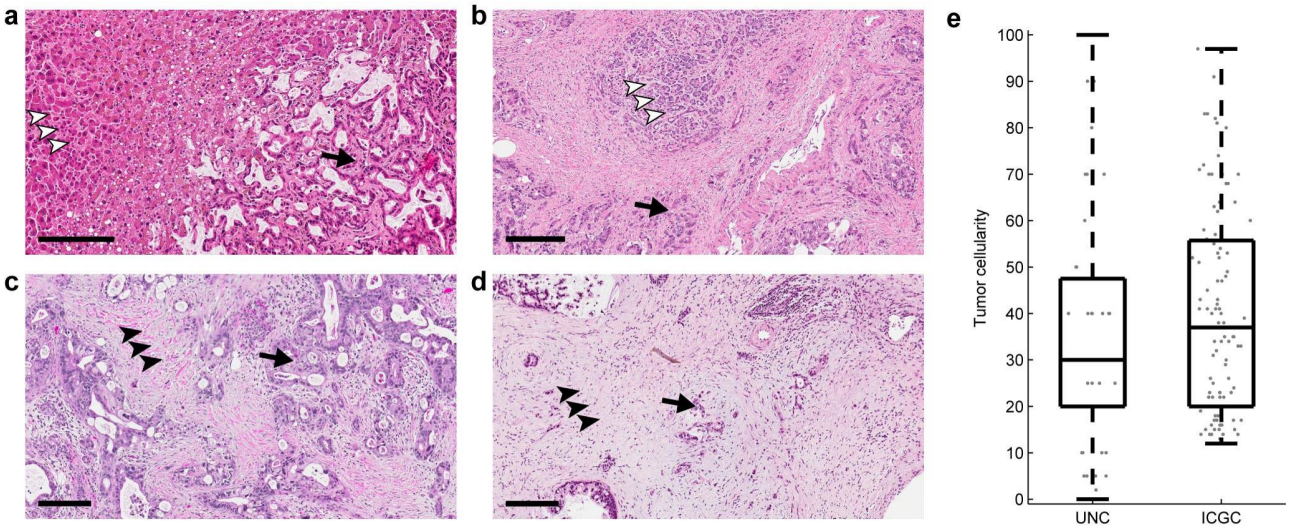
