Methods

Sample acquisition

Freshly frozen tissue samples of pancreatic ductal adenocarcinoma (PDA) (n = 129) were obtained from patients who underwent surgical resection at the Pancreas Center at Columbia University Medical Center. The clinical data of these patients are shown in Supplementary Tables 1 and 2. Prior to surgery, all patients had given surgical informed consent, which was approved by 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.

Sample extraction

Frozen tissue specimens were cut at 8 - 9 µm thickness and 2 - 3 sections were transferred onto a PEN membrane glass slide (Arcturus, Applied Biosystems). For initial histopathological review, immediate-adjacent sections were cut and stained using a standard H&E protocol to confirm the diagnosis and identify suitable areas for either laser capture microdissection or macrodissection.

Laser capture microdissection

Throughout the staining procedure, RNAse-free water was used. Sections were fixed in 95% ethanol and stained with cresyl violet acetate (1% in Tris-buffered 70% ethanol), followed by a brief washing step in 70% ethanol, and a final dehydration in 100% ethanol. Laser capture microdissection was performed on a PALM MicroBeam

microscope (Zeiss) to collect at least 1000 cells per compartment. Samples were microdissected from regions of frank carcinoma marked in advance by a GI Pathologist (A.C.I.). Captured cells were then transferred to RLT plus buffer (Qiagen) and lysed for 30 min at room temperature (RT).

RNA

RNA was extracted using the RNeasy Plus Micro Kit (Qiagen) following the manufacturer's instructions. Prior to further processing, RNA integrity and yield were determined using an Agilent 2100 Bioanalyzer (RNA 6000 Pico Kit for LCM and RNA 6000 Nano Kit for bulk samples, respectively). Yields ranged from 1 to 10 ng per LCM sample and several µg per bulk sample, respectively. Only samples with a RNA Integrity Number (RIN) of at least 7 were used for further processing.

RNA amplification and library preparation

For microdissected samples, 1 - 2 ng of RNA from LCM samples were amplified using the Ovation RNA-Seq System V2 Kit (NuGEN) following the manufacturer's instructions. The resulting cDNA libraries were fragmented using a Covaris S2 Sonicator. For samples that underwent macrodissection, a minimum of 200 ng of total RNA underwent a poly-A pull-down to enrich for mRNAs which then were used as input for the Illumina TruSeq RNA prep kit. Both types of samples were prepared for the Illumina HiSeq 2000 platform using a Beckmann-Coulter Roboter and the SPRIworks Fragment Library Kit I. Finally, a PCR using the KAPA PCR Amplification Kit was carried out. The libraries were then sequenced by the Columbia Genome Center to generate 30 million single-end reads of 100 bp length.

RNA-Seq analysis

Reads were mapped to the human reference genome (NCBI/build 37.2) using Tophat (Version 2.0.4)(1) and two methods of gene expression quantification were employed with standard settings: (i) HTSeq(2) to obtain raw read counts per gene and (ii) Cufflinks(3) (version 2.0.2) to obtain fragments per kilobase of exon per million reads mapped (FPKM) per gene and transcript, respectively. In addition to the NCBI RefSeq gene model, gene expression was quantified using Ensembl GRCh37 gene annotations

and the pipeline described, to allow the accurate evaluation of subtype-specific genes as described in Bailey et al. (4). [i]The RSeQC package (5) was used to evaluate the suitability of RNA-Seq libraries for further analysis.

Data processing. Raw counts from LCM RNA-Seq were normalized using the variance stabilizing transformation (VST) from the 'DESeq2(6) R(7) package after filtering out genes with fewer than 5 reads in 50% of samples, with the setting: varianceStabilizingTransformation(obj, blind = FALSE).

