Pancreatic Ductal Adenocarcinoma Comprises Coexisting Regulatory States with both Common and Distinct Dependencies

Pasquale Laise[1,2,†], Mikko Turunen[1,†], H. Carlo Maurer[3], Alvaro G. Curiel[4], Ela Elyada[5,6], Bernhard Schmierer[7], Lorenzo Tomassoni[1], Jeremy Worley[1], Mariano J. Alvarez[1,2], Jordan Kesner[1], Xiangtian Tan[1], Somnath Tagore[1], Alexander L. E. Wang[1], Sabrina Ge[8,9], Alina Cornelia Iuga[10], Aaron Griffin[1], Winston Wong[11,12] , Gulam Manji[11,12,13], Faiyaz Notta[8,9,14], David A. Tuveson[5,6], Kenneth P. Olive[4,11,13,*] and Andrea Califano[1,13,16,17,18,*]

[1] Department of Systems Biology, Columbia University, New York, NY, US

[2] DarwinHealth Inc., New York, NY US

[3] Klinikum rechts der Isar, II, Medizinische Klinik, Technische Universität München, 81675, Munich, Germany.

[4] Division of Digestive and Liver Diseases, Department of Medicine, Vagelos College of Physicians and Surgeons, Columbia University, New York, NY, USA.

[5] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, US

[6] Lustgarten Foundation Pancreatic Cancer Research Laboratory, Cold Spring Harbor, New York.

[7] Karolinska Institutet, Stockholm, Sweden

[8] Princess Margaret Cancer Centre, Toronto, Ontario, Canada.

[9] Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada

[10] Department of Pathology and Laboratory Medicine, Univ. of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[11] Department of Medicine, Vagelos College of Physicians and Surgeons, Columbia University, New York, NY, USA.

[12] Division of Hematology and Oncology, Columbia University Irving Medical Center and New York Presbyterian Hospital Herbert Irving Pavilion, New York, NY, USA.

[13] Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY, USA

[14] PanCuRx Translational Research Initiative, Ontario Institute for Cancer Research, Toronto, Ontario, Canada

[15] Department of Pathology and Cell Biology, Vagelos College of Physicians and Surgeons, Columbia University, New York, NY, USA.

[16] Department of Biomedical Informatics, Columbia University, New York, NY, USA.

[17] Department of Biochemistry and Molecular Biophysics, Columbia University, New York,

[18] J.P. Sulzberger Columbia Genome Center, New York, New York, USA.

[*] Correspondence to: kenolive@columbia.edu and ac2248@cumc.columbia.edu

Summary

Despite extensive efforts to characterize the transcriptional landscape of pancreatic ductal adenocarcinoma (PDA), reproducible assessment of subtypes with actionable dependencies remains challenging. Systematic, network-based analysis of regulatory protein activity stratified PDA tumours into novel functional subtypes that were highly conserved across multiple cohorts, including at the single cell level and in laser capture microdissected (LCM) samples. Identified subtypes were characterized by activation of master regulator proteins representing either gastrointestinal lineage markers or transcriptional effectors of morphogen pathways. Single cell analysis confirmed the existence of Lineage and Morphogenic states but also revealed a dominant population of more differentiated Oncogenic Precursor (OP) cells , present in all sampled patients, yet not apparent from bulk tumor analysis. Master regulators were validated by pooled, CRISPR/Cas9 screens, demonstrating both subtype-specific and universal dependencies. Conversely, ectopic expression of Lineage MRs, such as OVOL2, was sufficient to reprogram Morphogenic cells, thus providing a roadmap for the future targeting of patient-specific dependencies in PDA.

Pancreatic ductal adenocarcinoma (PDA) is the third-leading cause of cancer-related mortality and is highly resistant to cytotoxic, targeted, and immune therapies[1]. Compared to the heterogeneous mutational repertoire of other cancers, PDA is remarkable for its relatively uniform complement of DNA alterations, with frequent mutations in *KRAS*, *CDKN2A*, *TP53*, and *SMAD4*. Unfortunately, these hallmark events are not currently targetable and other mutations known to confer sensitivity to specific drugs are uncommon. Consequently, cytotoxic combinations remain the standard of care, with most patients quickly exhibiting primary or secondary chemoresistance.

Cellular heterogeneity has emerged as a major contributor to cancer chemoresistance, due to potential coexistence of malignant subpopulations with distinct transcriptional states (*cellular subtypes*)[2] and to the contribution of diverse stromal subpopulations[3]. The former may present with orthogonal drug sensitivity and remarkable plasticity, thus serving as chemoresistance reservoirs and frustrating efforts to delineate therapeutic vulnerabilities through bulk tissue analysis. Stromal cells further complicate matters as they often represent the dominant compartment in bulk PDA samples. Elucidating the repertoire of distinct, yet coexisting malignant subpopulations in PDA, independent of stromal contamination, thus represents a critical first step toward the rational design of more effective targeted and combination therapies.

While multiple studies agree on the presence of at least two transcriptional subtypes of PDA, the specific molecular signatures of these subtypes remain controversial and non-overlapping[4]. Overall, more differentiated tumors—corresponding to *Classical* or *Progenitor* subtypes in prior studies—are generally associated with better outcome, compared to poorly differentiated ones—termed *Quasi-mesenchymal*, *Basal-like*, or *Squamous* [5-10]. However, published classifiers produce limited overlap when applied across available cohorts[4]. While removing stromal

contributions from expression signatures helps harmonise discrepancies[10], a unifying molecular PDA classification schema is still elusive and an important next step for the field.

Regulatory network analysis can yield highly-multiplexed, tumor-specific gene reporter assays that accurately measure the activity of Transcriptional Regulator (TR) proteins—including transcription factors, co-factors, and chromatin remodeling enzymes—from gene expression profiles (see[11] for a comprehensive perspective). Specifically, the VIPER algorithm enables genome-wide quantification of TR activity with >80% accuracy. This helps identify the most aberrantly activated and repressed TRs between phenotypes as candidate Master Regulator (MR) proteins that may mechanistically implement and maintain a tumor cell's transcriptional state[12]. Protein activity measurement accuracy has been confirmed by silencing assays and targeted small-molecule perturbations and has helped identify MR proteins that were extensively validated[13-16]. VIPER calculates a protein's differential activity from the relative abundance of its target genes (*regulon*), as identified by the ARACNe algorithm[17] (see Supplementary Notes). More recently, the metaVIPER algorithm, which integrates VIPER predictions from multiple networks, was shown to further improve activity inference, including in single cell analyses[18], where it virtually eliminates gene dropout effects arising from low sequencing depth. The high reproducibility of VIPER and metaVIPER has led to NY State CLIA certification for two VIPER-based algorithms, OncoTarget[19] and OncoTreat[13] that are now routinely used in multiple clinical trials and studies.

VIPER-based analysis of gene expression profiles from 200 LCM human primary PDA epithelium samples identified three clusters, two of which corresponding to highly molecularly-distinct subtypes characterized by aberrant activity of either gastrointestinal lineage markers (*Lineage*

subtype) or proteins representing effectors of known morphogen signaling pathways (*Morphogenic subtype*); the third LCM cluster indicated a mixing of these two states. Single cell analysis of two independent cohorts confirmed coexistence of cellular subpopulations representing Lineage and Morphogenic transcriptional states at widely different ratios between individual tumors, and confirmed that the third LCM cluster lacks a corresponding cell state. Unexpectedly, single cell analysis revealed a third, molecularly-distinct cellular subtype, termed Oncogenic Progenitor (OP) that was less malignant and more differentiated than Lineage or Morphogenic cells. OP cells represented a large and relatively constant fraction of PDA sample cells, thus impeding their detection through differential analysis of bulk tumor tissues. RNA-velocity analysis suggested that OP cells may provide a reservoir for the replenishment of Lineage and Morphogenic cells, thus contributing to chemoresistance. Pooled, CRISPR/Cas9-mediated validation confirmed enrichment of candidate MRs in PDA essential proteins. Furthermore, bulk and single-cell based, pooled ectopic expression of candidate MR proteins confirmed the ability of subtype-specific MRs, such as OVOL2, to reprogram PDA cell state; the efficiency of reprogramming was even higher when OVOL2 was co-expressed with other MRs such as HNF1A. Taken together, these results provide a foundation for the development of novel therapies targeting the individual epithelial subpopulations that comprise bulk PDA and their molecular dependencies. A conceptual workflow of the analysis is depicted in Fig. 1a.

Regulatory, network-based classification of PDA: Tissue heterogeneity and stromal desmoplasia represent major confounding factors in PDA tissue analysis[10]. We thus generated a novel collection of RNASeq profiles from the microdissected epithelium of 200 PDA samples, 26 low-grade PanINs, and 19 IPMN adenomas (collectively, CUMC-E cohort). ARACNe analysis yielded a regulatory network (CUMC–E_{Net}) comprising >250,000 regulatory interactions for 1,795

Transcriptional Regulators (TRs). We then used VIPER to transform PDA RNASeq profiles into TR activity profiles, using the protein regulons identified by this network. Protein activity based clustering, using the Partitioning Around Medoids (PAM) algorithm[20], identified a $k = 3$ cluster solution as optimal, by AUC analysis (see Methods and Extended Data Fig.1a).

Based on functional analysis, we named the first cluster "Lineage Subtype," due to the presence of established gastrointestinal transcription factors (e.g., GATA6, PDX1, HNF1B, SOX9, GATA4, HNF1A, and CDX2) among the most aberrantly activated MRs, relative to the other clusters (Fig. 1b). However, activity of these TFs was even higher in PanIN and IPMN samples, suggesting that Lineage tumors may reflect partial loss of GI commitment. In sharp contrast, we named the third cluster "Morphogenic Subtype" to highlight inactivation of GI lineage markers, and aberrant activation of EMT regulators (e.g., ZBED2, ZEB1, ZEB2, SNAI2) and morphogen pathway transcription factors (e.g., GLI2, GLI3, NOTCH2, and SMAD3). This was confirmed by GSEA showing enrichment for hallmarks related to of WNT and TGF-β-mediated programs (Extended Data Fig.1b). Notably, blinded histopathological analysis of adjacent sections from source tissue blocks of CUMC-E samples showed that Morphogenic tumors were more likely to be poorly differentiated compared to Lineage tumors ($p = 8.9 \times 10^{-3}$) (Extended Data Fig. 1c), and an analysis of curated clinical data showed that the Morphogenic subtype was associated with shorter overall survival ($p = 9.1 \times 10^{-3}$) (Fig. 1c). Finally, the remaining cluster emerged as an *Intermediate Subtype* state, characterized by tumors showing a milder activation gradient of Lineage to Morphogenic MRs (Fig. 1b and Extended Data Fig. 1b). Such a gradient suggests that these samples may arise from a heterogeneous mixture of cells representative of the other two subtypes.

Lineage and Morphogenic transcriptional states are conserved across multiple cohorts: We next sought to assess consistency of this regulatory network-based classification in the more common and translatable setting of bulk tissue expression profiles, by leveraging two large-scale cohorts with available RNASeq profile data profiled by ICGC and TCGA[7,8]. For the latter, we used only samples annotated as "high purity" (n = 76). For this analysis, we leveraged the metaVIPER algorithm, which allows the integration of multiple regulatory networks, including those from these two cohorts, as well as the CUMC-E network. Network integration was previously shown to improve the detection of cancer driver genes in global cancer analyses[21]. Here we found that metaVIPER improved detection of PDA genetic dependencies, while preserving the regulatory programs of the epithelial compartment (Extended Data Fig. 2 a-b and Supplementary Notes).

Cluster analysis yielded k = 2 as the optimal cluster number in both cohorts. Yet, differentially active proteins between these clusters were strikingly enriched in Lineage and Morphogenic MRs ($p_{LCM \rightarrow ICGC} = 9.4 \times 10^{-42}$, $p_{LCM \rightarrow TCGA} = 2.0 \times 10^{-14}$) by 2-tails aREA test [12], as identified from the LCM dataset, respectively (Fig. 2a-b). Moreover, there was highly significant overlap between the MRs of ICGC and TCGA clusters ($p_{ICGC \rightarrow TCGA} \leq 1.4 \times 10^{-44}$) (Fig. 2c).

We have recently shown[22] that MR proteins representing the most differentially activated TRs form modular structures (*Tumor Checkpoints*) that canalize[23,24] the effect of genetic alterations in their upstream pathways[25]. In PDA, the 50 most activated and inactivated MRs were sufficient to account for ~75% of patient-specific alterations in the TCGA cohort. Consistent with this observation, a classifier trained on the top 50 MRs for each subtype (i.e., *Tumor Checkpoint MRs* for simplicity), from either ICGC or TCGA samples, produced almost perfect classification in the other cohort, as assessed by Area Under the Curve analysis ($AUC_{ICGC \rightarrow TCGA} = 0.95$ and

$AUC_{TCGA \rightarrow ICGC}$ = 0.92, respectively) (Fig. 2d). Similar to previous cross-cohort studies[14], we thus used the Stouffer's method[26] to integrate the p-value of each MR, as inferred from the two cohorts, yielding an integrated TCGA/ICGC MR signature (Fig. 2e). Differential activity of Tumor Checkpoint MRs, following integration, was highly conserved in the LCM cohort, as well as in two microarray-based cohorts (UNC[6] and Collison[5]) (Extended Data Fig. 2d). Consistently, a Tumor Checkpoint MR-based Random Forest classifier trained on TCGA and ICGC samples produced almost perfect classification of LCM samples (AUC = 0.97), confirming that bulk-samples analysis with ARACNe networks effectively recapitulates MR activity of pure epithelial samples (Fig. 2f).

Consistent with prior reports[5], genetic alterations did not co-segregate with either expression or activity-based subtypes (Extended Data Fig. 2e). In contrast, epigenetic analysis of TCGA samples revealed a strikingly distinct differential DNA methylation pattern in Lineage vs. Morphogenic samples (Fig. 2g-h), including for subtype-specific MRs. For instance, Lineage MRs (GATA6 and HNF1A) and Morphogenic MRs (ZEB1 and ZNF423) were differentially methylated between subtypes, demonstrating concordance between epigenetic and regulatory states. Of note, while clusters corresponding to Lineage and Morphogenic subtypes were consistently identified across all previously published PDA cohorts, there was only limited overlap with prior, gene expression-based subtypes (Fig. 2i). For instance, Morphogenic samples represented an almost equivalent fraction of samples previously classified as immunogenic or progenitor[7] (19% and 21%, respectively); the same was true for Lineage samples (34% and 43%, respectively). Only the squamous subtype[7] was found to comprise mostly Morphogenic rather than Lineage samples (46% vs. 2%). Taken together, these data demonstrate that the proposed classification, as confirmed by LCM epithelial samples analysis, is both distinct from prior ones and largely

independent of stromal contamination, owing to the integrative use of regulatory networks that recapitulate protein transcriptional targets in an epithelial setting.

Single-cell analysis identifies three subtypes: To test our earlier hypothesis that Intermediate Subtype samples from the CUMC-E cohort comprises an admixture of subtypes, we used metaVIPER to measure protein activity from single cell RNASeq profiles (scRNASeq), initially with 1,900 epithelial cells dissociated from 6 fresh human PDA samples[3]. Of these, 30 cells were removed as non-malignant, based on ploidy analysis[27] (see methods). A single-cell epithelial PDA network $scPDA_{Net}$ was then generated by ARACNe and included in the metaVIPER analysis (see methods). Three distinct clusters emerged (Fig. 3a and Extended Data Fig. 3a). Strikingly, Lineage and Morphogenic MRs from bulk samples were almost perfectly recapitulated by proteins differentially active in two of the single-cell clusters (SC_2 and SC_3). In contrast, SC_1 was not enriched in markers of either bulk sample subtype (Fig. 3b), but reflected an entirely novel transcriptional state characterized by high relative activity of early pan-GI lineage markers, such as NEUROD1 and RFX6 (Fig. 3c), chromatin remodeling enzymes such as PHF2, HOX genes such as HOXA2, and a large fraction of poorly characterized zinc-finger factors. We named these Oncogenic Precursor (OP) cells in light of the following observations. First, the SLICE algorithm[28]—which uses entropy analysis to infer stemness—assessed OP cells to be more differentiated than Lineage and Morphogenic cells (Fig. 3d). Second, we noted that established PDA-related stemness markers [29,30] were specifically overexpressed in Lineage ($MSI2^{High}$) and Morphogenic cells ($MSI2^{High}$ and $PROM1/CD133^{High}$) but not in OP cells. CXCR4, an established marker of drug resistance[31], was also specifically overexpressed in Morphogenic but not OP cells. Finally, RNA Velocity analysis[32] suggests that a subset of OP cells occupy a transient rather than stable state, consistent with reprogramming towards either a Lineage or Morphogenic states (Fig.

