Inferring tumour microenvironment ecosystems from scRNA-seq atlases

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. Here, we describe how ecosystems of cell populations can be robustly inferred from nonspatial single-cell RNA-seq (scRNA-seq) atlases. Leveraging a unique discovery-validation setup across eight scRNA-seq datasets profiling pancreatic ductal adenocarcinoma (PDAC), we show highly consistent co-occurrence of fine-grained cell states across patients and characterize the positive predictive value of such analyses. Building on this, we develop a novel probabilistic model to quantify multi-cellular ecosystems directly from such atlas-scale scRNA-seq datasets. By mapping these ecosystems to spatial transcriptomics data, we demonstrate that such ecosystems represent bona fide spatial variation within individual tumours despite being learned from nonspatial data. Importantly, through mapping these ecosystems across two large clinical cohorts, we show they are more predictive of therapy response than any individual cell state and are associated with specific tumour somatic mutations. Together, this work lays the foundation for inferring reproducible multicellular ecosystems directly from large nonspatial scRNA-seq atlases.

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 ecosystems)[16–18]. These ecosystems 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 TME.

However, reliable quantification of cell state co-occurrence within individual tumours leading to cellular ecosystems 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 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 can measure 100s of mRNA species or even the whole transcriptome simultaneously, but are currently too expensive to profile at scale and suffers resolution or RNA contamination problems[20]. In contrast, scRNA-seq can quantify single-cell whole transcriptome expression measurements with quantified cell state abundance largely reflecting tissue composition[21–23]. 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[24] and Human Tumour Analysis Network[25] has resulted in unprecedented data sizes, comprising more than 12 tumour atlases[4,26–31] by mid-2024 alone. This leads to the tantalizing possibility of quantifying cell state co-occurrence and cellular ecosystems 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) due to both the availability of multiple independent primary tumour scRNA-seq datasets studies[4,32–38], combined with its complex microenvironment that is characteristically difficult to target. Existing

studies of the PDAC TME have highlighted diverse cellular heterogeneity across all lineages[39–43]. Additionally, single-cell profiling of PDAC tumours has highlighted previous underappreciated heterogeneity within individual tumours cells themselves, with the basal-like/classical[44] and later hybrid subtypes[40] being shown to co-exist within single tumours[45,46]. 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[41], 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 multi-cellular ecosystems can be reproducibly inferred in the PDAC tumour microenvironment from atlas-scale scRNA-seq (**Fig. 1A**). We 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 cellular subtypes in the TME can be independently verified, along with their ability to co-occur within and across cell types and form cellular ecosystems. We map 50 cell type-specific gene programs across 8 cell types and confirm the existence of known subtypes 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 consistent co-occurrence across major cell types. We subsequently formalize a probabilistic model for quantifying multicellular ecosystems from atlas-scale scRNA-seq and find them to be reproducible across our datasets. Interestingly, by mapping these ecosystems to spatial transcriptomics data, we demonstrate that such ecosystems represent bona fide spatial variation within individual tumours despite being learned from nonspatial data. Finally, we map the validated gene programs and ecosystems to two bulk datasets and find that certain ecosystems are associated with patient outcome along with specific KRAS mutations.

Results

Identifying reproducible cellular subtype heterogeneity in the PDAC TME

To systematically quantify cell states from the PDAC tumour microenvironment, we compiled 8 previously generated scRNA-seq datasets (**Supplementary Fig. 1A** and **Methods**, **Supplementary Table 1**)[4,32–37,47] 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 patient 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). Raw count data preprocessing was performed in line with previous studies (**Supplementary Fig. 1B**, **Methods**). An automated workflow was developed to systematically annotate single cells using a public multi-cohort PDAC atlas as reference (**Supplementary Fig. 2A** and **Methods**)[47]. Eight major cell types were identified in our curated atlas (**Fig. 1B**). To validate the automatically assigned cell type labels, we examined known marker expressions in each dataset (**Supplementary Fig. 2B**).

To quantify fine-grained cellular subtypes in the TME, we extracted cell type-specific gene programs independently for the discovery and the validation datasets. Specifically, we applied integrative non-negative matrix factorization (iNMF)[48] 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. 3A**, **Methods**). The inherent plasticity of the cells of the TME motivated our use of NMF rather than sub-clustering[5,32,49,50], which allows each cell to have a continuous activation of multiple gene programs compared to clustering approaches that force each cell to belong to a distinct cluster, following recent work in this area[51].

Given concerns around whether granular cell subclusters from scRNA-seq reflect noise or overclustering[52], we set out to identify a subset of cell type-specific high-confidence gene programs found in both our discovery and validation datasets. We first employed the Hungarian algorithm[53] to obtain optimal one-to-one mapping between gene programs from the discovery and the validation datasets (**Supplementary Fig. 3B, C**). Next, we merged mapped gene programs with similar gene weights to reduce redundancy (**Methods: Program validation, Program merge**) (**Supplementary Figure 3D**). Lastly, we removed any gene programs likely corresponding to ambient RNA programs contaminating[54] by manual examination of the top program markers (**Methods**). After applying this procedure, we obtained 10 high-confidence gene programs in epithelial cells, 5 high-confidence gene programs in each of fibroblasts and dendritic cells, 11 in monocytes/macrophages, 8 in CD8 T/NK cells, 3 in CD4 T cells, 4 in B cells, and 4 in endothelial cells (**Fig. 1D**). In total, we obtained 50 cell type-specific gene programs that capture both the overall TME landscape and the granular phenotypic heterogeneity (**Supplementary Fig. 4**). Overall, our approach allowed us to identify high-confidence recurrent cellular subtypes in PDAC TMEs for downstream analysis.

We next contextualized the high-confidence cell type-specific gene programs to distinguish potentially novel PDAC TME cell states from known ones. We took a multi-faceted approach including marker gene inspection, GSEA, and literature review (**Methods**). Importantly, our unique discovery-validation setup that independently discovered reproducible programs in

148 multiple datasets confers high confidence on their manifestation in patients. While this
149 recapitulated many existing cell states as previously reported in single-cell profiling of
150 PDAC[32,39,44,55]—including basal and classical epithelial subtypes as well as inflammatory and
151 myofibroblastic cancer associated fibroblasts (CAFS)—we discovered certain high-confidence
152 gene programs previously undescribed.
153
154 In the epithelial compartment, we discovered a previously undescribed single-cell KRAS
155 amplification epithelial program marked by upregulation of *KRAS* and *BNIP3L*/*NIX*—previously
156 shown to be a driver of glycolysis and disease progression in pancreatic cancer[56] (**Fig. 1E**). In
157 the fibroblast compartment, we identified two distinct inflammatory cancer associated fibroblast
158 (iCAF) subsets and two myofibroblastic CAF (myCAF) subsets for the first time (**Fig. 1F**). The
159 *classical pathway (CCP) iCAF* program showed strong upregulation of *C7* and the classical
160 pathway genes *C1S* and *C1R*. The CCP-iCAF program was also enriched for a neurotropic
161 gene set linked to tumour growth, stress adaptation, and metastasis in PDAC[5,57]
162 (**Supplementary Fig. 5A**). In contrast the *adhesive iCAF* program was highly enriched for an
163 adhesion gene signature including strong upregulation of the gene *SERPINE1* that encodes for
164 the protein plasminogen activator inhibitor-1 (*PAI-1*) (**Fig. 1F**). The *ACSL3-PAI-1* signaling axis
165 supports PDAC progression, links to fibrosis, immunosuppression, and prognosis[58]. We found
166 the CCP iCAF program and the adhesive iCAF program to be present in a mutually exclusive
167 pattern across patients within the dataset (**Supplementary Fig. 5B**). Two myCAF programs
168 were identified based on CAF subtype gene loadings in fibroblast gene programs
169 (**Supplementary Fig. 5C**). The *TPM2*+ myCAF-1 program was highlighted by the high loading
170 of *MIF*[56] (**Fig. 1F**). The *VMP1*+ myCAF-2 program was identified as a potential novel myCAF
171 subtype. Although VMP1 has been associated with PanIN formation[59] and patient survival in
172 pancreatic neuroendocrine tumours[60], its role in PDAC CAFs is still unclear.
173
174 Finally, since intra- and inter-tumour heterogeneity is closely coupled with treatment resistance
175 and therapy response across cancers[6,30,61,62], we investigated whether heterogeneity in the
176 granular cell states depicted by our gene programs was conserved between our discovery and
177 validation datasets. We calculated gene program inter-patient heterogeneity and intra-patient
178 heterogeneity **(Methods)** and found inter- and intra-patient program loading heterogeneity
179 varied greatly across cell types and within cell types (**Supplementary Fig. 6A, B**). Gene
180 programs of CD4 T cells, CD8 T/NK cells, epithelial cells, and macrophages showed wide
181 ranges of intra-patient heterogeneity across patient cohorts (**Supplementary Fig. 6B**). Across
182 cell types, we found inter-patient heterogeneity tended to agree between the discovery and the
183 validation datasets, except for T cell programs and macrophage programs (**Supplementary**
184 **Fig. 6B**). In short, we showed in single-cell data for the first time that both intra- and inter-patient
185 cell state heterogeneity is conserved between discovery datasets and validation datasets (intra-
186 patient heterogeneity: $\rho = 0.67$, $p = 1\text{x}10^{-08}$, inter-patient heterogeneity: $\rho = 0.44$, $p = 0.00053$).

This suggests that the abundance of granular gene programs within the TME can serve as a robust estimate of heterogeneity and help guide target discovery efforts that specifically attempt to cover the entire phenotypic space.

