Decoding multicellular niche formation in the tumour microenvironment from nonspatial single-cell expression data

Chengxin Yu[1, 2], Michael J Geuenich[1, 2], Sabrina Ge[3, 8], Gun-Ho Jang[6], Tan Tiak Ju[1, 2], Amy Zhang[6], Grainne M O'Kane[6, 8, 9], Faiyaz Notta[3, 6, 8], Kieran R Campbell[1, 2, 4, 5, 6, 7, *]

[1] Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON M5G 1X5, Canada

[2] Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada

[3] Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 1L7, Canada

[4] Department of Statistical Sciences, University of Toronto, Toronto, ON M5S 3G3, Canada

[5] Department of Computer Science, University of Toronto, Toronto, ON M5T 3A1, Canada

[6] Ontario Institute of Cancer Research, Toronto, ON M5G 1M1, Canada

[7] Vector Institute, Toronto, ON M5G 1M1, Canada

[8] Princess Margaret Cancer Centre, Toronto, ON M5G 2M9, Canada

[9] Department of Medical Oncology, St. Vincents University Hospital and University College,Dublin

[*] Correspondence: kierancampbell@lunenfeld.ca

Accurate inference of granular cell states that co-occur within the tumour microenvironment (TME) is central to defining pro- and anti-tumour environments. However, to reliably identify recurrent coexisting cell populations it is fundamental to analyze datasets encompassing a substantial number of tumour samples with a resolution sufficient to capture granular cell states. Here, we leverage eight scRNA-seq datasets of pancreatic ductal adenocarcinoma (PDAC) in a unique discovery-validation setup and find reproducible cell states, gene programs, and cellular niches that are predictive of specific clinical outcomes. Across tumours, we show highly consistent co-occurrence of cell states within and between lineages, including those reflecting known and *de novo* cellular interactions alongside the formation of multi-cellular clusters such as tertiary lymphoid structures. In addition, we develop a novel probabilistic model to quantify multi-cellular communities directly from atlas-scale scRNA-seq datasets. This model identified cellular niches predictive of clinical outcomes including communities associated with response to therapy and with specific KRAS mutations. Together, this work lays the foundation for inferring reproducible multicellular niches directly from large nonspatial scRNA-seq atlases and linking their presence in individual patients to prognosis and therapy response.

Introduction

Recent advances in expression profiling technologies have significantly accelerated our ability to quantify cellular phenotypes at single-cell and spatial resolutions. Single cell RNA-seq[1,2] (scRNA-seq) enables transcriptome-wide gene expression measurements at single cell resolution and has led to multiple discoveries of cell types and transcriptional states associated with patient outcomes and therapy response[3–8]. In tandem, spatial expression profiling technologies[9–12] that quantify gene or protein expression at (sub-) cellular resolution while retaining the spatial location of each measurement have led to multiple studies that associate tumour spatial architecture with patient outcomes[4,5,13,14]. Of note, significant attention has been given to using these technologies for the *de novo* identification of cell types and states that co-occur within tissues[15] and form multi-cellular communities (or niches)[16–18]. These niches form multi-cellular networks whose presence in tissues has been linked to clinical outcomes such as treatment response across a range of pathologies[13,19]. Importantly, much of our understanding of tissue dynamics leading to target discovery comes from reasoning about cellular co-occurrence within the tumour microenvironment. For example, cell states believed to be responsible for an immunosuppressive microenvironment will co-occur with exhausted cytotoxic T cells[4,20,21], thereby providing a first list of potential therapeutic targets directly from patient data to convert the TME into an immunogenic state and improve patient outcomes.

However, reliable quantification of cell state co-occurrence within individual tumours leading to cellular niches remains challenging. In particular, the ability to robustly identify recurrent coexisting cell states or gene programs in specific clinical scenarios relies on having sufficient independent samples, particularly for rare or transient cell states. Spatial protein technologies such as Imaging Mass Cytometry[9] and Multiplexed Ion Beam Imaging[22] are sufficiently cost-efficient to profile the required number of tissue samples, but can only quantify 40-60 markers simultaneously, limiting the ability to infer granular cell states and signaling pathways. On the other hand, spatial transcriptomic technologies such as Nanostring CosMx and 10X Genomics Xenium can measure 100s of mRNA species simultaneously, but are currently too expensive to profile at scale; therefore, despite the potential of spatial approaches, obtaining a sufficient number of samples to confidently identify true co-occurring cell states remains a significant limiting factor. Moreover, whole-transcriptome technologies like the 10X Genomics Visium platform do not quantify true single-cell measurements and often suffer contamination caused by spot swapping[23]. In contrast, scRNA-seq can quantify single-cell whole transcriptome expression measurements with quantified cell state abundance largely reflecting tissue composition[24–26]. While scRNA-seq nominally suffers from the same expense per sample as spatial transcriptomics technologies, the push by the community over the past decade through initiatives such as the Human Cell Atlas[27] and Human Tumour Analysis Network[28] has resulted in unprecedented data sizes, comprising more than 12 tumour atlases[4,29–34] by mid-2024 alone.

This leads to the tantalizing possibility of quantifying cell state co-occurrence and cellular niches directly from large multi-cohort scRNA-seq cancer atlases. However, whether this can lead to reproducible, clinically relevant insights remains largely unknown.

To help answer this question, we focus on pancreatic ductal adenocarcinoma (PDAC) – a malignant neoplasm of the pancreas – due to both the availability of multiple independent primary tumour scRNA-seq datasets studies[4,35–41], combined with its complex microenvironment that is characteristically difficult to target. For example, targets in the PDAC TME that achieved strong preclinical validation in mouse models[42] ultimately performed worse than standard of care when tested in human clinical trials[43], highlighting the need to understand immune dynamics directly in patient tissues.

Existing studies of the PDAC TME have highlighted diverse cellular heterogeneity across all lineages[44–48]. Additionally, single-cell profiling of PDAC tumours has highlighted previous underappreciated heterogeneity within individual tumours cells themselves, with existence of the basal/classical Moffitt subtypes[49] and later basal A/B, classical A/B, and hybrid subtypes[35] being shown to co-exist within single tumours[50,51]. Despite this, few efforts have sought to explicitly quantify the co-occurrence of granular cell states within single tumours. Oh et al. curated scRNA-seq PDAC from four studies to infer cell type co-occurrence using clustering and interactome analysis[50], but due to having relatively few samples did not include validation or highly granular cell states. In contrast, Luca et al. built a machine learning framework EcoTyper to discover multicellular communities across solid tumours from bulk RNA-seq data, which is well-powered at the patient level but may lack the ability to uncover granular cell states *de novo*. Therefore, there is an outstanding question for whether granular PDAC cell states can be robustly inferred and validated across patient cohorts, and whether cell state co-occurrence provides insights into the immune dynamics in tumours.

In this work, we present a comprehensive platform to dissect the extent to which cell state co-occurrence and cellular niches can be reproducibly inferred in the PDAC tumour microenvironment from atlas-scale scRNA-seq (**Fig. 1A**). We leverage the vast quantity of PDAC scRNA-seq generated by the community to collate eight independent datasets encompassing >150 samples and >500,000 cells to create a unique dataset-level discovery-validation setup. In doing so, the existence of known and *de novo* granular cell states in the TME can be independently verified, along with their ability to co-occur within and across cell types and form reproducible cellular niches. We map 60 lineage-specific cell state gene programs across 8 cell types and confirm the existence of known cell states such as exhausted cytotoxic T cell while discovering novel cellular variation such as multiple reproducible inflammatory cancer associated fibroblast (CAF) subtypes. Next, we evaluate the extent to which co-occurrence of these granular cell state programs is reproducible within and between

cell types, finding highly consistent co-occurrence across all major lineages. We subsequently formalize a novel probabilistic model for quantifying cellular niches from atlas-scale scRNA-seq and find these are both significantly reproducible between our discovery and validation datasets and may represent continuous niche-driven microenvironment remodeling with disease progression. Finally, we map the discovered gene programs and niches to two bulk datasets to associate these TME dynamics with survival rates, and found that the presence of certain niches are associated with poorer or more favorable outcomes along with specific KRAS mutations. A reproducible Snakemake pipeline for this analysis along with STAN code implementing the probabilistic niche model is available on GitHub at https://github.com/camlab-bioml/PDAC_cellular_niche_analyses.

Results

Curation of PDAC scRNA-seq discovery-validation cell state resource

To systematically quantify cell states from the PDAC tumour microenvironment, we compiled 8 previously generated scRNA-seq datasets (**Fig. 1B** and **Methods**, **Supplementary Table 1**)[4,35–40,52] comprising 578,000 total cells. To exclude possible technology-induced bias, all selected scRNA-seq data were generated using the 10X Genomics platform. We selected primary PDAC tumour samples from each cohort and split the datasets into a discovery group (4 datasets, n=74 tumour samples) and a validation group (4 datasets, n=89 tumour samples). Starting with raw counts, quality control and normalization were performed in line with previous studies (**Supplementary Fig. 1A**, **Methods**). We developed an automated workflow to systematically annotate single cells using a previously generated multi-cohort PDAC atlas as reference (**Supplementary Fig. 1B** and **Methods**)[7]. Eight major cell type lineages were identified in our curated dataset (**Fig. 1C**) with all present in both the discovery and validation groups as expected (**Fig. 1D**). We also noted that cell type proportion variations exist between the discovery group and the validation group, highlighting possible sampling biases in different patient cohorts. Our gene program extraction and validation workflow were carefully designed to handle such biases and preserve intra-tumour heterogeneity (**Methods**). Specifically, our workflow extracted and validated cell state gene programs in each cell lineage independently, which ensured adequate number of programs were allocated to relatively rare cell lineages such as B cells. To validate the automatically assigned cell type labels, we examined known marker expression in each group to confirm cell identity (**Fig. 1E**). The workflow we present here could be readily applied to identify and validate rare and fine-grained cell states when comparing datasets from multiple sources, despite the presence of batch effects and sampling biases.

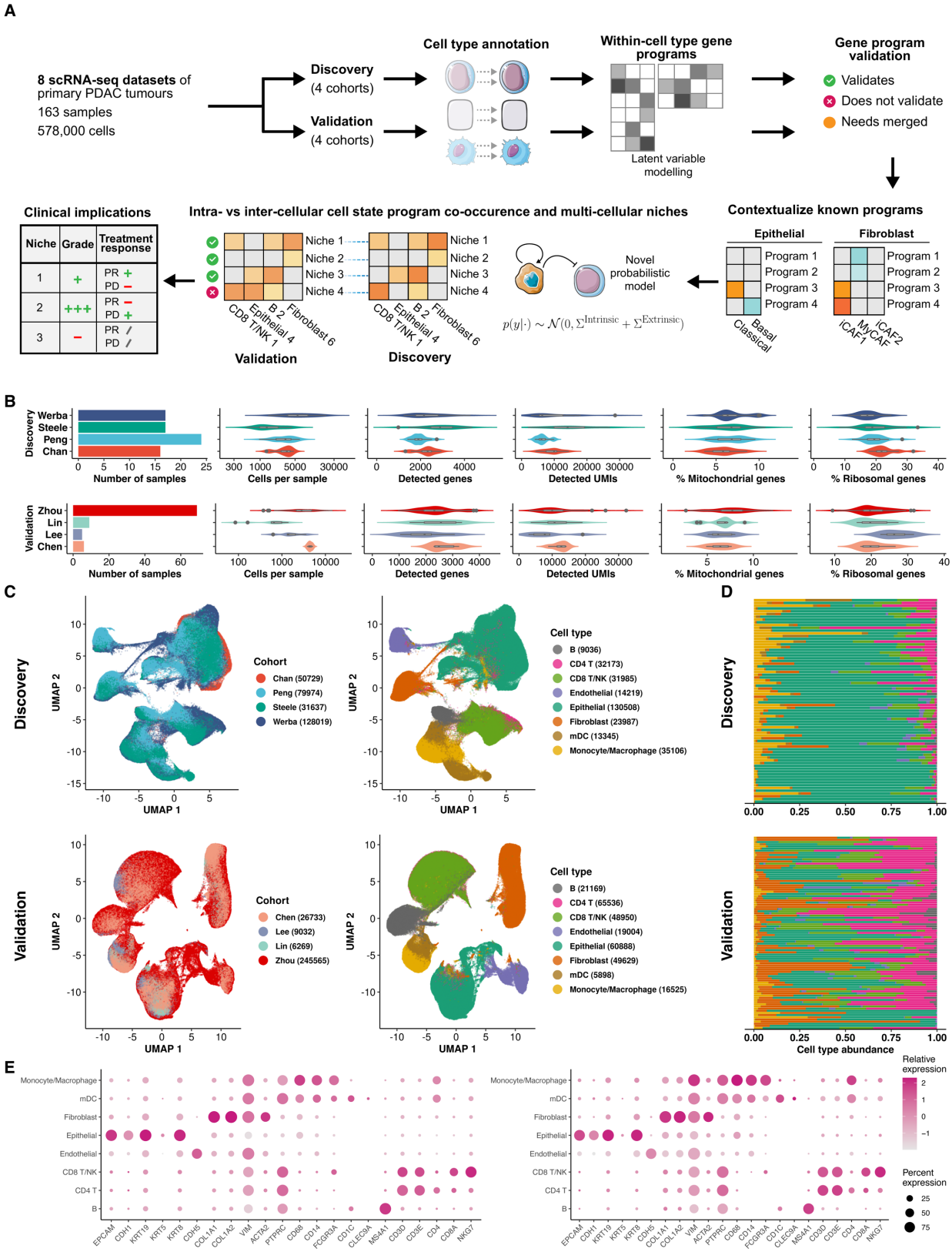


Figure 1: Overview of analytical workflow and discovery-validation datasets. A, Overview of discovery-validation platform for identifying co-occurrence and niches of granular cell states

in the PDAC TME. **B**, Dataset and QC metrics for discovery (top) and validation (bottom) datasets. **C**, UMAP representations for the discovery (top) and validation (bottom) datasets colored by dataset (left) and major annotated cell type (right). mDC, Myeloid Dendritic Cells. **D**, Overview of cell type composition for samples (in rows) in the Discovery group (top) and in the Validation group (bottom). **E**, Overview of selected lineage marker expression in each annotated cell type population for discovery (left) and validation (right).

Cell lineage specific gene programs capture cell state heterogeneity in the PDAC TME

To quantify granular cell states in the tumour micro-environment, we extracted cell lineage-specific gene expression programs independently for the discovery and the validation datasets. Specifically, we applied integrative non-negative matrix factorization (iNMF)[53] that learns a set of gene programs defined by the expression pattern of a set of genes along with their activities in each single-cell (**Supplementary Fig. 2A**, **Methods**). Our use of NMF rather than sub-clustering was motivated by the inherent plasticity of the cells of the tumour microenvironment[5,35,54,55], which is well-modelled by each cell having a continuous activation of multiple gene programs compared to clustering approaches that force each cell to belong to a distinct cluster. A similar computational approach to interpretable gene program discovery has been introduced by SPECTRA[56], which demonstrated the capability for T cell state deconvolution.

To identify whether specific cell types consistently activate certain gene programs across PDAC samples, we set out to identify a subset of lineage-specific high-confidence gene programs found in both our discovery and validation datasets. We first employed the Hungarian algorithm[57] on the gene program similarity matrix to match gene programs between the discovery and validation groups. For example, in fibroblasts, nine out of ten candidate gene programs were reproducible between discovery and validation, with strong consensus for the top weighted genes (**Fig. 2A** and **Supplementary Fig. 2B**). Next, to remove redundant programs, we merged validated reproducible gene programs with highly similar gene weights (**Methods: Program extraction, Program validation, Program merge**). For example, we merged two gene programs in B cells that were highly transcriptionally similar, both showing high expression of *CD37* and *CD52* (**Supplementary Figure 2C**). Following this, we updated gene program weights in single cells by fitting a linear relationship between the expression explained by the gene programs before merging and the average program gene weights after merging (**Methods**) and removed any gene programs likely corresponding to ambient RNA programs contaminating[58] by examination of the top program markers (**Methods**). After applying this procedure, we obtained 10 high-confidence gene programs present in both discovery and validation in each of B cells, CD4 T cells, CD8 T/NK cells, fibroblasts, and myeloid dendritic cells (mDCs), 20 in each of epithelial cells and macrophages/monocytes, and 5 in endothelial

cells (**Fig. 2B**). In total, we obtained 60 cell-lineage-specific gene programs that capture both the overall tumour microenvironment landscape and the granular phenotypic heterogeneity (**Fig. 2C**). Overall, our approach allowed us to discover and validate recurrent cell states in PDAC TMEs encompassing immune, epithelial, fibroblast and endothelial cells with high statistical confidence.

