated multivariate model. Nonetheless, we expect that this approach to subtyping will have immediate applications, for example, in interpreting the results of small-scale clinical trials where random inequalities of molecular subtypes could dramatically affect the expected survival between groups or relative to historical controls.

METHODS

The information provided here is a succinct summary of the experimental procedures. Detailed information is provided in the supplementary information.

Samples studied

Information is provided from a total of 122 patients with PDA who underwent surgery at the Columbia Pancreas Center. From these, an implementation of the ADVOCATE algorithm[18] was trained on 60 pairs of epithelial and stromal samples matched by the patient. Additional samples were used in unsupervised clustering analyses, as detailed in the online supplementary table S25. Patients provided surgical informed consent which was approved by a local ethics committee (IRB # AAAB2667). Samples were frozen intraoperatively by the Columbia University Tumor Bank. Clinical and pathological information on the 122 cases is provided in the online supplementary tables S1 and S2.

Laser capture microdissection and RNA sequencing

Cryosections of OCT-embedded tissue blocks were transferred to PEN membrane glass slides and stained with cresyl violet acetate. Adjacent sections were H&E stained for pathology review. Laser capture microdissection was performed on a PALM MicroBeam microscope (Zeiss), collecting at least 1000 cells per compartment. RNA was extracted and libraries prepared using the Ovation RNA-Seq System V2 kit (NuGEN). Libraries were sequenced to a depth of 30 million, 100 bp, single-end reads.

Computational modelling

This manuscript makes use of a novel computational model called ADVOCATE. A description of this approach is being prepared for submission in a separate manuscript. However, we have appended a 'Conceptual Approach' document describing the mathematical basis of this method of ADCOATE for the benefit of reviewers of this manuscript. The ADVOCATE software is publically available on Github[18] and a manuscript describing its development is in preparation.

Author affiliations
[1]Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York, New York, USA
[2]Department of Medicine, Division of Digestiveand Liver Diseases, Columbia University Medical Center, New York, New York, USA
[3]Department of Pathology and Cell Biology, Columbia University Medical Center, New York, New York, USA
[4]Department of Biomedical Informatics, Columbia Unicersity Medical Center, New York, New York, USA
[5]Department Systems Biology, Columbia University Medical Center, New York, New York, USA
[6]Department of Surgery, Division of GI/EndocrineSurgery, Columbia University Medical Center, New York, New York, USA
[7]Department of Medicine, Division of Hematology and Oncology, New York University Langone Medical Center, New York, New York, USA
[8]Department of Epidemiology, Mailman School of Public Health, New York, New York, USA

[9]Department of Computer Science and Engineering, University of California, San Diego, California, USA
[10]PsychoGenics Inc, Paramus, New Jersey, USA

Correction notice This article has been corrected since it published Online First. The funding statement has been corrected.

Acknowledgements The authors would like to thank Richard Moffitt for valuable critique of the manuscript.

Contributors KPO, AC and MB: conceptualisation. HCM and JH: computational analysis. JH, PL, JZ and MB: software. HCM, SRH, JH, JG, ACI, ARS, PEO and KPO: investigation. JAC, HH, AA, TS, KPO and AC: resources. HCM, JH and KPO: visualisation. HCM, KPO and AC: funding acquisition. KPO, AC and MB: project oversight and management. HCM and KPO: wrote the manuscript with feedback from SRH, JH and AC. All authors discussed the results and commented on the manuscript.

Funding This work was supported by the National Cancer Institute (NCI) Cancer Target Discovery and Development program (U01CA217858 to AC), NCI Research Centers for Cancer Systems Biology Consortium (1U54CA209997 to AC and KPO), NCI Outstanding Investigator Award (R35CA197745-02 to AC), NCI Cancer Center Support Grant (3 P30 CA13696-40) and NCI Research Project Grant (R01CA157980 to KPO). Financial support was also provided by the Columbia University Pancreas Center. HCM. received support from a Mildred Scheel Postdoctoral Fellowship (Deutsche Krebshilfe). PEO received support from the NIH NCATS (KL2TR001874).

Competing interests AC is a founder and shareholder of DarwinHealth Inc. and a member of the Tempus Inc. SAB and shareholder. Columbia University is a shareholder of DarwinHealth Inc. KPO is a member of the SAB for Elstar Therapeutics.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement All data from this study are included in the manuscript or publicly available.