3e). These cells were enriched in two of the three probability density peaks for the OP cells (OP-L and OP-M, Fig. 3a) and exhibited partial activation of either Lineage or Morphogenic markers, suggesting that a fraction of OP cells may already be committed to these endpoints.

Analysis of individual tumors showed a large fraction of OP cells in every sample, ranging from 39 – 78% of cells (Fig. 3f). In contrast, the Lineage to Morphogenic cell ratio was highly variable, with two out of six samples exhibiting only two dominant cellular states. The OP signature was minimally differentially activated between tumors, thus providing a rationale for why it is not identified as an independent subtype by bulk-sample analysis. Consistent with this observation, synthetic bulk samples—with half of the cells comprising OPCs and the other half comprising a variable ratio (0% to 100%) of Lineage vs. Morphogenic cells—recapitulated the three bulk subtypes detected in the LCM dataset (methods and Extended Data Fig. 3d) and confirmed the Intermediate LCM subtype as the likely byproduct of a heterogeneous mixture of cells in different states.

Analysis of an independent dataset with >8,300 PDA single cells from five PDA patients, experimentally enriched for tumor cells (see methods), recapitulated these findings and independently confirmed the same three subtypes (Supplementary Notes and Extended Data Fig 4a-c.). These findings were further recapitulated in single epithelial cells from a PDX model (Extended Data Fig. 4d) and from human PDA cell lines classified as Lineage (HPAFII, PATU) and Morphogenic (KP4) (Extended Data Fig. 4 e-f).

Finally, we used previously published PDA subtypes to classify single cells and to assess expression of previously reported hallmark genes, (methods and Supplementary Notes).

Surprisingly, the individual cells were often classified as belonging to mutually exclusive subtypes and expressing mutually-exclusive markers—e.g., FAM3D (Classical) vs. LEMD1 (Basal-like)[6] (Extended Data Fig. 5a-d). This suggests that prior classification systems may reflect emergent bulk-tissue properties rather than the biology of individual cells.

PDA master regulators are enriched in essential genes: To assess PDA cell dependency on VIPER-inferred MRs, we performed long-term (33 day) viability assays by pooled CRISPR/Cas9 screens in cell lines selected as optimal representatives of Lineage and Morphogenic subtypes. Specifically, Cancer Cell Line Encyclopedia (CCLE) analysis revealed two groups of PDA cell lines presenting aberrant activity of either Lineage or Morphogenic MRs, leading to selection of three Lineage (PATU-8988S, HPAFII, and CAPAN1) and three Morphogenic cells (Panc1, KP4, and PK45H) as optimal representatives for the screen. Their identity was confirmed by RNASeq profiling and analysis of isolates from each cell line (Extended Data Fig. 6a).

Pooled CRISPR/Cas9-mediated knockout screens were performed in each cell line, using guide RNAs (sgRNAs) targeting 3,179 genes including all TR proteins studied in the VIPER analysis[12], as well as selected core essential and non-essential genes, as positive and negative controls [33,34] (Extended Data Fig. 6b). Cells were harvested at 24 hours and 33 days following sgRNA transduction and sequenced. sgRNA counts were integrated across technical and biological replicates to reveal depletion associated with decreased cell viability (see methods). We first assessed subtype-specific dependencies by measuring the enrichment of Lineage and Morphogenic MRs among genes essential in the respective cell lines. The top 50 subtype specific MRs were significantly enriched in Lineage or Morphogenic cell line essential genes, respectively (Fig. 4a) ($p_{Lin} = 5.0 \times 10^{-3}$ and $p_{Mor} = 4.0 \times 10^{-2}$, respectively). This includes, for instance,

proteins such as CDX2, GATA6, and HNF1A (Lineage) and MYBL1, ZEB1, RUNX2, and SNAI2 (Morphogenic).

To assess essentiality of subtype-independent MRs, we then used VIPER to identify MRs regulating a signature of differentially expressed genes between each PDA sample and the average of all physiologic samples in GTEx[35]. MRs were then ranked by integrating their *p*-values across all cohort samples using Stouffer's method. Enrichment of subtype-independent MRs in genes whose essentiality was consistent across all six PDA cell lines (i.e. subtype-independent essentiality) produced an even stronger enrichment ($p = 9.4 \times 10^{-4}$) (Figure 4b). Furthermore, Achilles project analysis[36] (https://depmap.org/portal/) showed that subtype-dependent and independent MRs do not contain common essential genes (i.e., essential across a majority of the 739 cell lines) (Extended Data Fig. 6c). Thus, the analysis confirmed that PDA cells present both subtype-specific dependencies as well as more universal dependencies. Both may yield novel pharmacological targets.

***Ectopic Lineage MR Expression Reprograms Morphogenic Cells*:** Based on these results, we then tested whether ectopic expression of Lineage MRs in Morphogenic cells would reprogram them towards a Lineage state. To test this hypothesis, we used lentiviral transduction to overexpress each of the top 8 Lineage MRs (ELF3, ESSRA, ETV4, FOXA2, FOXA3, HNF1A, NR2F6, and OVOL2) in the Morphogenic KP4 cell line, via the tetracycline inducible M2rtTA system[37]. Strikingly, while most Lineage MRs were effective in inducing activation of positive Lineage MRs, only OVOL2 overexpression also repressed negative Lineage MRs, resulted in a near-complete recapitulation of the full Lineage MR signature ($p = 4.3 \times 10^{-15}$); this included downregulation of universal essential genes, such as UHRF1, and key regulators of proliferation,

such as PTTG1, FOXM1, and TOP2A (Fig. 4c and Extended data Fig. 7a-b). We next assessed whether overexpression of MR combinations could more effectively induce reprogramming than individual MRs. For this, we performed ectopic cDNAs expression in single cells at MOI = 1, followed by single-cell RNASeq profiling. At this MOI, most cells received either one or two cDNAs and there were 11 MR pairs, 9 triplets, 4 quadruplets and 1 quintuplet that were co-ectopically expressed in n ≥ 30 cells, thus allowing faithful assessment of gene expression signature. As shown in Figure 4d, and in line with the results obtained by ectopic expression individual TF in bulk data (Figure 4c), the top combinations included OVOL2. However, co-expression of OVOL2 and HNF1A significantly improved reprogramming efficiency by >13% (from 48% to 61.2%). These data provide functional confirmation of the mechanistic role of specific MRs in determining PDA cell state through regulation of their transcriptional targets. Finally, we assessed OVOL2's cDNA-mediated activation by WB assays of known OVOL2 downstream repressed targets representing EMT effectors, a hallmark of the Morphogenic subtype. Indeed, OVOL2 ectopic expression inhibited protein expression of ZEB1 and Vimentin[38,39], while inducing expression of the established epithelial marker E-cadherin (Extended Data Fig.7c). While OVOL2 was previously reported as a Mesenchymal To Epithelial (MET) transition inducer, when silenced in specific tissues[40-44], its ability to reprogram cells from a Morphogenic to a Lineage state in PDA when activated was not previously reported.

Discussion

The molecular classification of PDA has remained a stubbornly complex challenge, exacerbated by the contributions of dozens of stromal cell types intermixed with a minor fraction of heterogenous malignant epithelial cells. We navigated this complexity through both experimental

isolation of the PDA epithelial compartment and the application of regulatory network algorithms that confer key advantages over traditional gene expression profiling by allowing accurate protein activity measurements from mRNA profiles. A key limitation of gene expression analysis includes high measurement noise on an individual gene basis, which, combined with the high dimensionality of these data and the assumption of independence between each gene, can produce highly variable gene signatures, even when identical phenotypes are considered. In contrast, use of regulatory protein activity measurements, each of which is based on the expression of tens of target genes, dramatically increases measurement reproducibility[45]. By incorporating PDA epithelium-enriched networks using metaVIPER, these methodologies also mitigate the effects of stromal cell infiltration. These advantages are especially relevant in single-cell analyses, where the approach allows accurate quantification of regulatory protein activity independent of the number of reads of their encoding genes[18,46], including in PDA[3,18]. Most critically, mechanism-based insight is directly built into these approaches because prediction of regulatory protein activity is directly related to the genetic signature they physically regulate, thus supporting identification of transcriptional regulators that drive phenotypic states through physical regulation of their targets (see for instance[13,15,16,47] and recent reviews[10,11]).

The application of regulatory network analysis to bulk, microdissected, and single cell samples brings new insights to the molecular subtypes of PDA. Three clearly distinct cellular states (*cellular subtypes*) emerged from single cell analysis of PDA samples, whose coexistence within individual tumors and with other non-epithelial populations can produce inconsistent bulk-level subtype inference. Two of these (Lineage and Morphogenic) exhibit highly variable contributions across patients whereas the third (Oncogenic Precursor) appears to contribute in roughly similar proportions to all PDA tumors. Consequently, rather than six different potential combinations of

the three cell types, PDA resolves into either two predominant groups at the bulk tissue level, depending on the relative abundance of Lineage versus Morphogenic cells. In the most highly epithelial cohorts, such as our LCM dataset, it is also possible to resolve an intermediate group of tumors with roughly equal contributions of Lineage and Morphogenic cells, whereas higher stromal content obscures this third group. Intrinsically, classification schemes built solely from bulk or even microdissected PDA samples will reflect an averaging of the expression contributions across different cellular subtypes. Indeed, the intermediate subtype could be effectively recapitulated by creating synthetic epithelial bulk samples where about half of the cells were OP cells and the remaining ones had a relatively similar composition of Lineage and Morphogenic cells. This is also well illustrated by the fact that mutually exclusive hallmark genes from prior, bulk-tissue classifiers were expressed in the same individual single cells (Extended Data Fig. 6). In contrast, the regulatory protein activity subtypes mapped uniquely to individual cells across two independent scRNAseq datasets.

Nevertheless, some features are shared with prior classification systems, particularly an association of some subtypes with tumour differentiation, as well as a role for certain gastrointestinal lineage transcription factors. However, it is important to note that while high GI transcription factor activity relative to other PDA cells led to their designation as Lineage cells, the activity of these MRs is in fact lower than in low-grade PanINs and IPMNs. As such, Lineage cells are perhaps best defined by their incomplete loss of GI identity. By contrast, Morphogenic cells have completely shed their GI epithelial identity and have acquired new mesenchymal features based on the activity of MRs such as SNAI1, ZEB1, and ZEB2.

Oncogenic Precursor cells are of particular interest for their commonality across tumors. Based on the results of RNA Velocity analysis, one can postulate their potential role in repopulating the Lineage and Morphogenic compartments. However, they also bear features that defy the traditional concept of a “cancer stem cell”. Indeed, entropy analysis suggests that OP cells are more differentiated than Lineage and Morphogenic cells and they express comparatively low levels of the stemness markers MSI2, PROM1, and CXCR4. Curiously, OP cells exhibit relative activation of RFX6, NEUROD1, and other early GI/neuroendocrine transcription factors, which are understudied in the context of PDAC. One interpretation of these varied data points is that OP cells may occupy of stable equilibrium point within the GI committed program that is both rapidly proliferating and plastic, much in the way that early pancreatic bud cells both proliferate and differentiate. Lack of a pure OP cell line currently prevents direct assessment of MR essentiality in this population, though we infer from their distinct MR profile that they will present distinct drug sensitivities. However, the presence of a small fraction of OP cells in Lineage and Morphogenic cell lines suggests that once appropriate techniques for their isolation are available, such studies should be possible.

The intermixing of multiple cellular subtypes with distinct regulatory states provides a straightforward mechanism for chemoresistance, a phenomenon that has been demonstrated in multiple other tumor types[48,49]. This concept is emphasized by the distinct genetic dependencies observed in Lineage versus Morphogenic cell lines in both loss of function and overexpression studies. Critically, the overexpression of a few or even just one subtype MR was sufficient to reprogram Morphogenic cells towards a Lineage state, an intriguing result given the reduced survival of patients with Morphogenic tumours. Notably, cellular subtype heterogeneity was apparent even in long-established cell lines, a finding that has widespread implications for the

interpretation of drug perturbation and other *in vitro* studies. The core PDA dependencies comprising the experimentally validated MR proteins identified in our analyses represent attractive targets for future therapeutic development using protein activity reversal based methods, such as OncoTarget[19] and OncoTreat[13].

References

1 Rahib, L. *et al.* Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Res* **74**, 2913-2921, doi:10.1158/0008-5472.CAN-14-0155 (2014).

2 Neftel, C. *et al.* An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell* **178**, 835-849 e821, doi:10.1016/j.cell.2019.06.024 (2019).

3 Elyada, E. *et al.* Cross-Species Single-Cell Analysis of Pancreatic Ductal Adenocarcinoma Reveals Antigen-Presenting Cancer-Associated Fibroblasts. *Cancer discovery* **9**, 1102-1123, doi:10.1158/2159-8290.CD-19-0094 (2019).

4 Birnbaum, D. J., Finetti, P., Birnbaum, D., Mamessier, E. & Bertucci, F. Validation and comparison of the molecular classifications of pancreatic carcinomas. *Mol Cancer* **16**, 168, doi:10.1186/s12943-017-0739-z (2017).

5 Collisson, E. A. *et al.* Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat Med* **17**, 500-503, doi:10.1038/nm.2344 (2011).

6 Moffitt, R. A. *et al.* Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat Genet* **47**, 1168-1178, doi:10.1038/ng.3398 (2015).

7 Bailey, P. *et al.* Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* **531**, 47-52, doi:10.1038/nature16965 (2016).

8 Cancer Genome Atlas Research Network. Electronic address, a. a. d. h. e. & Cancer Genome Atlas Research, N. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell* **32**, 185-203 e113, doi:10.1016/j.ccell.2017.07.007 (2017).

9 Puleo, F. *et al.* Stratification of Pancreatic Ductal Adenocarcinomas Based on Tumor and Microenvironment Features. *Gastroenterology* **155**, 1999-2013 e1993, doi:10.1053/j.gastro.2018.08.033 (2018).

10 Maurer, C. *et al.* Experimental microdissection enables functional harmonisation of pancreatic cancer subtypes. *Gut* **68**, 1034-1043, doi:10.1136/gutjnl-2018-317706 (2019).

11 Califano, A. & Alvarez, M. J. The recurrent architecture of tumour initiation, progression and drug sensitivity. *Nat Rev Cancer* **17**, 116-130, doi:10.1038/nrc.2016.124 (2017).

12 Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat Genet* **48**, 838-847, doi:10.1038/ng.3593 (2016).

13 Alvarez, M. J. *et al.* A precision oncology approach to the pharmacological targeting of mechanistic dependencies in neuroendocrine tumors. *Nat Genet* **50**, 979-989, doi:10.1038/s41588-018-0138-4 (2018).

14 Rajbhandari, P. *et al.* Cross-Cohort Analysis Identifies a TEAD4-MYCN Positive Feedback Loop as the Core Regulatory Element of High-Risk Neuroblastoma. *Cancer discovery* **8**, 582-599, doi:10.1158/2159-8290.Cd-16-0861 (2018).

15 Aytes, A. *et al.* Cross-species regulatory network analysis identifies a synergistic interaction between FOXM1 and CENPF that drives prostate cancer malignancy. *Cancer Cell* **25**, 638-651, doi:10.1016/j.ccr.2014.03.017 (2014).

16 Carro, M. S. *et al.* The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**, 318-325, doi:nature08712 [pii] 10.1038/nature08712 (2010).