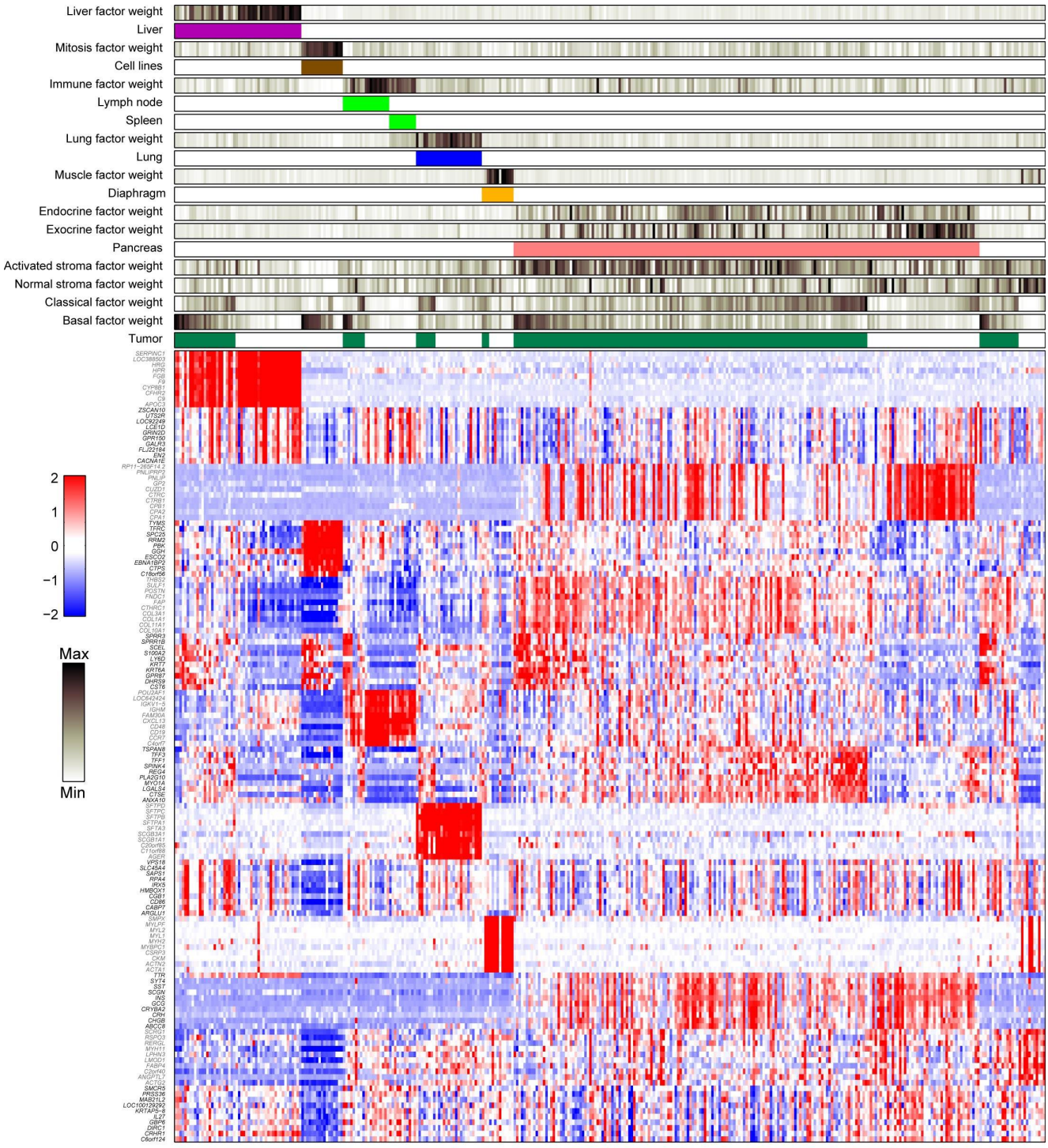