Co-occurring cell type-specific gene programs are highly reproducible in PDAC

We next evaluated the extent to which gene program co-occurrence within individual tumours could be reproducibly quantified to understand if fine-grained cellular subtype co-occurrence can be robustly estimated from scRNA-seq data. We computed co-occurrence of all possible gene program pairs (N=1770 total, **Methods**) separately in the discovery and the validation groups (**Fig. 2A**). To estimate this co-occurrence, we computed the Spearman correlation of the activation of each pair of gene programs across samples. This lets us estimate both the co-occurrence of gene programs within major cell types (such as if exhausted and proliferative CD8 T cells co-occur) as well as the co-occurrence of gene programs between major cell types (such as if exhausted CD8 T cells and inflammatory cancer associated fibroblasts co-occur). We quantified reproducibility of these estimates by correlating the co-occurrence in discovery and validation settings, finding highly concordant estimates ($\rho=0.46$, $p<2\times10^{-16}$) (**Fig. 2B**). We next used the aggregated program pair co-occurrence data to derive the positive predictive value (PPV) treating the validation co-occurrence as the ground truth and discovery as the prediction. We found that overall, the program co-occurrence PPV 80.6% for intra-cell-type co-occurrence and 74.7% for inter-cell-type co-occurrence (**Fig. 2B**). In addition, most programs saw a PPV above 50% when examined individually (**Fig. 2E**), allowing for high confidence interpretation of co-occurrence relationships.

Next, we examined the reproducibility of cell state co-occurrence within and between major cell types. For co-occurrence of programs from the same cell type, we found highest reproducibility in fibroblasts ($\rho=0.77$) followed by monocyte/macrophage ($\rho=0.59$) (**Fig. 2C**). Gene program co-occurrence within T cells, DCs, and epithelial cells exhibited lower reproducibility between discovery and validation ($\rho=0.37$ to 0.5), while endothelial cells and B cells saw the lowest reproducibility across datasets ($\rho=0.086$ and 0.22 respectively). Nevertheless, B cell programs had positive co-occurrence in discovery ($\rho=0.2$-0.81) and validation ($\rho=0.38$-0.8), demonstrating that the overall effect size was consistent (**Fig. 2C**). We found all cell types exhibit significantly reproducible inter-cell-type program co-occurrence, most notably once again in Fibroblasts ($\rho=0.47$) and Monocyte/Macrophage ($\rho=0.49$) (**Fig. 2D**). Indeed, on a program-by-program basis we found most to have significantly reproducible inter-cell type co-occurrence relationships (**Fig. 2E**).

We next focused on the co-occurrence of specific cellular subtypes that are either significantly reproducible between discovery and validation or *a priori* known to hold clinical significance. The co-occurrence of regulatory CD4 T cells and exhausted CD8 T cells in the PDAC

226 microenvironment was well known from mouse studies[63], but is yet to be demonstrated in
227 humans. We found that the CD4 T_{reg} program consistently co-occurred with the CD8 T cell
228 exhaustion program in patient tumours (**Fig. 2F**). Also, the CD4 T_{reg} program co-occurred with
229 the *CD163*+ alternatively activated macrophage program and the endothelial angiogenesis state
230 program marked by *TIE1*, *INSR*, and *NOTCH4*. This angiogenesis program has been found to
231 be an ECM remodeling and angiogenic signature in lung and breast cancers[64,65] and likely
232 interact with TAMs and regulatory T cells in breast cancer to promote an immunosuppressive
233 phenotype in the microenvironment[65]. This T_{reg}-T_{exh}-Mac_{alt} co-occurrence pattern is spatially
234 confirmed in multiple cancers[66], suggesting our subtype program co-occurrence may be
235 associated with spatial ecosystems in the PDAC.
236
237 Finally, we investigated whether known multi-cellular structures in the TME were consistent with
238 our co-occurrence results. Specifically, we focused on Tertiary lymphoid structures (TLSs),
239 localized immune aggregates within the TME that can lead to production of tumour-specific
240 antibodies. TLSs in PDAC are usually composed of proliferating, activated, germinal center
241 (GC) B cells, antigen presenting cells, and T cells and their presence is predictive of patient
242 survival[67]. Therefore, we examined the co-occurrence pattern for our germinal center B cell
243 program. Indeed, the B cell activated and cycling programs are among the top programs that
244 co-occur with the GC program (**Fig. 2G**). In addition, we found significant co-occurrence of the
245 activation and memory program of CD8 T cell programs with the GC B cells (ρ=0.3 in the
246 validation group) along with mature and plasmacytoid dendritic cells (ρ=0.45). Examination of
247 the patient sample level loadings of the TLS related cell state programs suggests variable
248 presence of TLS in PDAC tumours (**Supplementary Fig. 7A, B**). These results suggest that in
249 our datasets, TLS presence and phenotype could be read out from scRNA-seq. Our results
250 imply co-occurrence of granular cellular subtypes from scRNA-seq can be meaningfully
251 quantified from large scale atlases, and complex immune cell clusters can be identified and
252 characterized without the need for additional spatial transcriptomics or proteomics.
253

254 A probabilistic model to infer recurrent multi-cellular ecosystems in the TME

255 Beyond the pairwise co-occurrence of gene programs, multi-cellular ecosystems play a
256 fundamental role in the TME, as these systems often act in concert, and their function is highly
257 synergistic[68]. For this reason, we asked whether it was possible to infer ecosystems
258 reproducibly from scRNA-seq. However, there are no existing computational tools to do this
259 from scRNA-seq using a set of biologically meaningful gene programs as input. Therefore, we
260 developed a novel probabilistic model (**Methods**) that can take gene programs representing
261 cellular subtypes and infer multiple multi-cellular ecosystems, reporting each both in terms of
262 the subtypes that co-occur as well as the ecosystem abundance within each patient sample.
263 Importantly, our model carefully decouples the fact that programs can co-vary within the same
264 cell (intrinsic variation) but also between cells (extrinsic variation) (**Methods**). Using maximum a

265 posteriori inference (**Methods**), we fitted the ecosystem model separately on the discovery and
266 validation groups using all the 50 cell type specific gene programs. We quantified four
267 ecosystems, a number which is guided by sub-TMEs found with histology[46].
268
269 We first contrasted the ecosystems found in the discovery and validation datasets to both
270 validate the overall methodology and generalizability of the findings. We characterized the
271 ecosystems by examination of the abundance of gene programs within each (**Fig. 3A**). We
272 found significant positive correlation in 3/4 ecosystems between discovery and validation (**Fig.**
273 **3B**). We identified a Classical ecosystem enriched for corresponding epithelial programs,
274 endothelial cells, iCAFs, and tumour associated macrophages. A Basal-like ecosystem was
275 enriched for epithelial EMT programs, myCAF programs, and alternatively activated
276 macrophage programs. We termed the third ecosystem immune activated as it was enriched
277 with functional immune cell programs related to anti-tumour response, such as the memory CD4
278 T cell program and the effector CD8 T cell program. Notably, the TLS related programs were
279 also enriched in the immune activated ecosystem (**Fig. 3A**). For confirmation, we found the
280 classical and basal-like ecosystems in individual tumour samples were significantly anti-
281 correlated in both discovery and validation cohorts ($\rho = -0.43$ and $\rho = -0.51$ respectively)
282 (**Supplementary Fig. 8A**).
283
284 We next investigated whether our tissue ecosystems represent spatial variation within the
285 tumour rather than simply per-patient co-occurrence. Leveraging 24 PDAC 10x Visium spatial
286 transcriptomics slides across 16 patients[4], we scored individual spots for each ecosystem (**Fig.**
287 **3C, F, Methods**)**.** We computed Moran's I—a measure of spatial autocorrelation—for each
288 ecosystem, then shuffled gene program weights to create 100 randomized ecosystems. We
289 found the observed spatial autocorrelation from our scRNA-seq inferred ecosystems was higher
290 on average than random across all three ecosystems, and significantly so for two (Basal-like: p
291 $= 0.024$, immune activated: $p < 0.0001$, **Fig. 3D**). We also compared our ecosystems to an
292 unsupervised learning approach for TME community discovery by generating consensus multi-
293 program factors from spot-wise program scores using NMF. We compared the spot-wise spatial
294 loadings of the factors with our ecosystem by computing a correlation coefficient. We found the
295 NMF factors were not able to reproduce the Basal and immune activated ecosystem, while only
296 the NMF factor 3 was showing moderate agreement with the classical ecosystem (**Fig. 3E**).
297 Indeed, examining the spatial presence of the ecosystems inferred from scRNA-seq across four
298 Visium slides shows pronounced spatial heterogeneity (**Fig. 3F**).
299

300 Associating ecotypes with tumour genetics and clinical outcomes

301 To quantify the association between both our gene programs and ecosystems with tumour
302 genetics and clinical outcomes, we curated clinical and genome sequencing data for two
303 cohorts of PDAC tumour samples with paired bulk RNA-seq data, which we refer to as the Chan