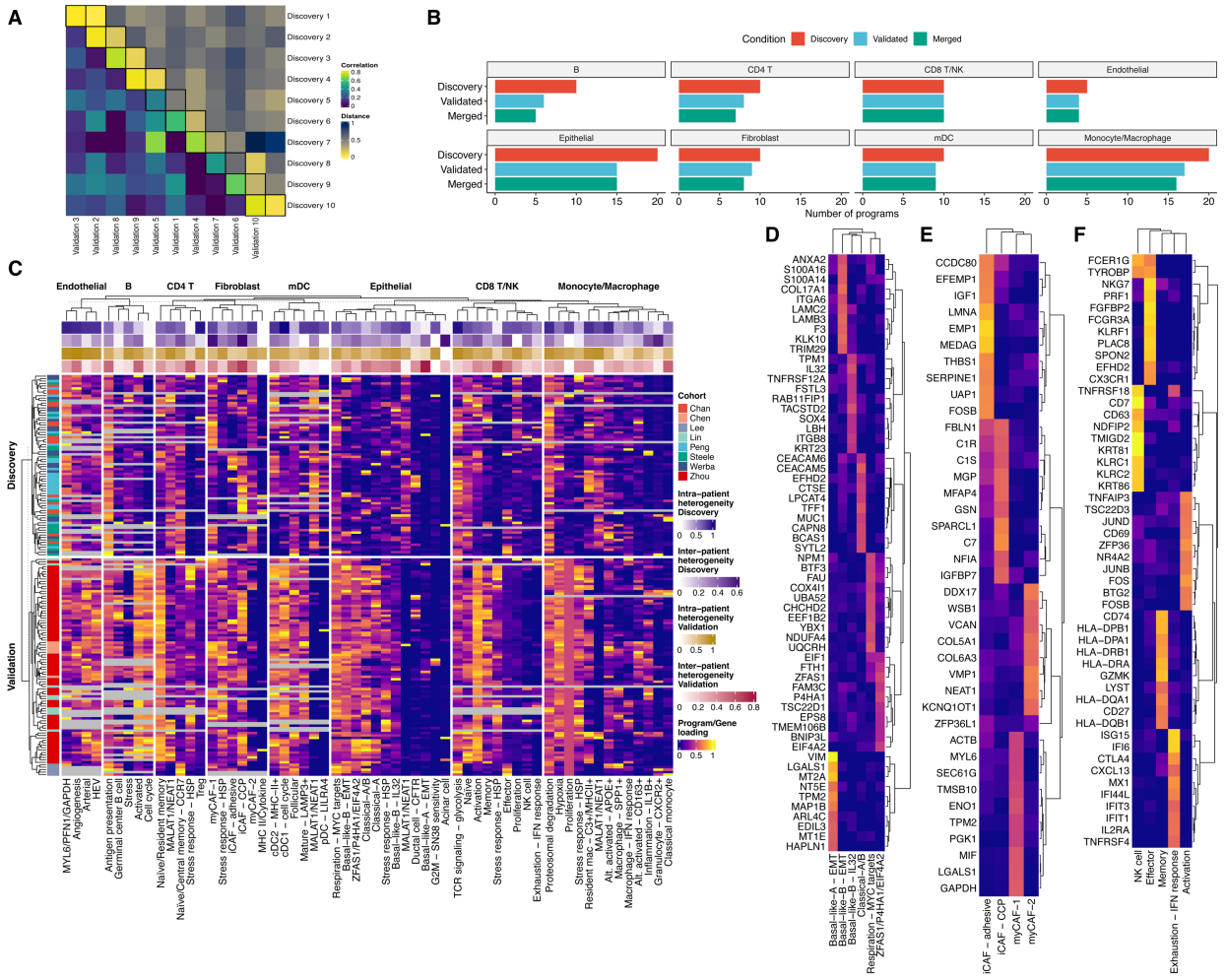


Fig. 2: Cell-lineage-specific gene programs reveal granular cell states present across PDAC samples. A, Gene weights correlations and Manhattan distances between candidate fibroblast specific gene programs extracted from the discovery group and the validation group. Columns (validation gene programs) are ordered according to best matching discovery-validation gene program pairs. **B,** Overview of numbers of candidate gene programs extracted from the discovery group (orange), numbers of gene programs validated between discovery and validation (blue), and numbers of consolidated gene programs after merging the redundant ones (green) in each cell lineage. **C,** Consolidated gene program sample-level loading weights. Each row is a different patient tumour sample; the left bar indicates patient cohort. The samples are grouped by discovery/validation. Each column is a cell lineage specific gene program labeled

206 according to cell state interpretation; the top bar and the third bar indicate sample-level gene
207 program loading variance means in discovery and validation respectively; the second bar and
208 the bottom bar indicate gene program loading mean variance between samples in discovery
209 and validation respectively (**Methods**). The heatmap denotes scaled loadings of each gene
210 program for each tumour sample, grouped by cell lineage indicated on the top. **D-F,** Heatmaps
211 showing weights of top ranked genes in selected gene programs from epithelial cells (**D**),
212 fibroblasts (**E**), and CD8 T/NK cells (**F**).
213

214 Interpretation of reproducible cell-lineage-specific programs as cellular states in 215 the PDAC microenvironment

216 To identify cell states positively associated with pro-tumorigenic or immune-suppressive TME
217 states, and distinguish new cell states from known ones, we contextualized the high-confidence
218 lineage-specific gene programs. For this we took a multi-faceted approach including manual
219 marker inspection, gene set enrichment analysis, and over-enrichment analysis (**Methods**).
220 Importantly, our unique discovery-validation setup that independently discovered reproducible
221 cell lineage-specific gene programs in multiple datasets confers high confidence on their
222 manifestation in patients.
223
224 In the epithelial compartment, we found gene programs that captured the refined existing PDAC
225 gene expression tumour subtypes (**Fig. 2D**). We found two distinct gene programs
226 corresponding to the Basal-like-B subtype, which from previous bulk analysis has been
227 considered a distinct subtype[35]. While enrichment analysis showed both Basal-like-B programs
228 upregulated epithelial-to-mesenchymal transition (EMT) genes (**Supplementary Fig. 3A-B**),
229 one program ("Basal-like-B - IL32") exhibited upregulation of *IL32*[59] which has previously been
230 found to be upregulated at the invasive tumour front. In contrast, the second Basal-B program
231 ("Basal-like-B - EMT") showed high regulation of *S100A14*[60,61] that is linked to gemcitabine
232 resistance in mouse models of PDAC[60]. Importantly, these two Basal-B programs are present in
233 distinct subpopulations of tumour cells (**Supplementary Fig. 3C**). Interestingly, a MYC-targets
234 program, which links to metastasis and chemoresistance[62,63], was present in cells that were
235 different from those of the Basal-like related programs (**Supplementary Fig. 3D-F**), though
236 MYC-target genes are usually enriched in Basal-like tumours. We further found one novel
237 epithelial program marked by the long-noncoding gene *ZFAS1* (**Fig. 2D**).
238
239 We next examined the fibroblast compartment where multiple previous studies have separated
240 cells into distinct inflammatory and myofibroblastic cancer associated fibroblast populations
241 (iCAF and myCAF respectively). Here, we identified two inflammatory cancer associated
242 fibroblast (iCAF) cell states and two myofibroblastic CAF (myCAF) cell states for the first time
243 (**Fig. 2E**). One iCAF program ("classical pathway (CCP) iCAF") was showed strong upregulation
244 of the terminal mediator of the complement system, *C7*, and the classical pathway genes *C1S*

245 and *C1R*. Interestingly, the CCP-iCAF program was enriched for a neurotropic gene set
246 previously described in PDAC[5] (**Supplementary Fig. 4A**), which is linked to tumour growth,
247 stress adaptation, and metastasis[64]. The second iCAF gene program ("adhesive iCAF") was
248 highly enriched for an adhesion gene signature including strong upregulation of the gene
249 *SERPINE1* that encodes for the protein plasminogen activator inhibitor-1 (*PAI-1*) (**Fig. 2E**). The
250 *ACSL3-PAI-1* signaling axis was found to support PDAC progression, linked to fibrosis and
251 immunosuppression, and have prognostic value[65]. Indeed, we found the CCP iCAF program
252 and the adhesive iCAF program to be present in a mutually exclusive pattern across patients
253 within the dataset (**Supplementary Fig. 4C**). Two myCAF programs were identified based on
254 CAF subtype gene loadings in fibroblast gene programs (**Supplementary Fig. 4B**). The *TPM2*+
255 myCAF-1 program was highlighted by the high loading of *MIF*[55] (**Fig. 2E**). The *VMP1*+ myCAF-2
256 program was further identified in addition to the myCAF-1 program. Although VMP1 has been
257 associated with PanIN formation[66] and patient survival in pancreatic neuroendocrine tumours[67],
258 its role in PDAC CAFs is still unclear.
259
260 Given the emerging importance of T cell dynamics in the immune response to PDAC[68,69], we
261 next examined the CD8 T/NK cell compartment where we identified multiple cell states reflecting
262 known T and NK cell heterogeneity and sub-lineage gene programs. For example, we found a
263 *CTLA4*+/*CXCL13*+ exhaustion cell state, a GZMK+/CD27+ memory/transitional exhausted
264 program, and a PRF1+/FCGR3A+ effector cell state (**Fig. 2F**). Notably, loadings of the
265 *CTLA4*+/*CXCL13*+ exhaustion program and the effector program with low exhaustion marker
266 loadings in CD8 T cells were mutually exclusive (**Supplementary Fig. 5B**), suggesting
267 heterogeneity in the degree of anti-tumour immune effector function and exhaustion across
268 PDAC tumours.
269
270 Next we examined how known and *de* novo macrophage phenotypes were captured by distinct
271 gene expression programs. We identified two alternatively activated macrophage gene
272 programs. One alternative activated program showed high loading of *APOE* (**Supplementary**
273 **Fig. 5C**). It was shown in mouse model that *ApoE* induced *Cxcl1* and *Cxcl5* expression in
274 pancreatic tumour cells, promoting immune suppression[70]. Notably, lysosomal genes *CTSD*,
275 *PSAP* and *LIPA*, and phagocytosis/apoptotic cell-clearance gene *GPNMB* were some of the top
276 markers for this gene program, linking efferocytosis to immune suppression[71]. The other
277 alternatively activated program was marked by M2 genes *CD163*, *STAB1*, and *FOLR2*
278 (**Supplementary Fig. 5C**). An inflammation related macrophage program marked by high *IL1B*
279 and *NFKB1* loading was identified. While traditionally denoted as the M1 macrophages, recent
280 research has revealed a new role in pathogenetic inflammation and poor patient prognosis for
281 *IL1B*+ macrophages[46]. We also found an *SPP1*+ macrophage program, which has been
282 previously described to play an important role in angiogenesis, a much needed function in the
283 hypoxic PDAC TME[72–74]. One macrophage program enriched for *ISG15* and *MX1* was identified

284 as the interferon response program and had high loadings in a small population of cells
285 (**Supplementary Fig. 5C-E**). In a recent mouse model study, the dynamics of the interferon
286 pathway in PDAC is closely related to *CXCL13* production in macrophage, which affects B cell
287 and NK cell recruitment[75]. In short, our gene programs comprehensively captured the
288 transcriptional heterogeneity of the macrophage/monocyte lineage, encompassing classic
289 (M1/M2) and novel functional annotations.
290
291 We obtained three notable gene programs in the B cell lineage. The germinal center (GC)
292 program is characterized by the specific and high loading of *BCL6* (**Supplementary Fig. 6A**), a
293 transcription factor critical for proliferation and the maintenance of the GC phenotype[76–78]. The B
294 cell activation program saw intermediate loading of *IL-6* and high loading activation marker
295 *CD69* (**Supplementary Fig. 6A**). The MHC class I genes (*HLA-A*, *HLA-B*, and *HLA-C*)
296 specifically and heavily loaded in the antigen presentation (AP) program, while the MHC class II
297 genes (*HLA-DP*, *HLA-DQ*, and *HLA-DR*) were heavily loaded in both the AP program and the
298 GC program (**Supplementary Fig. 6B**). We did not find distinct loading pattern of plasma
299 markers (*IRF4*, *XBP1*, *PRDM1*, *MZB1*)[78] in any B cell program (**Supplementary Fig. 6C**),
300 suggesting fully differentiated plasma cells are relatively rare in the PDAC TME. Lastly, there
301 was no B cell lineage gene program showing specific or high loading of regulatory B cell
302 markers such as *IL-35* (*EBI3*)[77,79] (**Supplementary Fig. 6A**).
303
304 Within the endothelial lineage, we identified a high endothelial venule (HEV) gene program, an
305 angiogenesis program, and an arterial gene program. The HEV program had specific and high
306 loadings of MHC class II genes[80] and vein-specifying transcription factor *NR2F2*[81]
307 (**Supplementary Fig. 6D, E**). The angiogenesis program had specific and high loadings of
308 *ANGPTL2*, *INSR*, and *TIE1*[82] (**Supplementary Fig. 6D**). The arterial program had high and/or
309 specific loadings of *CXCL12*[83], *VEGFC*, *HEY1*[84], arterial marker *EFNB2*, *DLL4*[85] and notch
310 signaling gene *NOTCH4*[86] (**Supplementary Fig. 6D, F**).
311
312 In all cell lineages except for endothelial cells and mDCs, we detected a stress response
313 program marked by upregulated heat shock response proteins (**Supplementary Table 2**).
314 Stress response gene sets have traditionally been linked to low quality cells in scRNA-seq.
315 However, in a recent pan cancer study, Chu et. al.[87] identified a unique stress response state in
316 the tumour microenvironment of multiple cancer types, highlighted by heat shock response
317 protein *HSPA1A* and *HSPA1A* expression, which was enriched in nonresponsive tumours after
318 ICB treatment. This result suggests more analysis is needed to fully understand the biological
319 meaning of stress responses in the tumour microenvironment. Our programs recapitulated the
320 epithelial basal and classical subtypes[35,49], the macrophage activation heterogeneity[37,74], and
321 the diverging CAF subtypes[5,88]. The presence of previously described cell populations reinforces
322 our confidence in the newly identified cell state programs, particularly the CCP-iCAF program,

323 the B cell GC phenotype program, and epithelial *ZFAS1* program, which might have a role in
324 treatment response and clinical outcomes.
325
326 **Intra- and inter-tumoural heterogeneity in gene programs is highly consistent**
327 Since intra- and inter-tumour heterogeneity is closely coupled with treatment resistance and
328 therapy response across cancers[6,33,89,90], we next investigated whether heterogeneity in the
329 granular cell states depicted by our gene programs was conserved between our discovery and
330 validation datasets. We calculated gene program inter-patient heterogeneity by taking the
331 variances for each program's single-cell loading medians in tumour samples. We summarized
332 gene program intra-patient heterogeneity by averaging loading variances in tumour samples for
333 each program. We found inter- and intra-patient program loading heterogeneity varied greatly
334 across cell types and within cell types (**Fig. 3A, B**). Gene programs of CD4 T cells, CD8 T/NK
335 cells, epithelial cells, and macrophages showed wide ranges of intra-patient heterogeneity
336 across patient cohorts (**Fig. 3B**). On the other hand, fibroblast, B cell, and mDC programs
337 showed more variable loadings in the tumours from the discovery datasets compared to those
338 from the validation datasets (**Fig. 3B**). Of note, endothelial cell programs were consistently and
339 highly variably loaded in tumours, suggesting the endothelial cell states are clearly
340 distinguished. Across cell types, we found inter-patient heterogeneity tended to agree between
341 the discovery and the validation datasets, except for CD4 T cell programs and macrophage
342 programs (**Fig. 3B**). In short, we showed in single-cell data for the first time that both intra- and
343 inter-patient cell state heterogeneity is highly conserved between discovery datasets and
344 validation datasets (intra-patient heterogeneity: $\rho = 0.67$, $p = 1\text{x}10^{-08}$, inter-patient heterogeneity:
345 $\rho = 0.44$, $p = 0.00053$). This suggests that the abundance of granular cell states within the TME
346 can serve as a robust estimate of heterogeneity and help guide target discovery efforts that
347 specifically attempt to cover the entire phenotypic space.
348

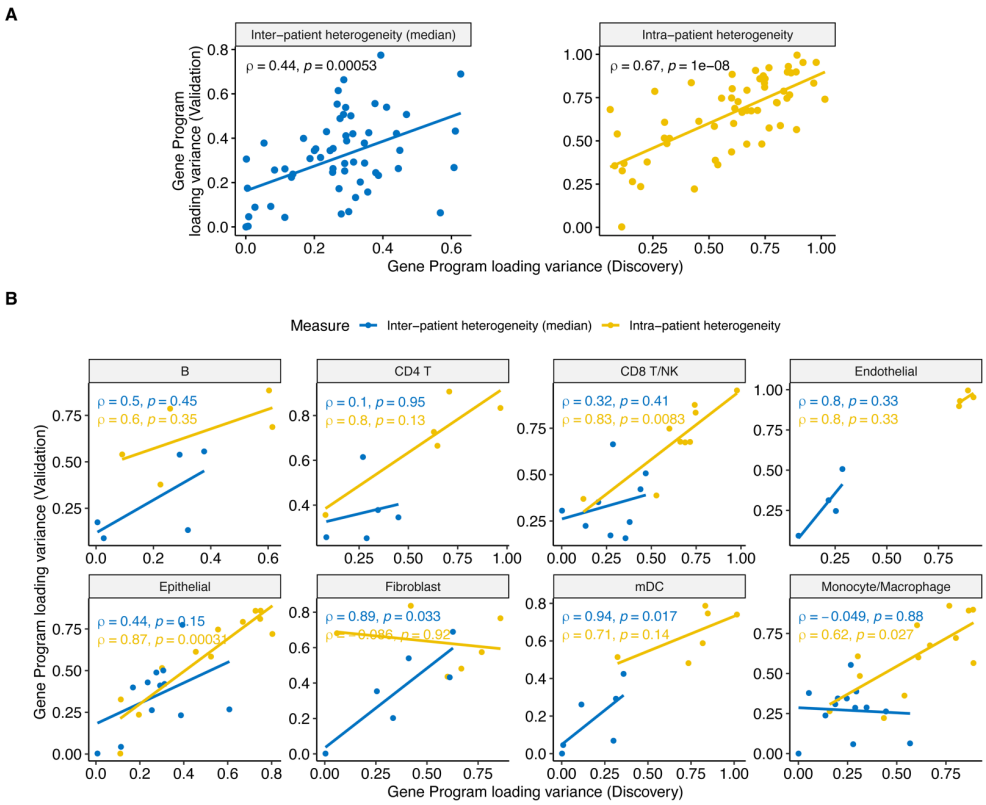


Figure 3: Heterogeneity of gene program loading is highly consistent between and within individual patients. A, Spearman correlations of overall gene program loading heterogeneity. Each dot is one cell type specific program. Inter-patient heterogeneity (left) was calculated by first getting the program loading means in each sample, then computing the variance of sample level loading means in Discovery and Validation, respectively. Intra-patient heterogeneity (right) was calculated by first computing the program loading variances in each sample, then getting the mean of sample level loading variances in discovery and validation, respectively. **B**, Spearman correlations of gene program loading heterogeneity grouped by cell lineage. Each dot is one gene program.