Derivation of a stromal subtype signature. Stromal LCM-RNA-Seq gene expression profiles from 110 patients were VST-normalized as described above and subset to stroma-specific genes as determined from the paired DEG between epithelium and stroma (t-statistic > 0 and FDR < 0.1, respectively). Furthermore, genes expressed from chromosome Y as well as from the XIST (X-inactive specific transcript) locus were excluded as their variance is likely to be related to gender-specific differences in gene expression, which we reasoned should not be considered during subtyping, leaving us with 4401 genes. The 1000 most variable genes were determined from 100 bootstraps of the 4401 genes x 110 samples input matrix based on their median absolute deviation (MAD). The 1000 genes occurring most often in these 100 bootstraps were used as input for nonnegative matrix factorization (NMF) using the NMF R package(8) with its standard algorithm (Brunet) and a random seeding method where entries of each factor are drawn from a uniform distribution over [0, max(A)], where A is the input matrix. The factorization rank was estimated in 50 runs of NMF from actual and randomized data for 2 through 10 ranks. Cophenetic correlation and silhouette scores of the consensus matrix decreased substantially already from 2 to 3 ranks and declined further for higher factorization ranks. NMF was then run with 200 iterations (*nmf(x, rank = 2, nrun = 200, seed = 1)*) and the samples were assigned a cluster by hierarchical clustering of the consensus matrix using complete linkage. In order to derive a signature distinguishing between the two subtypes, differential genes were identified using the siggenes R package (9). Genes with a q-value < 0.001 and either an above median expression or an effect size > 1.5 were retained. The signature genes and their

functional annotation using Gene Ontology Biological Processes and the DAVID Bioinformatics Resources (10) can be found in supplementary tables 22 – 24.

Effect size meta-analysis. In order to evaluate evidence from four PDA cohorts on the co-occurrence of certain compartment–specific molecular subtypes, a meta-analysis was carried out using the metafor (11) R package. The 2 x 2 table was set up in a way that Basal-like tumors and ECM-rich stromas represented the first row and column, respectively. The log odds ratio and its variance were calculated for each study using the *escalc(measure = "OR", …)* function. Random and fixed effect models were fit using the *rma* function (*method = "REML"* and *method = "FE"*, respectively). All necessary input data can be retrieved from supplementary table 6.

Experimental validation

Assessment of RNA amplification behavior. In order to test whether RNA amplification using the Ovation RNA-Seq System V2 (NuGEN) indeed occurs in a linear manner (12), we added ERCC Spike-In Mix 1 (Ambion) at increasing concentrations (1X, 2X, …, 32X - 2 replicates per concentration) to RNA samples before amplification. ERCC Spike-In Mix 1 contains a collection of 92 synthetic, polyadenylated transcripts between 250 and 2000 bp at defined concentrations. For quantification purposes, the ERCC Spike-In sequences were added to the reference genome and annotation file, respectively, and the same Tophat/HTSeq pipeline that was outlined above was used for raw read quantification. Read counts from libraries containing ERCC Spike-In mix were normalized by accounting for differences in library size using the DESeq2 package (*counts(obj, normalized = TRUE)*). ERCC Spike-In species that did not have at least one read in 80% of the spiked-in libraries were discarded from the analysis, which left 19 distinct ERCC transcripts, 16 of which could be categorized into 2 major length categories and 3 major concentration categories (see table below). Both normalized reads and Spike-In concentration were log2-transformed and Pearson correlation was determined at each level of length and concentration, respectively.

ERCC ID	Concentration log10(attomol/µl)	Concentration category	Length in bp	Length category
ERCC-00002	4.18	10^4 - 10^5	1061	1000 bp
ERCC-00003	2.97	10^2 - 10^3	1023	1000 bp
ERCC-00004	3.88	10^3 - 10^4	523	500 bp
ERCC-00009	2.97	10^2 - 10^3	984	1000 bp
ERCC-00042	2.67	10^2 - 10^3	1023	1000 bp
ERCC-00043	2.67	10^2 - 10^3	1023	1000 bp
ERCC-00046	3.58	10^3 - 10^4	522	500 bp
ERCC-00060	2.36	10^2 - 10^3	523	500 bp
ERCC-00074	4.18	10^4 - 10^5	522	500 bp
ERCC-00095	2.08	10^2 - 10^3	521	500 bp
ERCC-00108	2.97	10^2 - 10^3	1022	1000 bp
ERCC-00111	2.67	10^2 - 10^3	994	1000 bp
ERCC-00130	4.48	10^4 - 10^5	1059	1000 bp
ERCC-00136	3.28	10^3 - 10^4	1033	1000 bp
ERCC-00145	2.97	10^2 - 10^3	1042	1000 bp
ERCC-00171	3.58	10^3 - 10^4	505	500 bp