304 cohort[67] (N=183 patients) and TCGA pancreatic cancer cohort[69] (N=151 patients). We scored
305 each sample for each of the gene programs we identified (**Methods**)[70] and subsequently
306 associated each sample with each ecotype by quantifying weighted combinations of gene
307 programs (**Methods**). We first examined associations between the novel gene programs we
308 identified and overall patient survival. Interestingly, the KRAS amplification epithelial program
309 was significantly associated with poor patient survival in both the TCGA cohort and the Chan
310 cohort (**Fig. 4A**). Importantly, the presence of this program was more predictive of patient
311 survival than well the established classical and basal gene expression subtypes (**Fig. 4B**).
312 Together, this implies such workflows that exploit high-confidence gene programs identified
313 from single-cell RNA-seq data and validated via discovery-validation splits may be more
314 clinically relevant than existing subtypes identified from bulk profiling approaches.
315
316 Next, we examined links between the multicellular ecosystems and patient response to therapy.
317 Interestingly, the Basal-like ecosystem was positively associated with progressive disease after
318 chemotherapy in both the Chan cohort ($p<0.01$) and the TCGA cohort ($p=0.40$)
319 (**Supplementary Fig. 8B, 9**), and more significantly associated with progressive disease than
320 any individual gene program including all Basal epithelial programs in the Chan cohort (**Fig. 4C,**
321 **D**). Similarly, the immune activation ecosystem was positively associated with partial remission
322 and negatively with progressive disease in the TCGA cohort ($p<0.05$) more significantly than
323 any individual gene program (**Fig. 4C, D**). Similarly, this ecosystem was negatively associated
324 with progressive disease in the Chan cohort, albeit not statistically significantly, ($p=0.11$)
325 (**Supplementary Fig. 8B, 9**). These differences in statistical power between cohorts likely
326 reflect differences in sample preparation: the Chan cohort underwent laser capture
327 microdissection for the tumour epithelium, making it better placed to capture tumour associated
328 ecosystems (Basal-like), while the TCGA cohort was not tumour enriched, which given the low
329 cellularity of PDAC tumours (average = 35%) makes it more adept at capturing immune
330 populations.
331
332 Finally, we contextualized the link between multicellular ecosystems and tumour genetics.
333 PDAC tumours carrying *KRAS*-G12D mutations are associated with poorer prognosis[71], and
334 strategies targeting specific *KRAS* neoantigens are being developed[72]. However, the
335 mechanisms through which different mutations affect the microenvironment in human PDAC
336 remains unclear. We obtained *KRAS* variant calls for the HTAN Zhou (WUSTL) patient dataset[4]
337 in our validation cohort and fitted a linear mixed model between ecosystem presence and *KRAS*
338 variant status (**Methods**). We found that compared to *KRAS*-G12D, tumours with a *KRAS*-
339 G12V/G12R mutations had a significantly higher presence of the immune activated ecosystem,
340 while tumours with *KRAS*-G12R mutations had a reduced presence of the basal-like ecosystem
341 (**Fig. 4E**). Importantly, we retrieved similar *KRAS* variant calls in the bulk Chan patient
342 cohort[69,70] and found consistent effect sizes for all associations (**Supplementary Fig. 8C**,

343 **Supplementary Table 3**), validating the overall findings. The decreased presence of the
344 immune activated ecosystem in *KRAS*-G12D PDAC suggests differences in anti-tumour
345 immune response may underline prognosis and effectiveness of immunotherapy[71,72]. Together,
346 these results suggest that the specific driver mutations in PDAC have a significant effect on the
347 phenotypes and cell communities in the TME.
348

349 # Discussion

350 Transcriptome profiling of tumour samples has historically been used to identify molecular
351 markers that could facilitate cancer diagnosis, prognosis prediction and identification of
352 molecular candidates to target therapeutically[73–76]. The advent of single cell technologies has
353 facilitated the identification of the relative distribution of cell populations specific for certain
354 cancer types, predicting clinical outcomes and uncovering specific cellular subtypes as
355 targets[77–80]. However, it is becoming increasingly clear that focusing on the expression pattern
356 of a handful of genes or the presence/absence of certain cell populations is not sufficiently
357 informative to gain useful biological insights into the whole TME, even though this information
358 could be leveraged to design targeted therapies aimed at tackling the combined action of
359 multiple cell types. In this context, spatial information is extremely useful to identify co-
360 localization of different cell types, whose crosstalk requires a certain three-dimensional
361 organization to synergistically alter tumour growth and survival. However, a substantial number
362 of samples are needed to robustly identify co-occurrence of interacting cells across patients,
363 especially considering that the cellular subtypes of interest are potentially rare. Therefore, the
364 current cost and/or limitations of spatial transcriptomics and proteomics limit the identification of
365 reproducible co-existing cell states with prognostic value[79]. By contrast, the number of scRNA-
366 seq datasets from TMEs that contain transcriptome-wide expression quantification at single-cell
367 resolution is increasing exponentially, allowing, in theory, to identify co-occurrence of granular
368 cell states and associate them with clinical outcomes.
369

370 Here we have undertaken a unique discovery-validation analysis on the largest PDAC scRNA-
371 seq atlas to-date to quantify the extent to which granular cellular subtypes and ecosystems can
372 be reliably inferred. Our 50 reproducible cell type specific gene programs captured previously
373 known heterogeneity such as classical and basal tumour transcriptional but also novel subtypes
374 such as two iCAF subtypes as well as a novel *KRAS amplification* epithelial cell program
375 significantly associated with survival. We further revealed for the first time that co-occurrence of
376 gene programs can be reproducibly quantified from large scRNA-seq datasets. We further
377 identified a TLS pattern centered around the GC B cell program, recapitulating a spatial
378 structure important for anti-tumour immune response in multiple cancers including PDAC[81]. We
379 introduced a new model for inferring ecosystems in TME and found reproducible ecosystems,

380 which were validated in spatial transcriptomic data and associated with both driver gene
381 mutants as well as response to therapy.
382
383 Overall, we show that large scRNA-seq atlas can be employed to not only identify TME cellular
384 subtypes and ecosystems with high confidence, but that these have profound biological and
385 clinical significance, as the frequency of co-occurrence of certain programs and the presence of
386 certain ecosystems are associated with patient survival. These analyses suggest that in the
387 future, scRNA-seq datasets will be used to study the broad interplay among pro-tumorigenic cell
388 populations in the TME. This approach may eventually expand the clinical usability of single cell
389 data, offering a new lens to develop innovative therapeutic strategies targeting PDAC.
390
391 Despite these findings, our study had multiple limitations. While our discovery group was well
392 balanced between four datasets, the validation cohort was heavily weighted towards one cohort,
393 which may have limited the cell state program validation process though we note that, since
394 there was no overlap in samples, it still functions as a well-powered validation cohort. In
395 addition, the lack of well-standardized and curated clinical metadata across the single-cell
396 cohorts limited the ability to directly associate cell states and ecosystems with clinical
397 covariates, instead necessitating mapping them to bulk datasets with standardized metadata.
398
399 Moving forward, we identified several venues to extend the project. Future spatial profiling of the
400 PDAC microenvironment at cellular resolution in large cohorts will help to create a more
401 comprehensive understanding of the cell state co-occurrence related to therapy resistance and
402 disease progression. Integrating T cell receptor (TCR) or B cell receptor (BCR) sequencing
403 analysis will help to delineate relationships between lymphocyte clonality with functional cell
404 states, potentially revealing specific anti-tumour T cell clones and effects of therapy on the
405 tumour immune cell population. In the long term, our analysis workflow may be adapted to
406 multiple cancer types, leveraging the ever-increasing tumour single-cell data space to create a
407 pan-cancer microenvironment cell state atlas, and identifying clinically relevant cellular subtypes
408 across cancers. The increasing characterization of the TME at single-cell level will enable the
409 identification of concomitant presence of multicellular ecosystems in various types of cancer.
410 Our approach unlocks a new way to analyse scRNA-seq data that will not only facilitate
411 prognosis and assessment of therapeutic response, but will provide valuable information to
412 design specific therapies aimed at rewiring the milieu compatible with the co-existence of pro-
413 tumorigenic cellular subtypes and ecosystems.
414

Methods

Data curation

A total of 8 publicly available PDAC single-cell RNA-seq (scRNA-seq) datasets, which include 214 samples, were curated[4,32–37,47]. 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[82]. 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[83].

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[81]. 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[84]. 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

453 downstream analysis, endocrine cells and neural cells were removed from the curated dataset,
454 and acinar and ductal cells were both labeled as epithelial cells. The eight consolidated cell type
455 labels were then shortened for readability.
456

457 **Discovery-Validation data split**

458 We first subset the full curated dataset by cell annotation, splitting it into eight cell type subsets:
459 Epithelial cell, Endothelial cell, Fibroblast, Monocyte/Macrophage, mDC, CD4 T cell, CD8 T/NK
460 cell, and B cell. We then split each cell type subset into two 4-cohort groups, one as the
461 discovery group and the other as the validation group. The discovery group contained 74
462 samples from the Peng, Steele, Werba, and Chan patient cohorts, while the validation group
463 contains 89 samples from the Chen, Lee, Lin, and Zhou patient cohorts. In total, the discovery-
464 validation split resulted in eight discovery-validation pairs of cell type subsets, with each pair
465 containing single cells assigned to one cell type.
466

467 **Reproducible analysis pipeline**

468 We built a comprehensive analysis pipeline to generate reproducible results from our curated
469 dataset and reduce human-introduced variations in analyses across the cell types and between
470 the discovery group and the validation group. We used Snakemake v.7.25.0[85] for parameter and
471 input-output control for all analysis scripts and visualization steps.
472

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

474 We carried out integrative NMF (iNMF) using LIGER[82] for the discovery and validation group of
475 patient cohorts across cell types independently to generate cell lineage-specific gene programs
476 that capture the phenotypic heterogeneity within each cell type in the PDAC microenvironment.
477 LIGER runs iNMF to decompose the expression matrices from multiple batches (patient
478 cohorts) and remove batch effect at the same time by maximizing a basis matrix shared
479 between cohorts while minimizing the multiple cohort-specific basis matrices that represent
480 patient cohort level biases. We defined the batches to be patient cohorts to focus on capturing
481 heterogeneity present across tumours through cell lineage-specific programs. We followed the
482 suggested workflow of LIGER to perform iNMF. Importantly, the following steps were performed
483 after the discovery-validation split as well as an 8-way cell type split within the discovery and the
484 validation group, leading to 16 independent gene program extraction steps, 8 independent
485 program validation steps, and 8 independent program merging steps.
486