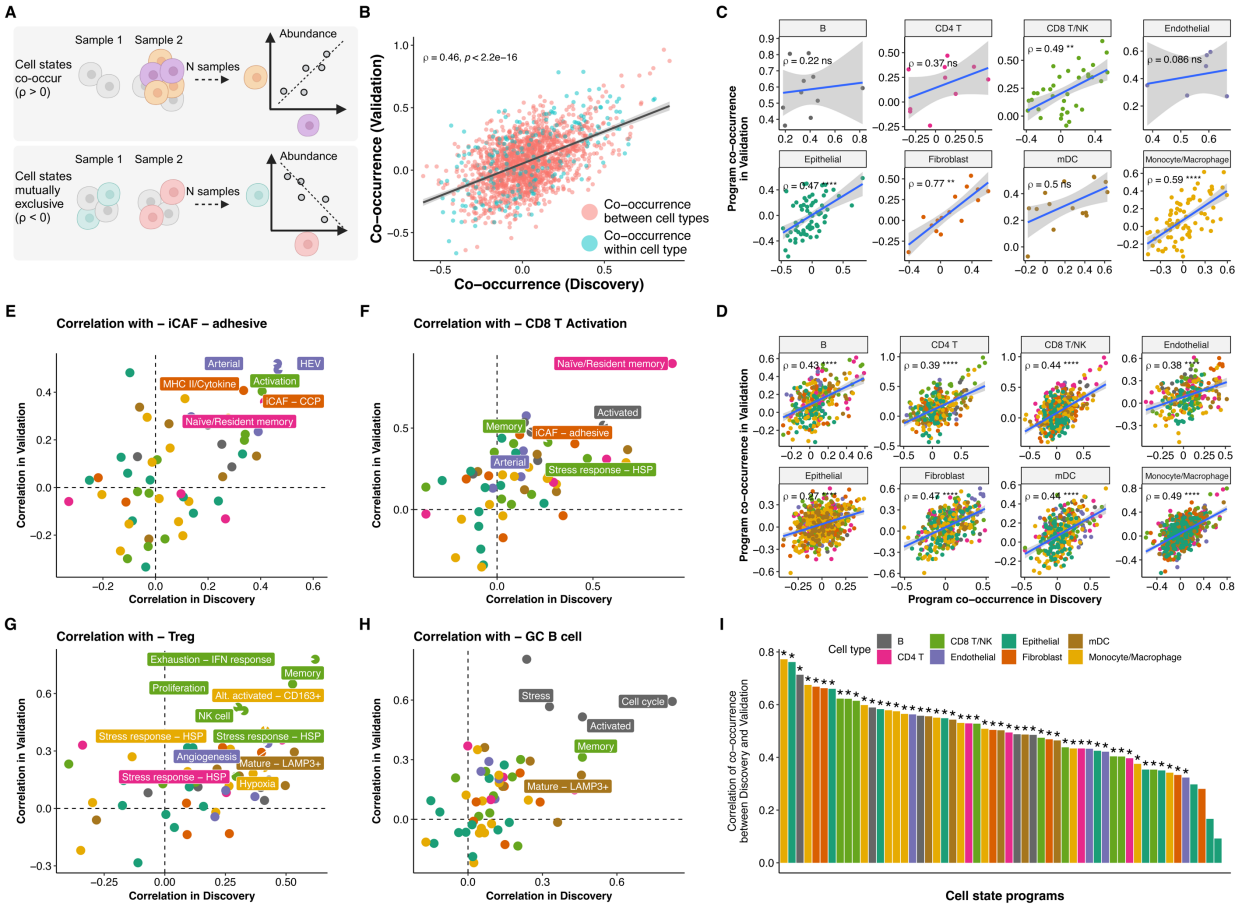


Figure 4: Pairwise co-occurrences of cell states are highly reproducible across PDAC samples. A, Two examples of cell state program co-occurrence patterns. The program co-occurrence scores for each program pair are calculated in the discovery and the validation group separately for comparison between datasets. **B**, Program co-occurrence agreement between discovery and validation. Each dot denotes one gene program pair. Co-occurrence per program pair was measured as Spearman correlations between sample level loading means. **C**, Within cell type gene program co-occurrence agreement between discovery and validation. **D**, Across cell type gene program co-occurrence agreement between discovery and validation. **E-H,** Tumour sample level loading mean correlations between the adhesive iCAF gene program and other programs (**E**). Tumour sample level loading mean correlations between the CD8 T cell activation program and other programs (**F**). Tumour sample level loading mean correlations between the CD4 regulatory T cell gene program and other programs (**G**). Tumour sample level loading mean correlations between the GC B cell activation program and other programs (**H**). **I**, Gene programs ranked by co-occurrence agreement between discovery and validation. Programs whose co-occurrence patterns are more similar between discovery and validation have higher correlation of co-occurrence between discovery and validation. Asterisk signs denote gene programs whose co-occurrence with other programs is significantly correlated between discovery and validation.

381
382 **Co-occurring granular cell states are highly reproducible in PDAC datasets**
383
384 Next, to understand if granular cell-state co-occurrence can be robustly estimated from scRNA-
385 seq of the TME, we evaluated the extent to which both between and within cell type gene
386 program co-occurrence within individual tumours could be reproducibly quantified. We
387 computed co-occurrence of all possible gene program pairs at the patient level (N=1770 total,
388 **Methods**) using Spearman correlation separately in the discovery and the validation groups
389 (**Fig. 4A**). We quantified how consistent this co-occurrence by comparing the co-occurrence (i.e.
390 correlation in gene program presence) in discovery and validation settings, finding highly
391 concordant estimates (ρ=0.46, $p<2x10^{-16}$) both gene programs within and between cell lineages
392 (**Fig. 4B**).
393
394 Next, we examined to what extent the reproducibility of cell state co-occurrence varies within
395 and between major cell lineages. We sought to compare co-occurrence of a cell state program
396 pair between discovery and validation. By filtering down to program pairs that are from the same
397 lineage or from different lineages, we were able to obtain an overview for the reproducibility
398 landscape. We found highest agreement between discovery-validation on within lineage co-
399 occurrence in fibroblasts (ρ=0.77) followed by monocyte/macrophage gene programs (ρ=0.59)
400 (**Fig. 4C**). Gene program co-occurrence within T cells, mDCs, and epithelial cells exhibited
401 lower but still highly significant co-occurrence between discovery and validation (ρ=0.37 to 0.5),
402 while endothelial cells and B cells lower (ρ=0.086 and 0.22 respectively). However, we noted
403 that while B cells showed lower correlation (ρ=0.22, p=0.54) between discovery and validation
404 patient cohorts, they both exhibited positive correlation in discovery (ρ=0.2-0.81) and validation
405 (ρ=0.38-0.8), demonstrating that the overall effect size was consistent (**Fig. 4C**).
406
407 Next, we examined the reproducibility of the co-occurrence of cell states across different cell
408 lineages. For example, for a given B cell gene program, we compute the average per patient,
409 and quantify the co-occurrence of this with the average-per-patient for a gene program in a
410 different lineage such as CD8 T cells. We found gene programs across all cell types exhibit
411 significantly reproducible co-occurrence, most notably once again in Fibroblasts (ρ=0.47),
412 Monocyte/Macrophage (ρ=0.49), and CD8 T cells (ρ=0.44) (**Fig. 4D**). Indeed, on a program-by-
413 program basis we found most have significantly reproducible inter-cell type co-occurrence
414 relationships (**Fig. 4I**).
415
416 We next focused on the co-occurrence of specific cell states that are either significantly
417 reproducible between discovery and validation or *a priori* known to hold clinical significance. The
418 co-occurrence of regulatory CD4 T cells and exhausted CD8 T cells in the PDAC
419 microenvironment was well known from mouse studies[69], but is yet to be elucidated in humans.

420 We show here for the first time that the CD4 T_{reg} program consistently and strongly co-occurred
421 with the CD8 T cell exhaustion program in patient tumours (**Fig. 4G**). Also, the CD4 T_{reg}
422 program co-occurred with the *CD163*+ alternatively activated macrophage program and the
423 endothelial angiogenesis state program marked by *TIE1*, *INSR*, and *NOTCH4*. This
424 angiogenesis program has been found to be an ECM remodeling and angiogenic signature in
425 lung and breast cancers[91,92] and likely interact with TAMs and regulatory T cells in breast cancer
426 to promote an immunosuppressive phenotype in the microenvironment[92]. This T_{reg}-T_{exh}-Mac_{alt}
427 co-occurrence pattern is spatially confirmed in multiple cancers[93], suggesting our cell state
428 program co-occurrence may be associated with spatial relationships in the PDAC.
429
430 The adhesive iCAF program significantly co-occurred with the endothelial HEV state and the
431 arterial state programs, as well as the CD8 T cell activation state program in the
432 microenvironment (**Fig. 4E**). We also found co-occurrence of the CCP iCAF program and the
433 MHC class II/cytokine program with the adhesive iCAF program. Together, this result suggests
434 multiple cell types in different states participate in the recruitment and activation of CD8 T cells
435 in the tumour (**Fig. 4F**). We further noted that the Naïve and memory CD4 T cell programs also
436 co-occurred with the adhesive iCAF program.
437
438 Finally, we investigated whether known multi-cellular structures in the TME were consistent with
439 our co-occurrence results. Specifically, we focused on Tertiary lymphoid structures (TLSs)
440 clusters of immune cells in the TME that when mature can lead to production of tumour-specific
441 antibodies and facilitate anti-tumour immune response. TLSs in PDAC are usually composed of
442 proliferating, activated, germinal center (GC) B cells, antigen presenting cells such as mature
443 dendritic cells, and T cells and have previously shown to be predictive of patient survival[77].
444 Therefore, we examined the co-occurrence pattern for our germinal center B cell program.
445 Indeed, we found the B cell activated and cycling programs are among the top within-lineage
446 programs that co-occur with the GC program (**Fig. 4H**). In addition, we found significant co-
447 occurrence (ρ=0.3 in the validation group) of the activation and memory program of CD8 T cell
448 programs with the GC B cells both in discovery and validation, along with mature and
449 plasmacytoid dendritic cells (ρ=0.45). These results suggest that in both our discovery and
450 validation datasets, TLS presence and phenotype could be read out from scRNA-seq.
451 Examination of the patient sample level loadings of the TLS related cell state programs
452 suggests variable presence of TLS in PDAC tumours (**Supplementary Fig. 7B, D**).
453 Overall, our results imply co-occurrence of granular cell states from scRNA-seq can be
454 meaningfully quantified from large scale atlases, and complex immune cell clusters can be
455 identified and characterized without the need for additional spatial transcriptomics or
456 proteomics.
457

A novel probabilistic model can infer recurrent cellular niches in PDAC microenvironments

Beyond the pairwise co-occurrence of cell states, cell niches play a fundamental role in the tumor microenvironment, as these multicellular populations often act in concert, and their function is highly synergistic[94]. For this reason, we asked whether it was possible to infer niches reproducibly from scRNA-seq. However, there are no existing computational tools to infer niches from scRNA-seq using a set of NMF factors with interpretations as input. Therefore, we developed a novel probabilistic model (see **Methods**) that can take annotated NMF factors representing cell states and infer a pre-defined number of niches, reporting each both in terms of the cell states that co-occur as well as the niche abundance within each patient sample. Importantly, our model carefully decouples the fact that cell states can co-vary within the same cell (intrinsic variation) but also between cells (extrinsic variation) (**Fig. 5A**). Using maximum a posteriori inference (**Methods**), we fitted the niche model separately on the discovery and validation groups using all the 60 cell state programs. We quantified n=4 niches, though found stable performance varying the number of niches from 2-5.

We first contrasted the niches found in the discovery and validation datasets to both validate the overall methodology and generalizability of the findings. We found significant positive correlation in 3/4 niches between discovery and validation (**Supplementary Fig. 7A**) and subsequently renamed the validated niche pairs as niches 1-3. Next, we interpreted the validated niches by examination of the abundance of cell states within each (**Fig. 5B**). Niche 1 contained an over-representation of Classical A and ductal epithelial programs, along with HEV, arterial, and angiogenesis endothelial cell state programs, the CCP and adhesive iCAF programs, the proliferation, resident, and *CD163*+ alternatively activated macrophage programs, and the memory CD8 T cell program, (**Fig. 5B**). Niche 2 is a Basal-like subtype related niche. Notable core members of Niche 2 include the myCAF-1 program, the epithelial MYC targets program, the epithelial Basal-like-A/EMT and Basal-like-B/EMT program, the *APOE*+ alternatively activated macrophage program, and the *SPP1*+ macrophage program (**Fig. 5B**). The *LAMP3*+ mature dendritic cell program was likely a core member of Niche 2 as well, albeit being slightly higher in the validation group. Niche 3 represented an immune activated niche, including an over-representation of cell states corresponding to classical monocytes, granulocytes, follicular mDCs, antigen presentation B cells, *CCR7*+ naïve/memory CD4 T cells, and CD8 effector T cells (**Fig. 5B**). Notably, the TLS related cell state programs discussed above were over-represented in Niche 3 as well.

We next examined whether niches corresponding to the basal/classical transcriptional subtypes showed overall mutual-exclusivity. We found the presence of classical and basal-like niches in individual tumour samples were significantly anti-correlated in both discovery and validation cohorts (ρ= -0.28 and -0.56 respectively) (**Fig. 5C**). We then scored 174 bulk RNA-seq PDAC

497 primary tumour samples from the Chan cohort[35] for our validated niches (**Methods**). The
498 presence of the Classical and the Basal-like niches were likewise anti-correlated in this sample
499 (ρ=-0.14, p=0.065) (**Fig. 5D**), though to a lesser degree likely reflecting the difficulty of scoring
500 bulk samples for niches of granular cell states. Notably, this anti-correlation was not simply
501 driven by basal-classical bulk calls (**Fig. 5D**). Together, these results hint at a gradual re-
502 modelling of the tumour microenvironment in line with disease progression, in later stage PDAC
503 the balance of expression subtypes converts towards basal[35].
504
505 **Microenvironments from PDAC carrying specific KRAS mutations are associated with**
506 **distinct immunological niches**
507 *KRAS* mutation is commonly used as a molecular marker to subdivide PDAC types. PDAC
508 carrying *KRAS*-G12D mutations are associated with poorer prognosis when circulating tumour
509 DNA is detected in patient blood[95], and strategies targeting specific *KRAS* neoantigens are
510 being developed[96]. However, the mechanisms through which different mutations affect the
511 microenvironment in human PDAC remains unclear. We investigated whether genomic driver
512 mutation status was associated with TME niche. We obtained *KRAS* variant calls for the HTAN
513 Zhou (WUSTL) patient dataset[4] in our validation cohort and fitted a linear mixed model between
514 niche presence and *KRAS* variant status (**Methods**). We found that compared to *KRAS*-G12D,
515 tumours with a *KRAS*-G12V/G12R mutations had a significantly higher presence of the immune
516 activated niche, while tumours with *KRAS*-G12R mutations had a reduced presence of the
517 basal-like niche (**Fig. 5E**). Importantly, we retrieved similar *KRAS* variant calls in the bulk Chan
518 patient cohort[35] and found consistent effect sizes for all associations (**Supplementary Fig. 7C**,
519 **Supplementary Table 3**), validating the overall findings. The decreased presence of the
520 immune activated niche in *KRAS*-G12D PDAC compared to other *KRAS* variants suggests
521 differences in anti-tumour immune response may underlie prognosis and effectiveness of
522 immunotherapy[95,96]. Together, these results suggest that the specific driver mutations in PDAC
523 have a significant effect on the phenotypes and cell communities in the tumour
524 microenvironment.