Human Protein Atlas (HPA). Using the HPA search algorithm, we subset the list of potential proteins to those that had highest quality antibodies for immunohistochemistry (IHC) by applying the filter 'ih_tissue_reliability:Supportive'. Next, we evaluated IHC staining patterns for PDA tumor sections in the HPA pathology database for the top and bottom 50 genes of our subset DEG list which represented mRNA-predicted stromal and epithelial genes, respectively. The observed IHC staining patterns were categorized as follows: (i) 'strongly supportive' for cases where IHC staining aligned with mRNA prediction with both high signal intensity and compartment-specificity, (ii) 'weakly supportive' for cases where IHC staining aligned with mRNA prediction with either moderate signal intensity or moderate compartment-specificity (i.e. differential expression can be appreciated on the protein level, but is not bimodal), (iii) 'indeterminate' for cases with absent compartment-specificity (i.e. no appreciable differential expression on the protein level), (iv) 'not expressed' for cases where low or absent signal intensity precluded meaningful analysis, and (v) 'opposing' for cases where IHC staining contradicted mRNA prediction. The full list of genes that are used as compartment-specific genes by ADVOCATE, the list of genes for which antibodies with

high quality for IHC are available at the HPA and the results from the described analysis can be found in Supplementary Tables 3 - 5.

External PDA datasets

Dataset preparation. The UNC dataset was downloaded using the R package 'GEOquery'(13), and accession number GSE71729. Only pancreatic primary tumor samples with survival data were retained for our analysis (n = 125). For the ICGC cohort, normalized expression and clinical data were extracted from the supplementary information provided by Bailey et al(4). After matching sample IDs with phenotype data, we kept 93 samples for our analysis, excluding the two acinar cell carcinomas. RNA-Seq V2 data was downloaded from the TCGA data portal on 5/31/2016 together with clinical and biospecimen information - including a report on 27 cases that were retracted after review by the PAAD EPC. In addition, we excluded cases with a diagnosis of 'other malignancy'. Raw read counts per gene were extracted from the RSEM output files for all samples, normalized to account for different library sizes, and the variance was stabilized as implemented in the DESeq2(6) package. Details on the epithelial and stromal molecular subtypes per case as well as the survival data used can be found in supplementary table 6.

Gene signatures used for classification

Clustering of both bulk expression profiles and virtually purified compartment–specific profiles was performed by sub-selecting the normalized expression data of epithelial signature genes described by Moffitt et al (14) for epithelial subtypes and genes distinguishing the two major stromal subtypes in LCM data for stromal subtypes. Both signature gene lists can be found in supplementary tables 21 and 22, respectively.

Derivation of subtype-specific genes in the ICGC cohort

Subtype-specific classifier genes for 'ADEX', 'Immunogenic', 'Pancreatic Progenitor' and 'Squamous' tumors were derived as follows from the list of 613 genes found to be differentially expressed between all subtypes by Bailey et al. (= SAM genes).

Normalized expression data and information on subtype and silhouette score for 96 tumors from the ICGC cohort were extracted from the supplemental data provided by Bailey et al. Only tumors with positive silhouette scores were retained for the analysis (n = 83). First, 'subtype vs. rest' comparisons were performed for each subtype using the *limma* R package. As an initial filtering step to identify subtype-specific genes, only those genes were kept that were significantly upregulated in the respective subtype while not being significantly upregulated in any of the other comparisons.

If this included genes for which there were non-significant trends (i.e. positive t-statistics) in other comparisons, an additional comparison was made between 'subtype A and B' specifically. This strategy yielded 48 ADEX-specific, 92 Immunogenic-specific, 3 Progenitor-specific and 70 Squamous-specific genes.

Input data, results from the DEG analyses and all classifier lists used in this manuscript can be found in Supplementary Tables 17-22.

Clustering

A Pearson dissimilarity distance matrix was used as the input for k-means consensus clustering for each dataset with the same preset random seed value. Each clustering was run 500 times. Subsequently, the cluster output from all k-means clustering was hierarchically clustered and the number of clusters was decided based on the *cutree* function in R 'stats', thereby yielding the final results for each dataset and each compartment. Using the average silhouette score, most robust results were obtained for 2 epithelial and stromal clusters, respectively, in each data set. For the virtual epithelial expression profiles in the UNC cohort, both 2 and 3 clusters were equally feasible. We elected to use 2 clusters for this dataset for the sake of consistency. In order to be consistent with the stromal signature derivation process, stromal class assignment was carried out using the same NMF and hierarchical clustering pipeline described above. As expected some CUMC stromal signature genes were less informative, i.e. showed low variance as determined by their interquartile range (IQR), in the external cohorts. Therefore, those stromal signature genes exhibiting a below global variance in a given cohort were removed before clustering.