17 Basso, K. *et al.* Reverse engineering of regulatory networks in human B cells. *Nat Genet* **37**, 382-390, doi:10.1038/ng1532 (2005).
18 Ding, H. *et al.* Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. *Nature communications* **9**, 1471, doi:10.1038/s41467-018-03843-3 (2018).
19 Zeleke, T. *et al.* Network-based assessment of HDAC6 activity is highly predictive of pre-clinical and clinical responses to the HDAC6 inhibitor ricolinostat. *medRxiv* (2020).
20 Reynolds, A., Richards, G., de la Iglesia, B. and Rayward-Smith, V. Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms* **5**, 475–504 (1992).
21 Cava, C. *et al.* Integration of multiple networks and pathways identifies cancer driver genes in pan-cancer analysis. *BMC Genomics* **19**, 25, doi:10.1186/s12864-017-4423-x (2018).
22 Paull, E. O. *et al.* A Modular Master Regulator Landscape Determines the Impact of Genetic Alterations on the Transcriptional Identity of Cancer Cells. *Cell, in press* (2020).
23 Cannon, W. B. Organization For Physiological Homeostasis. *Physiol Rev* **9**, 399-431 (1929).
24 Waddington, C. H. Canalization of development and genetic assimilation of acquired characters. *Nature* **183**, 1654-1655 (1959).
25 Evan O. Paull, A. A., Prem Subramaniam, Federico M. Giorgi, Eugene F. Douglass, Brennan Chu, Sunny J. Jones, Siyuan Zheng, Roel Verhaak, Cory AbateShen, Mariano J. Alvarez, and Andrea Califano. A Modular Master Regulator Landscape Determines the Impact of Genetic Alterations on the Transcriptional Identity of Cancer Cells. *BioRxiv*, doi:https://doi.org/10.1101/758268 (2019).
26 Stouffer, S. A., Suchman, E.A., DeVinney, L.C., Star, S.A.;,Williams, R.M. Jr. *The American soldier: Adjustment during Army life*. Vol. Vol. I (NJ: Princeton University Press, 1949).
27 Yuan, J. *et al.* Single-cell transcriptome analysis of lineage diversity in high-grade glioma. *Genome Med* **10**, 57, doi:10.1186/s13073-018-0567-9 (2018).
28 Guo, M., Bao, E. L., Wagner, M., Whitsett, J. A. & Xu, Y. SLICE: determining cell differentiation and lineage based on single cell entropy. *Nucleic Acids Res* **45**, e54, doi:10.1093/nar/gkw1278 (2017).
29 Fox, R. G. *et al.* Image-based detection and targeting of therapy resistance in pancreatic adenocarcinoma. *Nature* **534**, 407-411, doi:10.1038/nature17988 (2016).
30 Hermann, P. C. *et al.* Distinct populations of cancer stem cells determine tumor growth and metastatic activity in human pancreatic cancer. *Cell Stem Cell* **1**, 313-323, doi:10.1016/j.stem.2007.06.002 (2007).
31 Singh, S., Srivastava, S. K., Bhardwaj, A., Owen, L. B. & Singh, A. P. CXCL12-CXCR4 signalling axis confers gemcitabine resistance to pancreatic cancer cells: a novel target for therapy. *Br J Cancer* **103**, 1671-1679, doi:10.1038/sj.bjc.6605968 (2010).
32 La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494-498, doi:10.1038/s41586-018-0414-6 (2018).
33 Hart, T. *et al.* Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3 (Bethesda)* **7**, 2719-2727, doi:10.1534/g3.117.041277 (2017).
34 Palin, K. *et al.* Contribution of allelic imbalance to colorectal cancer. *Nat Commun* **9**, 3664, doi:10.1038/s41467-018-06132-1 (2018).

35 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585, doi:10.1038/ng.2653 (2013).
36 Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576 e516, doi:10.1016/j.cell.2017.06.010 (2017).
37 Hockemeyer, D. *et al.* A drug-inducible system for direct reprogramming of human somatic cells to pluripotency. *Cell Stem Cell* **3**, 346-353, doi:10.1016/j.stem.2008.08.014 (2008).
38 Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol* **36**, 89-94, doi:10.1038/nbt.4042 (2018).
39 Zeisberg, M. & Neilson, E. G. Biomarkers for epithelial-mesenchymal transitions. *J Clin Invest* **119**, 1429-1437, doi:10.1172/JCI36183 (2009).
40 Roca, H. *et al.* Transcription factors OVOL1 and OVOL2 induce the mesenchymal to epithelial transition in human cancer. *PLoS One* **8**, e76773, doi:10.1371/journal.pone.0076773 (2013).
41 Qi, X. K. *et al.* OVOL2 links stemness and metastasis via fine-tuning epithelial-mesenchymal transition in nasopharyngeal carcinoma. *Theranostics* **8**, 2202-2216, doi:10.7150/thno.24003 (2018).
42 Ye, G. D. *et al.* OVOL2, an Inhibitor of WNT Signaling, Reduces Invasive Activities of Human and Mouse Cancer Cells and Is Down-regulated in Human Colorectal Tumors. *Gastroenterology* **150**, 659-671 e616, doi:10.1053/j.gastro.2015.11.041 (2016).
43 Murata, M. *et al.* OVOL2-Mediated ZEB1 Downregulation May Prevent Promotion of Actinic Keratosis to Cutaneous Squamous Cell Carcinoma. *J Clin Med* **9**, doi:10.3390/jcm9030618 (2020).
44 Liu, J. *et al.* Ovol2 induces mesenchymal-epithelial transition via targeting ZEB1 in osteosarcoma. *Onco Targets Ther* **11**, 2963-2973, doi:10.2147/OTT.S157119 (2018).
45 Rajbhandari, P. *et al.* Cross-Cohort Analysis Identifies a TEAD4-MYCN Positive Feedback Loop as the Core Regulatory Element of High-Risk Neuroblastoma. *Cancer Discov* **8**, 582-599, doi:10.1158/2159-8290.CD-16-0861 (2018).
46 al., D. e. Single-cell based elucidation of molecularly-distinct glioblastoma states and drug sensitivity. . *bioRxiv* (2019).
47 Piovan, E. *et al.* Direct reversal of glucocorticoid resistance by AKT inhibition in acute lymphoblastic leukemia. *Cancer Cell* **24**, 766-776, doi:10.1016/j.ccr.2013.10.022 (2013).
48 Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401, doi:10.1126/science.1254257 (2014).
49 Yeo, S. K. *et al.* Single-cell RNA-sequencing reveals distinct patterns of cell state heterogeneity in mouse models of breast cancer. *Elife* **9**, doi:10.7554/eLife.58810 (2020).

Figure legends

Figure 1: (a) Schematic workflow of the analysis. **(b)** Heatmap showing the activity of 1795 TRs (rows) across 200 PDA samples (columns) in the three clusters. **(c)** Kaplan Meier showing the difference in survival between the three clusters. The p-value was computed using the log-rank test.

Figure 2: (a-b) GSEA plots showing enrichment of Tumor Checkpoint MRs in TCGA and ICGC (i.e., top 25 most activated and inactivated) in the CUMC protein activity signature (reference). P-values and normalized enriched scores (NES) were computed by aREA algorithm. **(c)** GSEA plot showing enrichment of TCGA Tumor Checkpoint MRs in ICGC differentially active proteins (reference). **(d)** ROC-curve showing the performance of a random forest classifier, based on Tumor Checkpoint MRs, trained on each ICCG clusters and tested on the corresponding TCGA clusters (red curve) or vice-versa (black curve). **(e)** Heatmaps showing most differentially active Tumor Checkpoint MRs for the Lineage and Morphogenic subtypes, following integration of the ICGC and TCGA analyses. **(f)** ROC-curves for a random forest classifier based on the most differentially active Lineage and Morphogenic Tumor Checkpoint MRs, following integration of ICGC and TCGA analyses, and tested on: (*i*) CUMC lineage and morphogenic clusters (green curve); (*ii*) UNC viper- based clusters (brown curve); and (*iii*) Collison viper-based clusters (gold curve). For UNC and Collisson cohorts (microarray data sets) the performance was evaluated on 109/125 samples and on 23/27 samples that clustered with a significant Silhouette Score >0.25. **(g)** Heatmap showing differentially methylated sites between Linage and Morphogenic TCGA

samples. **(h)** Scatterplot showing correlation between differential methylation and differential activation of HNF1A. **(i)** PieDonuts plots showing the overlap between the VIPER-based clusters corresponding to Lineage and Morphogenic subtypes and previously published classification schemes.

Figure3: (a) Diffusion map of the PDA epithelial cells (n=1900), showing the three different clusters identified by fuzzy clustering. Cells with a membership probability lower than 0.5, computed by Fuzzy clustering, across the three clusters are shown in grey. The black line indicates the psudo-trajectory computed by principal curve analysis. **(b)** GSEA plots showing the enrichments of the top 50 activate MRs inferred by VIPER SC1, SC2 and SC3 on the bulk linage-morphogenic protein activity signature. The p-values and the NES were estimated by GSEA with 1000 permutations. **(c)** Heatmap showing the differential activity of the top 200 MRs of each single-cell cluster sorted by their median activity. Bottom annotation shows the expression of CD133, MSI2 and CXCR4 computed by metacell integration. **(d)** Scatterplot showing the single cell entropy computed on protein activity using the SLICE algorithm (y-axis) in single cells ordered according to the pseudo trajectory computed by principal curve. **(e)** PCA of the PDA epithelial cells showing the transitions predicted by RNA-velocity analysis. Large arrows were manually added to highlight the major transitions compute by velocity analysis (small arrows) **(f)** Pie charts showing the distribution of PDA epithelial subtypes in each patient.

Figure 4: (a) Scatter plot showing the differential essentiality signature between lineage and morphogenic cell lines (L/M_{Ess}). Genes are ranked according to the differential essentiality score (z-score) from the top lineage essential genes (left) to the top morphogenic essential genes (right). Representative lineage and morphogenic MRs are shown in red and blue, respectively. The p-

values and the NES were estimated by GSEA (1-tail) with 1000 permutations. **(b)** Scatter plot showing all the essential genes ranked according to their essentiality score (z-score) computed across all the PDA cell lines. The p-value and the NES were estimated by GSEA (1-tail) with 1000 permutations. **(c)** Heatmap showing the activity of the lineage and morphogenic TRs in the morphogenic cell line (KP4) after the ectopic expression of individual top Lineage TFs (n=8) The barplots represent the $-\log_{10}$ p-value relative to the Morphogenic-To-Lineage reprogramming score (NES) computed by the aREA algorithm. **(d)** *Left*, box plots showing the distributions of the reprogramming scores generated for each single-cell upon the ectopic expression of individual or combinations of the top 8 lineage TF. The p-values were computed using the aREA algorithm. *Right,* Heatmap showing the median activity of the lineage and morphogenic TRs in the morphogenic single cells (KP4) after the ectopic expression (individual or combinations) of the top 8 lineage TFs.

Extended Data Figure 1: (a) Scatter plot showing the optimal number of clusters based on conservation of master regulators evaluated by AUC analysis. **(b)** Heatmap showing the top biological hallmarks enriched in each cluster. Columns represent samples and rows represent hallmarks. The purple color represents the Normalized Enrichment Score estimated by GSEA with 200 permutations. **(c)** Confusion matrix showing the overlap of the poor differentiated and well differentiated tumors with Lineage and Morphogenic subtypes. The p-values were computed using the X^2 test.

Extended Data Figure 2: (a) Plot showing the enrichment for PDA essential genes in the PDA TRs computed using either LCM EPDA-NET only or metaVIPER. The x-axis represents the TFs ranked according to their differential activity between PDA epithelial samples and the pool of all physiologic tissues in GTEx. The y-axis represents the $-\log_{10}$ p-values relative to the enrichment of the TRs in PDA essential genes. The p-values were estimated by GSEA with 1000 permutations. **(b)** Enrichment plot showing the conservation of the PDA epithelial regulons in the 3 PDA bulk-derived networks used in this study and in 8 non PDA tumor networks. All the networks were inferred by ARACNe algorithm. The top bar represents the activity of the LCM EPDA-NET regulons computed by VIPER on the differential gene expression signature of the Lineage subtype. The regulons are ranked from the most inactivated (blue) to the most activated (red). The other bars represent the rank of the top 100 most differentially activated regulons (top 50 activated and top 50 inactivated) of each bulk network computed on the same differential expression signature (i.e. the signature of Lineage subtype). Red vertical lines indicate regulon activation, blue vertical lines indicate regulon inactivation. The conservation of the epithelial regulons for each network was assessed by performing a two-tails enrichment analysis using the LCM EPDA-NET regulons as reference. The p-values were computed by aREA algorithm. **(c)** Plots showing the optimal number of clusters based on conservation of master regulators evaluated by AUC analysis (methods). For Collisson et al., 2011, given the limited sample size (n=27) only two clusters solutions (k=2 and k=3) were tested as optimal. **(d)** Oncoprint plot showing the genetic alterations in TCGA Lineage and Morphogenic samples. **(e)** Heatmaps showing activated Tumor Checkpoint MRs for the Lineage and Morphogenic subtypes obtained by the integration of ICGC and TCGA, in the UNC and in Collisson et al cohorts.

Extended Data Figure 3: (a) Scatter plot showing the optimal number of clusters based on conservation of master regulators evaluated by AUC analysis in PDA epithelial single cells. **(b)** PCA based on chromosomal expression (ploidy analysis) of PDA epithelial single cells. **(c)** Differentially expressed genes between the two groups of epithelial cells predicted by PCA on chromosomal expression. **(d)** Dot plot showing the classification of pseudo bulk samples based on neural network classifier trained on the CUMC epithelial clusters.

Extended Data Figure 4: (a) Diffusion maps showing the PDA epithelial single cells (n=8304) from an independent cohort of 5 PDA patients (second cohort). Cells are colored according to the 3 clusters identified by Fuzzy clustering analysis. **(b)** Roc curve showing the performance of a random forest classifier trained on the 3 single cell subtypes (OP, scLineage and scMorphogenic) of the first cohort and tested on the 3 clusters of second cohort. The AUC was computed on >80% of the cells (n=6653/8304) with a membership probability >0.51 computed by fuzzy clustering. **(c)** Diffusion maps showing the activity of TRs associated to OP, Lineage and Morphogenic subtypes. Red indicates activation, blue indicates inactivation. **(d)** Diffusion map showing the clusters of PDA epithelial cells identified in the PDX model and the activity f TRs associated to OP, Lineage and Morphogenic subtypes. **(e)** Diffusion maps of single cells derived from morphogenic (KP4) and lineage (PATU and HPAFII) PDA cancer cell lines. Top left, annotation showing the cell line from which each single cell was derived. Top right, cluster annotation. Bottom, diffusion maps showing the activity of lineage and morphogenic TRs . **(f)** Diffusion maps showing the activity of TRs associated to OP in single cell derived from PDA cell lines.

Extended Data Figure 5: (a-c) Diffusion maps showing the enrichment of previously published the gene expression signatures associated to 3 main PDA classification schemes. The enrichment was evaluated by GSEA on the single-cell differential gene expression signatures computed using the robust z-score method on metacells gene expression profiles. **(d)** Diffusion maps showing the co-expression in the lineage cells (without meta-cell integration) of a basal-like and a classical gene in both single cell cohorts.

Extended Data Figure 6: (a) Heatmap showing the enrichment for top 50 (top 25 activate and top 25 inactivate) Lineage and Morphogenic in PDA cell lines. Cell lines labeled in red represent the profiles of Lineage cell lines selected for the CRISPRcas9 pool screening re-sequenced at CUMC. Cell lines labeled in blue represent the profiles of morphogenic cell lines selected for the CRISPRcas9 pool screening re-sequenced at CUMC. **(b)** ECDF plot showing distributions of z-scores the core-essential genes (positive controls) , non-core essential genes (negative controls) and the all the other TRs profiled by CRISPRcas9 pool screening. **(c)** Violin plots showing the distribution of essentiality scores of the top essential master regulators associated to Lineage subtype, Morphogenic subtype and universal PDA reported in Achilles data base (https://depmap.org/portal/achilles/) computed by CrisprCas9 across 739 cell lines. The red dashed line corresponds to the threshold below of which a gene is considered a common essential.

Extended Data Figure 7: (a) GSEA plot showing the activation of the top 50 lineage TRs and inactivation of the top 50 morphogenic TRs in KP4 cells by OVOL2. The NES and p.value were

estimated by GSEA with 1000 permutations. **(b)** GSEA plot showing the upregulation of the top 200 lineage gene and down-regulation of the top 200 morphogenic genes in KP4 cells by OVOL2. The NES and p.value were estimated by GSEA with 1000 permutations. **(c)** Western blot showing the inhibition of mesenchymal markers (ZEB1 and Vimentin) and the induced expression of the E-cadherin (Epithelial marker) in KP4 cells overexpressing either mCherry or OVOL2 (+/- transcriptional activator M2RTTA).