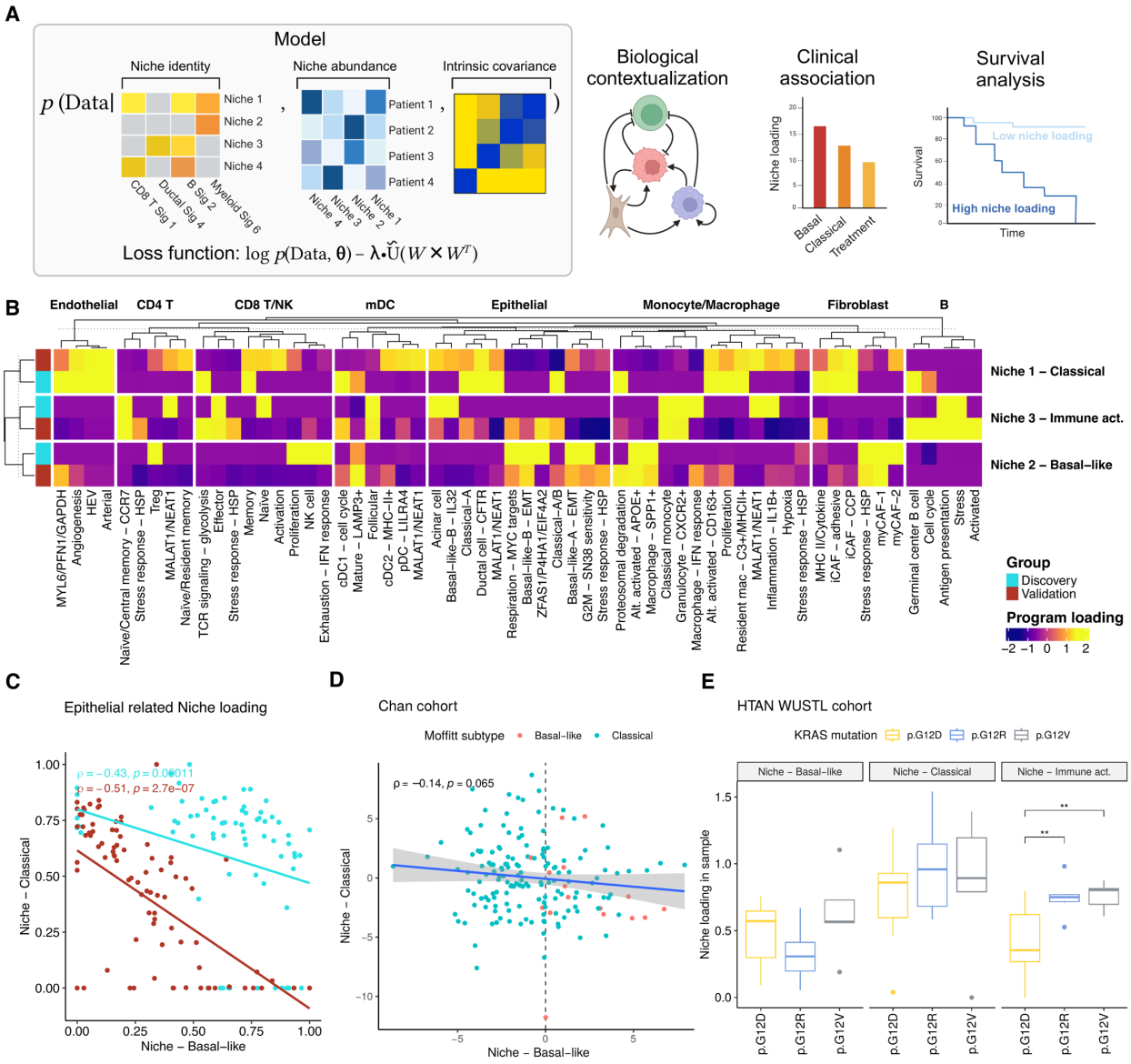


Figure 5: Novel cellular niches identified by a probabilistic model are associated with PDAC subtypes. A, Schematic of the Bayesian model for identifying multi-cell state niches in the tumour microenvironment and taking within-cell intrinsic cell state program covariance into consideration. The loss function includes an niche covariance term to encourage discovery of distinct niches. **B**, Cell state program loadings in niches identified in the discovery group and the validation group. The program loadings in the heatmap are normalized. **C**, Loadings of the Basal-like subtype related niche vs. loadings of the Classical subtype related niche in discovery (light blue) and validation (red). **D**, Niche enrichment scores in the RNA-seq samples of the Chan cohort. Each dot is one patient sample, colored by Moffitt subtypes. **E**, The relationships between KRAS variants and niche loadings in tumours.

537 **Specific PDAC TME cell states and niches are correlated with clinical outcomes**
538 After validating that cell states, gene programs and multicellular populations (niches) co-occur
539 frequently across PDAC samples, we sought to investigate whether certain combinations of co-
540 occurrent populations were more likely to be present in patients with positive or negative
541 outcomes. To examine whether there was an association between our discovered cell states
542 and niches with clinical covariates, we curated clinical and metadata for two cohorts of PDAC
543 tumour samples with paired bulk RNA-seq data, which we refer to as the Chan cohort[35] (N=183
544 patients) and TCGA pancreatic cancer cohort[97] (N=151 patients). We scored each sample in
545 each cohort by using gene set enrichment analysis on cell state program gene sets (**Methods**)[98]
546 and subsequently associated each clinical variable (**Supplementary Table 4**) with the presence
547 of each cell state in each patient.
548
549 We first confirmed that cell states corresponding to Basal-like and Classical expression
550 subtypes were associated with their related niches (**Fig. 6A**). We further confirmed the Basal-
551 like-A epithelial cell state and the Classical-A and Classical-A/B epithelial cell states were
552 associated with the Basal-like and Classical subtypes[99] respectively (**Methods**) in both the
553 Chan cohort and the TCGA cohort, with $p < 0.001$ (**Fig. 6E**).
554
555 In the Chan cohort, the B cell antigen presentation program, most T cell programs except the
556 naïve and activation programs, alternatively activated macrophage programs, as well as the
557 CAF programs, were significantly associated with higher histological grade (**Fig. 6E**). Such
558 associations were however less clear in the TCGA cohort. In the epithelial lineage, we noted the
559 Basal-like programs were linked to liver metastasis in TCGA, while the Basal-like-A/EMT
560 program was associated with higher tumour grade in both TCGA and Chan. We further found
561 that the Classical programs were associated with more differentiated tumours in Chan, while
562 strongly negatively associated with liver metastasis in TCGA.
563
564 Next, we examined cell state program association with clinical outcomes. In the Chan cohort,
565 the Basal-like-A/B/EMT programs, the myCAF-1/2 programs, the adhesive iCAF program, and
566 the macrophage *SPP1*+ program conferred significant survival risk (HR>1, $p < 0.05$). In the
567 TCGA cohort, a previously undescribed epithelial program upregulating *ZFAS1*, *P4HA1*, and
568 *EIF4A2*, the macrophage *SPP1*+ program, and Basal-like-B/EMT and G2M epithelial programs
569 conferred significant survival risk (HR>1, $p<0.05$) (**Fig. 6E**). Upon closer examination, the novel
570 *ZFAS1*/*P4HA1* epithelial program predicted patient survival in both the TCGA cohort and the
571 Chan cohort (**Fig. 6B**). The long non-coding RNA *ZFAS1* had been found to regulate EMT in
572 colon adenocarcinoma[100] and promote metastasis and tumour cell proliferation in pancreatic
573 cancer[101]. In breast cancer, *P4HA1* was shown to modulate cancer cell stemness and promote
574 chemoresistance[102]. This novel epithelial cell state program had lower correlation with the

575 Basal-like niche (ρ =0.29, **Fig. 6C**) compared to the Basal-like-A program, while better
576 dichotomizing patient in terms of survival compared to the expression subtypes (**Fig. 6D**).
577
578 Finally, using the cell state program enrichment scores, we computed weighted enrichment
579 scores for the niches by summing the program scores weighted by respective program loading
580 in each niche. As both the Chan and the TCGA cohort recorded patient response to adjuvant
581 therapy, we examined the relationships between our niches and therapy responses.
582 Interestingly, the Basal-like niche was positively associated with progressive disease after
583 adjuvant therapy in both the Chan cohort ($p<0.01$) and the TCGA cohort (**Fig. 6E**), and more
584 associated with progressive disease than any individual cell state including all Basal related
585 tumour epithelial states. Similarly, the Basal-like niche was associated with higher tumour grade
586 in both patient cohorts. On the other hand, the immune activation niche was associated with
587 complete and partial remission in the TCGA cohort, as well as lower tumour grade and lower
588 likelihood for distant metastasis (**Fig. 6E**). Overall, we show that large scRNA-seq datasets can
589 be employed to not only identify cell populations and cell niches with high confidence, but that
590 these have profound biological and clinical significance, as the frequency of co-occurrence of
591 certain cell states and the presence of certain cell niches are associated with patient survival
592 rates. These analyses suggest that in the future, scRNA-seq datasets will be used to study the
593 broad interplay among pro-tumorigenic cell populations in the tumour microenvironment. This
594 approach may eventually expand the clinical usability of single cell data, offering a new lens to
595 develop innovative therapeutic strategies targeting PDAC.
596
597
598

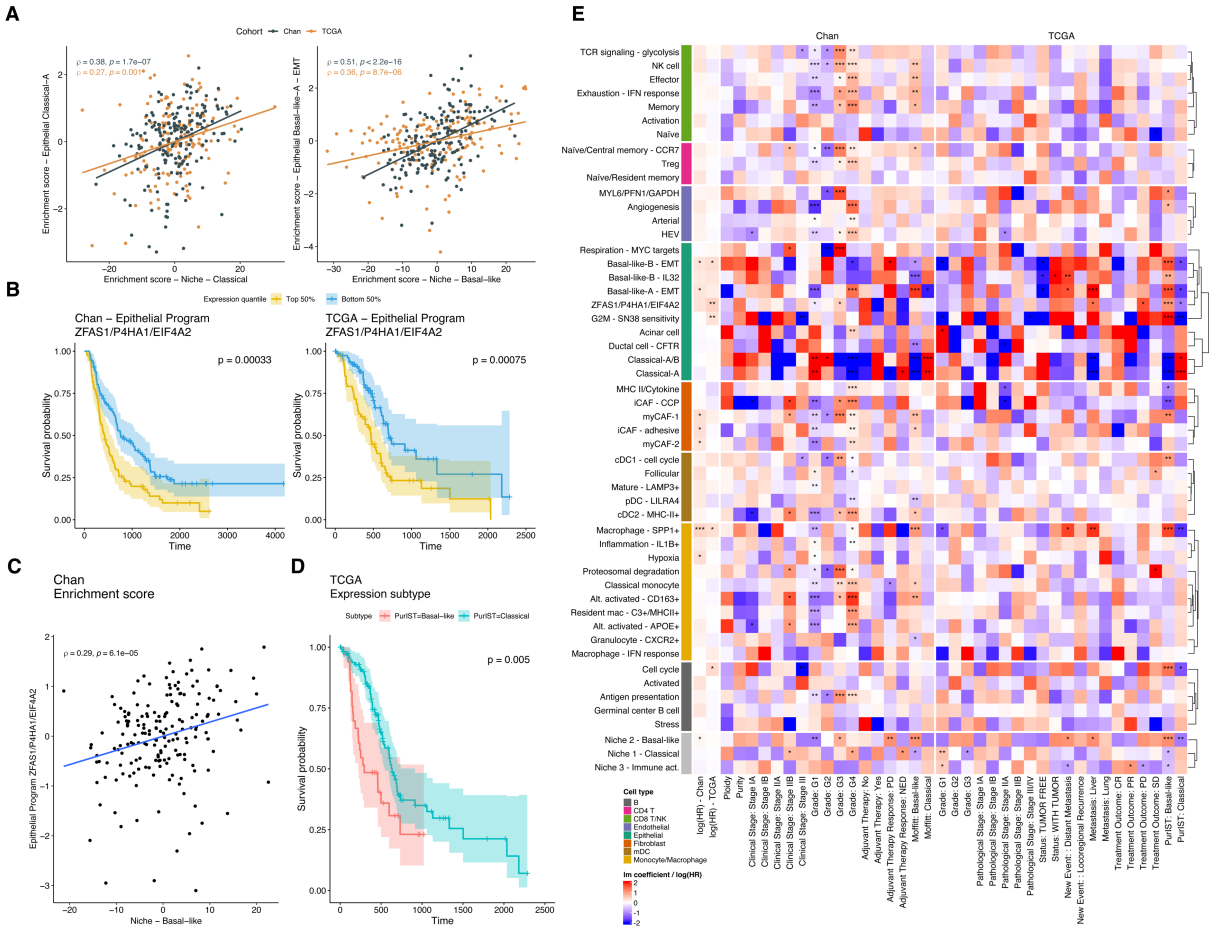


Figure 6: A novel epithelial program predicts shorter survival time. A, Niche vs. cell state program enrichment scores in tumour RNA-seq samples. Correlation coefficients were calculated by patient cohort. **B**, The *ZFAS1*/*P4HA1* epithelial program dichotomized patients in the Chan cohort (left) and the TCGA cohort (right). Program associations to clinical variables in the Toronto patient cohort and the TCGA cohort. **C**, Enrichment scores of the *ZFAS1*/*P4HA1* epithelial program vs. the Basal-like niche in Chan patients. **D**, Dichotomization of patient survival by expression subtypes in the TCGA PDAC cohort. **E**, Clinical associations of the lineage specific cell state programs and the niches. The two leftmost columns of the heatmap show (Hazard Ratio – 1) in the Chan and the TCGA cohort respectively for each program and niche from fitting univariate CoxPH models. The other columns of the heatmap show fitted coefficients for each program and niche using linear regression. '*': p-value < 0.05; '**': p-value < 0.01, '***': p-value < 0.001 (all Benjamini-Hochberg FDR corrected).

Discussion

Transcriptome profiling of tumour samples has historically been used to identify molecular markers that could facilitate cancer diagnosis, prognosis prediction and identification of molecular candidates to target therapeutically[103–106]. The advent of single cell technologies has

617 facilitated the identification of the relative distribution of cell populations specific for certain
618 cancer types, predicting clinical outcomes and uncovering specific cellular subtypes as
619 targets[107–110]. However, it is becoming increasingly clear that focusing on the expression pattern
620 of a handful of genes or the presence/absence of certain cell populations is not sufficiently
621 informative to gain useful biological insights into the whole tumour microenvironments, even
622 though this information could be leveraged to design targeted therapies aimed at tackling the
623 combined action of multiple cell types. In this context, spatial information is extremely useful to
624 identify co-localization of different cell types, whose crosstalk requires a certain three-
625 dimensional organization to synergistically alter tumour growth and survival. However, a
626 substantial number of samples are needed to robustly identify co-occurrence of interacting cells
627 across patients, especially considering that the cell states of interest are potentially rare.
628 Therefore, the current cost and/or limitations of spatial transcriptomics and proteomics limit the
629 identification of reproducible co-existing cell states with prognostic value[23]. By contrast, the
630 number of scRNAseq datasets from tumor microenvironments that contain transcriptome-wide
631 expression quantification at single-cell resolution is increasing exponentially, allowing, in theory,
632 to identify co-occurrence of granular cell states and associate them with clinical outcomes.
633
634 Here we have undertaken a unique discovery-validation analysis on the largest PDAC scRNA-
635 seq atlas to-date to quantify the extent to which granular cell states and cellular niches can be
636 reliably inferred. Our 60 reproducible cell lineage specific cell states captured previously known
637 heterogeneity such as classical and basal tumour transcriptional but also novel cell states such
638 as two iCAF subtypes as well as a novel *ZFAS1*/*P4HA1* epithelial cell state significantly
639 associated with survival. We further revealed for the first time that co-occurrence of granular cell
640 states can be reproducibly quantified from large scRNA-seq datasets. We further identified a
641 TLS pattern centered around the GC B cell program, recapitulating a spatial structure important
642 for anti-tumour immune response in multiple cancers including PDAC[111]. Finally, we introduced
643 a new model for inferring niches in TME and found reproducible niches associated with both
644 driver gene status as well as patient response to therapy.
645
646 Despite these findings, our study had multiple limitations. The lack of paired spatial proteomic
647 data prevented association of the inferred cell type co-occurrence with specific spatial tumour
648 architectures that have previously been shown to be predictive of survival in other cancer
649 types[13,19,112]. We are confident that future technological advances in proteomics and
650 transcriptomics[113] will lead to an increase in spatially profiled samples with high resolution that
651 will first be used to validate our approach based on scRNAseq data, and will then serve as
652 discovery platforms. In addition, while our discovery group was well balanced between four
653 datasets, the validation cohort was heavily weighted towards one patient cohort, which may
654 have limited the cell state program validation process though we note that, since there was no
655 overlap in samples, it still functions as a well-powered validation cohort. Our pipeline would

anyway accommodate for the inclusion of additional scRNAseq atlases of PDAC samples in the future that could be used as discovery or validation cohorts. Finally, the lack of well-standardized and curated clinical metadata across the single-cell cohorts limited the ability to directly associate cell states and niches with clinical covariates, instead necessitating mapping them to bulk datasets with standardized metadata.

Moving forward, we identified several venues to extend the project. Future spatial profiling of the PDAC microenvironment at cellular resolution in large cohorts will help to create a more comprehensive understanding of the cell state co-occurrence related to therapy resistance and disease progression. Integrating T cell receptor (TCR) or B cell receptor (BCR) sequencing analysis will help to delineate relationships between lymphocyte clonality with functional cell states, potentially revealing specific anti-tumour T cell clones and effects of therapy on the tumour immune cell population. In the long term, our analysis workflow may be adapted to multiple cancer types, leveraging the ever-increasing tumour single-cell data space to create a pan-cancer microenvironment cell state atlas, and identifying clinically relevant cell states across cancers. The increasing characterization of the tumour microenvironment at single-cell level will enable the identification of concomitant presence of multicellular niches in various types of cancer. Our approach unlocks a new way to analyse scRNAseq data that will not only facilitate prognosis and assessment of therapeutic response, but will provide valuable information to design specific therapies aimed at rewiring the milieu compatible with the co-existence of pro-tumorigenic cell states and niches.

Methods

Data curation

A total of 8 publicly available PDAC single-cell RNA-seq (scRNA-seq) datasets, which include 214 samples, were curated[4,35–40,52]. Most datasets were available for downloading through the Gene Expression Omnibus database. One dataset was available on author consent, in which case we contacted the authors for permission. Each dataset contains varying types of samples. We excluded samples which are of the following nature: metastasis, normal pancreas, peripheral blood, and adjacent normal tissue. In total, 163 samples labeled as primary PDAC tumour were retained for downstream analysis.

After sample filtering, we performed in house pre-processing and quality control given all datasets were available in the CellRanger output format. Doublet scores for single cells were calculated and their doublet status was called using doubletFinder_v3 from DoubletFinder v.2.0.3[114]. Cells labeled as doublets were removed from each sample. After doublet removal, cells with mitochondrial RNA content percentage greater than 20%, detected number of genes lower than 500, and total transcript counts lower than 1000 were removed as low quality cells. Finally, the UMI counts were normalized by per-cell size factor and log2-transformed with logNormCounts from scuttle v.1.8.4[115].