Gene set enrichment analysis

We used the R implementation of single sample Gene Set Enrichment analysis: GSVA (gene set variation analysis) with default parameters (15). The input expression matrix was filtered for the most variable 50% of the genes as determined by interquartile range (IQR) in case of the epithelial annotation and to stromal genes (log_2 fold change > 0.5 in the paired epithelium vs. stroma differential expression analysis) in case of the stromal annotation. For annotation of epithelial subtypes, we tested a set of gene sets shown previously to be discriminating between classical and basal-like tumors (4,14). For the annotation of stromal subtypes, we first determined stroma-specific pathways among the c2 canonical pathways module from the MSigDB(4) by comparing the ADVOCATE epithelial and stromal training samples. Using these stroma-specific pathways (FDR in epithelial vs. stromal GSVA comparison < 0.05), we performed differential enrichment analysis between the prominent stromal subtypes in the LCM cohort using the R package 'limma'(16) and selected a similar number of illustrative gene sets as for the epithelium. All tested gene sets and their description can be found in supplementary tables 7 and 8.

Silhouette score and Differential Gene Expression (DEG)

Silhouette scores were calculated using the silhouette function from the cluster (17) R package with the k-means clustering results and the Pearson dissimilarity matrix as input. Differential gene expression analysis was calculated out using the 'limma'(16) package. Among all differentially expressed genes between compartment-specific subtypes (adjusted p-value < 0.05) the top 30 genes per subtype were shown in the heatmaps.

Survival analysis

The association of molecular subtypes with disease outcome was evaluated using Kaplan-Meier survival analysis calculated along with log-log p value using the *surv* and *coxph* functions from the 'survival'(18) package. Furthermore, fitting and visualization of

Cox proportional hazards models were carried out using the *cph* function from the R package 'rms' with all ties handled using the Breslow method.

References

1. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics **2009**;25(9):1105-11 doi 10.1093/bioinformatics/btp120.
2. Simon Anders P, Wolfgang H. HTSeq—A Python framework to work with high-throughput sequencing data. Bioinformatics **2014**.
3. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ*, et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol **2010**;28(5):511-5 doi 10.1038/nbt.1621.
4. Bailey P, Chang DK, Nones K, Johns AL, Patch A-M, Gingras M-C*, et al.* Genomic analyses identify molecular subtypes of pancreatic cancer. Nature **2016**;531(7592):47-52 doi 10.1038/nature16965.
5. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics **2012**;28(16):2184-5 doi 10.1093/bioinformatics/bts356.
6. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol **2014**;15(12):550 doi 10.1186/s13059-014-0550-8.
7. Team RDC. R: A Language and Environment for Statistical Computing. Vienna, Austria2016.
8. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics **2010**;11(1):367 doi 10.1186/1471-2105-11-367.
9. Schwender H. siggenes: Multiple testing using SAM and Efron's empirical Bayes approaches 2012.

10. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protocols **2008**;4(1):44-57 doi 10.1038/nprot.2008.211.
11. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. Journal of Statistical Software, Articles **2010**;36(3):1--48 doi 10.18637/jss.v036.i03.
12. Kurn N, Chen P, Heath JD, Kopf-Sill A, Stephens KM, Wang S. Novel isothermal, linear nucleic acid amplification systems for highly multiplexed applications. Clinical chemistry **2005**;51(10):1973-81 doi 10.1373/clinchem.2005.053694.
13. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics **2007**;23(14):1846-7 doi 10.1093/bioinformatics/btm254.
14. Moffitt RA, Marayati R, Flate EL, Volmar KE, Loeza SGH, Hoadley KA*, et al.* Virtual microdissection identifies distinct tumor-and stroma-specific subtypes of pancreatic ductal adenocarcinoma. Nat Genet **2015**;47(10):1168-78.
15. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics **2013**;14(1):1-15 doi 10.1186/1471-2105-14-7.
16. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W*, et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res **2015**;43(7):e47 doi 10.1093/nar/gkv007.
17. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. cluster: Cluster Analysis Basics and Extensions. **2017**.
18. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. Springer New York; 2013.