REFERENCES

1 Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. *Nat Commun* 2015;6:8971.
2 Oberstein PE, Olive KP. Pancreatic cancer: why is it so hard to treat? *Therap Adv Gastroenterol* 2013;6:321–37.
3 Siegel RL, Miller KD, Jemal A, *et al*. Cancer Statistics, 2017. *CA Cancer J Clin* 2017;67:7–30.
4 Rhim AD, Oberstein PE, Thomas DH, *et al*. Stromal elements act to restrain, rather than support, pancreatic ductal adenocarcinoma. *Cancer Cell* 2014;25:735–47.
5 Provenzano PP, Cuevas C, Chang AE, *et al*. Enzymatic targeting of the stroma ablates physical barriers to treatment of pancreatic ductal adenocarcinoma. *Cancer Cell* 2012;21:418–29.
6 Jacobetz MA, Chan DS, Neesse A, *et al*. Hyaluronan impairs vascular function and drug delivery in a mouse model of pancreatic cancer. *Gut* 2013;62:112–20.
7 Olive KP, Jacobetz MA, Davidson CJ, *et al*. Inhibition of Hedgehog signaling enhances delivery of chemotherapy in a mouse model of pancreatic cancer. *Science* 2009;324:1457–61.
8 Vonderheide RH, Bayne LJ. Inflammatory networks and immune surveillance of pancreatic carcinoma. *Curr Opin Immunol* 2013;25:200–5.
9 Jones S, Zhang X, Parsons DW, *et al*. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* 2008;321:1801–6.
10 Biankin AV, Waddell N, Kassahn KS, *et al*. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature* 2012;491:399–405.
11 Witkiewicz AK, McMillan EA, Balaji U, *et al*. Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nat Commun* 2015;6:6744.
12 Waddell N, Pajic M, Patch AM, *et al*. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* 2015;518:495–501.
13 Bailey P, Chang DK, Nones K, *et al*. Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* 2016;531:47–52.
14 Collisson EA, Sadanandam A, Olson P, *et al*. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat Med* 2011;17:500–3.
15 Moffitt RA, Marayati R, Flate EL, *et al*. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat Genet* 2015;47:1168–78.
16 Cancer Genome Atlas Research Network.. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell* 2017;32:185–203.
17 Hruban RHP, M.B.; Klimstra DS. *Tumor of the pancreas*. Washington DC: American Registry of Pathology, 2007.
18 Laise PH, Bansal J;, M; Califano A. ADVOCATE. 2018 https://github.com/califano-lab/ADVOCATE
19 Adiconis X, Borges-Rivera D, Satija R, *et al*. Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat Methods* 2013;10:623–9.
20 Shanker S, Paulson A, Edenberg HJ, *et al*. Evaluation of commercially available RNA amplification kits for RNA sequencing using very low input amounts of total RNA. *J Biomol Tech* 2015;26:jbt.15-2601-001–18.
21 Pontén F, Jirström K, Uhlen M. The Human Protein Atlas-a tool for pathology. *J Pathol* 2008;216:387–93.
22 JM H, Maurer HC, Holmstrom SR, *et al*. Transcriptional deconvolution reveals consistent functional subtypes of pancreatic cancer epithelium and stroma. *Bioarxiv* 2018.
23 Laklai H, Miroshnikova YA, Pickup MW, *et al*. Genotype tunes pancreatic ductal adenocarcinoma tissue tension to induce matricellular fibrosis and tumor progression. *Nat Med* 2016;22:497–505.
24 Neesse A, Algül H, Tuveson DA, *et al*. Stromal biology and therapy in pancreatic cancer: a changing paradigm. *Gut* 2015;64:1476–84.
25 Nicolle R, Blum Y, Marisa L, *et al*. Pancreatic adenocarcinoma therapeutic targets revealed by tumor-stroma cross-talk analyses in patient-derived xenografts. *Cell Rep* 2017;21:2458–70.
26 Yachida S, White CM, Naito Y, *et al*. Clinical significance of the genetic landscape of pancreatic cancer and implications for identification of potential long-term survivors. *Clin Cancer Res* 2012;18:6339–47.
27 Berger AC, Garcia M, Hoffman JP, *et al*. Postresection CA 19-9 predicts overall survival in patients with pancreatic cancer treated with adjuvant chemoradiation: a prospective validation by RTOG 9704. *J Clin Oncol* 2008;26:5918–22.