487 **Program extraction**

488 First, we selected variable genes from non-ribosomal/mitochondrial protein-coding genes using
489 selectGenes from rliger v.1.0.0[83] with default settings. We then created a liger object using the
490 raw counts matrix, removed genes with zero expression, and normalized and scaled gene
491 expression following suggested workflow from rliger documentation before performing a subset

492 to the variable genes. Finally, after filtering to variable genes, we fitted iNMF on the expression
493 matrix using optimizeALS. We set rank (k) to 20, 10, 5, 10, 10, 20, 10, and 10 for program
494 extraction from epithelial cell, fibroblast, endothelial cell, CD4 T cell, CD8 T/NK cell,
495 monocyte/macrophage, mDC, and B cell respectively. We set lambda to 5 for all cell types as
496 we did not find it to be a major factor affecting extraction results. The gene loadings in each
497 program were obtained as the W matrix from the fitted liger object, and the program loadings in
498 single-cells were obtained as the H.norm matrix from the fitted liger object.
499

500 **Program validation**

501 Our program validation process was performed between discovery and validation for each cell
502 type. To start, we transformed the discovery and validation gene loading (W) matrix with log1p
503 in R. We then computed the spearman correlation and Minkowski distance (with p = 1) between
504 all possible pairs of discovery and validation programs within the cell type. The Hungarian
505 algorithm was used to generate a one-to-one relationship between discovery and validation
506 programs by permuting the rows of the distance and correlation matrices to minimize and
507 maximize the trace of the matrices respectively. After assigning the one-to-one partner of each
508 discovery program, we rejected any discovery programs whose gene loading correlation with
509 their validation program partner was below 0.5, or if the gene loading distance with their
510 validation program partner was above 25,000. We further confirm validated programs by
511 permutation test. Gene loadings in each validated program were permuted to generate a 1000-
512 sample test set and the correlation and distance to the validation program partner were
513 calculated for fitting a null distribution. A discovery program is finalized as a validated program if
514 the real distance and correlation are laid beyond 1% of the null distribution.
515

516 **Program merge**

517 We collapsed validated programs which were too similar to each other given NMF does not
518 guarantee orthogonality between the extracted factors. We looked at all possible pairs of
519 validated programs within each cell type and marked any pair whose gene loading correlation
520 was above 0.4 for merging. The merging step leveraged simple linear models. For a pair of
521 validated programs, e.g. sig-1 and sig-2, marked for merging, we first computed their explained
522 expression y by multiplying program loading in cell and gene loading in program respectively.
523 Then, we obtained the merged program's gene loading x by taking the mean of gene loadings in
524 sig-1 and sig-2. Finally, the single-cell loading of the merged program was obtained by fitting
525 lm(y ~ 0 + x) and pulling the coefficients. After merging, the newly merged program was
526 included for downstream analysis while sig-1 and sig-2 were excluded.
527

528 **Program rescoring in the validation group**

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

531 program loadings in the discovery group in downstream analysis. The rescoring step was
532 performed for each cell type and each patient cohort independently, leading to 32 rescoring
533 steps.
534
535 To start, gene expression in a validation group patient cohort was normalized using
536 runSeuratNormalizeData from singleCellTK v.2.8.0[86] with default settings. Then, for rescoring
537 for programs of, for example, cell type A, we subset an expression matrix to cells annotated as
538 cell type A and variable genes used in the program extraction step. Finally, we performed non-
539 negative least squares to fit a simple linear model to the subset validation expression with cell
540 type A program gene loadings, and pulled out the estimated optimal coefficient matrix as the
541 rescored program loading in single-cells in validation. We used fcnnls from NMF v.0.26[87] with
542 default settings for this step.
543

544 **Program contextualization with known cell states**

545 We combined multiple approaches to get a comprehensive understanding of the biological
546 meaning of our validated cell type-specific programs. We curated multiple PDAC focused
547 malignant, stromal, and immune cell state marker lists from literature (**Supplementary Table 4**).
548 To associate the programs with the known cell states, we selected the top 50 loaded genes of
549 each program and performed a hypergeometric test with the cell state marker lists from its
550 respective cell type, using GeneOverlap v.1.34.0[88] in R. The human reference hg19 genome[89]
551 was used as the background for significance testing. We made a preliminary biological
552 interpretation of the programs based on significance (P-value < 1e-3) and high Jaccard index
553 with cell state marker lists. Then, the programs were further subjected to GO and KEGG
554 overrepresentation analysis using the clusterProfiler v.4.6.2 package[90]. Specifically, the
555 biological process subcategory of GO and KEGG PATHWAY gene sets were chosen for the
556 analysis. The top ten enriched gene sets in each category were used to help correct or confirm
557 our initial program interpretation. After the first correction, we also applied GSEA to the
558 programs with gene sets mined from the Curated Cancer Cell Atlas (3CA)[91] to resolve any
559 remaining uncertainty. We ran fgsea from fgsea v.1.24.0[92] with default settings, while setting
560 nPermSimple to 10,000, as suggested by the documentation. We used the top 10 positively
561 enriched 3CA gene programs to guide a second program interpretation correction step.
562 Programs that were not associated with known cell states previously were further characterized
563 after this analysis. Finally, we completed a manual check across programs by looking at the top
564 10 loaded genes, focusing on the programs that were yet to be contextualized and searching in
565 the literature for novel evidence of their biological meaning in PDAC. Any novel programs
566 remained were named by their top loaded genes.
567

568 **Ambient RNA program removal**

569 Due to technical limitation of droplet-based single-cell RNA-seq experiments, some reads
570 associated with a cell barcode may originate from ambient RNA[54], contributing to background
571 noises that are picked up by the granular cell state gene programs. Therefore, we denoted gene
572 programs in a cell lineage that are marked by lineage markers from other lineages to be
573 ambient RNA programs and removed such ambient programs from downstream analysis.
574

575 **Cell state co-occurrence**

576 To study co-occurrence between programs in the PDAC microenvironment, we constructed
577 sample level program loading profiles by averaging program loadings across single-cells. We
578 then calculated the correlation across samples between programs both within cell type and
579 across cell types. By comparing program loading correlations between the discovery group and
580 the validation group, we define program pairs whose loading correlation was above 0.4 and
581 differed by less than 0.25 in discovery and validation as high confidence pairs that tend to co-
582 occur with each other.
583

584 **Scoring RNA-seq samples for cell state programs and ecosystems**

585 To score bulk PDAC tumour samples for our cell state gene programs, we performed gene set
586 variation analysis (GSVA) on each bulk sample using gsva from GSVA[93] ver. 1.46.0, with the
587 top-50 loaded genes in each gene programs, and used the program enrichment scores or
588 downstream analysis. To score bulk PDAC tumour samples for our ecosystems, we performed
589 multiplication between the sample-by-program enrichment score matrix and the program-by-
590 ecosystem abundance matrix, to obtain a sample-by-ecosystem enrichment score matrix.
591

592 **Association between ecosystems and KRAS variants**

593 KRAS variant information was obtained from published data and from collaborators. To account
594 for the patient-repeated measurements in the Chan cohort, we fitted a linear mixed model with
595 the *lmerTest* package in R with the KRAS mutation status as the fixed effect and the patient as
596 the random effect. For the TCGA cohort to estimate effect of KRAS variant on ecosystem
597 enrichment scores we used a simple linear model while controlling for the expression subtype
598 (Basal-like or Classical) of the tumour.
599

600 **Association between gene programs, ecosystems, and clinical outcomes**

601 To associate the cell type-specific gene programs and multicellular ecosystems with clinical
602 outcomes, we mapped scored RNA-seq samples in the Chan cohort and the TCGA cohort to
603 metadata and clinical data. We fitted univariate linear regression models to each gene program-
604 clinical outcome pair and each ecosystem-clinical outcome pair, and retrieved the t-statistics as
605 estimate of association. We further fitted univariate Cox Proportional-Hazard ratio models to

obtain hazard ratios for each program and ecosystem. For survival analysis, we used consolidated survival data from each cohort to plot KM curves to stratify patients.

Cell state co-occurrence ecosystems discovery model

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

When considering why programs are (anti-) correlated in single-cells, there can be two reasons. 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 ecosystems 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:

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

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

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

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 ecosystem parameters by defining:

1. The ecosystem loading matrix $\boldsymbol{H} = (h_{ij})_{P \times L}$
2. The ecosystem 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 ecosystem l, and each row h_p of $\boldsymbol{H}$ is an L-dimensional vector of ecosystem weights in sample p.

With the ecosystem 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:

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

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

To encourage the model to find more distinct ecosystems, we add an orthogonality term to the model target function. The upper triangular matrix of the transpose cross product of the ecosystem basis matrix $\boldsymbol{W}$ is used to calculate a mean as a proxy for ecosystem 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 ecosystem 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 ecosystem discovery model using the statistical programing language Stan. We used cmdstan_model from the R package cmdStanR ver. 0.5.3[94] 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 ecosystem matrices. Covariance between program loadings within respective cell types were input as the initial values of $\boldsymbol{\Sigma}_{intrinsic}$.