Supplementary Figure 1

Representative H&E staining of patient tumor samples.

(**a**) Liver metastases showing regions of tumor and normal tissue. (**b**) Primary pancreatic tumor sample showing normal pancreatic tissue and tumor cells in the same field. (**c**) Primary pancreatic tumor with high tumor cellularity. (**d**) Primary pancreatic tumor with abundant tumor stroma. (**e**) Percentage of tumor in primary pancreatic tumors in the UNC and ICGC cohorts. Black arrowheads show areas of tumor stroma. Black arrows show areas of tumor. White arrowheads show normal tissue. Scale bars, 200 μm.

Nature Genetics: doi:10.1038/ng.3398

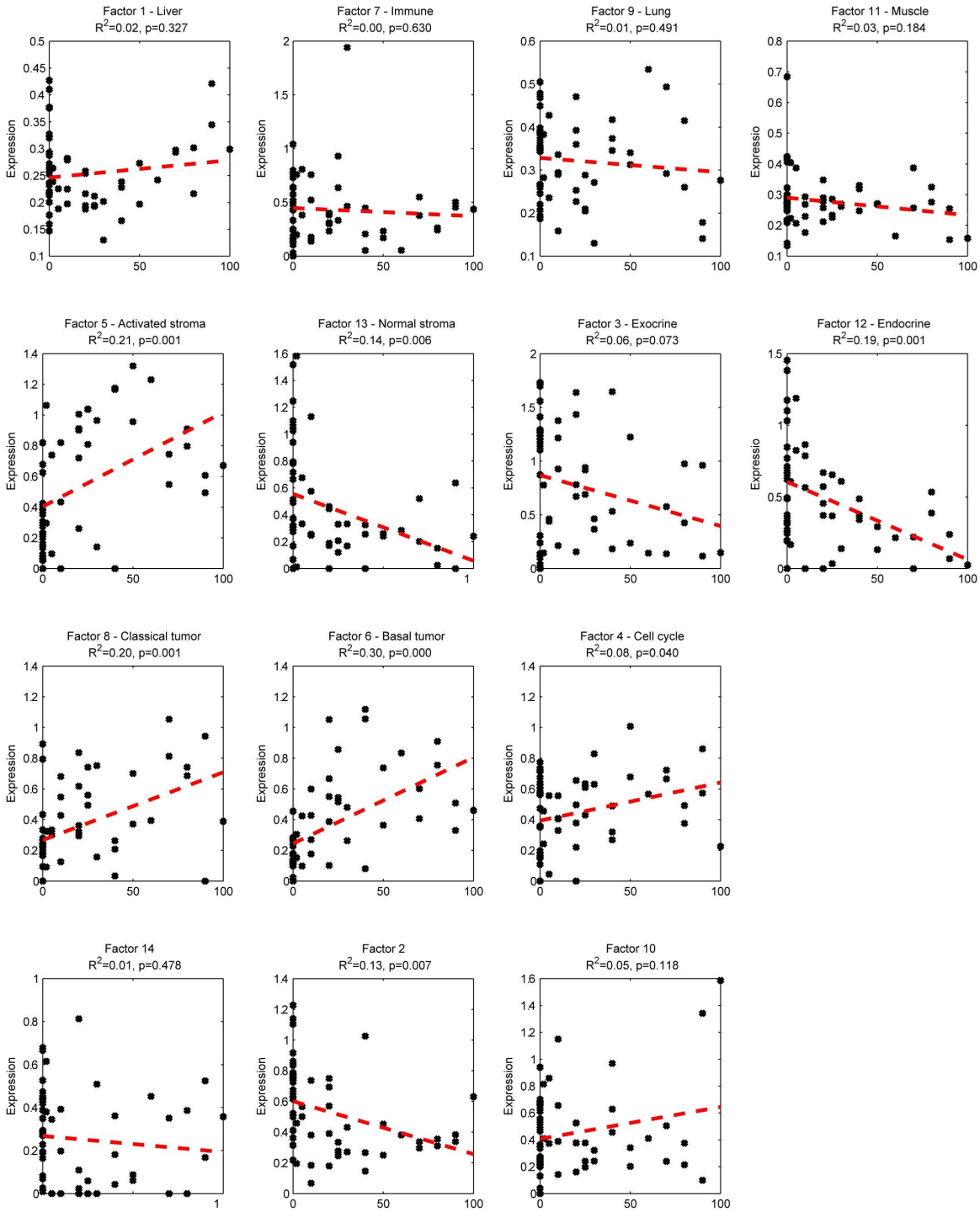


Supplementary Figure 2

Deconvolution of a large cohort of PDAC reveals distinct gene expression patterns from multiple tissue types.

Solid color bars above the heat map show the tissue of origin and tumor status of the samples, which were used to order the samples horizontally. Factor weights derived by NMF for selected factors are shown as grayscale bars. Heat maps show the *z*-normalized gene expression of exemplar genes from each factor. All tumors, cell lines and adjacent normal tissues from our cohort are shown.