Author's Contributions

A.C., K.P.O. and P.L. conceived the project. P.L. designed, performed and oversaw the computational analyses. M.T. designed, performed and oversaw the experimental analyses. C.H.M. was responsible for the LCM. A.G.C., E.E., B.S., J.W., J.K., X.T, S.G. and F.N. contributed to experimental execution, data acquisition and data generation. L.T., M.J.A., S.T., A.W. and A.G. contributed to the computational analyses. A.I. reviewed histopathology for PDA samples in the CUMC cohort. G.A.M. and W.W. aided in the curation of human outcomes data. A.C., K.P.O. and P.L. wrote the manuscript with feedback from C.H.M, M.T and M.J.A.. K.P.O., D.A.T. and A.C. supervised the study.

Acknowledgment

This work was supported by a Clinical Translational Program Grant from the Lustgarten Foundation for Pancreatic Cancer Research (KPO) and a Precision Medicine Pilot Award, Irving Institute for Clinical and Translational Research (KPO), NCI Research Centers for Cancer Systems Biology Consortium (1U54CA209997 to AC and KPO), and NCI Cancer Center Support Grant (2P30 CA013696-45) for the Herbert Irving Comprehensive Cancer Center (including the High Throughput Sequencing, Single Cell Analysis, OPTIC, Molecular Pathology, and Database Shared Resources). Funding was also provided by The Pancreas Center at Columbia/NY Presbyterian Hospital. This study was also supported by Sigrid Juselius Foundation (MT).
We acknowledge support from the Swedish National Genomics Infrastructure, SNIC (project SNIC 2017-7-265) and the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

Disclosure of Potential Conflicts of Interest

P.L. is Director of Single-Cell Systems Biology at DarwinHealth, Inc., a company that has licensed some of the algorithms used in this manuscript from Columbia University.
M.J.A. is CSO and equity holder of DarwinHealth, Inc.
Columbia University is also an equity holder in DarwinHealth Inc.
A.C. is founder, equity holder, consultant, and director of DarwinHealth Inc.

Methods

LCM-RNAseq

Freshly frozen tissue samples were obtained from patients who underwent surgical resection at the Pancreas Center at Columbia University Medical Center as described previously[1].Prior to surgery, all patients had given surgical informed consent, which was approved by the institutional review board. Immediately after surgical removal, the specimens were cryopreserved, sectioned and microscopically evaluated by the Columbia University Tumor Bank (IRB AAAB2667). Suitable samples were transferred into OCT medium (Tissue Tek) and snap frozen in a 2-methylbutane dry ice slurry. The tissue blocks were stored at -80°C until further processing. H&E stained sections of frozen PDA samples from the Tumor Bank were initially screened to confirm diagnosis and overall sample RNA quality was assessed by the Pancreas Center supported Next Generation Tumor Banking program using gel electrophoresis, with samples exhibiting high RNA quality utilized for subsequent analyses. Laser Capture Microdissection (LCM), RNA sequencing and gene expression quantification. LCM-RNA-Seq was performed as described previously[1,2]. Briefly, Cryosections of OCT-embedded tissue blocks were transferred to PEN membrane glass slides and stained with cresyl violet acetate. Adjacent sections were H&E stained for pathology review. Laser capture microdissection was performed on a PALM MicroBeam microscope (Zeiss), collecting at least 1000 cells per compartment. RNA was extracted and libraries prepared using the Ovation RNA-Seq System V2 kit (NuGEN). Libraries were sequenced to a depth of 30 million, 100bp, single-end reads on an Illumina HiSeq 2000 platform.

VIPER analysis of LCM and Bulk RNA-seq cohorts

Network-based protein activity inference was performed by applying either VIPER[3] or metaVIPER [4] algorithms using context-specific pancreatic cancer (PDA) transcriptional networks.

PDA transcriptional networks were generated from CUMC, ICGC and UNC cohorts by the ARACNe algorithm[5] with 100 bootstrap iterations and MI (mutual information) *P*-value threshold of 10^{-8}. TCGA networks, including PDA TCGA network, were downloaded from the aracne.networks package [6]. ARACNe networks included TF, co-TFs and signaling molecules. A subset of transcriptional regulators enriched for TFs, co-TFs and chromatin regulators were considered in this study (Supplementary Table1).

RNAseq gene counts from CUMC, ICGC (Bailey et al., [7]) and TCGA (https://www.cancer.gov/tcga) cohorts were normalized using the variance stabilization transformation (VST) procedure as implemented in DESeq2 package [8].

From TCGA we selected only samples annotated as "high purity".

Microarray data from Collisson et al.,[9] and Moffitt et al.,[10] were downloaded as normalized gene expression profiles. Single sample differential gene expression signatures were generated independently for each cohort from the normalized gene expression profiles using the "*scale"* method *(z-score)* implemented in the viper package[3].

CUMC-LCM epithelial differential gene expression signatures were transformed into protein activity profiles by the VIPER algorithm[3] using the LCM-PDAnet epithelial network. Differential gene expression signatures from the other PDA cohorts and PDA cell lines, were transformed into protein activity profiles using the metaVIPER algorithm[4] by integrating the four PDA networks (LCM-PDAnet, ICGC-net, TCGA-net, UNC-net).

Cluster analysis was performed independently in each cohort by applying the Partitioning Around Medoids algorithm (PAM)[11] on the viper-distance matrix, computed using reciprocal enrichment analysis[12] of the top 25 most activated and top 25 most inactivated proteins in each signature, as implemented by the *viperSimilarity* function in the viper package[3].

The optimal number of clusters (K) was identified by assessing the conservation of cluster-specific master regulators using the Area Under the Curve (AUC) metric[13].

Briefly, for each value of K we assessed the conservation of the top 50 cluster-specific regulators (top 25 over-activated and top 25 under-activated) in each sample using a leaving one out cross validation procedure (LOOCV). The cluster-specific regulator signatures were computed by integrating the activity score (NES) of the samples in the same cluster using the Stouffer's method, excluding the test sample used for the LOOCV. The conservation of the cluster-specific master regulators was assessed by performing a two tails GSEA. Each sample was then assigned to a given cluster based on the highest NES computed by GSEA. This generated a membership vector predicted by GSEA. Finally, we compared memberships predicted by GSEA with the memberships computed by PAM algorithm and computed the ROC-AUC using the multiclass AUC approach[13] as implemented in the pROC package[14] (see Supplementary Notes for more details).
After cluster analysis, subtype-specific protein activity signatures were computed by VIPER comparing the samples of each cluster against all the other samples (one *vs.* the rest). Specifically, we first computed a differential gene expression signature for each cluster (one *vs.* the rest) using a bootstrap Student's t-test with 100 bootstraps iterations; then, we applied either VIPER or MetaVIPER on the subtype-specific differential expression signatures. For the CUMC cohort we applied VIPER using the LCM network only; for all the other cohorts we applied metaviper[4], with the integration of the four PDA networks. For the CUMC cohort a Lineage vs. Morphogenic signature was computed through a direct comparison between the Lineage and Morphogenic samples, excluding the samples in the intermediate cluster.

Survival analysis
Survival analysis was performed by comparing patients between protein activity-based clusters using the Kaplan-Meier method as implemented in the "survival" software package for R[15] . The p-values were computed using the log-rank test. The Kaplan-Meier curve were generated using the "survminer" software package[16]

Cross-cohorts analysis of bulk samples
Cross-cohorts conservation analysis of master regulators was performed by assessing the enrichment for the top Lineage (n=25) and top Morphogenic (n=25) master regulator proteins in the cluster-specific protein activity signatures of the other cohorts. The normalized enrichment scores and the p-values were computed using the aREA algorithm[3].
The random forest classification was performed using the top 50 candidate master regulators (top 25 Lineage and top 25 morphogenic) with RandomForest package[17].
The integration between TCGA and ICGC protein activity signatures was performed by integrating activity scores (NES) of each regulon using the Stouffer's method. Previous to applying the RF classifier and performing the integration of the ICGC and TCGA results, we assessed the distribution of the activity scores of each regulator between the two cohorts by performing a Kolmogorov-Smirnov's test. This analysis revealed 99.95% (1834/1835) of the regulators showed no significant differences in their distribution between the cohorts. We report that one regulator, "SCX", showed significantly different distributions (Bonferroni adjusted p-value<0.05) between the two datasets. SCX was not identified among differentially activated proteins between Lineage and Morphogenic subtypes and was not used for the random forest classifier.
The same procedure based on the Kolmogorov-Smirnov's test was applied for all the cross-cohorts comparisons and all the regulators showing significantly different distributions (Bonferroni adjusted p-value<0.05) between the train set and test set were excluded for the random forest classification procedures. For the microarray-based cohorts, the AUC scores were computed on the samples that clustered with a silhouette score > 0.25. Specifically, the AUC for Collisson et al.,[9] was computed on 23/27 (85%) of samples and the AUC for UNC[10] was computed 102/125 (82%) of the samples.

DNA methylation analysis

450K DNA methylation profiles were downloaded from TCGA using TCGAbiolinks package [18]. The beta values were converted to M-values using the "beta2m" function implemented in the Minfi package[19]. Differential methylation analysis between Lineage and Morphogenic samples was performed on M-values using the limma package[20]. All probes with a FDR<0.05 were considered differentially methylated. Cluster analysis was performed using the PAM algorithm[11].

Single-cell analysis of Elyada et al.[21]

Single cell UMI-count matrix from Elyada et al.[21] was filtered to remove low quality cells with less than 1000 UMI-counts and genes with zero counts across all the cells. UMI counts were then normalized to counts per million (CPM). Epithelial cells were computationally selected for each patient using a GSEA clustering procedure (see Supplementary Notes) based on the enrichment of cell type specific markers used in Elyada et al.[21], including markers for epithelial, endothelial, immune, fibroblasts and pericytes cells (see Supplementary Methods). A total of 1900 cells were predicted as epithelial across all the six patients. A subset of 500 cells randomly selected from the predicted PDA epithelial cells (n=1,900) were used to build a single-cell ARACNe network (scNET). ARACNe was applied to CPM normalized counts with 100 bootstrap iterations and MI *P*-value threshold of 10^{-8}. ARACNe inferred 506/1,835 regulons from single-cell data. The activity of the remaining transcriptional regulators (n = 1,329) was then inferred by metaVIPER through the integration four PDA networks (CUMC-net, TCGA-net, ICGC-net and UNC-net). All the ARACNe regulons were pruned to the top 50 targets before metaVIPER analysis and protein activity inference.

metaVIPER was applied on single-cell differential gene expression signatures computed using the "mad" method, equivalent to a robust z-score[22], implemented by the VIPER package on rank normalized single-cell gene expression profiles. Cluster analysis was performed using the fuzzy clustering algorithm implemented in the cluster package[11] The optimal number of clusters was estimated based on the conservation of cluster-specific master regulators as previously described for bulk analysis. Unlike bulk analysis, for single-cell data sets the conservation of the cluster-specific master regulators was assessed by Monte Carlo Cross Validation instead of leave-one-out cross validation. The protein activity signature of each cluster was computed by performing a bootstrap Student's t-test between the single cell clusters (one *vs* the rest) on 1,870 putative malignant cells identified by aneuploidy analysis (see chromosome expression analysis).

Single-cell analysis of PDX model

PDA tumors from PDX mice were dissociated using the protocol described in Peng et al.,[23]. Briefly, a digestion buffer that contained trypsin, DNAse and enzymatic cocktail and the gentleMACS Octo Dissociator (Milteny Biotec, Cat. No. 130-095-937) were made for initial tumor disruption using manufacture's protocol. Cell suspensions were then filtered using a 40 µm cell strainer (Falcon, Cat. No. 352340) and red blood cells (RBC) were removed by RBC lysis buffer (Invitrogen, Cat. No. 1966634). Dissociated cells were washed twice with PBS with step by step descending centrifuging speed and increasing time. Finally, cells were stained with 0.4% Trypan blue (Invitrogen, Cat. No. T10282) to check the viability, and diluted with PBS to about 1×10^6 cells/ml for single cell sequencing.

Cell Ranger pipeline (v3.3, 10X Genomics) was used to process single-cell sequencing data, align the FASTQ files on GRCh38-3.0.0 transcriptome reference and produce the UMI count matrix. The count matrix was then filtered for low quality cells by removing cells with more than 10% of UMIs in the mitochondrial genes and cells with less 1×10^3 counts or more than 1×10^5 counts. We also removed genes with zero counts across all the cells. Putative epithelial cells were

selected based on the enrichment for epithelial genes using the GSEA clustering procedure as previously described of Elyada et al[21] (see Supplementary Notes for details). A total of 900 cells were predicted as epithelial. Gene expression profiles of predicted epithelial cells (n = 900) were transformed into differential gene expression signatures using "mad" method and the median expression of the 1,900 predicted epithelial cells from Elyada et al.[21] as reference. Differential gene expression signatures were transformed into protein activity profiles with metaVIPER using the scNET and four bulk networks.

Single-cell analysis of a second cohort of PDA epithelial cells

Single cells samples were processed as described in Chan-Seng-Yue et al.,[24].
Briefly, freshly resected tumors or fresh core biopsies were minced finely then dissociated at 4°C overnight. Tumour cells were enriched through depletion of CD45+/CD90+/GlyA+ populations using MS columns (Miltenyi Biotec). Enriched tumour cells were loaded and separated into droplet emulsion using the Chromium Single Cell 3′ v.2 kit (10x Genomics) and subsequent libraries were sequenced on the Illumina HiSeq 2500 platform. Single-cell sequencing data were processed using CellRanger v.2.1.1 (10x Genomics), aligned to hg19.
Single-cell UMI count matrices from five different patients were filtered for low quality cells, of these 8,900 cells passed quality controls. 596/8,900 (~6%) cells showing no expression for EPCAM were also removed. A total of 8,304 cells were used for the downstream analysis. The aneuploidy analysis based on chromosome expression performed on the 8,304 epithelial cells did not show separate clusters of cells as putative non-transformed cells (see chromosome expression analysis). All the 8,304 cells were then considered "*bona fide*" tumor cells.
Single cell gene expression profiles were transformed into differential gene expression profiles using the "mad" method implemented in the VIPER package. The single-cell epithelial network integrated by metaVIPER with the four PDA bulk networks was applied to transform differential gene expression signatures to protein activity profiles. Cluster analysis and cluster-specific protein activity signatures were computed as previously described in the single-cell analysis of Elyada et al., 2019. The optimal number of clusters was estimated on a representative subset of 1,000 cells randomly selected from the 8,304 PDA epithelial cells.
Cross cohort analysis with Elyada et al., 2019 was performed by applying random forest classifier trained on the single-cell subtypes identified in Elyada et al cohort. The random forest classifier was trained on the 218 proteins showing no significant differences in the distribution between the two single-cell cohorts by Kolmogorov-Smirnov test. The AUC was computed on 6,653/8,304 (>80%) cells with a membership probability >0.51 computed by fuzzy cluster analysis.

Chromosome expression analysis

Chromosome expression analysis was performed to identify putatively non transformed cells similar to Yuan et al.[25] Single cell gene expression matrix was rank normalized and scaled using the robust z-score procedure[22].
The scaled matrix was used to compute the average chromosomal expression in each cell. A principal component analysis (PCA) was applied the chromosomal expression matrix to identify clusters of putatively transformed and untransformed cells.
The differential expression analysis between putatively transformed and untransformed cells was performed by applying the Scanpy toolkit[26].

Single-cell entropy analysis and RNA velocity analysis

Single cell entropy analysis was performed on protein activity profiles using the SLICE algorithm [27]. Velocity analysis was performed on 1840/1900 cells using velocyto pipeline[28]. At level of bam

files showed we noticed that 60/1900 cells were assigned to the same barcode, these cells were removed from this analysis. Bam files were converted to loom files using a genome annotation and repeat masker file. The input bam file was sorted by mapping position, representing either a single sample or a single cell, containing error corrected cell barcodes as a TAG named CB or XC and error corrected molecular barcodes as a TAG named UB or XM. Moreover, the masker file was needed to mask expressed repetitive elements, as those counts constitute a confounding factor. The masker file required for this conversion was downloaded from UCSC genome browser website[29]. The output of this analysis generated a 4-layered loom file, which is an HDF5 file containing specific groups representing the main matrix as well as rows and columns attributes. The conversion from .bam to .loom was performed on each patient independently. Finally, all the loom files were merged in a single loom file for the downstream pseudo-time analysis that was performed with scVelo software package[30].