PDAC Visium data

Visium data was downloaded from Synapse[95] with author permission. Tumour sample slides from primary tumours were selected. Preprocessing was done using the SpatialData[96] library. Low quality spots containing more than 20% hemoglobin genes and fewer than 1000 detected genes were filtered. Non-protein coding genes, mitochondrial genes, ribosomal genes, and hemoglobin genes were excluded from downstream analysis. Raw UMI counts were log-normalized by calling computePooledFactors followed by logNormCounts with default parameters.

Scoring programs and ecosystems in Visium

Top 20 genes ranked by weight in each gene program were used to compute spot-wise scores in Visium PDAC samples. Since each spot on the slide is effectively a mini bulk sample, we leveraged gene set variation analysis (GSVA) to obtain an enrichment score for each program in each spot of each Visium slide. Next, program scores in each spot were scaled to 0-1 range. Consensus weighing factors for each ecosystem was obtained by removing programs whose weights had opposite signs in discovery vs. validation. We then obtained spot-wise ecosystem scores as the sum of the program scores weighted by the consensus weighing factors. Randomized ecosystem baselines were generated by shuffling the consensus weighing factors.

Concatenated NMF

To obtained consensus NMF factors. All Visium spots were concatenated to create one spot x program score matrix. A single NMF model was fitted on the pooled matrix, and the expression was decomposed into three factors common across images.

Spatial autocorrelation

To measure spatial autocorrelation, we computed the Moran's Index for the inferred ecosystems, the randomized ecosystems, and the concatenated NMF factors in each Visium image. We used Squidpy[97] graph algorithms with default parameters for Moran's I computation.

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.

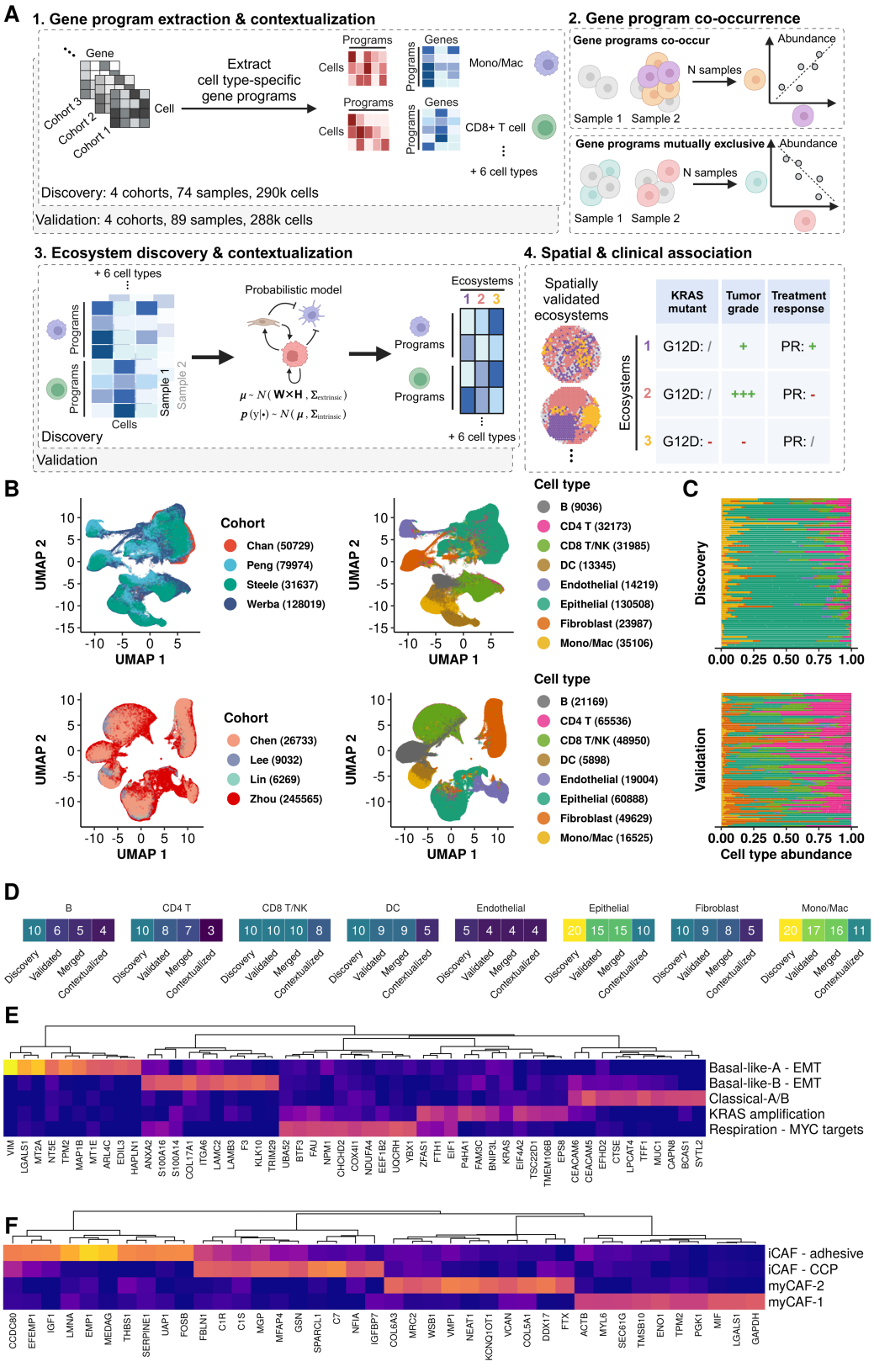
Contributions

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).

Figures


A
1. Gene program extraction & contextualization
Gene
Cell
Cohort 3
Cohort 2
Cohort 1
Extract cell type-specific gene programs
Programs
Cells
Genes
Mono/Mac
CD8+ T cell
+ 6 cell types
Discovery: 4 cohorts, 74 samples, 290k cells
Validation: 4 cohorts, 89 samples, 288k cells
2. Gene program co-occurrence
Gene programs co-occur
Abundance
N samples
Sample 1
Sample 2
Gene programs mutually exclusive
3. Ecosystem discovery & contextualization
+ 6 cell types
Probabilistic model
Ecosystems
1 2 3
Sample 1
Sample 2
Discovery
Validation
4. Spatial & clinical association
Spatially validated ecosystems
KRAS mutant
Tumor grade
Treatment response
Ecosystems
1 G12D: / + PR: +
2 G12D: / +++ PR: -
3 G12D: - - PR: /
B
UMAP 1
UMAP 2
Cohort
Chan (50729)
Peng (79974)
Steele (31637)
Werba (128019)
Cell type
B (9036)
CD4 T (32173)
CD8 T/NK (31985)
DC (13345)
Endothelial (14219)
Epithelial (130508)
Fibroblast (23987)
Mono/Mac (35106)
Chen (26733)
Lee (9032)
Lin (6269)
Zhou (245565)
B (21169)
CD4 T (65536)
CD8 T/NK (48950)
DC (5898)
Endothelial (19004)
Epithelial (60888)
Fibroblast (49629)
Mono/Mac (16525)
C
Discovery
Validation
0.00 0.25 0.50 0.75 1.00
Cell type abundance
D
B
CD4 T
CD8 T/NK
DC
Endothelial
Epithelial
Fibroblast
Mono/Mac
Discovery
Validated
Merged
Contextualized
E
Basal-like-A - EMT
Basal-like-B - EMT
Classical-A/B
KRAS amplification
Respiration - MYC targets
F
iCAF - adhesive
iCAF - CCP
myCAF-2
myCAF-1

789 **Figure 1: Learning multicellular ecosystems directly from scRNA-seq atlases. A**. Overview
790 of discovery-validation platform for identifying co-occurrence and ecosystems of cell type
791 specific gene programs in the PDAC TME. **B**. UMAP representations for the discovery (top) and
792 validation (bottom) datasets colored by cohorts (left) and annotated cell type (right). DC, Myeloid
793 Dendritic Cells. Mono/Mac, Monocyte/Macrophage. **C**. Overview of cell type composition for
794 samples (in rows) in discovery group (top) and in validation group (bottom). **D**, Overview of
795 numbers of candidate gene programs extracted from the discovery group, numbers of gene
796 programs validated between discovery and validation, and numbers of consolidated gene
797 programs after merging the redundant ones in each cell type. **E-F**. Heatmaps showing weights
798 of top ranked genes in selected gene programs from epithelial cells (**E**) and fibroblasts (**F**).
799