Cell annotation unification

We performed re-annotation to unify the cell annotation in each patient cohort in our curated dataset. First, we obtained a single-cell PDAC atlas built from divergent and heterogeneous datasets[7]. The 'celltype2' annotation from the atlas was selected as the reference labels after comparison of resolution and quality of all annotation levels in the atlas. Then, we performed unbiased cell type recognition at the single cell level using SingleR from SingleR v.2.0.0[116]. Briefly, SingleR first trains a classifier on the PDAC atlas reference dataset with reference labels by leveraging differential expression between labels, then the labels were assigned to each single cell in the query dataset using the classifier combined with iterative fine-tuning. We ran SingleR for each patient cohort in the curated dataset independently with the following settings: excluding ribosomal and mitochondrial genes in the reference, selecting top $500 * (2/3)^{log_2(N)}$ (N = # unique genes) DE genes with ranking by log-fold change, aggregating the reference into pseudo-bulk samples with default settings, and enabling fine-tuning with default settings. After annotation by SingleR, quality checks were performed by examining the SingleR predicted scores for each label in single cells in each cohort, the distribution of delta between assigned label score and median score across labels, and the distribution of scores for each label (**Supplementary Fig. 1C**). Final validation was done by examining known cell type marker expressions in each assigned cell type. Finally, the SingleR pruned.label column was selected and cell type labels were assigned as the unified annotation for single cells. To facilitate

716 downstream analysis, endocrine cells and neural cells were removed from the curated dataset,
717 and acinar and ductal cells were both labeled as epithelial cells. The eight consolidated cell type
718 labels were then shortened for readability.
719

720 **Discovery-Validation data split**

721 We first subset the full curated dataset by cell annotation, splitting it into eight cell type subsets:
722 Epithelial cell, Endothelial cell, Fibroblast, Monocyte/Macrophage, mDC, CD4 T cell, CD8 T/NK
723 cell, and B cell. We then split each cell type subset into two 4-cohort groups, one as the
724 discovery group and the other as the validation group. The discovery group contained 74
725 samples from the Peng, Steele, Werba, and Chan patient cohorts, while the validation group
726 contains 89 samples from the Chen, Lee, Lin, and Zhou patient cohorts. In total, the discovery-
727 validation split resulted in eight discovery-validation pairs of cell type subsets, with each pair
728 containing single cells assigned to one cell type.
729

730 **Reproducible analysis pipeline**

731 We built a comprehensive analysis pipeline to generate reproducible results from our curated
732 dataset and reduce human-introduced variations in analyses across the cell types and between
733 the discovery group and the validation group. We used Snakemake v.7.25.0[117] for parameter
734 and input-output control for all analysis scripts and visualization steps.
735

736 **Defining a set of robust lineage-specific cell state gene programs**

737 We carried out integrative NMF (iNMF) using LIGER[53] for the discovery and validation group of
738 patient cohorts across cell types independently to generate cell lineage-specific gene programs
739 that capture the phenotypic heterogeneity within each cell type in the PDAC microenvironment.
740 LIGER runs iNMF to decompose the expression matrices from multiple batches (patient
741 cohorts) and remove batch effect at the same time by maximizing a basis matrix shared
742 between cohorts while minimizing the multiple cohort-specific basis matrices that represent
743 patient cohort level biases. We defined the batches to be patient cohorts to focus on capturing
744 heterogeneity present across tumours through cell lineage-specific programs. We followed the
745 suggested workflow of LIGER to perform iNMF. Importantly, the following steps were performed
746 after the discovery-validation split as well as an 8-way cell type split within the discovery and the
747 validation group, leading to 16 independent gene program extraction steps, 8 independent
748 program validation steps, and 8 independent program merging steps.
749

750 **Program extraction**

751 First, we selected variable genes from non-ribosomal/mitochondrial protein-coding genes using
752 selectGenes from rliger v.1.0.0[53] with default settings. We then created a liger object using the
753 raw counts matrix, removed genes with zero expression, and normalized and scaled gene
754 expression following suggested workflow from rliger documentation before performing a subset

755 to the variable genes. Finally, after subsetting variable genes, we fitted iNMF on the expression
756 matrix using optimizeALS. We set rank (k) to 20, 10, 5, 10, 10, 20, 10, and 10 for program
757 extraction from epithelial cell, fibroblast, endothelial cell, CD4 T cell, CD8 T/NK cell,
758 monocyte/macrophage, mDC, and B cell respectively. We set lambda to 5 for all cell types as
759 we did not find it to be a major factor affecting extraction results. The gene loadings in each
760 program were obtained as the W matrix from the fitted liger object, and the program loadings in
761 single-cells were obtained as the H.norm matrix from the fitted liger object.
762

763 **Program validation**

764 Our program validation process was performed between discovery and validation for each cell
765 type. To start, we transformed the discovery and validation gene loading (W) matrix with log1p
766 in R. We then computed the spearman correlation and Minkowski distance (with p = 1) between
767 all possible pairs of discovery and validation programs within the cell type. The Hungarian
768 algorithm was used to generate a one-to-one relationship between discovery and validation
769 programs by permuting the rows of the distance and correlation matrices to minimize and
770 maximize the trace of the matrices respectively. After assigning the one-to-one partner of each
771 discovery program, we rejected any discovery programs whose gene loading correlation with
772 their validation program partner was below 0.5, or if the gene loading distance with their
773 validation program partner was above 25,000. We further confirm validated programs by
774 permutation test. Gene loadings in each validated program were permuted to generate a 1000-
775 sample test set and the correlation and distance to the validation program partner were
776 calculated for fitting a null distribution. A discovery program is finalized as a validated program if
777 the real distance and correlation are laid beyond 1% of the null distribution.
778

779 **Program merge**

780 We collapsed validated programs which were too similar to each other given NMF does not
781 guarantee orthogonality between the extracted factors. We looked at all possible pairs of
782 validated programs within each cell type and marked any pair whose gene loading correlation
783 was above 0.4 for merging. The merging step leveraged simple linear models. For a pair of
784 validated programs, e.g. sig-1 and sig-2, marked for merging, we first computed their explained
785 expression y by multiplying program loading in cell and gene loading in program respectively.
786 Then, we obtained the merged program's gene loading x by taking the mean of gene loadings in
787 sig-1 and sig-2. Finally, the single-cell loading of the merged program was obtained by fitting
788 lm(y ~ 0 + x) and pulling the coefficients. After merging, the newly merged program was
789 included for downstream analysis while sig-1 and sig-2 were excluded.
790

791 **Program rescoring in the validation group**

792 Using the robust set of validated and merged cell type specific programs, we rescored the
793 validation group single cells to get single-cell program loadings that can be compared with

794 program loadings in the discovery group in downstream analysis. The rescoring step was
795 performed for each cell type and each patient cohort independently, leading to 32 rescoring
796 steps.
797
798 To start, gene expression in a validation group patient cohort was normalized using
799 runSeuratNormalizeData from singleCellTK v.2.8.0[118] with default settings. Then, for rescoring
800 for programs of, for example, cell type A, we subset an expression matrix to cells annotated as
801 cell type A and variable genes used in the program extraction step. Finally, we performed non-
802 negative least squares to fit a simple linear model to the subset validation expression with cell
803 type A program gene loadings, and pulled out the estimated optimal coefficient matrix as the
804 rescored program loading in single-cells in validation. We used fcnnls from NMF v.0.26[119] with
805 default settings for this step.
806

807 **Program contextualization with known cell states**

808 We combined multiple approaches to get a comprehensive understanding of the biological
809 meaning of our validated cell type-specific programs. We curated multiple PDAC focused
810 malignant, stromal, and immune cell state marker lists from literature (**Supplementary Table 4**).
811 To associate the programs with the known cell states, we selected the top 50 loaded genes of
812 each program and performed a hypergeometric test with the cell state marker lists from its
813 respective cell type, using GeneOverlap v.1.34.0[120] in R. The human reference hg19 genome[121]
814 was used as the background for significance testing. We made a preliminary biological
815 interpretation of the programs based on significance (P-value < 1e-3) and high Jaccard index
816 with cell state marker lists. Then, the programs were further subjected to GO and KEGG
817 overrepresentation analysis using the clusterProfiler v.4.6.2 package[122]. Specifically, the
818 biological process subcategory of GO and KEGG PATHWAY gene sets were chosen for the
819 analysis. The top ten enriched gene sets in each category were used to help correct or confirm
820 our initial program interpretation. After the first correction, we also applied GSEA to the
821 programs with gene sets mined from the Curated Cancer Cell Atlas (3CA)[123] to resolve any
822 remaining uncertainty. We ran fgsea from fgsea v.1.24.0[124] with default settings, while setting
823 nPermSimple to 10,000, as suggested by the documentation. We used the top 10 positively
824 enriched 3CA gene programs to guide a second program interpretation correction step.
825 Programs that were not associated with known cell states previously were further characterized
826 after this analysis. Finally, we completed a manual check across programs by looking at the top
827 10 loaded genes, focusing on the programs that were yet to be contextualized and searching in
828 the literature for novel evidence of their biological meaning in PDAC. Any novel programs
829 remained were named by their top loaded genes.
830

831 **Ambient RNA program removal**

832 Due to technical limitation of droplet-based single-cell RNA-seq experiments, some reads
833 associated with a cell barcode may originate from ambient RNA[58], contributing to background
834 noises that are picked up by the granular cell state gene programs. Therefore, we denoted gene
835 programs in a cell lineage that are marked by lineage markers from other lineages to be
836 ambient RNA programs and removed such ambient programs from downstream analysis.
837

838 **Cell state co-occurrence**

839 To study co-occurrence between programs in the PDAC microenvironment, we constructed
840 sample level program loading profiles by averaging program loadings across single-cells. We
841 then calculated the correlation across samples between programs both within cell type and
842 across cell types. By comparing program loading correlations between the discovery group and
843 the validation group, we define program pairs whose loading correlation was above 0.4 and
844 differed by less than 0.25 in discovery and validation as high confidence pairs that tend to co-
845 occur with each other.
846

847 **Scoring RNA-seq samples for cell state programs and niches**

848 To score bulk PDAC tumour samples for our cell state gene programs, we performed gene set
849 variation analysis (GSVA) on each bulk sample using gsva from GSVA[125] ver. 1.46.0, with the
850 top-50 loaded genes in each gene programs, and used the program enrichment scores or
851 downstream analysis. To score bulk PDAC tumour samples for our niches, we performed
852 multiplication between the sample-by-program enrichment score matrix and the program-by-
853 niche niche abundance matrix, to obtain a sample-by-niche enrichment score matrix.
854

855 **Association between niches and KRAS variants**

856 KRAS variant information was obtained from published data and from collaborators. To account
857 for the patient-repeated measurements in the Chan cohort, we fitted a linear mixed model with
858 the *lmerTest* package in R with the KRAS mutation status as the fixed effect and the patient as
859 the random effect. For the TCGA cohort to estimate effect of KRAS variant on niche enrichment
860 scores we used a simple linear model while controlling for the expression subtype (Basal-like or
861 Classical) of the tumour.
862

863 **Cell state co-occurrence niches discovery model**

864 To further dissect the relationships between programs, we created a statistical model to
865 deconvolve program co-occurrence and discover multi-cellular population niches.
866

867 When considering why programs are (anti-) correlated in single-cells, there can be two reasons.
868 Firstly, within single cells, programs may be correlated due to gene regulation, epigenetics, etc.

We defined correlations resulting from these mechanisms cell intrinsic program correlations. Secondly, within patients, cells with a high program of one type may be present with (or exclude) cells with a different program. This could be due to paracrine signaling. We defined these correlations cell extrinsic program correlations. If, for example, there is positive correlation between loading of two programs only at the patient level, but not at the single cell level, then the correlation is dominated by cell extrinsic program correlation, and vice versa for cell intrinsic program correlation. We note that in both scenarios, the overall correlation between the programs is the same, even if the mechanisms and interpretation are different. So, it is important to distinguish the two types of correlation in our model.

Model data definition

We start by defining L as the number of niches we want to get, C as the number of cell types, N as the total number of cells, K as the total number of cell state programs, and P as the total number of patient samples. Therefore, we have:

$$\text{The set of all cell types: } \boldsymbol{C} = \{C_1, C_2, C_3, \dots, C_c\}$$

$$\sum_{i=1}^{C} K_{C_i} = K$$

$$\sum_{i=1}^{C} N_{C_i} = N$$

$$\text{The set of all patients: } \boldsymbol{P} = \{1, 2, 3, \dots, P\}$$

$$\sum_{i=1}^{C} P_{C_i} = P$$

$$0 < P_{C_i} < P \quad \forall\, C_i \in \boldsymbol{C}$$

Then, for each cell type C_i, we have K_{C_i} programs across N_{C_i} single-cells, and each single-cell belongs to one of P_{C_i} patients.

Next, we define some cell-type-specific variables. For each cell type C_i, we have a matrix $\boldsymbol{Y}_{C_i}$ storing the cell state program loadings in single-cells:

$$\boldsymbol{Y}_{C_i} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1K_{C_i}} \\ y_{21} & y_{22} & \cdots & y_{2K_{C_i}} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N_{C_i}1} & y_{N_{C_i}2} & \cdots & y_{N_{C_i}K_{C_i}} \end{pmatrix}$$

In $\boldsymbol{Y}_{C_i}$ each row represent one K_{C_i}-dimensional vector for a single-cell. This vector stores the loadings of each cell state program for cell type C_i in the cell. Of note, each single-cell is assigned to one cell type only, and thus only shows up in one of the $\boldsymbol{Y}_{C_i}$ matrices.

To further indicate the patient sample each single-cell belongs to, we introduce a sample indicator vector for each cell type C_i:

$$\boldsymbol{x}_{C_i} = \left(x_1, x_2, x_3, \dots, x_{N_{C_i}}\right) \text{ where } x_i \in \{1, 2, 3, \dots, P_{C_i}\} \ \forall i \in \{1, 2, 3, \dots, N_{C_i}\}$$

Model parameter definition

We first introduce the niche parameters by defining:

1. The niche loading matrix $\boldsymbol{H} = (h_{ij})_{P \times L}$
2. The niche basis matrix $\boldsymbol{W} = (w_{ij})_{L \times K}$
3. The patient level program loading estimate matrix considering extrinsic covariance $\boldsymbol{\Phi} = (\phi_{ij})_{P \times K} = \boldsymbol{H} \times \boldsymbol{W}$

Where each row w_l of $\boldsymbol{W}$ is a K-dimensional vector of program weights in niche l, and each row h_p of $\boldsymbol{H}$ is an L-dimensional vector of niche weights in sample p.

With the niche related parameters defined, we introduce:

1. The patient level program loading estimate matrix considering intrinsic + extrinsic covariance $\boldsymbol{M} = (\mu_{ij})_{P \times K}$
2. The cell-type-specific patient level program loading sub-matrices $\boldsymbol{M}_{C_i} = \boldsymbol{M}_{1:P, j_{C_i}:j_{C_i+K_{C_i}}}$

Where j_{C_i} is the index for the first program of cell type C_i in $\boldsymbol{M}$, specifically:

$$j_{C_1} = 0$$

$$j_{C_i} = j_{C_{i-1}} + K_{C_{i-1}} \ \forall i \in \{2, 3, 4, \dots, C\}$$

We lastly explicitly define, for each cell type C_i, the cell intrinsic program covariance matrix as:

$$\boldsymbol{\Sigma}_{intrinsic_{C_i}} = \begin{pmatrix} Var(program_1) & \cdots & Cov(program_1, program_{K_{C_i}}) \\ \vdots & \ddots & \vdots \\ Cov(program_{K_{C_i}}, program_1) & \cdots & Var(program_{K_{C_i}}) \end{pmatrix}$$

Model definition

First, we let

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad where$$
$$\boldsymbol{y} \in \mathbb{R}^k, \boldsymbol{\mu} \in \mathbb{R}^k, \boldsymbol{\Sigma} = (\sigma_{ij})_{k \times k}$$

We then consider cell extrinsic and intrinsic program correlations separately.

Cell extrinsic program correlations are correlations in the means of the program loadings for each patient. We thus introduce P mean vectors $\boldsymbol{\mu}_p$ (one $\boldsymbol{\mu}_p$ per patient) that are each K-dimensional and encode the mean program loading of patient p. Of note, Then, we can write:

$$\boldsymbol{\mu}_p \sim \mathcal{N}(\boldsymbol{\phi}_p, \boldsymbol{\Sigma}_{extrinsic})$$

Here, $\boldsymbol{\phi}_p$ is the prior on the patient mean program loading, and $\boldsymbol{\Sigma}_{extrinsic}$ is the program by program $K \times K$ covariance matrix of cell extrinsic program correlations.

Cell intrinsic program correlations are the correlation of the programs conditioned on the mean program loadings of each patient. We thus introduce N cell level program loading vectors $\boldsymbol{y}_{n_{C_i}}$ that are each K_{C_i}-dimensional and encode the program loadings for cell n_{C_i} where $C_i \in \boldsymbol{C}$. Therefore, we can write:

$$\boldsymbol{y}_{n_{C_i}} \sim \mathcal{N}(\boldsymbol{\mu}_{x_{n_{C_i}}}, \boldsymbol{\Sigma}_{intrinsic_{C_i}})$$

Here, $x_{n_{C_i}}$ selects the patient sample corresponding to cell n, so $\boldsymbol{\mu}_{x_{n_{C_i}}}$ corresponds to the mean program loadings of the patient corresponding to cell n_{C_i}. $\boldsymbol{\Sigma}_{intrinsic_{C_i}}$ is the $K \times K$ covariance matrix that denotes the program correlation after conditioning on the patient means. Using Gaussians allows us to interpret the total covariance as the sum of the intrinsic and extrinsic covariance, i.e. $\boldsymbol{\Sigma}_{total_{C_i}} = \boldsymbol{\Sigma}_{extrinsic} + \boldsymbol{\Sigma}_{intrinsic_{C_i}}$.