Metacell inference

Single cells gene expression profiles were transformed into Metacells gene expression profiles by integrating the UMI counts of the 50 nearest cells identified by applying a KNN algorithm[31] on the normalized gene expression profiles. Metacell counts were then normalized to CPM and used for downstream analysis. The transformation of single cell gene expression profiles to Metacell profiles is a smoothing procedure aimed at mitigating the severe dropout effects that affects single cell gene expression. Similar procedures have been extensively and successfully used in single cell studies [32-35]. In this work we use metacell profiles only to estimate the expression of CD133, MSI2 and CXCR4 and to evaluate the gene expression signatures related to bulk classification schemes. Metacell integration was not used to show the co-expression of single basal-like and classic marker genes in the same Lineage cells.

Pathology and Immunohistochemistry

Tissue samples were fixed in 10% formalin, paraffin-embedded and cut in 5 µm sections on a Leica RM 2235, which were mounted in superfrost plus microscope slides and dried. Tissues were deparaffinized in xylene and re-hydrated through a series of graded ethanol until water. For histopathological analysis, sections were stained with hematoxylin and eosin using the standard protocols. For immunohistochemistry staining, paraffin sections were first subjected to standard rehydration and antigen retrieval was carried out for five minutes in boiling 10 mM sodium citrate buffer pH 6, .05% Tween-20 using a pressure cooker. They were then blocked by endogenous peroxidase and incubated in blocking solution (2% animal-free blocker – Vector Laboratories, #SP5030 -, and 1.5% horse serum – Vector Laboratories, #S-2000-, in PBS .1% Tween). Using the following antibodies, primary incubation was carried out overnight at 4C: CDX2 (Cell Signaling Technology, #12306S, 1:1000), YBX2 (Abcam, #ab33164, 1:100). After that, all slides were incubated with appropriate secondary antibodies for 60 minutes at room temperature and then developed using the DAB reagent (Vector Labs, VV-93951085). Finally, slides were dehydrated, cleared and mounted with a permanent mounting medium.

Cell culture:

All the cells used in this study were obtained from the American Type Culture Collection (ATCC) and cultured at 37 °C in a humidified incubator (5% CO2) with the following medias:
HEK293T: DMEM + 10% FBS,
CAPAN1: Iscove's Modified Dulbecco's Medium + 20% FBS,
HPAFII: EMEM + 10% FBS,
KP4: Iscove's Modified Dulbecco's Medium + 20% FBS,
PANC1: DMEM + 10% FBS,

PATU8988S: DMEM + 5% FBS + 5% horse serum,
PK45H: RPMI-1640 + 10%FBS.
1% L-Glutamine and 1% Penicillin/Streptomycin was used with all the cell lines.
All the cell lines used were tested for mycoplasma status before and during the screening

CRISPR/cas9 screening

sgRNA containing lentiviruses were transduced into Cas9 expressing PDA cell lines in duplicates (in presence of 8ug/ml polybrene) with estimated MOI 0.2 - 0.3. Next day the lentivirus containing media was removed, cells washed with PBS and puromycin containing media (2ug/ml) was added to the cells for 48-72h until the control cells (no virus) were dead. The first timepoint of the screen was done immediately after this. At all times the cells were maintained at >1500 cells per guide through the 33-day screens (see Supplementary methods for more details).

Computational analysis of CRISPR/cas9 data

FASTQ files were counted and analyzed using MAGeCK version 0.5.6[36] , using RRA and total read count normalization, with default settings.
Each replicate was analyzed independently by comparing the 33 days (late time point) against day 0 (first time point).
A CRISPR/cas9 essentiality signature for each replicate was computed by transforming the p-value of each gene to z-score and multiplied for the sign of the fold change. Lineage and a morphogenic essentiality signatures were computed by integrating the z-scores using the Stouffer's method of lineage and morphogenic cell lines, respectively.
A differential essentiality signature between lineage and morphogenic cell lines was computed by subtracting the morphogenic essentiality signature from the lineage signature.
To define the PDA subtype-independent essentiality signature we integrated the essentiality signatures across all the cell lines, including lineage and morphogenic, using the Stouffer 's z-score method [37].

RNA extractions and RNA sequencing of PDA cell lines

RNA extractions and RNA sequencing
At the time of RNA extractions PDA cells were cultured in 6 well plates so that the cell confluency was < 50% and cells were seeded to the well at least 48-72h ago. Total RNA was extracted by using RNeasy Plus mini kit (Qiagen) and samples were sequenced by using NovaSeq 6000 (PE 20million reads).
Reads were processed with Kallisto pipeline using GRCh38 as reference. RNASeq counts were VST normalized. Single sample differential gene expression signatures were generated from the normalized gene expression profiles with the *scale* method *(z-score)* using the centroid of all CCLE as reference. Differential gene expression signatures were transformed to protein activity profiles using metaVIPER with the four bulk PDA networks. The selection of the lineage and morphogenic cell lines was performed by assessing the enrichment of the top 25 lineage and top 25 morphogenic differentially activated proteins in the protein activity profile of each cell line. The enrichment analysis was performed using the aREA algorithm.

Transcription factor overexpression assay (Bulk RNAseq)

The full-length open reading frame clones of the top 8 PDA lineage MR (full length ORFs) were ordered from CloneID (Harvard Medical School), and cloned into modified Tet-O-FUW lentiviral expression plasmid (addgene #30130), which included the puromycin resistance gene. mCherry and EGFP ORFs were also used for negative controls for the assay. All the clones were sequence verified. For each ORF we introduced a unique 20 bp barcode sequence located 200 bp upstream of the lentiviral 3'-long terminal repeat (LTR) region, similarly as in Parekh, U., et al [38]. This results a polyadenylated transcript including the barcode close to the 3' end.
All the viruses were produced and viral titers were measured individually for each virus. For the assay, each ORF containing lentivirus was transduced to KP4 morphogenic cell line with MOI 2 in triplicates (6 well format in the presence of polybrene). In another set of triplicates lentiviral ORFs were co-transduced with M2rtTA (FUW-M2rtTA, addgene #20342), which acts as transcriptional amplifier in tetracyclin inducible systems, enabling us to monitor each MR overexpression in higher and lower levels. The following day after the viral transductions, the media was changed and puromycin (2.5ug/ml) and doxocycline (0.6ug/ml) was added to the cells. Cells were incubated in total 5 days before total RNA was collected by Direct-zol RNA MiniPrep Plus kit (Zymo Research).
A total of 69 RNAseq profiles corresponding to 23 different conditions were generated by PlateSeq[39] using 100ng of total RNA as template in each well.
Single-end PLATE-Seq reads were pseudoaligned to the GRCh38 transcriptome (mRNA and ncRNA) and quantified using kallisto version 0.44.0 [40] with sequence-specific bias correction. Transcript-level counts were collapsed to entrez-id gene-level counts using the tximport package in R[41]. The biomaRt package in R[42,43] was used to map transcript-level Ensembl-ids to gene-level entrez-ids.
To measure the effect of the TFs ectopic expression we first computed the protein activity signature between the unperturbed lineage cell lines (PATU and HPAFII) against the unperturbed morphogenic cell lines (KP4) and compared this signature with the Lineage-Morhpogenic signature inferred from patients. This was done by first computing a differential gene expression signature between the two lineage an morphogenic cell lines using a Student's t.test as implemented in the VIPER package, and then by applying metaVIPER to compute the protein activity signature between unperturbed lineage and unperturbed morphogenic cell lines. After confirmed the conservation of the Lineage-Morphogenic signature in the unperturbed cell lines we then measured the effect of each perturbation.
We then computed the reprogramming effect of induced by each TF.
This was done by computing a differential gene expression signature for each experimental condition (perturbation) using the assay of negative controls in the same experimental background as reference. Specifically, the cell lines transduced with mCherry and EGFP only (negative controls without M2rtTA) were used as reference for the cell lines transduced with the TFs only (without M2rtTA). The cell lines transduced with mCherry and EGFP together with M2rtTA (negative controls with M2rtTA, i.e. mCherry+ M2rtTA and EGFP + M2rtTA) were used as reference for the cell lines transduced with the TFs together with M2rtTA. Differential gene expression signatures were then transformed into protein activity signatures by metaVIPER.
The reprogramming effect induced by the over expression of each TF and negative controls was computed using the aREA algorithm by comparing the top 100 differentially activated proteins (top 50 activate and top 50 inactivate) of each experimental condition with the Lineage-Morphogenic protein activity signature computed between lineage (HPAFII and PATU) and morphogenic (KP4) cell lines.

Pooled TFs overexpression assay (single-cell RNAseq)
The same barcoded top 8 lineage MRs (and mCherry as neg. control), used in plate-seq over-expression assay , were used in the single cell over expression assay. All the ORF viruses were pooled into two viral pools so that on average 2-3 ORFs enter each cell (MOI 0.288 / each virus) in a random fashion. M2rtTA was added to the other pool to increase the transcriptional output of the ORFs in that pool. KP4 cells were transduced with these two viral pools in 6 well format with the presence of polybrene. The following day media was changed and puromycin (2.5ug/ml) and doxocycline (0.6ug/ml) was added to the cells.
The cells were incubated in total 11 days, before trypsinization, addition Multiseq barcodes[44] and chromium-run. Non-transduced KP4, HPAFII and PATU8988S cells were also included into this stage of the protocol to act as assay starting population (morphogenic cell line KP4) and assay target end point population (lineage cell lines HPAFII and PATU8988S). All the cell lines (normal KP4, HPAFII, PATU8988S and KP4 assay pools +/- M2rtTA were MultiSeq barcoded and mixed prior the Chromium-run.

Single-cells Demultiplexing Analysis
Single-cell bam files were generated with the cell ranger pipeline (version 3.0.2) using the GRCh38 as reference genome.
Variant calling was performed with SamTools[45] and generate a .vcf file containing the genomic variations of three pancreatic cell lines processed by RNASeq (HP, KP4 and PATU).
Bam files and vcf files were used as input for Demuxlet[46] for demuxlet analysis .

Protein activity analysis of Pooled TFs overexpression assay (single-cell RNAseq)
Single-cells UMI-counts were filtered using the standard QC-filtering steps as previously described and normalized to CPM. A differential gene expression signature was computed between lineage (PATU and HPAFII) and morphogenic (KP4) unperturbed single cells. This differential gene expression signature was transformed into a protein activity signature using metaVIPER approach by integrating the single-cell epithelial network and the 4 bulk PDA networks. This protein activity signature was compared with the bulk, patient-derived, lineage and morphogenic protein activity signature to confirm the conservation of the PDA subtypes in the unperturbed lineage and morphogenic single cells. Single-cells were grouped based on perturbation (i.e. the ectopic expression of individual TFs or combinations of TFs).
In order to robustly asses the perturbation effect, group of perturbed cells with less than n=30 cells were not considered for downstream analyses.
To compute the reprogramming effect induced by each TF, or combination of TFs in each single-cell, we computed the differential gene expression signature of each perturbed single cell against the negative controls (mCherry +/- M2rtTA and EGFP +/- M2rtTA) and applied MetaVIPER (using the scNET and the bulk PDA networks) to transform the single-cell differential gene expression signatures in to protein activity signatures.
The reprogramming effect in each single cell was assessed using the aREA algorithm by comparing the top 100 differentially activated proteins (top 50 activate and top 50 inactivate) with the Lineage-Morphogenic protein activity signature computed between lineage (HPAFII and PATU) and morphogenic (KP4) unperturbed single-cells. The groups of all the perturbed cells were sorted based on the fraction of cells significantly reprogrammed (FDR<0.05, 2 tails aREA test).

Western Blots

The constructs of OVOL2 or mCherry (+/- M2RTTA) used for the overexpression assay were used for Western Blots. OVOL2 or mCherry (+/- M2RTTA) were lentivirally transduced in triplicates with MOI2 to KP4 cells followed by puromycin selection and 5 days incubation in the presence of doxocycline (0.6ug/ml)to maximize the effect of potential PDA lineage transitions. After 5 days incubation, the cells were lysed, total protein levels were measured with BCA Protein Assay Kit (Pierce) and samples were Western Blotted with following antibodies:
OVOL2 #PA5-31665 (Thermo Fisher)
mCherry ab167453 (Abcam)
ZEB1 ab203829 (Abcam)
Vimentin PIMA511880 (Thermo Fisher)
E-Cadherin 3195S (Cell Signaling Technology)
Beta-Actin sc-47778 (Santa Cruz Biotechnology)

1 Maurer, C. *et al.* Experimental microdissection enables functional harmonisation of pancreatic cancer subtypes. *Gut* **68**, 1034-1043, doi:10.1136/gutjnl-2018-317706 (2019).

2 Maurer, H. C. & Olive, K. P. Laser Capture Microdissection on Frozen Sections for Extraction of High-Quality Nucleic Acids. *Methods Mol Biol* **1882**, 253-259, doi:10.1007/978-1-4939-8879-2_23 (2019).

3 Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat Genet* **48**, 838-847, doi:10.1038/ng.3593 (2016).

4 Ding, H. *et al.* Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. *Nat Commun* **9**, 1471, doi:10.1038/s41467-018-03843-3 (2018).

5 Lachmann, A., Giorgi, F. M., Lopez, G. & Califano, A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* **32**, 2233-2235, doi:10.1093/bioinformatics/btw216 (2016).

6 Federico M. Giorgi, M. J. A. a. A. C. aracne.networks: ARACNe-inferred gene networks from TCGA tumor datasets. . *R package version 1.14.0*.

7 Bailey, P. *et al.* Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* **531**, 47-52, doi:10.1038/nature16965 (2016).

8 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

9 Collisson, E. A. *et al.* Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat Med* **17**, 500-503, doi:10.1038/nm.2344 (2011).

10 Moffitt, R. A. *et al.* Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat Genet* **47**, 1168-1178, doi:10.1038/ng.3398 (2015).

11 Martin Maechler, P. R., Anja Struyf, Mia Hubert and Kurt Hornik. cluster: Cluster Analysis Basics and Extensions. *R package version 2.1.0* (2019).

12 Kruithof-de Julio, M. *et al.* Regulation of extra-embryonic endoderm stem cell differentiation by Nodal and Cripto signaling. *Development* **138**, 3885-3895, doi:10.1242/dev.065656 (2011).
13 Till, D. J. H. R. J. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning* **45**, doi:10.1023/A:1010920819831 (2001).
14 Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77, doi:10.1186/1471-2105-12-77 (2011).
15 Therneau, T. M. A Package for Survival Analysis in R. (2020).
16 Alboukadel Kassambara, M. K., Przemyslaw Biecek. survminer: Drawing Survival Curves using "ggplot2". (2019).
17 Wiener, A. L. a. M. Classification and Regression by randomForest. *R News* **2**, 18-22 (2002).
18 Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* **44**, e71, doi:10.1093/nar/gkv1507 (2016).
19 Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363-1369, doi:10.1093/bioinformatics/btu049 (2014).
20 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47, doi:10.1093/nar/gkv007 (2015).
21 Elyada, E. *et al.* Cross-Species Single-Cell Analysis of Pancreatic Ductal Adenocarcinoma Reveals Antigen-Presenting Cancer-Associated Fibroblasts. *Cancer Discov* **9**, 1102-1123, doi:10.1158/2159-8290.CD-19-0094 (2019).
22 Malo, N., Hanley, J. A., Cerquozzi, S., Pelletier, J. & Nadon, R. Statistical practice in high-throughput screening data analysis. *Nat Biotechnol* **24**, 167-175, doi:10.1038/nbt1186 (2006).
23 Peng, J. *et al.* Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res* **29**, 725-738, doi:10.1038/s41422-019-0195-y (2019).
24 Chan-Seng-Yue, M. *et al.* Transcription phenotypes of pancreatic cancer are driven by genomic events during tumor evolution. *Nat Genet* **52**, 231-240, doi:10.1038/s41588-019-0566-9 (2020).
25 Yuan, J. *et al.* Single-cell transcriptome analysis of lineage diversity in high-grade glioma. *Genome Med* **10**, 57, doi:10.1186/s13073-018-0567-9 (2018).
26 Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15, doi:10.1186/s13059-017-1382-0 (2018).
27 Guo, M., Bao, E. L., Wagner, M., Whitsett, J. A. & Xu, Y. SLICE: determining cell differentiation and lineage based on single cell entropy. *Nucleic Acids Res* **45**, e54, doi:10.1093/nar/gkw1278 (2017).
28 La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494-498, doi:10.1038/s41586-018-0414-6 (2018).
29 Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res* **47**, D853-D858, doi:10.1093/nar/gky1095 (2019).