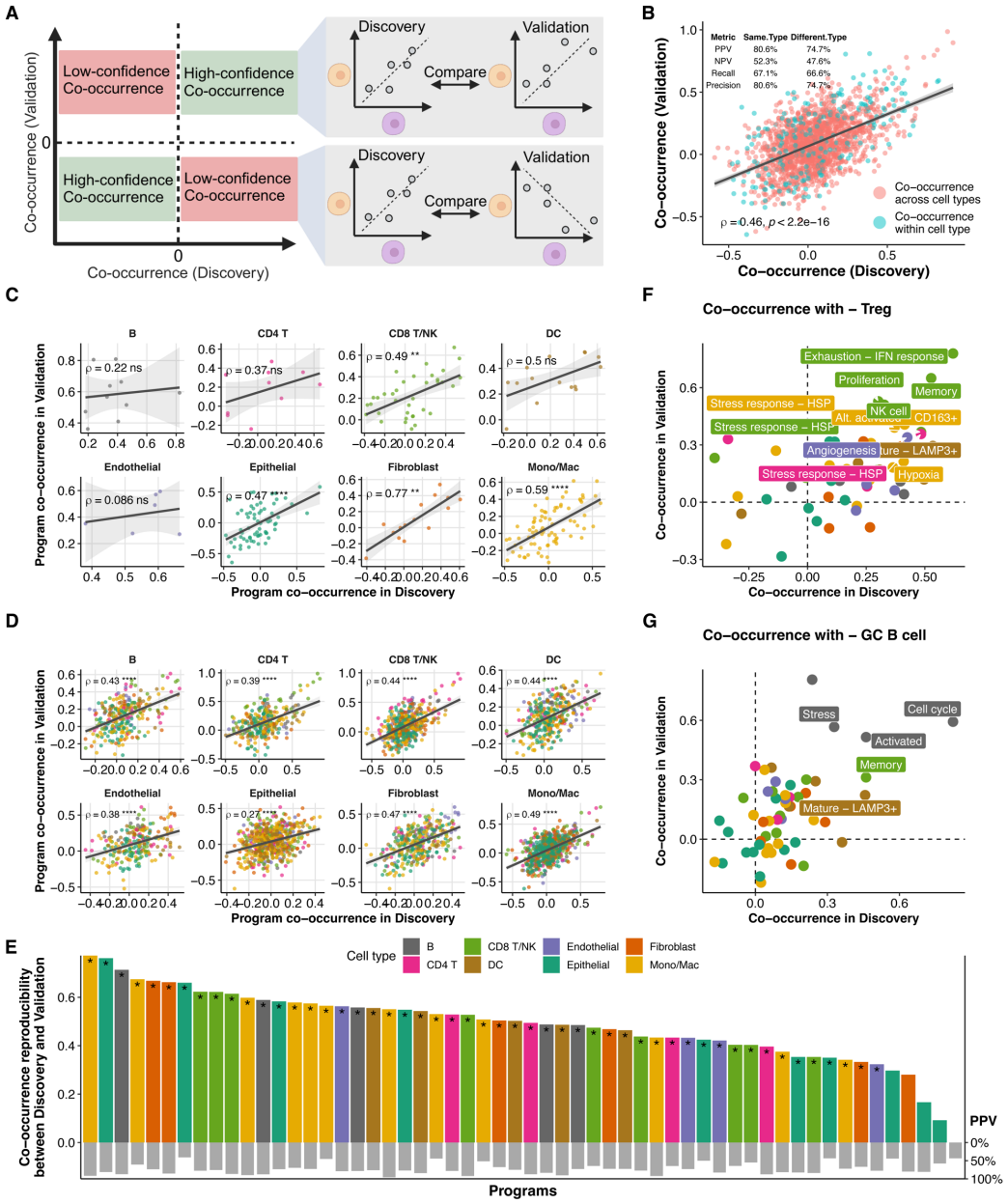


Figure 2: Pairwise co-occurrences of gene programs are highly reproducible. A. Two examples of gene program co-occurrence patterns. The program co-occurrence scores for each program pair are calculated in the discovery and the validation group separately for comparison between datasets. **B.** Program 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.** 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 co-occurrence

811 reproducibility. Asterisk signs denote gene programs whose co-occurrence with other programs
812 is significantly reproducible between discovery and validation. The grey bars below indicate
813 empirical positive predictive value for co-occurrence of the respective gene programs. **F-G,**
814 Sample level co-occurrence between the CD4 regulatory T cell gene program and other
815 programs (**F**), and between the germinal center B cell program and other programs (**G**).
816

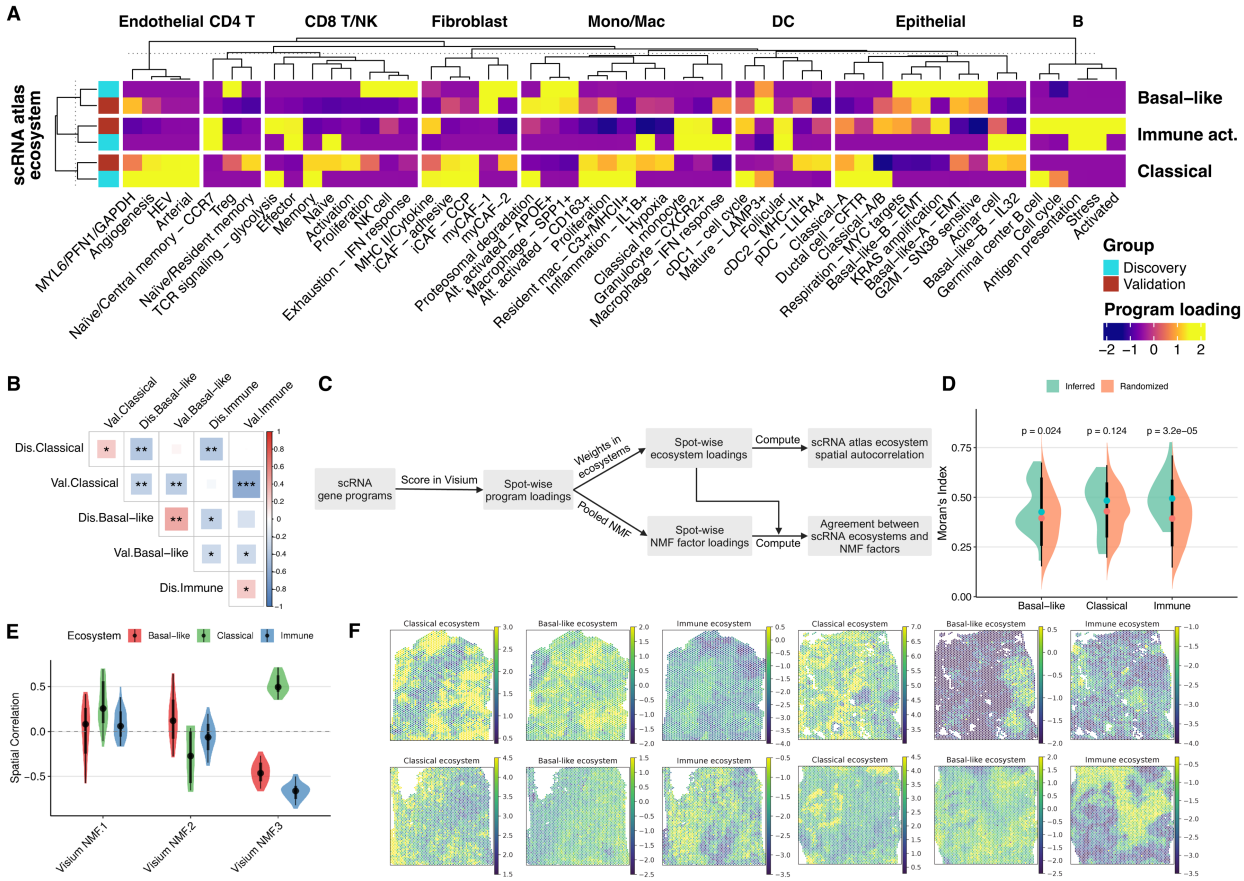


Figure 3: Multicellular ecosystems discovered by a probabilistic model reflect spatial variation. A. Gene program presence in validated ecosystems. **B**. Correlations of ecosystems between discovery and validation datasets. **C.** Workflow to validate scRNA atlas ecosystems using Visium data. **D**. Spatial autocorrelation of validated scRNA-seq inferred ecosystems compared to randomly generated ecosystems. **E**. Spot-wise correlation between scRNA ecosystems and NMF factors computed using spot-wise gene program loadings. **F**. Examples of ecosystem loadings in tumor samples.