Together, we can write the hierarchical model:

$$\boldsymbol{\phi}_p \sim \mathcal{N}(0, 10)\ \forall p \in \boldsymbol{P}$$

$$\boldsymbol{\Sigma}_{intrinsic_{C_i}} \sim InverseWishart(\cdot)\ \forall C_i \in \boldsymbol{C}$$

$$\boldsymbol{\Sigma}_{extrinsic} \sim InverseWishart(\cdot)$$

$$\boldsymbol{\mu}_p \sim \mathcal{N}(\boldsymbol{\phi}_p, \boldsymbol{\Sigma}_{extrinsic})$$

$$\boldsymbol{y}_{n_{C_i}} \sim \mathcal{N}(\boldsymbol{\mu}_{x_{n_{C_i}}}, \boldsymbol{\Sigma}_{intrinsic_{C_i}})$$

Here,

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \\ \vdots \\ \boldsymbol{\phi}_P \end{bmatrix} \quad \boldsymbol{M} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_P \end{bmatrix} \quad \boldsymbol{Y}_{C_i} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_{n_{C_i}} \end{bmatrix}$$

Target function

The target function of the model:

$$\mathbb{P}(\boldsymbol{\Phi}, \boldsymbol{M} \mid \boldsymbol{Y}) \propto \mathbb{P}(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{M}, \boldsymbol{\Sigma} \mid \boldsymbol{Y}) \propto \mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{\Sigma})\mathbb{P}(\boldsymbol{M} \mid \boldsymbol{W}, \boldsymbol{H})\mathbb{P}(\boldsymbol{W})\mathbb{P}(\boldsymbol{H})\mathbb{P}(\boldsymbol{\Sigma})$$

The joint posterior is then:

$$\mathbb{P}(\boldsymbol{\Phi}, \boldsymbol{M} \mid \boldsymbol{Y}) \propto \left(\prod_{i=1}^{C} \mathbb{P}(\boldsymbol{\Sigma_i})\mathbb{P}(\boldsymbol{Y_i} \mid \boldsymbol{M}, \boldsymbol{\Sigma_i})\right) \cdot \left(\prod_{i=1}^{P} \mathbb{P}(\boldsymbol{M} \mid \boldsymbol{W}, \boldsymbol{H_i}) \cdot \mathbb{P}(\boldsymbol{H_i})\right) \cdot \left(\prod_{i=1}^{L} \mathbb{P}(\boldsymbol{W_i})\right) \propto$$

$$\left(\prod_{i=1}^{C}\prod_{n=1}^{N_i} \mathbb{P}(\boldsymbol{\Sigma_i})\mathbb{P}(\boldsymbol{Y}_n \mid \boldsymbol{M}_{x_n}, \boldsymbol{\Sigma_i})\right) \cdot \left(\prod_{i=1}^{P} \mathbb{P}(\boldsymbol{M} \mid \boldsymbol{W}, \boldsymbol{H_i}) \cdot \mathbb{P}(\boldsymbol{H_i})\right) \cdot \left(\prod_{i=1}^{L} \mathbb{P}(\boldsymbol{W_i})\right)$$

To encourage the model to find more distinct niches, we add an orthogonality term to the model target function. The upper triangular matrix of the transpose cross product of the niche basis matrix $\boldsymbol{W}$ is used to calculate a mean as a proxy for niche similarity. This term is weighted by a hyperparameter λ and minimized in the target function. Therefore, using maximum a posteriori inference, we maximize the new target function including the log joint posterior and the orthogonality term:

$$\sum_{i=1}^{C}\left(log\mathbb{P}(\boldsymbol{\Sigma_i}) + \sum_{n=1}^{N_i} log\mathbb{P}(\boldsymbol{Y}_n \mid \boldsymbol{M}_{\boldsymbol{x_n}}, \boldsymbol{\Sigma_i})\right) + \sum_{i=1}^{P}(log\mathbb{P}(\boldsymbol{M} \mid \boldsymbol{W}, \boldsymbol{H_i}) + log\mathbb{P}(\boldsymbol{H_i})) + \sum_{i=1}^{L} log\mathbb{P}(\boldsymbol{W_i}) + \lambda \cdot Mean\big(upper_tri(\boldsymbol{W}^{\intercal}\boldsymbol{W})\big)$$

Data preparation for the niche discovery model

To increase the power of our model, we removed patient samples which contained fewer than 20 single-cells. To speed up the inference process, we sampled down each patient sample to 1500 single-cells. For each cell state programs, we scaled the program loadings in single-cells across all cells to get z-scores as input to the model.

Model initialization and optimization

We coded the niche discovery model using the statistical programing language Stan. We used cmdstan_model from the R package cmdStanR ver. 0.5.3[126] to compile the model. Then, we used optimize from cmdStanR to initialize and fit the model to our data, with settings: algorithm="lbfgs", tol_param=$1e^{-7}$, and iter=8000. The model parameters were initialized as specified below.

Model initialization was provided for $\boldsymbol{\Sigma}_{intrinsic}$ for each cell type and for $\boldsymbol{W}$ and $\boldsymbol{H}$. We performed NMF on the sample level cell state program loading mean matrix with rank=L to obtain the initial values of the niche matrices. Covariance between program loadings within respective cell types were input as the initial values of $\boldsymbol{\Sigma}_{intrinsic}$.

Funding

This work was supported by funding from CIHR project grant PJT175270 (KC), NSERC Discovery grant RGPIN-2020-04083 (KC), and support from the Data Science Institute at University of Toronto (CY). This research was undertaken, in part, thanks to funding from the Canada Research Chairs Program.

Data and code availability

All data and code required to reproduce this study can be found at https://zenodo.org/records/13332977 and https://github.com/camlab-bioml/PDAC_cellular_niche_analyses.

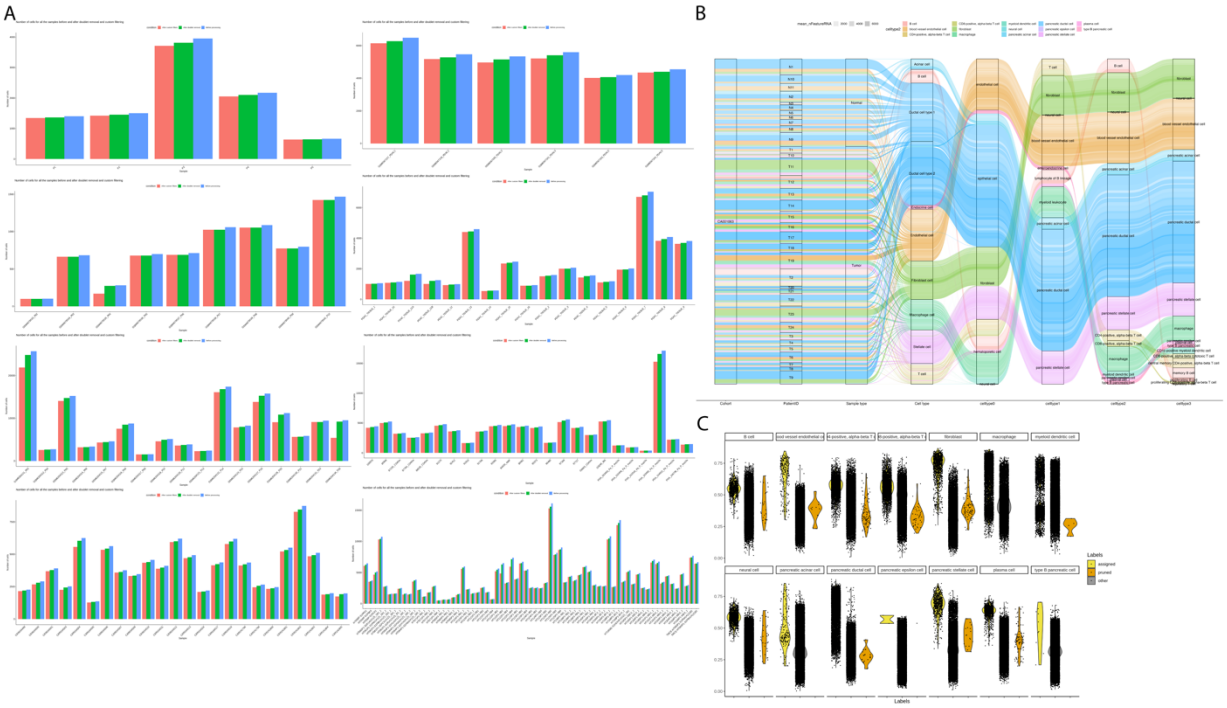
Contributions

Project conception: CY, KC. Data curation: CY, SG, TTJ, AZ. Data analysis and coding: CY, MG, SG, GHJ, TTJ, AZ. Result interpretation and manuscript writing: CY, KC. Project direction: GO, FN, KC.

Acknowledgements

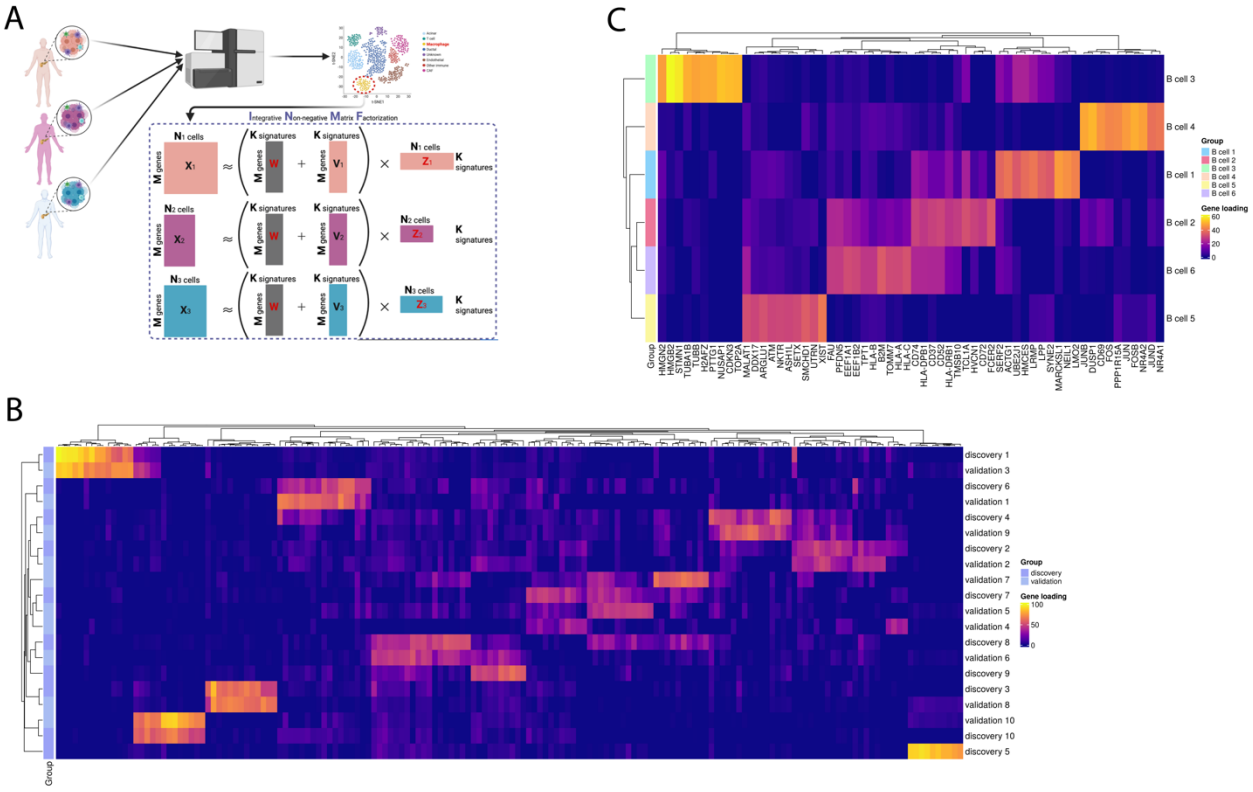
We thank Life Science Editors for editing services (www.lifescienceeditors.com).

Supplemental figures

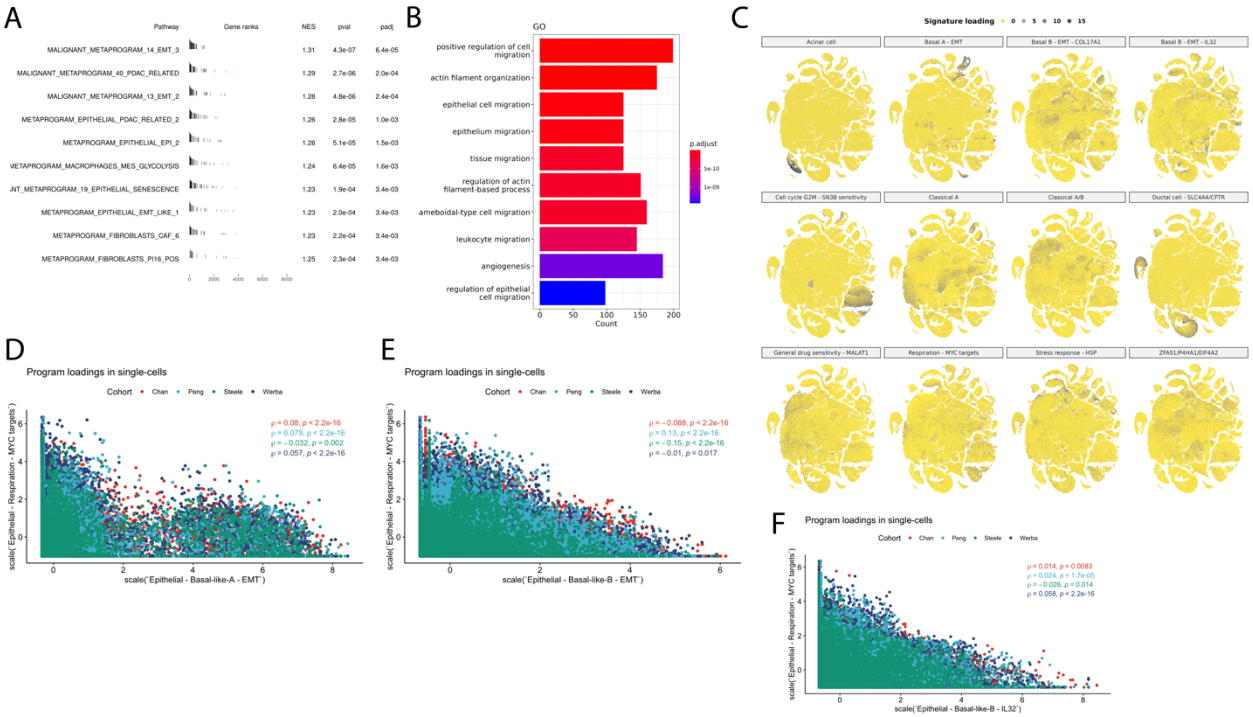


Supplementary Figure 1. A, Bar plots showing the number of cells in each primary tumour sample before quality control (blue), after doublet removal (green), and after doublet removal and quality filtering (red). Each panel shows samples in one patient cohort. **B**, Cell type

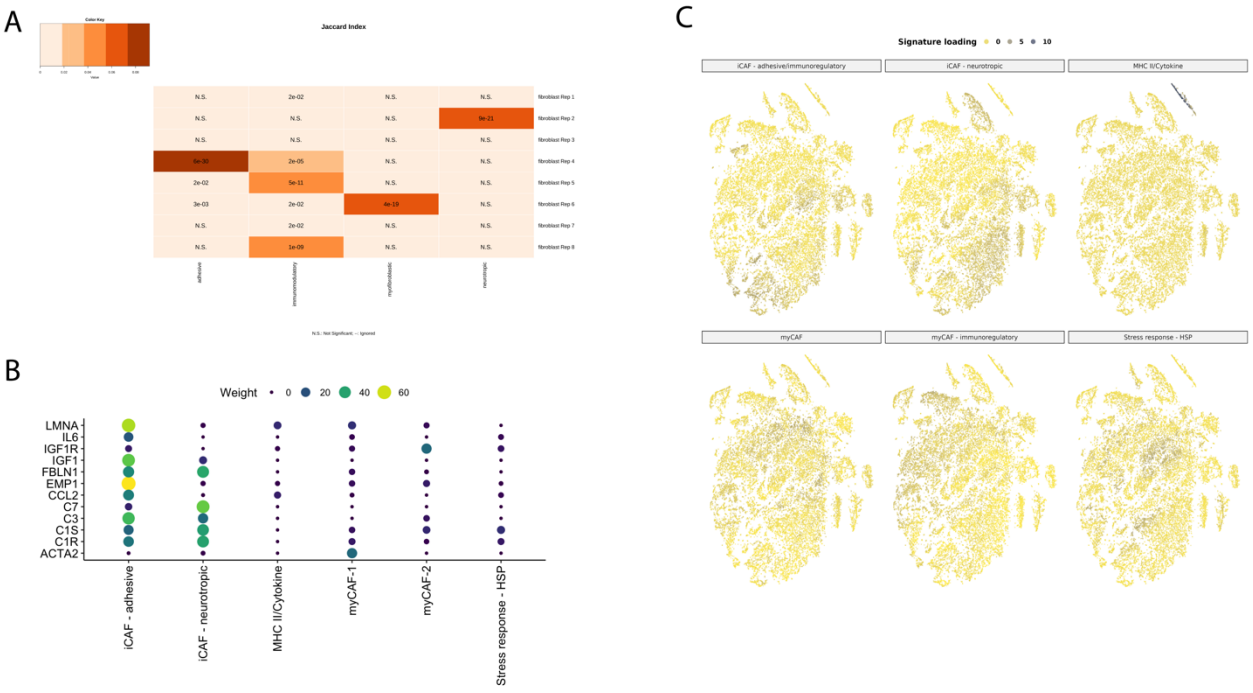
information fields in the PDAC single cell atlas used as reference for in house annotation. **C**, Violin plots showing score distributions for each cell type label in the Steele patient cohort.