30 Volker Bergen, M. L., Stefan Peidli, F. Alexander Wolf, Fabian J. Theis. Generalizing RNA velocity to transient cell states through dynamical modeling. *bioRxiv*, doi:10.1101/820936} (2019).
31 Li, A. B. a. S. K. a. J. L. a. S. A. a. D. M. a. S. FNN: Fast Nearest Neighbor Search Algorithms and Applications. *R package version 1.1.3* (2019).
32 van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716-729 e727, doi:10.1016/j.cell.2018.05.061 (2018).
33 Ronen, J. & Akalin, A. netSmooth: Network-smoothing based imputation for single cell RNA-seq. *F1000Res* **7**, 8, doi:10.12688/f1000research.13511.3 (2018).
34 Huang, M. *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nat Methods* **15**, 539-542, doi:10.1038/s41592-018-0033-z (2018).
35 Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat Commun* **9**, 997, doi:10.1038/s41467-018-03405-7 (2018).
36 Li, W. *et al.* MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol* **15**, 554, doi:10.1186/s13059-014-0554-4 (2014).
37 Stouffer, S. A., Suchman, E.A., DeVinney, L.C., Star, S.A.;,Williams, R.M. Jr. *The American soldier: Adjustment during Army life*. Vol. Vol. I (NJ: Princeton University Press, 1949).
38 Parekh, U. *et al.* Mapping Cellular Reprogramming via Pooled Overexpression Screens with Paired Fitness and Single-Cell RNA-Sequencing Readout. *Cell Syst* **7**, 548-555 e548, doi:10.1016/j.cels.2018.10.008 (2018).
39 Bush, E. C. *et al.* PLATE-Seq for genome-wide regulatory network analysis of high-throughput screens. *Nat Commun* **8**, 105, doi:10.1038/s41467-017-00136-z (2017).
40 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-527, doi:10.1038/nbt.3519 (2016).
41 Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521, doi:10.12688/f1000research.7563.2 (2015).
42 Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**, 1184-1191, doi:10.1038/nprot.2009.97 (2009).
43 Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439-3440, doi:10.1093/bioinformatics/bti525 (2005).
44 McGinnis, C. S. *et al.* MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat Methods* **16**, 619-626, doi:10.1038/s41592-019-0433-8 (2019).
45 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
46 Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol* **36**, 89-94, doi:10.1038/nbt.4042 (2018).

Figure 1

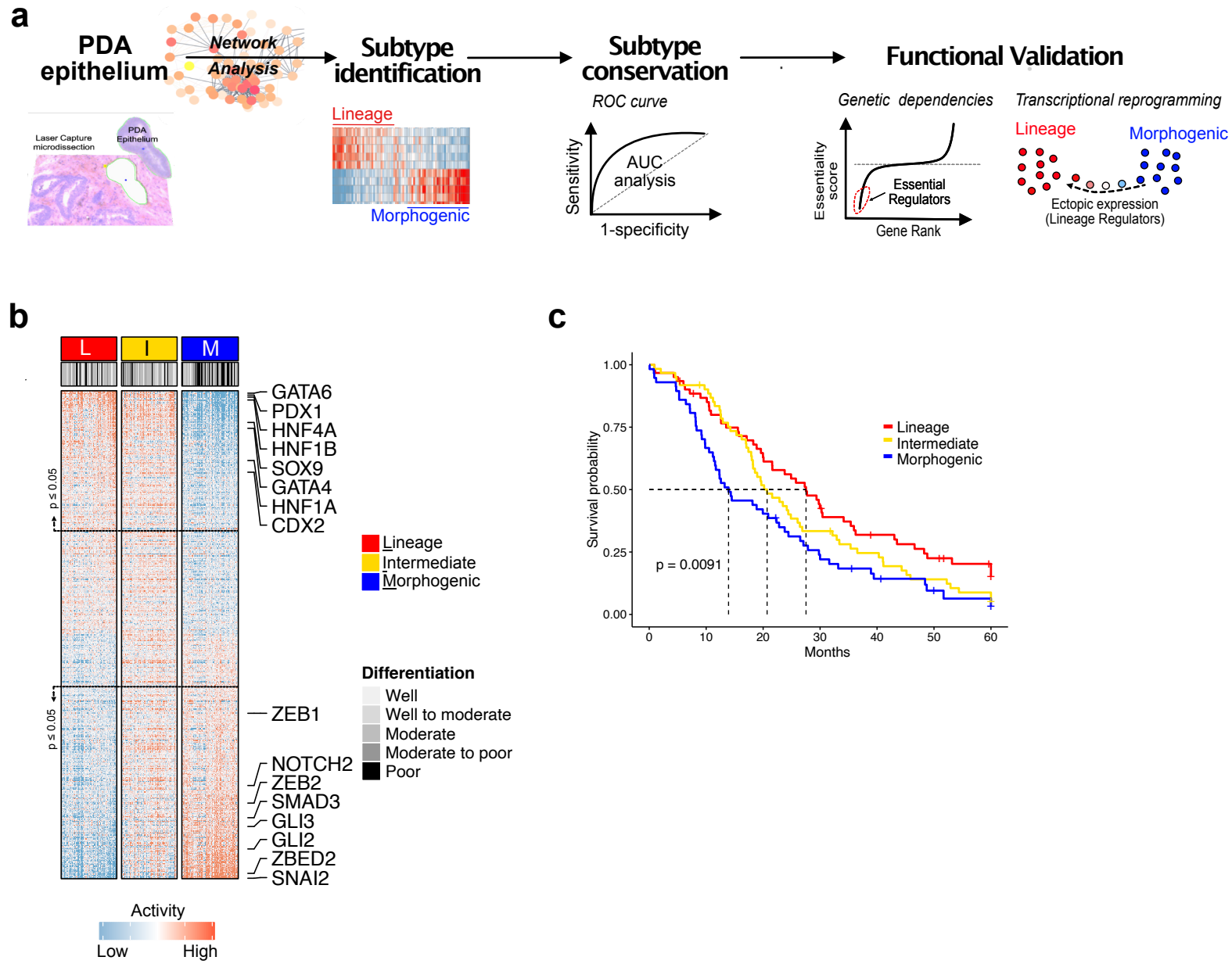

a
PDA epithelium
Network Analysis
Laser Capture microdissection
PDA Epithelium
Subtype identification
Lineage
Morphogenic
Subtype conservation
ROC curve
Sensitivity
AUC analysis
1-specificity
Functional Validation
Genetic dependencies
Essentiality score
Essential Regulators
Gene Rank
Transcriptional reprogramming
Lineage
Morphogenic
Ectopic expression (Lineage Regulators)
b
L
I
M
p ≤ 0.05
p ≤ 0.05
GATA6
PDX1
HNF4A
HNF1B
SOX9
GATA4
HNF1A
CDX2
ZEB1
NOTCH2
ZEB2
SMAD3
GLI3
GLI2
ZBED2
SNAI2
Lineage
Intermediate
Morphogenic
Differentiation
Well
Well to moderate
Moderate
Moderate to poor
Poor
Activity
Low
High
c
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.0091
Lineage
Intermediate
Morphogenic
0
10
20
30
40
50
60
Months

Figure 2

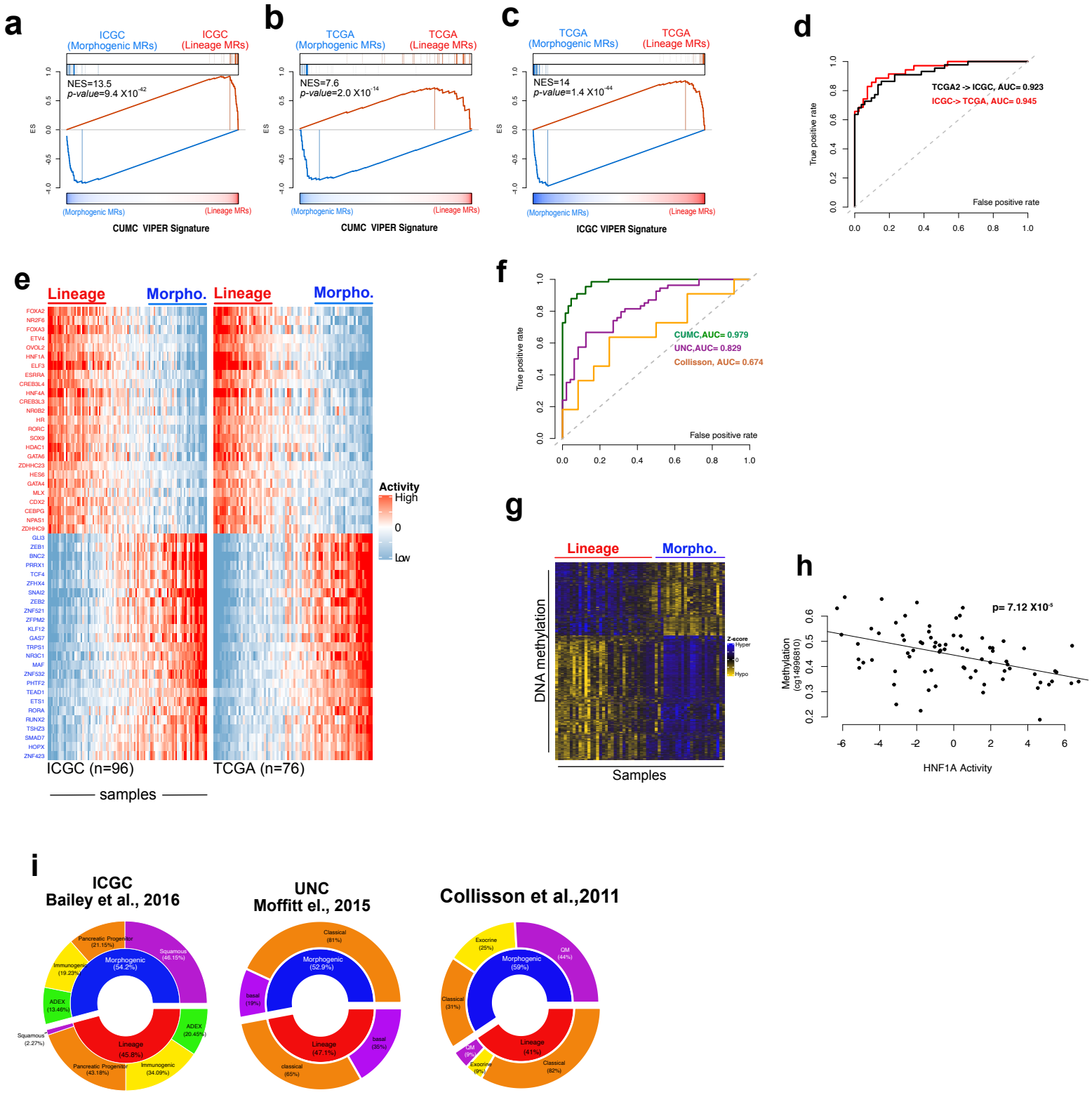

a
ICGC (Morphogenic MRs)
ICGC (Lineage MRs)
NES=13.5
p-value=9.4 X10-42
ES
(Morphogenic MRs)
(Lineage MRs)
CUMC VIPER Signature
b
TCGA (Morphogenic MRs)
TCGA (Lineage MRs)
NES=7.6
p-value=2.0 X10-14
CUMC VIPER Signature
c
TCGA (Morphogenic MRs)
TCGA (Lineage MRs)
NES=14
p-value=1.4 X10-44
ICGC VIPER Signature
d
True positive rate
False positive rate
TCGA2 -> ICGC, AUC= 0.923
ICGC-> TCGA, AUC= 0.945
e
Lineage
Morpho.
Activity
High
0
Low
ICGC (n=96)
TCGA (n=76)
samples
f
True positive rate
False positive rate
CUMC,AUC= 0.979
UNC,AUC= 0.829
Collisson, AUC= 0.674
g
Lineage
Morpho.
DNA methylation
Z-score
Hyper
Hypo
Samples
h
Methylation (cg14996810)
p= 7.12 X10-5
HNF1A Activity
i
ICGC
Bailey et al., 2016
UNC
Moffitt el., 2015
Collisson et al.,2011