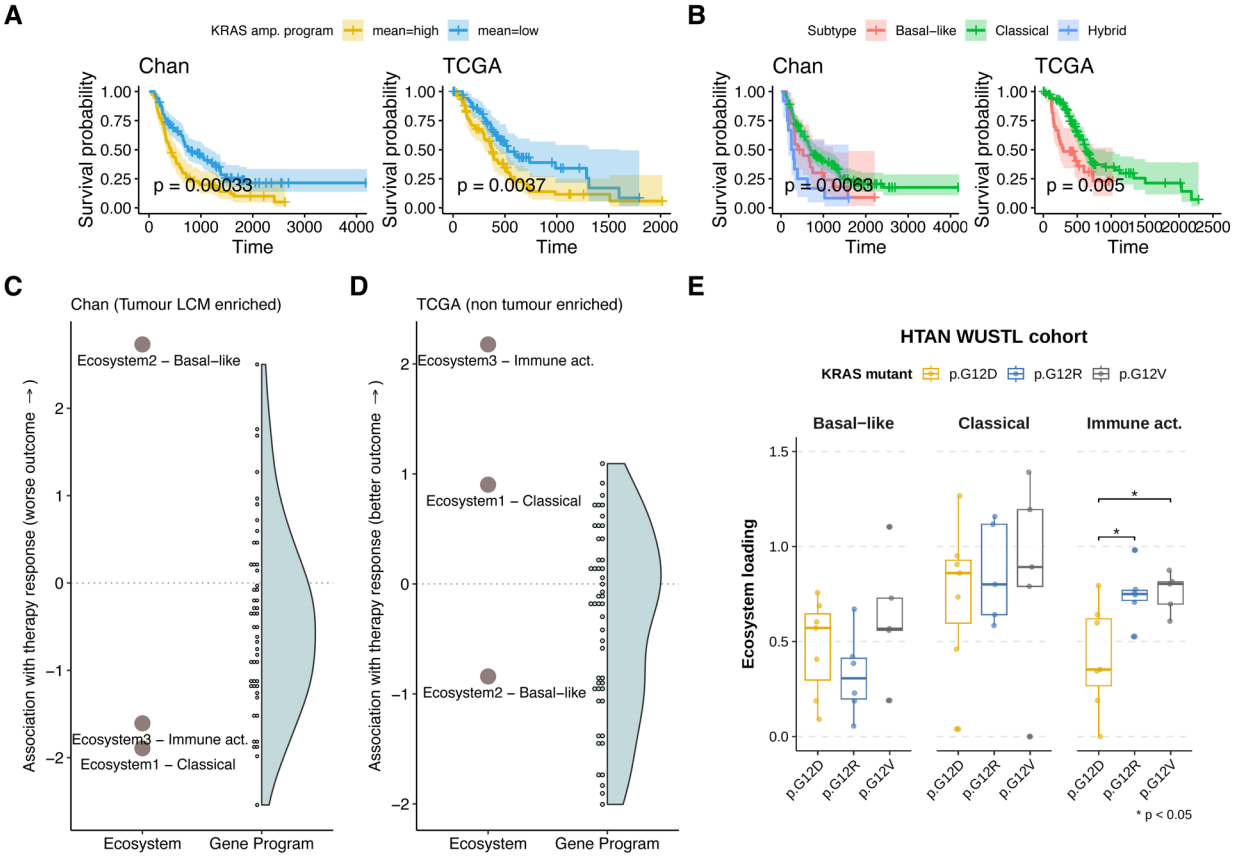


Figure 4: Multicellular ecosystems and cell type-specific programs are associated with patient outcomes and tumour genetics. **A**. The KRAS amplification epithelial program predicts patient survival in the Chan cohort (left) and the TCGA cohort (right). **B**. Dichotomization of patient survival by expression subtypes in the Chan and TCGA PDAC cohort. **C**. Association between ecosystems/programs and therapy outcome in the Chan cohort. **D.** Association between ecosystems/programs and therapy outcome in the TCGA cohort. **E**. Relationships between KRAS mutant and ecosystem loading in tumour samples. '*': p-value < 0.05; '**': p-value < 0.01, '***': p-value < 0.001 (all Benjamini-Hochberg FDR corrected).

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 Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
3. Zhang, S. *et al.* Single cell transcriptomic analyses implicate an immunosuppressive tumor microenvironment in pancreatic cancer liver metastasis. *Nat. Commun.* **14**, 5123 (2023).
4. Cui Zhou, D. *et al.* Spatially restricted drivers and transitional cell populations cooperate with the microenvironment in untreated and chemo-resistant pancreatic cancer. *Nat. Genet.* **54**, 1390–1405 (2022).
5. Hwang, W. L. *et al.* Single-nucleus and spatial transcriptome profiling of pancreatic cancer identifies multicellular dynamics associated with neoadjuvant treatment. *Nat. Genet.* **54**, 1178–1191 (2022).
6. McAndrews, K. M. *et al.* Identification of functional heterogeneity of carcinoma-associated fibroblasts with distinct IL6-mediated therapy resistance in pancreatic cancer. *Cancer Discov.* **12**, 1580–1597 (2022).
7. Chijimatsu, R. *et al.* Establishment of a reference single-cell RNA sequencing dataset for human pancreatic adenocarcinoma. *iScience* **25**, 104659 (2022).
8. Falcomatà, C. *et al.* Selective multi-kinase inhibition sensitizes mesenchymal pancreatic cancer to immune checkpoint blockade by remodeling the tumor microenvironment. *Nat. Cancer* **3**, 318–336 (2022).
9. Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods* **11**, 417–422 (2014).
10. Angelo, M. *et al.* Multiplexed ion beam imaging of human breast tumors. *Nat. Med.* **20**,

862 436–442 (2014).
863 11. Janesick, A. *et al.* High resolution mapping of the tumor microenvironment using integrated
864 single-cell, spatial and in situ analysis. *Nat. Commun.* **14**, 8353 (2023).
865 12. He, S. *et al.* High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue
866 by spatial molecular imaging. *Nat. Biotechnol.* **40**, 1794–1806 (2022).
867 13. Jackson, H. W. *et al.* The single-cell pathology landscape of breast cancer. *Nature* **578**,
868 615–620 (2020).
869 14. Kim, S. *et al.* Integrative analysis of spatial and single-cell transcriptome data from human
870 pancreatic cancer reveals an intermediate cancer cell population associated with poor
871 prognosis. *Genome Med.* **16**, 20 (2024).
872 15. Moncada, R. *et al.* Integrating microarray-based spatial transcriptomics and single-cell
873 RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat.*
874 *Biotechnol.* **38**, 333–342 (2020).
875 16. Storrs, E. P. *et al.* High-dimensional deconstruction of pancreatic cancer identifies tumor
876 microenvironmental and developmental stemness features that predict survival. *NPJ*
877 *Precis. Oncol.* **7**, 105 (2023).
878 17. Luca, B. A. *et al.* Atlas of clinically distinct cell states and ecosystems across human solid
879 tumors. *Cell* **184**, 5482-5496.e28 (2021).
880 18. Wu, S. Z. *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat.*
881 *Genet.* **53**, 1334–1347 (2021).
882 19. Sorin, M. *et al.* Single-cell spatial landscapes of the lung tumour immune microenvironment.
883 *Nature* **614**, 548–554 (2023).
884 20. Ni, Z. *et al.* SpotClean adjusts for spot swapping in spatial transcriptomics data. *Nat.*
885 *Commun.* **13**, 2971 (2022).
886 21. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early
887 organogenesis. *Nature* **566**, 490–495 (2019).

888 22. Smillie, C. S. *et al.* Intra- and inter-cellular rewiring of the human colon during ulcerative
889 colitis. *Cell* **178**, 714-730.e22 (2019).
890 23. Büttner, M., Ostner, J., Müller, C. L., Theis, F. J. & Schubert, B. scCODA is a Bayesian
891 model for compositional single-cell data analysis. *Nat. Commun.* **12**, 6876 (2021).
892 24. Rozenblatt-Rosen, O. *et al.* Building a high-quality Human Cell Atlas. *Nat. Biotechnol.* **39**,
893 149–153 (2021).
894 25. Rozenblatt-Rosen, O. *et al.* The Human Tumor Atlas Network: Charting tumor transitions
895 across space and time at single-cell resolution. *Cell* **181**, 236–249 (2020).
896 26. Terekhanova, N. V. *et al.* Epigenetic regulation during cancer transitions across 11 tumour
897 types. *Nature* **623**, 432–441 (2023).
898 27. Heiser, C. N. *et al.* Molecular cartography uncovers evolutionary and microenvironmental
899 dynamics in sporadic colorectal tumors. *Cell* **186**, 5620-5637.e16 (2023).
900 28. Johnson, B. E. *et al.* An omic and multidimensional spatial atlas from serial biopsies of an
901 evolving metastatic breast cancer. *Cell Rep. Med.* **3**, 100525 (2022).
902 29. Yanagawa, J. *et al.* Single-cell characterization of pulmonary nodules implicates
903 suppression of immunosurveillance across early stages of lung adenocarcinoma. *Cancer*
904 *Res.* **83**, 3305–3319 (2023).
905 30. Glasner, A. *et al.* Conserved transcriptional connectivity of regulatory T cells in the tumor
906 microenvironment informs new combination cancer therapy strategies. *Nat. Immunol.* **24**,
907 1020–1035 (2023).
908 31. Chan, J. M. *et al.* Signatures of plasticity, metastasis, and immunosuppression in an atlas
909 of human small cell lung cancer. *Cancer Cell* **39**, 1479-1496.e18 (2021).
910 32. Chan-Seng-Yue, M. *et al.* Transcription phenotypes of pancreatic cancer are driven by
911 genomic events during tumor evolution. *Nat. Genet.* **52**, 231–240 (2020).
912 33. Peng, J. *et al.* Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant
913 progression in pancreatic ductal adenocarcinoma. *Cell Res.* **29**, 725–738 (2019).

914 34. Steele, N. G. *et al.* Multimodal mapping of the tumor and peripheral blood immune
915 landscape in human pancreatic cancer. *Nature Cancer* **1**, 1097–1112 (2020).
916 35. Werba, G. *et al.* Single-cell RNA sequencing reveals the effects of chemotherapy on human
917 pancreatic adenocarcinoma and its tumor microenvironment. *Nat. Commun.* **14**, 797
918 (2023).
919 36. Lee, J. J. *et al.* Elucidation of tumor-stromal heterogeneity and the ligand-receptor
920 interactome by single-cell transcriptomics in real-world pancreatic cancer biopsies. *Clin.*
921 *Cancer Res.* **27**, 5912–5921 (2021).
922 37. Lin, W. *et al.* Single-cell transcriptome analysis of tumor and stromal compartments of
923 pancreatic ductal adenocarcinoma primary tumors and metastatic lesions. *Genome Med.*
924 **12**, 80 (2020).
925 38. Chen, K. *et al.* Single-cell RNA-seq reveals dynamic change in tumor microenvironment
926 during pancreatic ductal adenocarcinoma malignant progression. *eBioMedicine* **66**, (2021).
927 39. Öhlund, D. *et al.* Distinct populations of inflammatory fibroblasts and myofibroblasts in
928 pancreatic cancer. *J. Exp. Med.* **214**, 579–596 (2017).
929 40. Zhang, T., Ren, Y., Yang, P., Wang, J. & Zhou, H. Cancer-associated fibroblasts in
930 pancreatic ductal adenocarcinoma. *Cell Death Dis.* **13**, 1–11 (2022).
931 41. Caronni, N. *et al.* IL-1β+ macrophages fuel pathogenic inflammation in pancreatic cancer.
932 *Nature* **623**, 415–422 (2023).
933 42. LaRue, M. M. *et al.* Metabolic reprogramming of tumor-associated macrophages by
934 collagen turnover promotes fibrosis in pancreatic cancer. *Proceedings of the National*
935 *Academy of Sciences* **119**, e2119168119 (2022).
936 43. Cassetta, L. & Pollard, J. W. Targeting macrophages: therapeutic approaches in cancer.
937 *Nat. Rev. Drug Discov.* **17**, 887–904 (2018).
938 44. Moffitt, R. A. *et al.* Virtual microdissection identifies distinct tumor- and stroma-specific
939 subtypes of pancreatic ductal adenocarcinoma. *Nat. Genet.* **47**, 1168–1178 (2015).