Nature Genetics: doi:10.1038/ng.3398

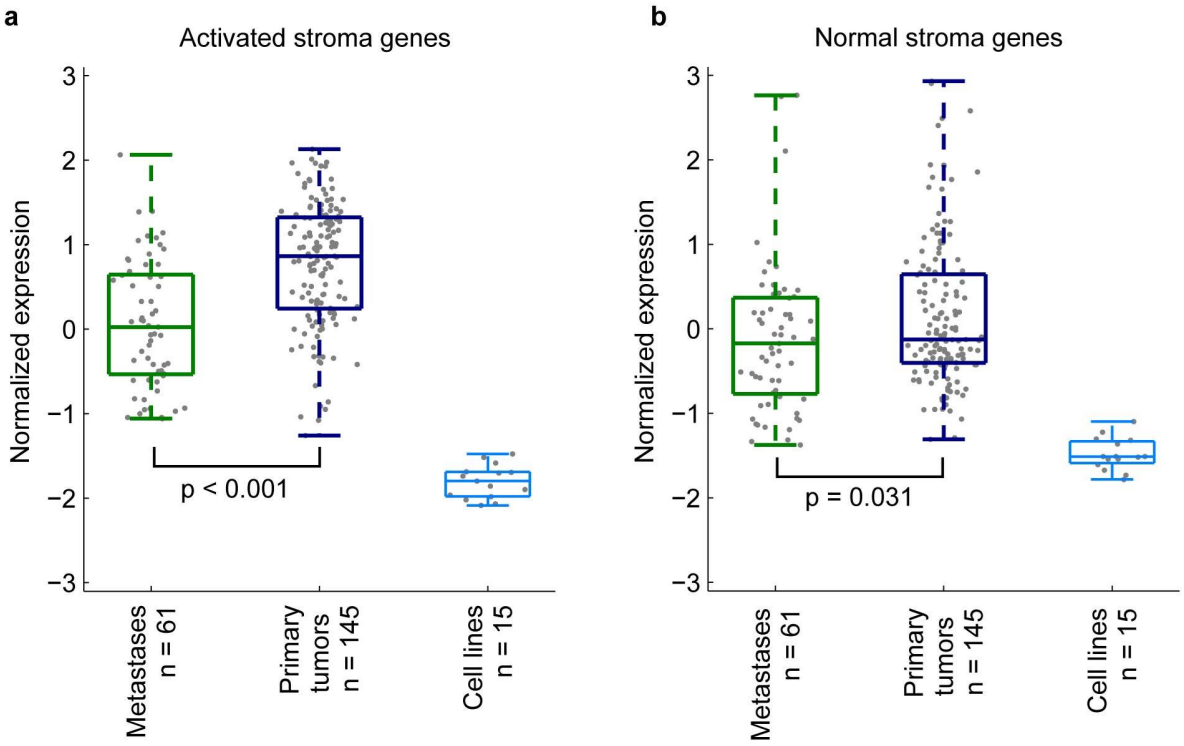


Supplementary Figure 3

Correlation of pathology assessments of tumor with factor weights in normal pancreas and primary tumors.

Horizontal axes all show tumor cellularity, while vertical axes show factor weight. Red dashed lines show best linear fits. *P* values are given for each R^2.