Figure 3

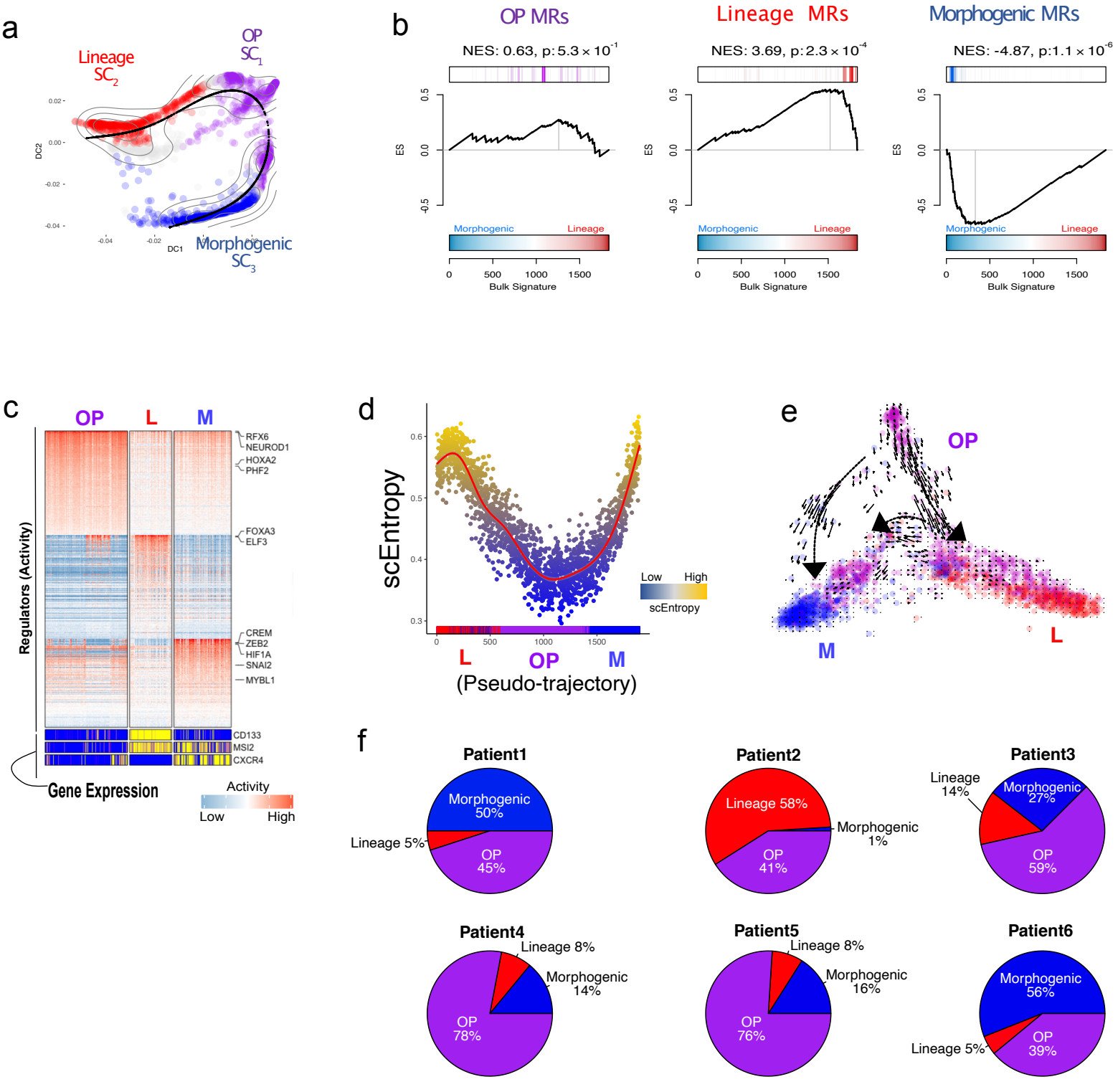
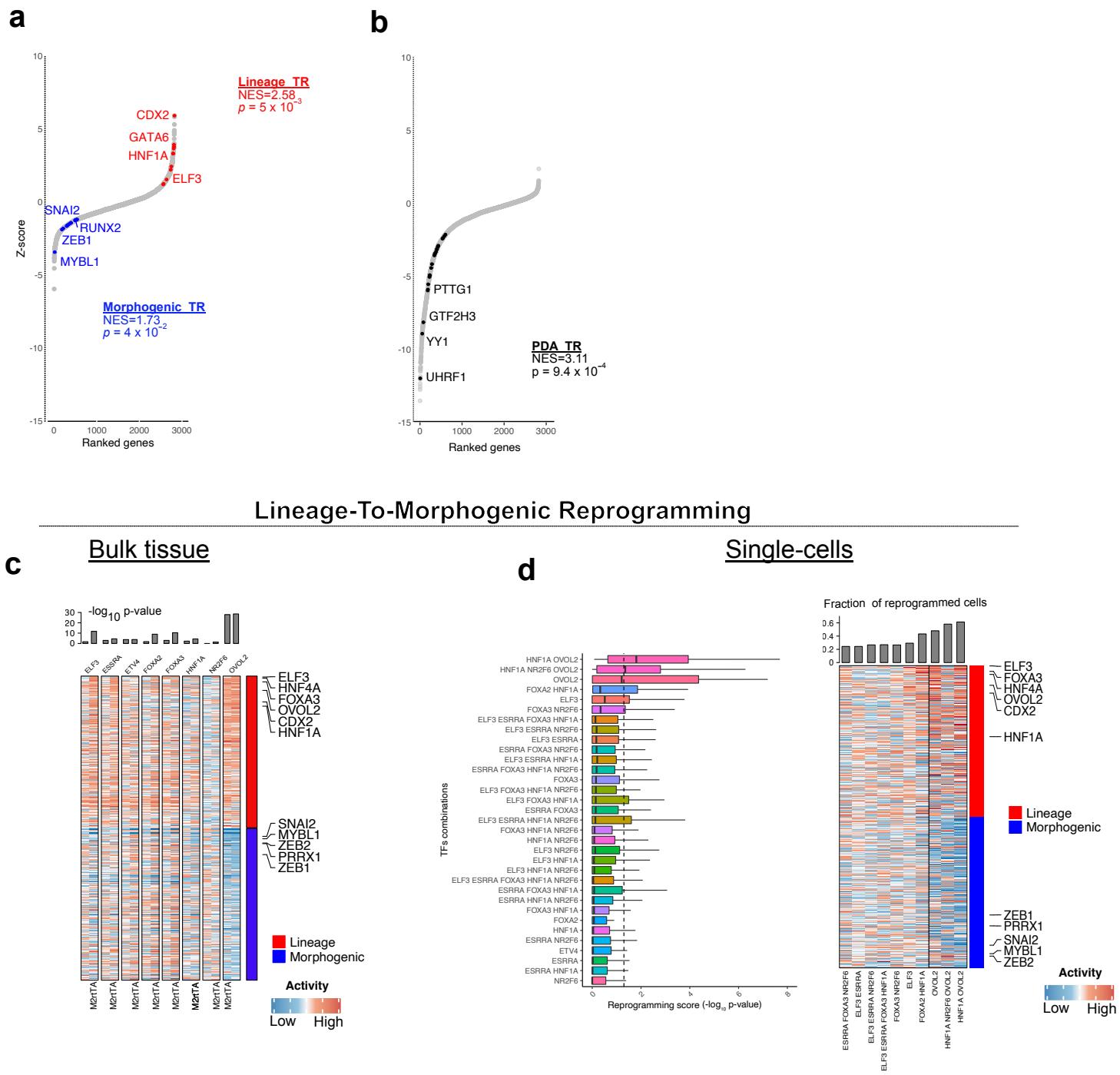
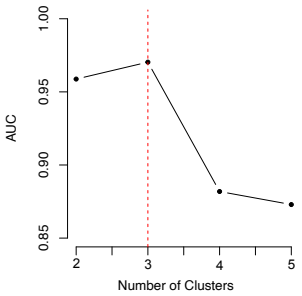

a
Lineage
SC2
OP
SC1
Morphogenic
SC3
DC1
DC2
b
OP MRs
NES: 0.63, p: 5.3 × 10-1
Lineage MRs
NES: 3.69, p: 2.3 × 10-4
Morphogenic MRs
NES: -4.87, p: 1.1 × 10-6
ES
Morphogenic
Lineage
Bulk Signature
c
OP
L
M
Regulators (Activity)
RFX6
NEUROD1
HOXA2
PHF2
FOXA3
ELF3
CREM
ZEB2
HIF1A
SNAI2
MYBL1
CD133
MSI2
CXCR4
Gene Expression
Activity
Low
High
d
scEntropy
Low
High
L
OP
M
(Pseudo-trajectory)
e
OP
M
L
f
Patient1
Morphogenic 50%
Lineage 5%
OP 45%
Patient2
Lineage 58%
Morphogenic 1%
OP 41%
Patient3
Lineage 14%
Morphogenic 27%
OP 59%
Patient4
Lineage 8%
Morphogenic 14%
OP 78%
Patient5
Lineage 8%
Morphogenic 16%
OP 76%
Patient6
Morphogenic 56%
Lineage 5%
OP 39%

Figure 4

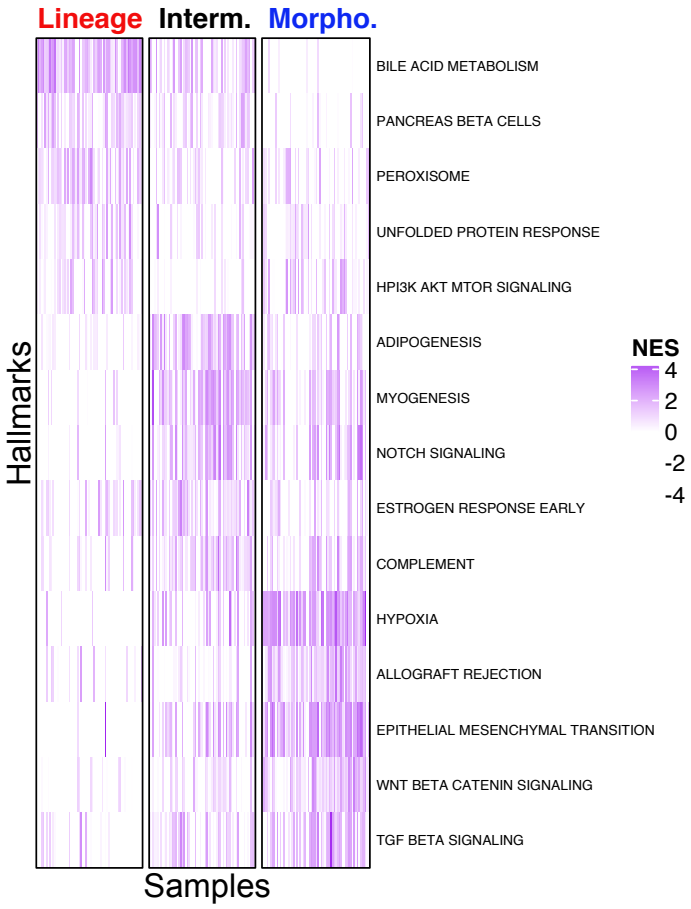

a
Lineage_TR
NES=2.58
p = 5 x 10^-3
CDX2
GATA6
HNF1A
ELF3
SNAI2
RUNX2
ZEB1
MYBL1
Morphogenic_TR
NES=1.73
p = 4 x 10^-2
Z-score
Ranked genes
b
PTTG1
GTF2H3
YY1
UHRF1
PDA_TR
NES=3.11
p = 9.4 x 10^-4
Ranked genes
Lineage-To-Morphogenic Reprogramming
c
Bulk tissue
-log10 p-value
ELF3
ESRRA
ETV4
FOXA2
FOXA3
HNF1A
NR2F6
OVOL2
ELF3
HNF4A
FOXA3
OVOL2
CDX2
HNF1A
SNAI2
MYBL1
ZEB2
PRRX1
ZEB1
Lineage
Morphogenic
M2rtTA
Activity
Low
High
d
Single-cells
TFs combinations
Reprogramming score (-log10 p-value)
Fraction of reprogrammed cells
ELF3
FOXA3
HNF4A
OVOL2
CDX2
HNF1A
Lineage
Morphogenic
ZEB1
PRRX1
SNAI2
MYBL1
ZEB2
Activity
Low
High