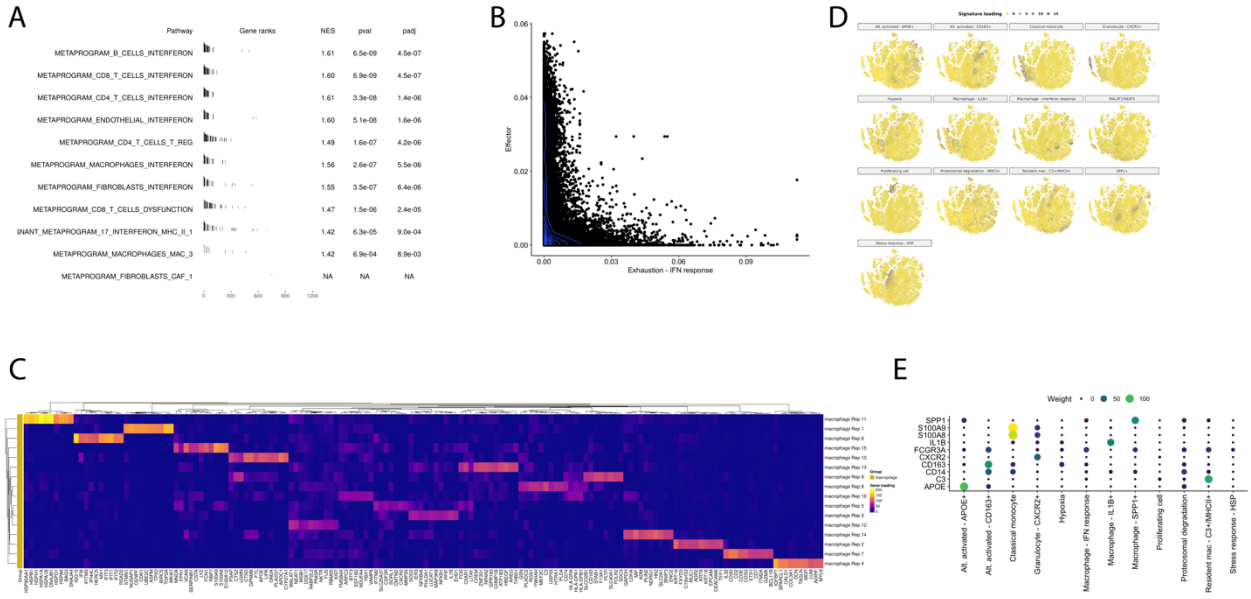
Supplementary Figure 2. A, Cell lineage specific gene programs extraction by iNMF. The schematic shows the extraction process. **B**, Top 10 loaded genes in the discovery and validation fibroblast programs. **C**, Top 10 loaded genes in the validated B cell programs extracted from B cells in the discovery group.



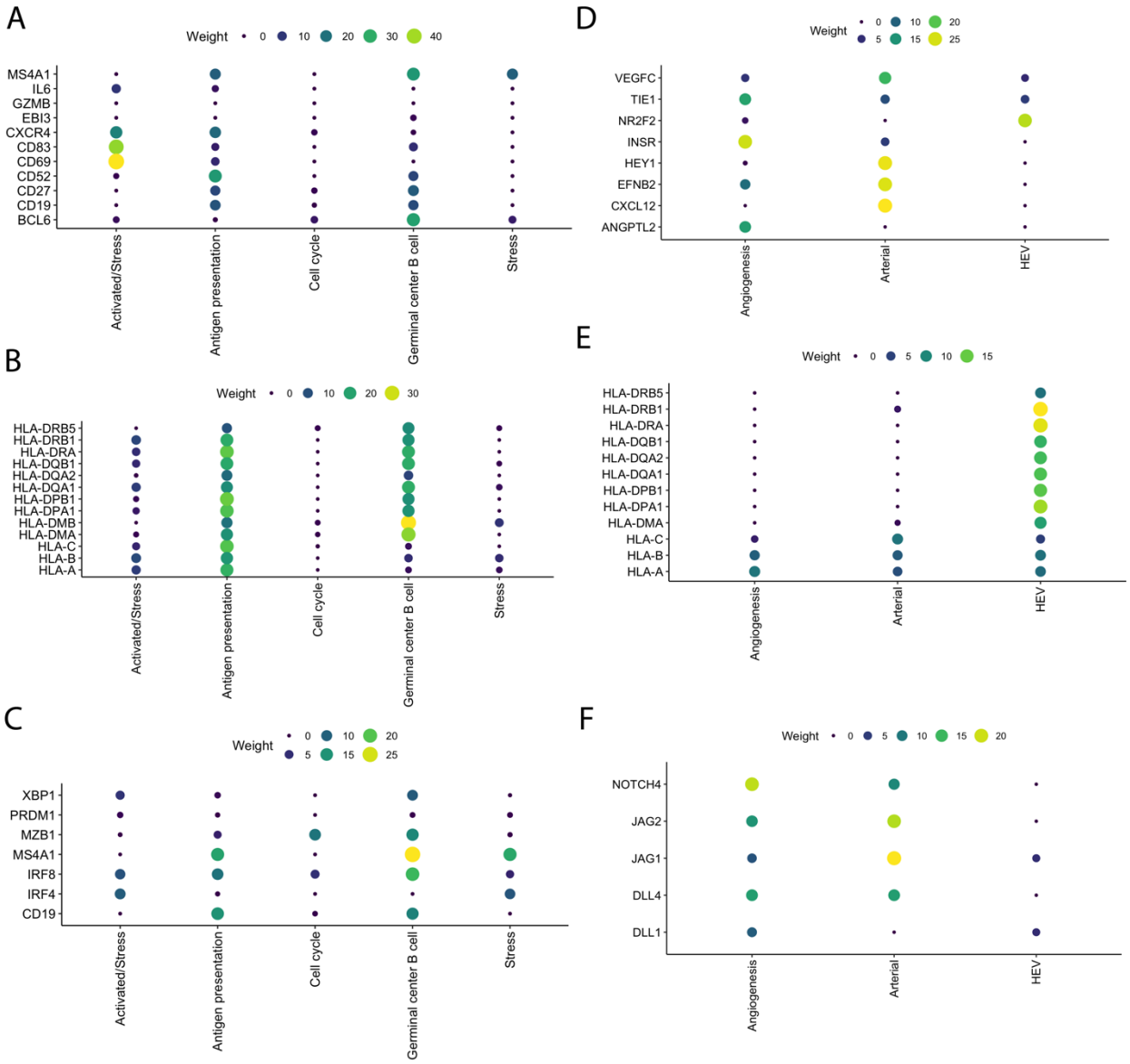
Supplementary Figure 3. A, GSEA of 3CA gene programs in the epithelial cell gene program characterized as Basal-like-A/EMT. **B**, GO enrichment analysis in the epithelial cell gene program characterized as Basal-like-A/EMT. **C**, Raw loadings of epithelial gene programs in single cells in the discovery group. **D-F**, Loadings of selected epithelial cell state programs in single-cells from the discovery group.



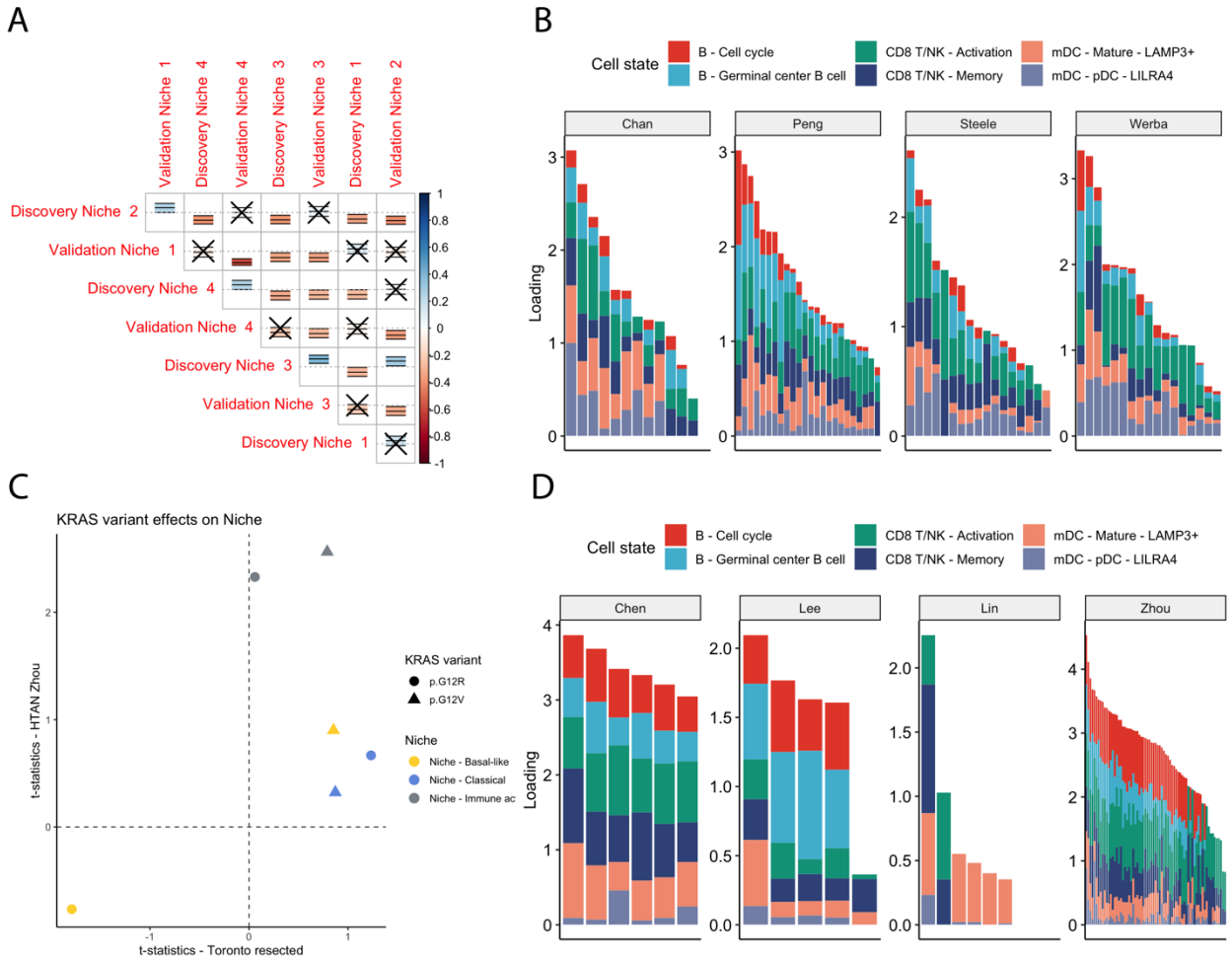
Supplementary Figure 4. A, Gene set overlap analysis between the top 50 genes of fibroblast programs and published fibroblast cell state marker gene lists for PDAC. **B**, Selected gene weights in fibroblast programs. **C**, Raw loadings of fibroblast gene programs in single cells in the Discovery group.



Supplementary Figure 5. A, GSEA of 3CA gene programs in the CD8 T cell gene program characterized as exhaustion/interferon response. **B**, Program loadings in the CD8 T cells. **C**, Loadings of the top 10 weighted genes in each macrophage program. **D**, Raw loadings of macrophage gene programs in single cells in the Discovery group. **E**, Selected gene weights in macrophage programs.



Supplementary Figure 6. A-C, Selected gene weights in B cell programs. **D-F**, Selected gene weights in endothelial cell programs.



Supplementary Figure 7. A, Correlations between the program weights in each niche between the discovery and the validation. Each cell shows the correlation between the niche pair. The dotted lines in each cell denote 0 correlation. The boxes in each cell shows the correlation (middle line) and 95% confidence interval (top and bottom of box). **B and D**, Loadings of the TLS related programs in the discovery patient samples (**B**) and in the validation patient samples (**D**). **C**, T-statistics from linear regression models fitting relationships between KRAS variants and niche loadings in bulk RNA-seq tumour samples.

References

1. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

2. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells

1068 Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
1069 3. Zhang, S. *et al.* Single cell transcriptomic analyses implicate an immunosuppressive tumor
1070 microenvironment in pancreatic cancer liver metastasis. *Nat. Commun.* **14**, 5123 (2023).
1071 4. Cui Zhou, D. *et al.* Spatially restricted drivers and transitional cell populations cooperate
1072 with the microenvironment in untreated and chemo-resistant pancreatic cancer. *Nat. Genet.*
1073 **54**, 1390–1405 (2022).
1074 5. Hwang, W. L. *et al.* Single-nucleus and spatial transcriptome profiling of pancreatic cancer
1075 identifies multicellular dynamics associated with neoadjuvant treatment. *Nat. Genet.* **54**,
1076 1178–1191 (2022).
1077 6. McAndrews, K. M. *et al.* Identification of functional heterogeneity of carcinoma-associated
1078 fibroblasts with distinct IL6-mediated therapy resistance in pancreatic cancer. *Cancer*
1079 *Discov.* **12**, 1580–1597 (2022).
1080 7. Chijimatsu, R. *et al.* Establishment of a reference single-cell RNA sequencing dataset for
1081 human pancreatic adenocarcinoma. *iScience* **25**, 104659 (2022).
1082 8. Falcomatà, C. *et al.* Selective multi-kinase inhibition sensitizes mesenchymal pancreatic
1083 cancer to immune checkpoint blockade by remodeling the tumor microenvironment. *Nat.*
1084 *Cancer* **3**, 318–336 (2022).
1085 9. Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by
1086 mass cytometry. *Nat. Methods* **11**, 417–422 (2014).
1087 10. Angelo, M. *et al.* Multiplexed ion beam imaging of human breast tumors. *Nat. Med.* **20**,
1088 436–442 (2014).
1089 11. Janesick, A. *et al.* High resolution mapping of the tumor microenvironment using integrated
1090 single-cell, spatial and in situ analysis. *Nat. Commun.* **14**, 8353 (2023).
1091 12. He, S. *et al.* High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue
1092 by spatial molecular imaging. *Nat. Biotechnol.* **40**, 1794–1806 (2022).
1093 13. Jackson, H. W. *et al.* The single-cell pathology landscape of breast cancer. *Nature* **578**,

1094 615–620 (2020).
1095 14. Kim, S. *et al.* Integrative analysis of spatial and single-cell transcriptome data from human
1096 pancreatic cancer reveals an intermediate cancer cell population associated with poor
1097 prognosis. *Genome Med.* **16**, 20 (2024).
1098 15. Moncada, R. *et al.* Integrating microarray-based spatial transcriptomics and single-cell
1099 RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat.*
1100 *Biotechnol.* **38**, 333–342 (2020).
1101 16. Storrs, E. P. *et al.* High-dimensional deconstruction of pancreatic cancer identifies tumor
1102 microenvironmental and developmental stemness features that predict survival. *NPJ*
1103 *Precis. Oncol.* **7**, 105 (2023).
1104 17. Luca, B. A. *et al.* Atlas of clinically distinct cell states and ecosystems across human solid
1105 tumors. *Cell* **184**, 5482-5496.e28 (2021).
1106 18. Wu, S. Z. *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat.*
1107 *Genet.* **53**, 1334–1347 (2021).
1108 19. Sorin, M. *et al.* Single-cell spatial landscapes of the lung tumour immune microenvironment.
1109 *Nature* **614**, 548–554 (2023).
1110 20. Freed-Pastor, W. A. *et al.* The CD155/TIGIT axis promotes and maintains immune evasion
1111 in neoantigen-expressing pancreatic cancer. *Cancer Cell* **39**, 1342-1360.e14 (2021).
1112 21. Kersten, K. *et al.* Spatiotemporal co-dependency between macrophages and exhausted
1113 CD8+ T cells in cancer. *Cancer Cell* **40**, 624-638.e9 (2022).
1114 22. Keren, L. *et al.* MIBI-TOF: A multiplexed imaging platform relates cellular phenotypes and
1115 tissue structure. *Sci. Adv.* **5**, eaax5851 (2019).
1116 23. Ni, Z. *et al.* SpotClean adjusts for spot swapping in spatial transcriptomics data. *Nat.*
1117 *Commun.* **13**, 2971 (2022).
1118 24. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early
1119 organogenesis. *Nature* **566**, 490–495 (2019).

1120 25. Smillie, C. S. *et al.* Intra- and inter-cellular rewiring of the human colon during ulcerative
1121 colitis. *Cell* **178**, 714-730.e22 (2019).
1122 26. Büttner, M., Ostner, J., Müller, C. L., Theis, F. J. & Schubert, B. scCODA is a Bayesian
1123 model for compositional single-cell data analysis. *Nat. Commun.* **12**, 6876 (2021).
1124 27. Rozenblatt-Rosen, O. *et al.* Building a high-quality Human Cell Atlas. *Nat. Biotechnol.* **39**,
1125 149–153 (2021).
1126 28. Rozenblatt-Rosen, O. *et al.* The Human Tumor Atlas Network: Charting tumor transitions
1127 across space and time at single-cell resolution. *Cell* **181**, 236–249 (2020).
1128 29. Terekhanova, N. V. *et al.* Epigenetic regulation during cancer transitions across 11 tumour
1129 types. *Nature* **623**, 432–441 (2023).
1130 30. Heiser, C. N. *et al.* Molecular cartography uncovers evolutionary and microenvironmental
1131 dynamics in sporadic colorectal tumors. *Cell* **186**, 5620-5637.e16 (2023).
1132 31. Johnson, B. E. *et al.* An omic and multidimensional spatial atlas from serial biopsies of an
1133 evolving metastatic breast cancer. *Cell Rep. Med.* **3**, 100525 (2022).
1134 32. Yanagawa, J. *et al.* Single-cell characterization of pulmonary nodules implicates
1135 suppression of immunosurveillance across early stages of lung adenocarcinoma. *Cancer*
1136 *Res.* **83**, 3305–3319 (2023).
1137 33. Glasner, A. *et al.* Conserved transcriptional connectivity of regulatory T cells in the tumor
1138 microenvironment informs new combination cancer therapy strategies. *Nat. Immunol.* **24**,
1139 1020–1035 (2023).
1140 34. Chan, J. M. *et al.* Signatures of plasticity, metastasis, and immunosuppression in an atlas
1141 of human small cell lung cancer. *Cancer Cell* **39**, 1479-1496.e18 (2021).
1142 35. Chan-Seng-Yue, M. *et al.* Transcription phenotypes of pancreatic cancer are driven by
1143 genomic events during tumor evolution. *Nat. Genet.* **52**, 231–240 (2020).
1144 36. Peng, J. *et al.* Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant
1145 progression in pancreatic ductal adenocarcinoma. *Cell Res.* **29**, 725–738 (2019).