Nature Genetics: doi:10.1038/ng.3398

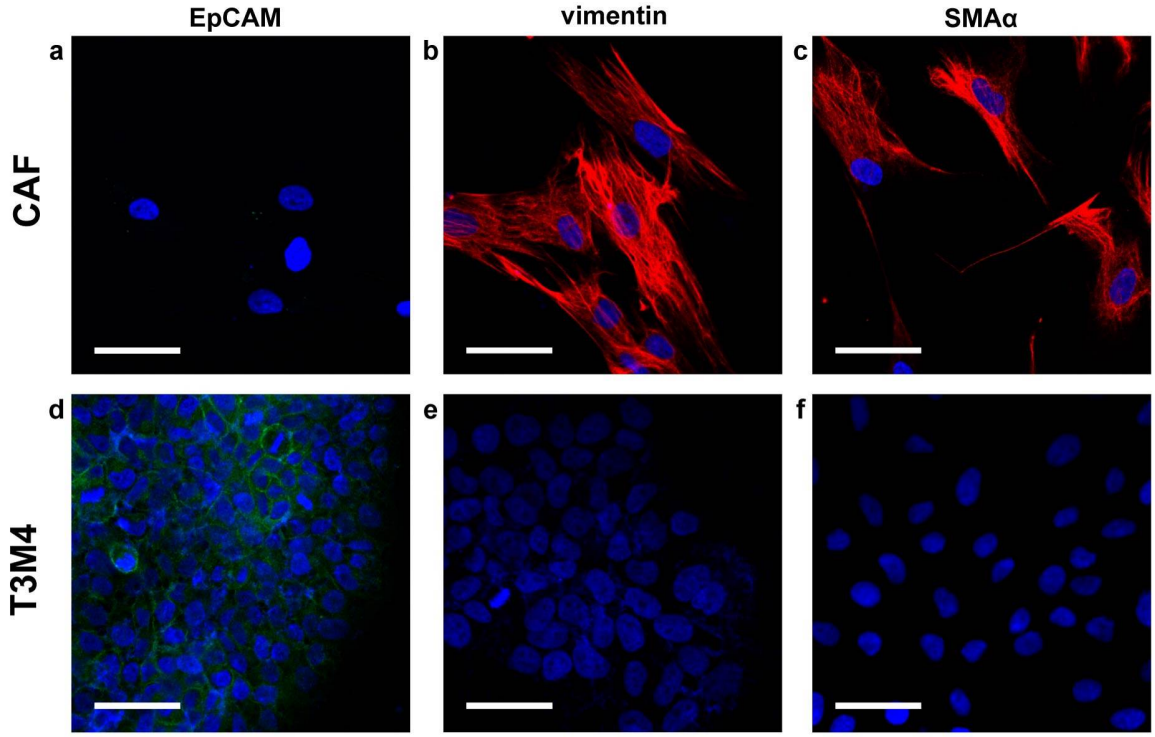


Supplementary Figure 4

Primary tumors have higher expression of stroma genes than metastatic samples.

(**a**) Normalized average gene expression of the 25 exemplar activated stroma genes is higher in primary tumor samples than in metastases ($P < 0.001$). (**b**) The average expression of the 25 exemplar normal stroma genes is slightly higher in primary tumors than in metastases ($P = 0.031$).

Nature Genetics: doi:10.1038/ng.3398

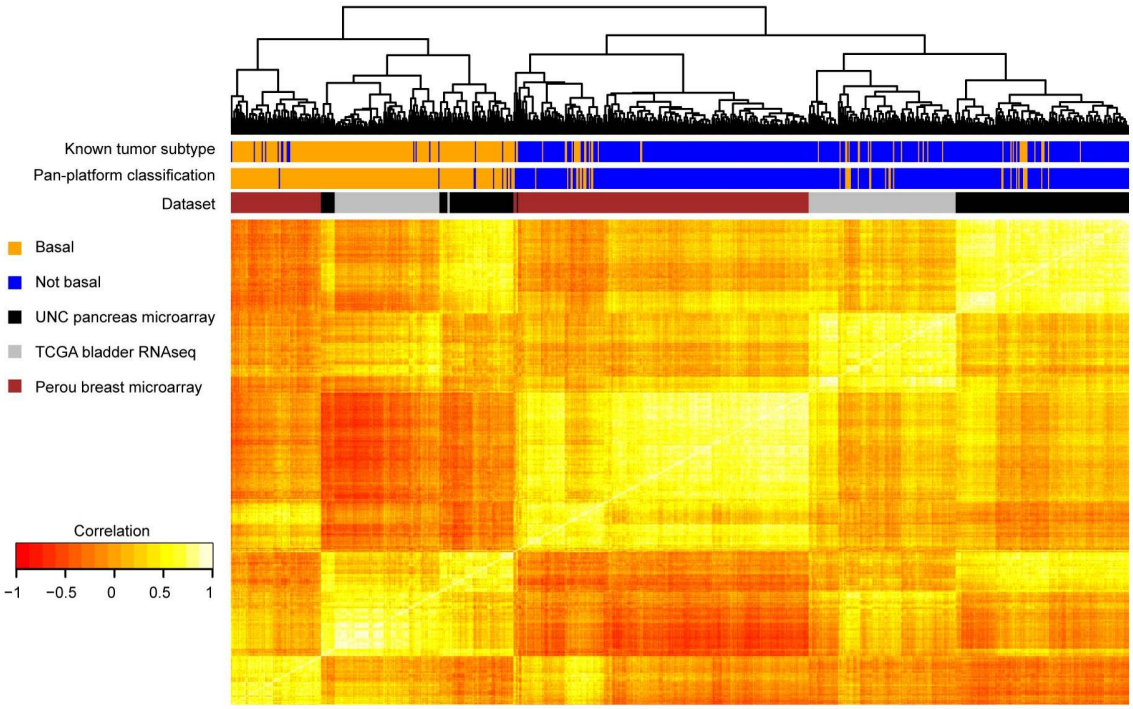


Supplementary Figure 5

Immunofluorescence staining of cancer-associated fibroblasts (CAFs).