940 45. Oh, K. *et al.* Coordinated single-cell tumor microenvironment dynamics reinforce pancreatic
941 cancer subtype. *Nat. Commun.* **14**, 1–13 (2023).
942 46. Grünwald, B. T. *et al.* Spatially confined sub-tumor microenvironments in pancreatic cancer.
943 *Cell* **184**, 5577-5592.e18 (2021).
944 47. Chen, K. *et al.* Immune profiling and prognostic model of pancreatic cancer using
945 quantitative pathology and single-cell RNA sequencing. *J. Transl. Med.* **21**, 210 (2023).
946 48. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of
947 Brain Cell Identity. *Cell* **177**, 1873-1887.e17 (2019).
948 49. Raghavan, S. *et al.* Microenvironment drives cell state, plasticity, and drug response in
949 pancreatic cancer. *Cell* **184**, 6119-6137.e26 (2021).
950 50. Elyada, E. *et al.* Cross-species single-cell analysis of pancreatic ductal adenocarcinoma
951 reveals antigen-presenting cancer-associated fibroblasts. *Cancer Discov.* **9**, 1102–1123
952 (2019).
953 51. Kunes, R. Z., Walle, T., Land, M., Nawy, T. & Pe'er, D. Supervised discovery of
954 interpretable gene programs from single-cell data. *Nat. Biotechnol.* **42**, 1084–1095 (2024).
955 52. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of
956 single-cell RNA-seq data. *Nat. Rev. Genet.* **20**, 273–282 (2019).
957 53. Kuhn, H. W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **2**,
958 83–97 (1955).
959 54. Janssen, P. *et al.* The effect of background noise and its removal on the analysis of single-
960 cell expression data. *Genome Biol.* **24**, 140 (2023).
961 55. Helms, E., Onate, M. K. & Sherman, M. H. Fibroblast heterogeneity in the pancreatic tumor
962 microenvironment. *Cancer Discov.* **10**, 648–656 (2020).
963 56. Humpton, T. J. *et al.* Oncogenic KRAS induces NIX-mediated mitophagy to promote
964 pancreatic cancer. *Cancer Discov.* **9**, 1268–1287 (2019).
965 57. Shi, D. D. *et al.* Therapeutic avenues for cancer neuroscience: translational frontiers and

966 clinical opportunities. *Lancet Oncol.* **23**, e62–e74 (2022).
967 58. Rossi Sebastiano, M. *et al.* ACSL3-PAI-1 signaling axis mediates tumor-stroma cross-talk
968 promoting pancreatic cancer progression. *Sci. Adv.* **6**, eabb9200 (2020).
969 59. Loncle, C. *et al.* The pancreatitis-associated protein VMP1, a key regulator of inducible
970 autophagy, promotes KrasG12D-mediated pancreatic cancer initiation. *Cell Death Dis.* **7**,
971 e2295–e2295 (2016).
972 60. Gao, H. *et al.* The expression of versican and its role in pancreatic neuroendocrine tumors.
973 *Pancreatology* **20**, 142–147 (2020).
974 61. Li, K. *et al.* Myeloid-derived suppressor cells as immunosuppressive regulators and
975 therapeutic targets in cancer. *Signal Transduction and Targeted Therapy* **6**, 1–25 (2021).
976 62. Bärthel, S., Falcomatà, C., Rad, R., Theis, F. J. & Saur, D. Single-cell profiling to explore
977 pancreatic cancer heterogeneity, plasticity and response to therapy. *Nat. Cancer* **4**, 454–
978 467 (2023).
979 63. Gulhati, P. *et al.* Targeting T cell checkpoints 41BB and LAG3 and myeloid cell
980 CXCR1/CXCR2 results in antitumor immunity and durable response in pancreatic cancer.
981 *Nat. Cancer* **4**, 62–80 (2023).
982 64. Goveia, J. *et al.* An Integrated Gene Expression Landscape Profiling Approach to Identify
983 Lung Tumor Endothelial Cell Heterogeneity and Angiogenic Candidates. *Cancer Cell* **37**,
984 421 (2020).
985 65. Geldhof, V. *et al.* Single cell atlas identifies lipid-processing and immunomodulatory
986 endothelial cells in healthy and malignant breast. *Nat. Commun.* **13**, 1–19 (2022).
987 66. Nieto, P. *et al.* A single-cell tumor immune atlas for precision oncology. *Genome Res.* **31**,
988 1913–1926 (2021).
989 67. Fridman, W. H. *et al.* B cells and tertiary lymphoid structures as determinants of tumour
990 immune contexture and clinical outcome. *Nat. Rev. Clin. Oncol.* **19**, 441–457 (2022).
991 68. Gorgun, C. *et al.* Synergistic role of three dimensional niche and hypoxia on conservation of

992 cancer stem cell phenotype. *Int. J. Biol. Macromol.* **90**, 20–26 (2016).
993 69. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality
994 Survival Outcome Analytics. *Cell* **173**, 400-416.e11 (2018).
995 70. Cancer Genome Atlas Research Network. Electronic address:
996 andrew_aguirre@dfci.harvard.edu & Cancer Genome Atlas Research Network. Integrated
997 Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell* **32**, 185-
998 203.e13 (2017).
999 71. Till, J. E. *et al.* Circulating KRAS G12D but not G12V is associated with survival in
1000 metastatic pancreatic ductal adenocarcinoma. *Nat. Commun.* **15**, 1–12 (2024).
1001 72. Leidner, R. *et al.* Neoantigen T-cell receptor gene therapy in pancreatic cancer. *N. Engl. J.*
1002 *Med.* **386**, 2112–2119 (2022).
1003 73. Uhlen, M. *et al.* A pathology atlas of the human cancer transcriptome. *Science* **357**, (2017).
1004 74. Marisa, L. *et al.* Gene expression classification of colon cancer into molecular subtypes:
1005 Characterization, validation, and prognostic value. *PLoS Med.* **10**, e1001453 (2013).
1006 75. Győrffy, B., Surowiak, P., Budczies, J. & Lánczky, A. Online survival analysis software to
1007 assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung
1008 cancer. *PLoS One* **8**, e82241 (2013).
1009 76. Calon, A. *et al.* Stromal gene expression defines poor-prognosis subtypes in colorectal
1010 cancer. *Nat. Genet.* **47**, 320–329 (2015).
1011 77. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary
1012 glioblastoma. *Science* **344**, 1396–1401 (2014).
1013 78. Zilionis, R. *et al.* Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals
1014 Conserved Myeloid Populations across Individuals and Species. *Immunity* **50**, 1317-
1015 1334.e10 (2019).
1016 79. Savas, P. *et al.* Single-cell profiling of breast cancer T cells reveals a tissue-resident
1017 memory subset associated with improved prognosis. *Nat. Med.* **24**, 986–993 (2018).

1018 80. van der Leun, A. M., Thommen, D. S. & Schumacher, T. N. CD8+ T cell states in human
1019 cancer: insights from single-cell analysis. *Nat. Rev. Cancer* **20**, 218–232 (2020).
1020 81. Kang, J. *et al.* Systematic dissection of tumor-normal single-cell ecosystems across a
1021 thousand tumors of 30 cancer types. *Nat. Commun.* **15**, 1–17 (2024).
1022 82. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-
1023 Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337.e4
1024 (2019).
1025 83. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality
1026 control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**,
1027 1179–1186 (2017).
1028 84. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a
1029 transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
1030 85. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Res.* **10**, 33 (2021).
1031 86. Wang, Y. *et al.* Interactive analysis of single-cell data using flexible workflows with SCTK2.
1032 *Patterns (N Y)* **4**, 100814 (2023).
1033 87. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC*
1034 *Bioinformatics* **11**, 367 (2010).
1035 88. Shen, L. GeneOverlap: An R package to test and visualize gene overlaps. *R Package*
1036 (2016).
1037 89. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
1038 90. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.
1039 *Innovation (Camb.)* **2**, 100141 (2021).
1040 91. Gavish, A. *et al.* Hallmarks of transcriptional intratumour heterogeneity across a thousand
1041 tumours. *Nature* **618**, 598–606 (2023).
1042 92. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* 060012 (2021)
1043 doi:10.1101/060012.

1044 93. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for
1045 microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
1046 94. Jonah Gabry and Rok Češnovar and Andrew Johnson and Steve Bronder. cmdstanr: R
1047 Interface to “CmdStan.” https://mc-stan.org/cmdstanr/ (2023).
1048 95. de Bruijn, I. *et al.* Sharing data from the Human Tumor Atlas Network through standards,
1049 infrastructure and community engagement. *Nat. Methods* **22**, 664–671 (2025).
1050 96. Marconato, L. *et al.* SpatialData: an open and universal data framework for spatial omics.
1051 *Nat. Methods* **22**, 58–62 (2025).
1052 97. Palla, G. *et al.* Squidpy: a scalable framework for spatial omics analysis. *Nat. Methods* **19**,
1053 171–178 (2022).
1054