Extended Data Figure 1

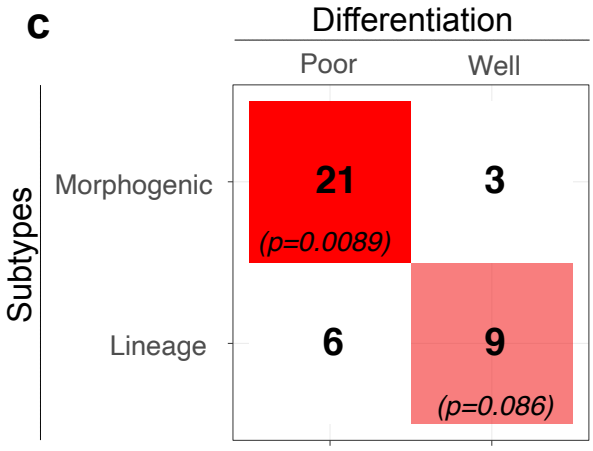
a



b

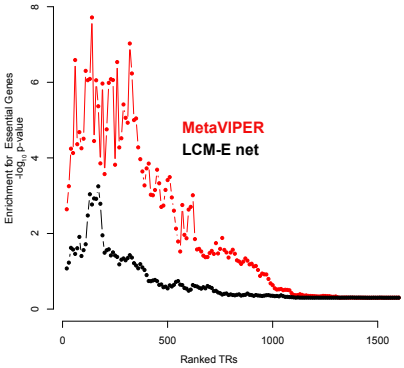


c

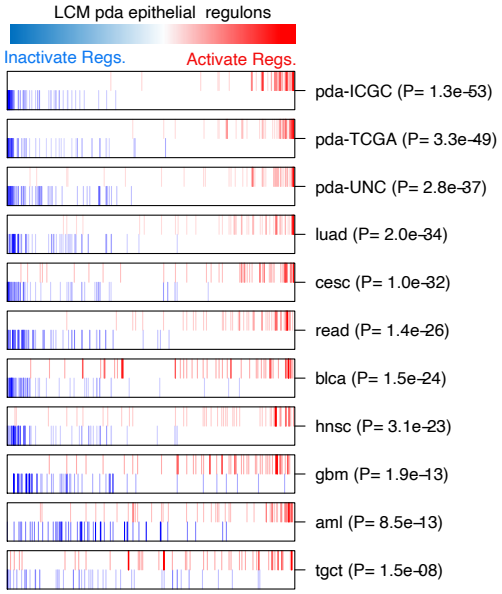


Extended Data Figure 2

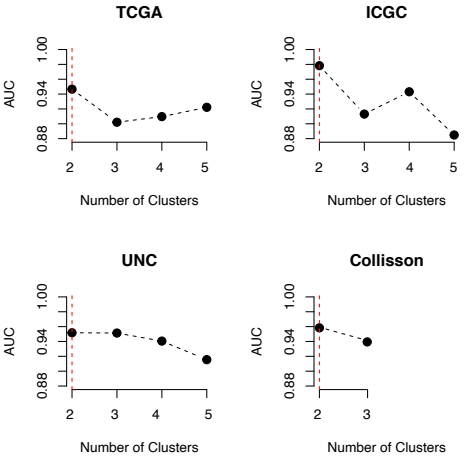
a



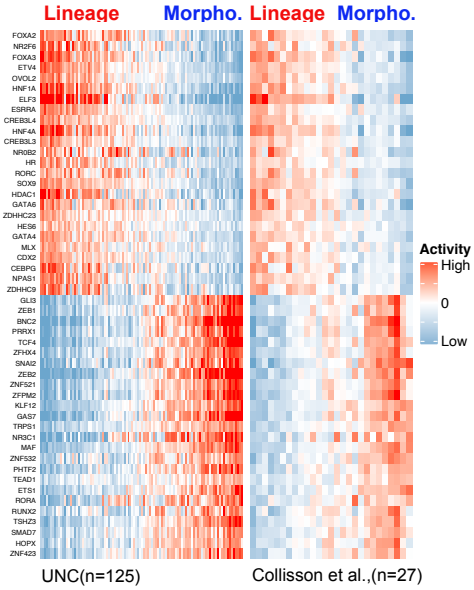
b



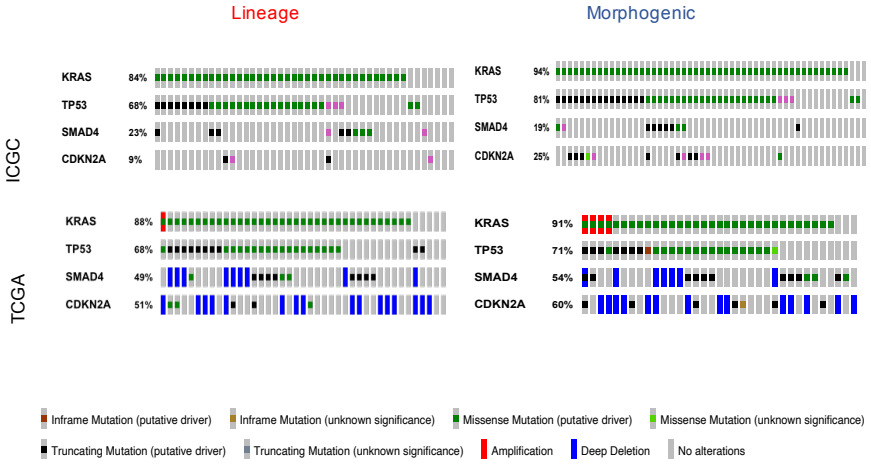
c



e

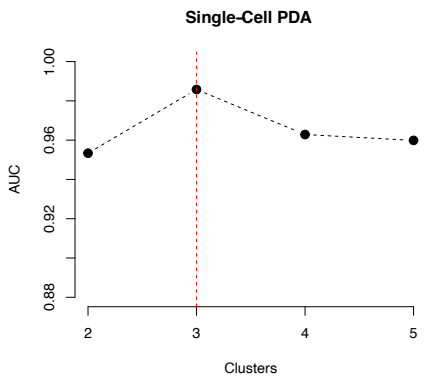


d

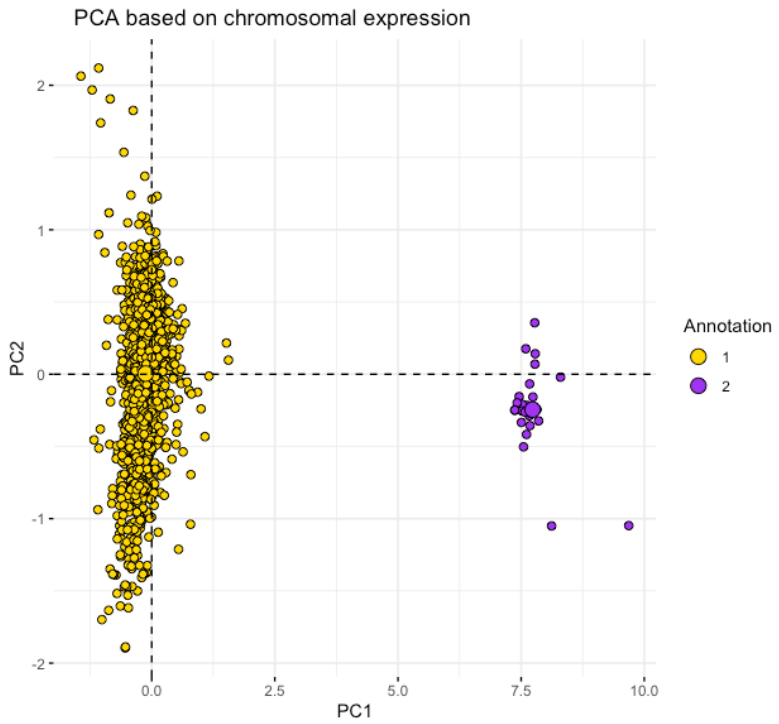


Extended Data Figure 3

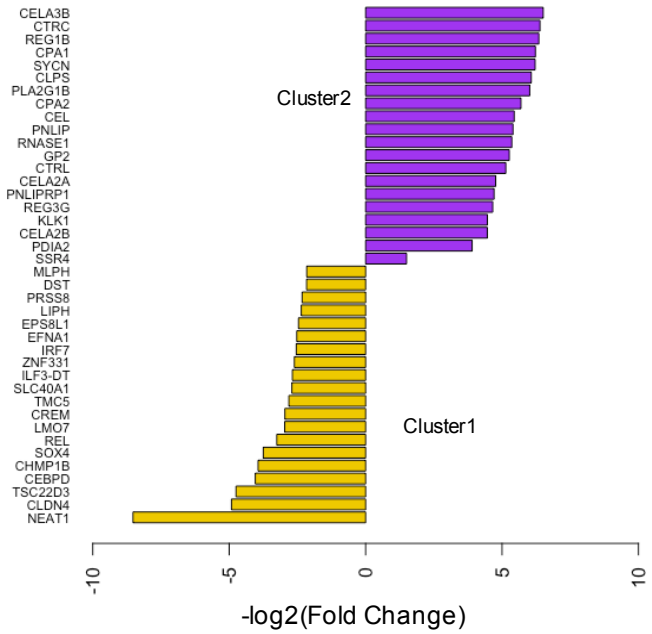
a



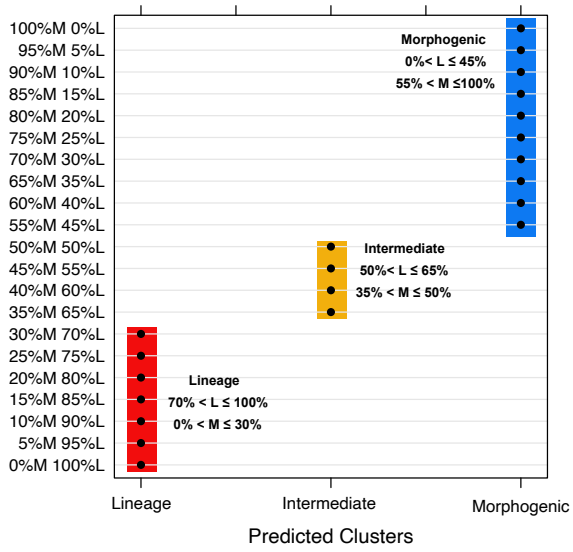
b



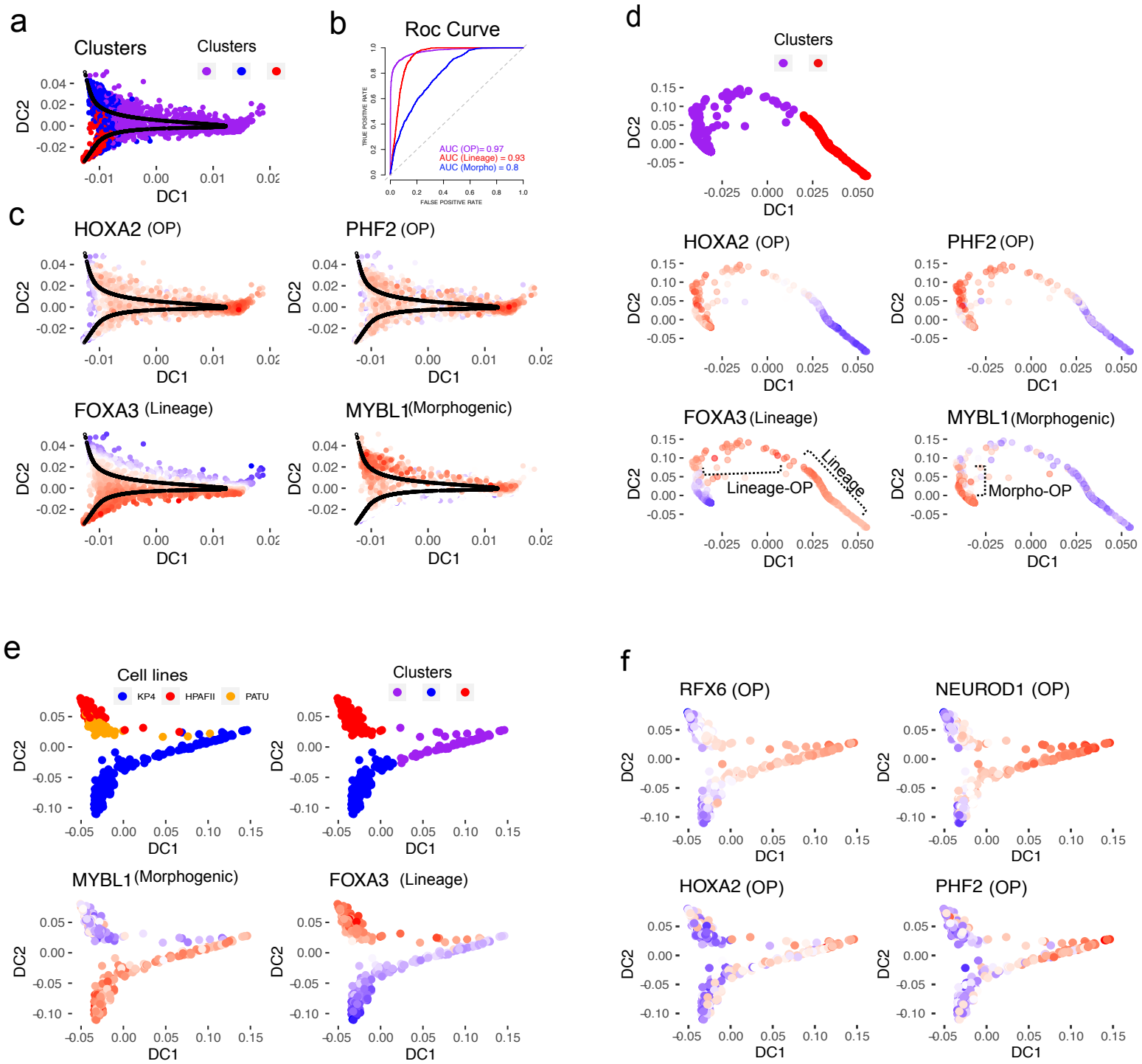
c



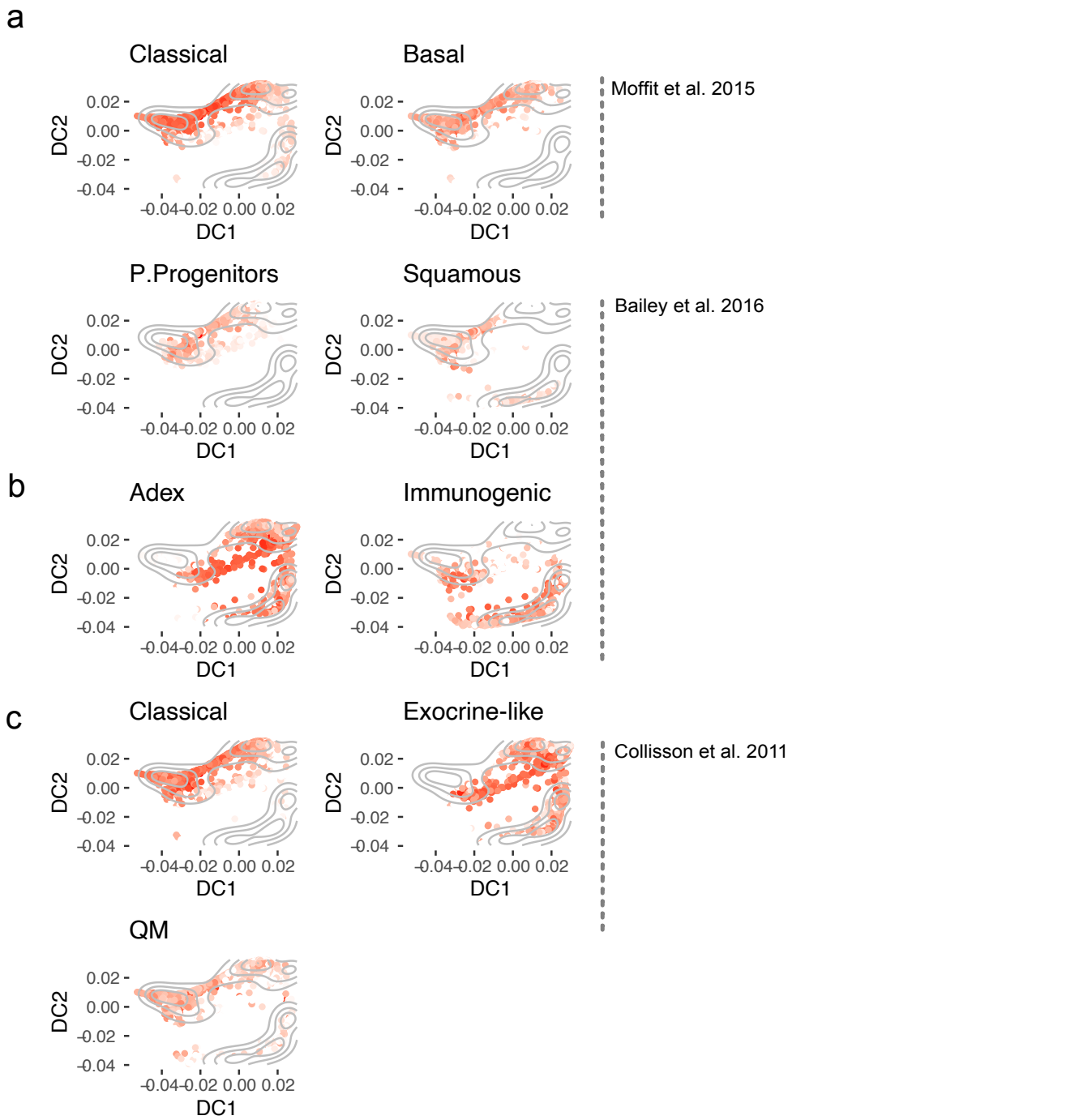
d

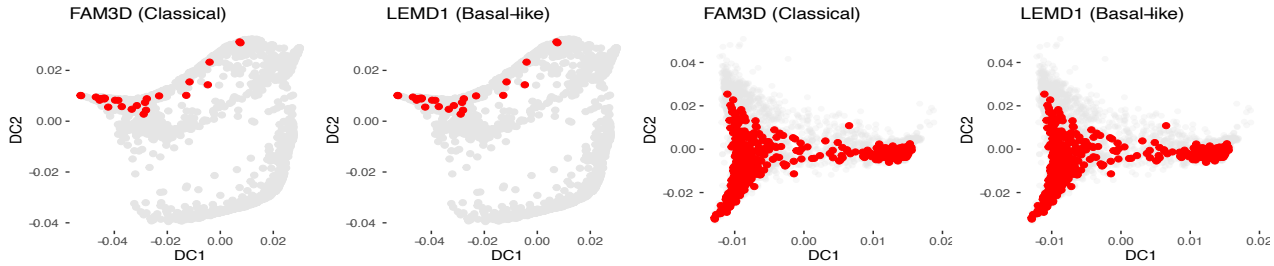


Extended Data Figure 4



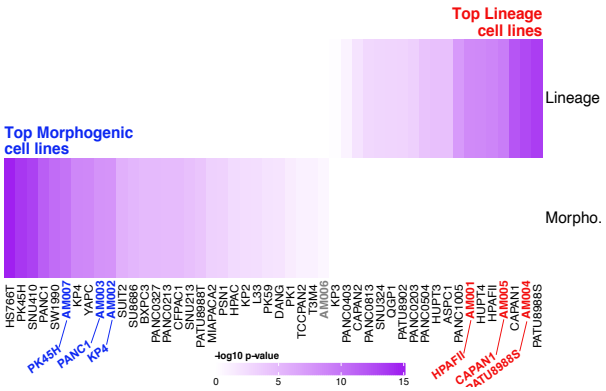
Extended Data Figure 5



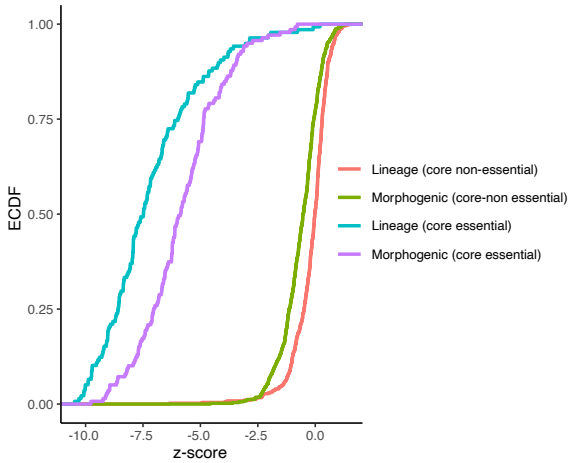


Extended Data Figure 6

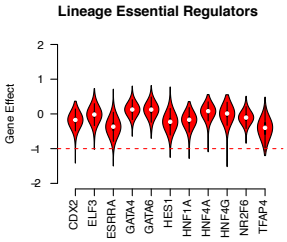
a

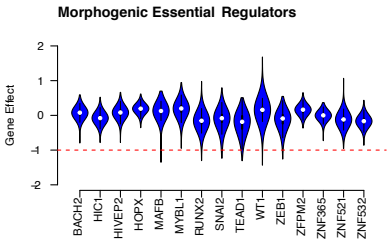


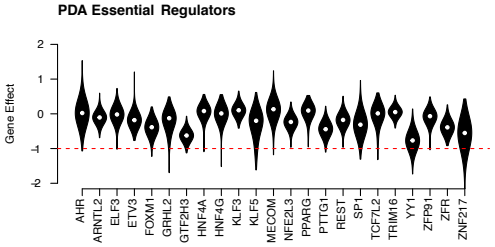
b



c

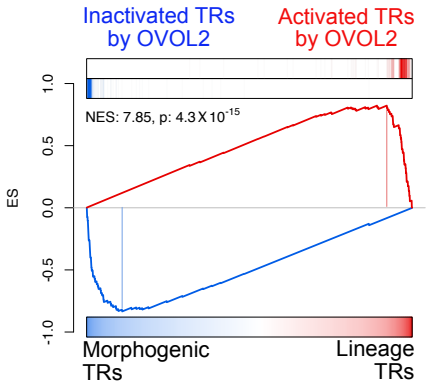




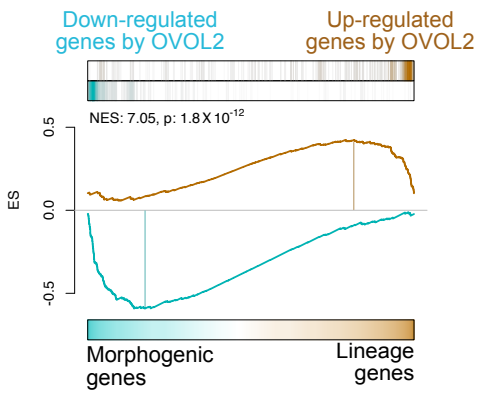


Extended Data Figure 7

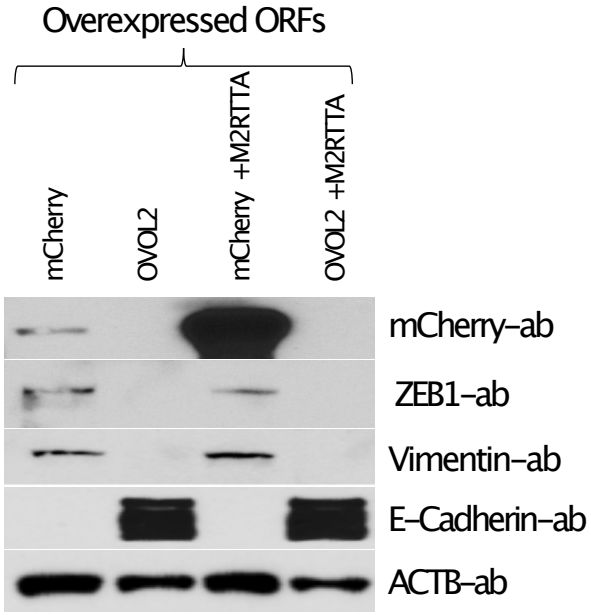
a



b



c



Western Blots