(**a**) EpCAM, (**b**) vimentin and (**c**) SMAα. Staining of T3M4 pancreatic cancer cells as a (**d**) positive control for EpCAM, (**e**) negative control for vimentin and (**f**) negative control for SMAα. Scale bars, 50 μm.

Nature Genetics: doi:10.1038/ng.3398

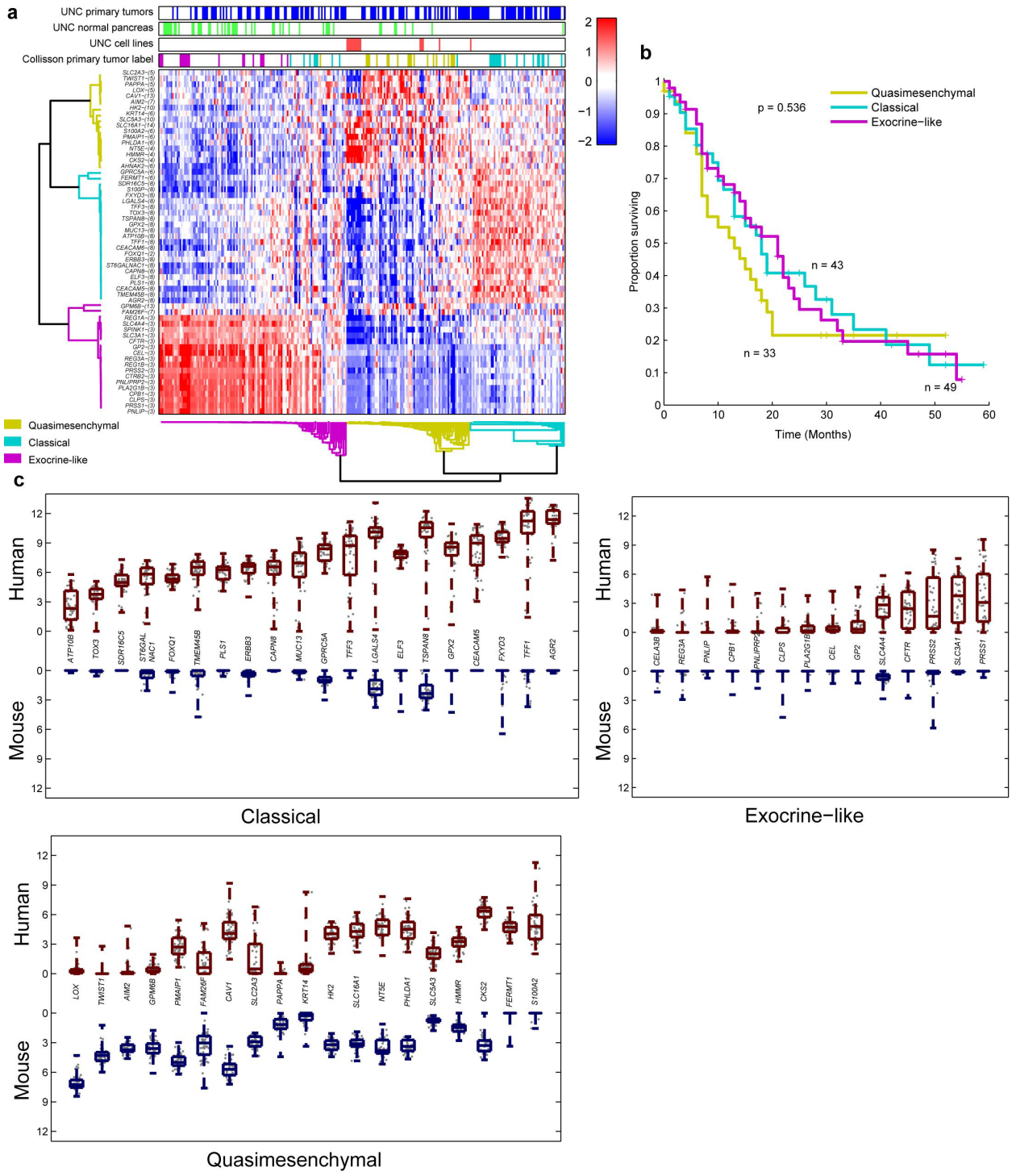


Supplementary Figure 6

Hierarchical clustering of the Spearman correlation of samples from the UNC, TCGA bladder cancer and Perou data sets shows similarities among basal-like samples.