1146 37. Steele, N. G. *et al.* Multimodal mapping of the tumor and peripheral blood immune
1147 landscape in human pancreatic cancer. *Nature Cancer* **1**, 1097–1112 (2020).
1148 38. Werba, G. *et al.* Single-cell RNA sequencing reveals the effects of chemotherapy on human
1149 pancreatic adenocarcinoma and its tumor microenvironment. *Nat. Commun.* **14**, 797
1150 (2023).
1151 39. Lee, J. J. *et al.* Elucidation of tumor-stromal heterogeneity and the ligand-receptor
1152 interactome by single-cell transcriptomics in real-world pancreatic cancer biopsies. *Clin.*
1153 *Cancer Res.* **27**, 5912–5921 (2021).
1154 40. Lin, W. *et al.* Single-cell transcriptome analysis of tumor and stromal compartments of
1155 pancreatic ductal adenocarcinoma primary tumors and metastatic lesions. *Genome Med.*
1156 **12**, 80 (2020).
1157 41. Chen, K. *et al.* Single-cell RNA-seq reveals dynamic change in tumor microenvironment
1158 during pancreatic ductal adenocarcinoma malignant progression. *eBioMedicine* **66**, (2021).
1159 42. Provenzano, P. P. *et al.* Enzymatic targeting of the stroma ablates physical barriers to
1160 treatment of pancreatic ductal adenocarcinoma. *Cancer Cell* **21**, 418–429 (2012).
1161 43. Ramanathan, R. K. *et al.* Phase IB/II Randomized Study of FOLFIRINOX Plus Pegylated
1162 Recombinant Human Hyaluronidase Versus FOLFIRINOX Alone in Patients With
1163 Metastatic Pancreatic Adenocarcinoma: SWOG S1313. *J. Clin. Oncol.* **37**, 1062–1069
1164 (2019).
1165 44. Öhlund, D. *et al.* Distinct populations of inflammatory fibroblasts and myofibroblasts in
1166 pancreatic cancer. *J. Exp. Med.* **214**, 579–596 (2017).
1167 45. Zhang, T., Ren, Y., Yang, P., Wang, J. & Zhou, H. Cancer-associated fibroblasts in
1168 pancreatic ductal adenocarcinoma. *Cell Death Dis.* **13**, 1–11 (2022).
1169 46. Caronni, N. *et al.* IL-1β+ macrophages fuel pathogenic inflammation in pancreatic cancer.
1170 *Nature* **623**, 415–422 (2023).
1171 47. LaRue, M. M. *et al.* Metabolic reprogramming of tumor-associated macrophages by

1172 collagen turnover promotes fibrosis in pancreatic cancer. *Proceedings of the National*
1173 *Academy of Sciences* **119**, e2119168119 (2022).
1174 48. Cassetta, L. & Pollard, J. W. Targeting macrophages: therapeutic approaches in cancer.
1175 *Nat. Rev. Drug Discov.* **17**, 887–904 (2018).
1176 49. Moffitt, R. A. *et al.* Virtual microdissection identifies distinct tumor- and stroma-specific
1177 subtypes of pancreatic ductal adenocarcinoma. *Nat. Genet.* **47**, 1168–1178 (2015).
1178 50. Oh, K. *et al.* Coordinated single-cell tumor microenvironment dynamics reinforce pancreatic
1179 cancer subtype. *Nat. Commun.* **14**, 1–13 (2023).
1180 51. Grünwald, B. T. *et al.* Spatially confined sub-tumor microenvironments in pancreatic cancer.
1181 *Cell* **184**, 5577-5592.e18 (2021).
1182 52. Chen, K. *et al.* Immune profiling and prognostic model of pancreatic cancer using
1183 quantitative pathology and single-cell RNA sequencing. *J. Transl. Med.* **21**, 210 (2023).
1184 53. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of
1185 Brain Cell Identity. *Cell* **177**, 1873-1887.e17 (2019).
1186 54. Raghavan, S. *et al.* Microenvironment drives cell state, plasticity, and drug response in
1187 pancreatic cancer. *Cell* **184**, 6119-6137.e26 (2021).
1188 55. Elyada, E. *et al.* Cross-species single-cell analysis of pancreatic ductal adenocarcinoma
1189 reveals antigen-presenting cancer-associated fibroblasts. *Cancer Discov.* **9**, 1102–1123
1190 (2019).
1191 56. Kunes, R. Z., Walle, T., Land, M., Nawy, T. & Pe'er, D. Supervised discovery of
1192 interpretable gene programs from single-cell data. *Nat. Biotechnol.* **42**, 1084–1095 (2024).
1193 57. Kuhn, H. W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **2**,
1194 83–97 (1955).
1195 58. Janssen, P. *et al.* The effect of background noise and its removal on the analysis of single-
1196 cell expression data. *Genome Biol.* **24**, 140 (2023).
1197 59. Nishida, A., Andoh, A., Inatomi, O. & Fujiyama, Y. Interleukin-32 expression in the

1198 pancreas. *J. Biol. Chem.* **284**, 17868–17876 (2009).
1199 60. Zhu, H. *et al.* S100A14 promotes progression and gemcitabine resistance in pancreatic
1200 cancer. *Pancreatology* **21**, 589–598 (2021).
1201 61. Low, R. R. J. *et al.* S100 family proteins are linked to organoid morphology and EMT in
1202 pancreatic cancer. *Cell Death Differ.* **30**, 1155–1165 (2023).
1203 62. Farrell, A. S. *et al.* MYC regulates ductal-neuroendocrine lineage plasticity in pancreatic
1204 ductal adenocarcinoma associated with poor outcome and chemoresistance. *Nat.*
1205 *Commun.* **8**, 1–12 (2017).
1206 63. Ala, M. Target c-Myc to treat pancreatic cancer. *Cancer Biol. Ther.* **23**, 34–50 (2022).
1207 64. Shi, D. D. *et al.* Therapeutic avenues for cancer neuroscience: translational frontiers and
1208 clinical opportunities. *Lancet Oncol.* **23**, e62–e74 (2022).
1209 65. Rossi Sebastiano, M. *et al.* ACSL3-PAI-1 signaling axis mediates tumor-stroma cross-talk
1210 promoting pancreatic cancer progression. *Sci. Adv.* **6**, eabb9200 (2020).
1211 66. Loncle, C. *et al.* The pancreatitis-associated protein VMP1, a key regulator of inducible
1212 autophagy, promotes KrasG12D-mediated pancreatic cancer initiation. *Cell Death Dis.* **7**,
1213 e2295–e2295 (2016).
1214 67. Gao, H. *et al.* The expression of versican and its role in pancreatic neuroendocrine tumors.
1215 *Pancreatology* **20**, 142–147 (2020).
1216 68. Rojas, L. A. *et al.* Personalized RNA neoantigen vaccines stimulate T cells in pancreatic
1217 cancer. *Nature* **618**, 144–150 (2023).
1218 69. Gulhati, P. *et al.* Targeting T cell checkpoints 41BB and LAG3 and myeloid cell
1219 CXCR1/CXCR2 results in antitumor immunity and durable response in pancreatic cancer.
1220 *Nat. Cancer* **4**, 62–80 (2023).
1221 70. Kemp, S. B. *et al.* Apolipoprotein E Promotes Immune Suppression in Pancreatic Cancer
1222 through NF-κB-Mediated Production of CXCL1. *Cancer Res.* **81**, 4305–4318 (2021).
1223 71. Astuti, Y. *et al.* Efferocytosis reprograms the tumor microenvironment to promote pancreatic

1224 cancer liver metastasis. *Nat. Cancer* **5**, 774–790 (2024).
1225 72. Nallasamy, P. *et al.* Pancreatic Tumor Microenvironment Factor Promotes Cancer
1226 Stemness via SPP1–CD44 Axis. *Gastroenterology* **161**, 1998-2013.e7 (2021).
1227 73. Bill, R. *et al.* CXCL9:SPP1 macrophage polarity identifies a network of cellular programs
1228 that control human cancers. *Science* **381**, 515–524 (2023).
1229 74. Cheng, S. *et al.* A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid
1230 cells. *Cell* **184**, 792-809.e23 (2021).
1231 75. Muthalagu, N. *et al.* Repression of the Type I Interferon Pathway Underlies MYC- and
1232 KRAS-Dependent Evasion of NK and B Cells in Pancreatic Ductal Adenocarcinoma.
1233 *Cancer Discov.* **10**, 872–887 (2020).
1234 76. Morgan, D. & Tergaonkar, V. Unraveling B cell trajectories at single cell resolution. *Trends*
1235 *Immunol.* **43**, 210–229 (2022).
1236 77. Fridman, W. H. *et al.* B cells and tertiary lymphoid structures as determinants of tumour
1237 immune contexture and clinical outcome. *Nat. Rev. Clin. Oncol.* **19**, 441–457 (2022).
1238 78. Young, C. & Brink, R. The unique biology of germinal center B cells. *Immunity* **54**, 1652–
1239 1664 (2021).
1240 79. Senturk, Z. N., Akdag, I., Deniz, B. & Sayi-Yazgan, A. Pancreatic cancer: Emerging field of
1241 regulatory B-cell-targeted immunotherapies. *Front. Immunol.* **14**, 1152551 (2023).
1242 80. Hua, Y. *et al.* Cancer immunotherapies transition endothelial cells into HEVs that generate
1243 TCF1+ T lymphocyte niches through a feed-forward loop. *Cancer Cell* **40**, 1600-1618.e10
1244 (2022).
1245 81. Su, T. *et al.* Single-cell analysis of early progenitor cells that build coronary arteries. *Nature*
1246 **559**, 356–362 (2018).
1247 82. Saharinen, P., Eklund, L. & Alitalo, K. Therapeutic targeting of the angiopoietin-TIE
1248 pathway. *Nat. Rev. Drug Discov.* **16**, 635–661 (2017).
1249 83. Ivins, S. *et al.* The CXCL12/CXCR4 axis plays a critical role in coronary artery

1250 development. *Dev. Cell* **33**, 455–468 (2015).
1251 84. Fischer, A., Schumacher, N., Maier, M., Sendtner, M. & Gessler, M. The Notch target
1252 genes Hey1 and Hey2 are required for embryonic vascular development. *Genes Dev.* **18**,
1253 901–911 (2004).
1254 85. Fish, J. E. & Wythe, J. D. The molecular regulation of arteriovenous specification and
1255 maintenance. *Dev. Dyn.* **244**, 391–409 (2015).
1256 86. Lawson, N. D., Vogel, A. M. & Weinstein, B. M. sonic hedgehog and vascular endothelial
1257 growth factor act upstream of the Notch pathway during arterial endothelial differentiation.
1258 *Dev. Cell* **3**, 127–136 (2002).
1259 87. Chu, Y. *et al.* Pan-cancer T cell atlas links a cellular stress response state to
1260 immunotherapy resistance. *Nat. Med.* **29**, 1550–1562 (2023).
1261 88. Elyada, E. *et al.* Cross-Species Single-Cell Analysis of Pancreatic Ductal Adenocarcinoma
1262 Reveals Antigen-Presenting Cancer-Associated Fibroblasts. *Cancer Discovery* vol. 9 1102–
1263 1123 Preprint at https://doi.org/10.1158/2159-8290.cd-19-0094 (2019).
1264 89. Li, K. *et al.* Myeloid-derived suppressor cells as immunosuppressive regulators and
1265 therapeutic targets in cancer. *Signal Transduction and Targeted Therapy* **6**, 1–25 (2021).
1266 90. Bärthel, S., Falcomatà, C., Rad, R., Theis, F. J. & Saur, D. Single-cell profiling to explore
1267 pancreatic cancer heterogeneity, plasticity and response to therapy. *Nat. Cancer* **4**, 454–
1268 467 (2023).
1269 91. Goveia, J. *et al.* An Integrated Gene Expression Landscape Profiling Approach to Identify
1270 Lung Tumor Endothelial Cell Heterogeneity and Angiogenic Candidates. *Cancer Cell* **37**,
1271 421 (2020).
1272 92. Geldhof, V. *et al.* Single cell atlas identifies lipid-processing and immunomodulatory
1273 endothelial cells in healthy and malignant breast. *Nat. Commun.* **13**, 1–19 (2022).
1274 93. Nieto, P. *et al.* A single-cell tumor immune atlas for precision oncology. *Genome Res.* **31**,
1275 1913–1926 (2021).

1276 94. Gorgun, C. *et al.* Synergistic role of three dimensional niche and hypoxia on conservation of
1277 cancer stem cell phenotype. *Int. J. Biol. Macromol.* **90**, 20–26 (2016).
1278 95. Till, J. E. *et al.* Circulating KRAS G12D but not G12V is associated with survival in
1279 metastatic pancreatic ductal adenocarcinoma. *Nat. Commun.* **15**, 1–12 (2024).
1280 96. Leidner, R. *et al.* Neoantigen T-cell receptor gene therapy in pancreatic cancer. *N. Engl. J.*
1281 *Med.* **386**, 2112–2119 (2022).
1282 97. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality
1283 Survival Outcome Analytics. *Cell* **173**, 400-416.e11 (2018).
1284 98. Cancer Genome Atlas Research Network. Electronic address:
1285 andrew_aguirre@dfci.harvard.edu & Cancer Genome Atlas Research Network. Integrated
1286 Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell* **32**, 185-
1287 203.e13 (2017).
1288 99. Rashid, N. U. *et al.* Purity Independent Subtyping of Tumors (PurIST), A clinically robust,
1289 single-sample classifier for tumor subtyping in pancreatic cancer. *Clin. Cancer Res.* **26**, 82–
1290 92 (2020).
1291 100. O'Brien, S. J. *et al.* Long non-coding RNA ZFAS1 is a major regulator of epithelial-
1292 mesenchymal transition through miR-200/ZEB1/E-cadherin, vimentin signaling in colon
1293 adenocarcinoma. *Cell Death Discov.* **7**, 61 (2021).
1294 101. Rao, M. *et al.* Long non-coding RNA ZFAS1 promotes pancreatic cancer proliferation and
1295 metastasis by sponging miR-497-5p to regulate HMGA2 expression. *Cell Death Dis.* **12**,
1296 859 (2021).
1297 102. Xiong, G. *et al.* Collagen prolyl 4-hydroxylase 1 is essential for HIF-1α stabilization and
1298 TNBC chemoresistance. *Nat. Commun.* **9**, 4456 (2018).
1299 103. Uhlen, M. *et al.* A pathology atlas of the human cancer transcriptome. *Science* **357**, (2017).
1300 104. Marisa, L. *et al.* Gene expression classification of colon cancer into molecular subtypes:
1301 Characterization, validation, and prognostic value. *PLoS Med.* **10**, e1001453 (2013).

1302 105. Győrffy, B., Surowiak, P., Budczies, J. & Lánczky, A. Online survival analysis software to
1303 assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung
1304 cancer. *PLoS One* **8**, e82241 (2013).
1305 106. Calon, A. *et al.* Stromal gene expression defines poor-prognosis subtypes in colorectal
1306 cancer. *Nat. Genet.* **47**, 320–329 (2015).
1307 107. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary
1308 glioblastoma. *Science* **344**, 1396–1401 (2014).
1309 108. Zilionis, R. *et al.* Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals
1310 Conserved Myeloid Populations across Individuals and Species. *Immunity* **50**, 1317-
1311 1334.e10 (2019).
1312 109. Savas, P. *et al.* Single-cell profiling of breast cancer T cells reveals a tissue-resident
1313 memory subset associated with improved prognosis. *Nat. Med.* **24**, 986–993 (2018).
1314 110. van der Leun, A. M., Thommen, D. S. & Schumacher, T. N. CD8+ T cell states in human
1315 cancer: insights from single-cell analysis. *Nat. Rev. Cancer* **20**, 218–232 (2020).
1316 111. Kang, J. *et al.* Systematic dissection of tumor-normal single-cell ecosystems across a
1317 thousand tumors of 30 cancer types. *Nat. Commun.* **15**, 1–17 (2024).
1318 112. Karimi, E. *et al.* Single-cell spatial immune landscapes of primary and metastatic brain
1319 tumours. *Nature* **614**, 555–563 (2023).
1320 113. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. Methods and applications for single-
1321 cell and spatial multi-omics. *Nat. Rev. Genet.* **24**, 494–515 (2023).
1322 114. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-
1323 Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337.e4
1324 (2019).
1325 115. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality
1326 control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**,
1327 1179–1186 (2017).

116. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
117. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Res.* **10**, 33 (2021).
118. Wang, Y. *et al.* Interactive analysis of single-cell data using flexible workflows with SCTK2. *Patterns (N Y)* **4**, 100814 (2023).
119. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367 (2010).
120. Shen, L. GeneOverlap: An R package to test and visualize gene overlaps. *R Package* (2016).
121. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
122. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb.)* **2**, 100141 (2021).
123. Gavish, A. *et al.* Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. *Nature* **618**, 598–606 (2023).
124. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* 060012 (2021) doi:10.1101/060012.
125. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
126. Jonah Gabry and Rok Češnovar and Andrew Johnson and Steve Bronder. cmdstanr: R Interface to “CmdStan.” https://mc-stan.org/cmdstanr/ (2023).