The color bars above the heat map show subtype, either from the original publication (known tumor subtype) or from our cross-platform classifier (pan-platform classification).

Nature Genetics: doi:10.1038/ng.3398



Supplementary Figure 7

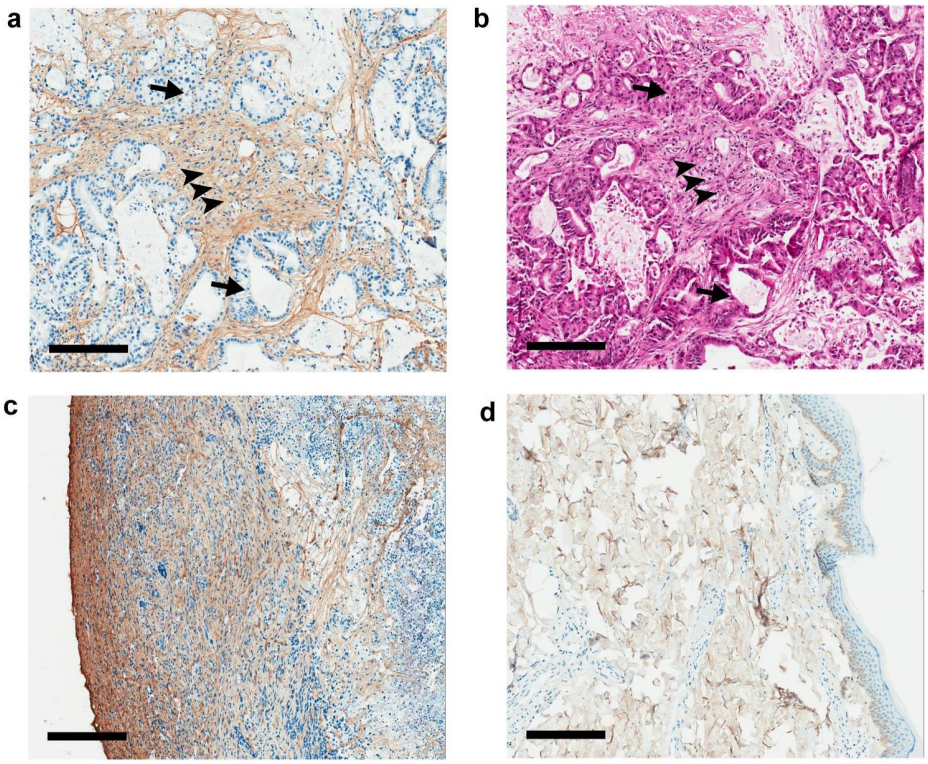
Comparison to the subtypes from Collisson *et al.*

(**a**) Consensus-clustered heat map of normalized data from UNC and Collisson *et al.* using the gene sets from Collisson *et al.* Primary tumors, normal pancreas and cell lines are shown. Samples from Collisson *et al.* were previously classified as exocrine-like (magenta), classical (cyan) and quasimesenchymal (yellow). (**b**) Kaplan-Meier plots of UNC samples classified by PAM into the subtypes from

Nature Genetics: doi:10.1038/ng.3398

Collisson *et al.* (**c**) Mouse- and human-specific gene expression of the gene lists from Collisson *et al.* in PDXs shown in $\log_2$ (1 + RPKM). Genes from the classical subtype are expressed by tumor cells, genes from the quasimesenchymal subtype are expressed by a mixture of human and mouse cells, and genes from the exocrine-like subtype are weakly expressed throughout.

Nature Genetics: doi:10.1038/ng.3398

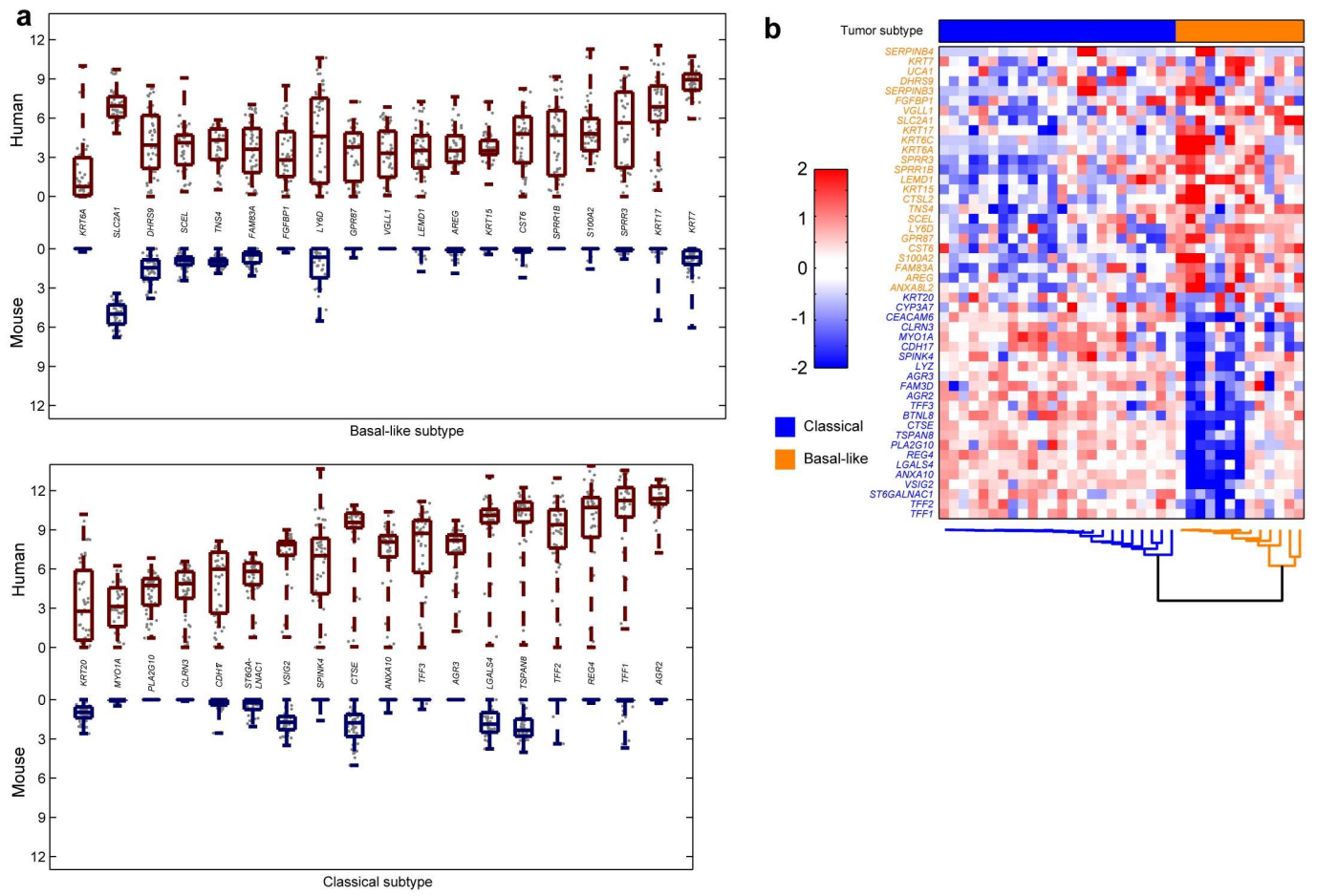


Supplementary Figure 8

Collagen I staining of mouse stroma in PDX tumors.

(**a**) Staining for mouse collagen I of stroma in a representative PDX tumor. (**b**) Corresponding H&E staining of an adjacent section. Staining for mouse collagen I of (**c**) mouse skin and (**d**) human skin. Black arrowheads show areas of tumor stroma. Black arrows show areas of tumor. Scale bars, 200 μm.

Nature Genetics: doi:10.1038/ng.3398

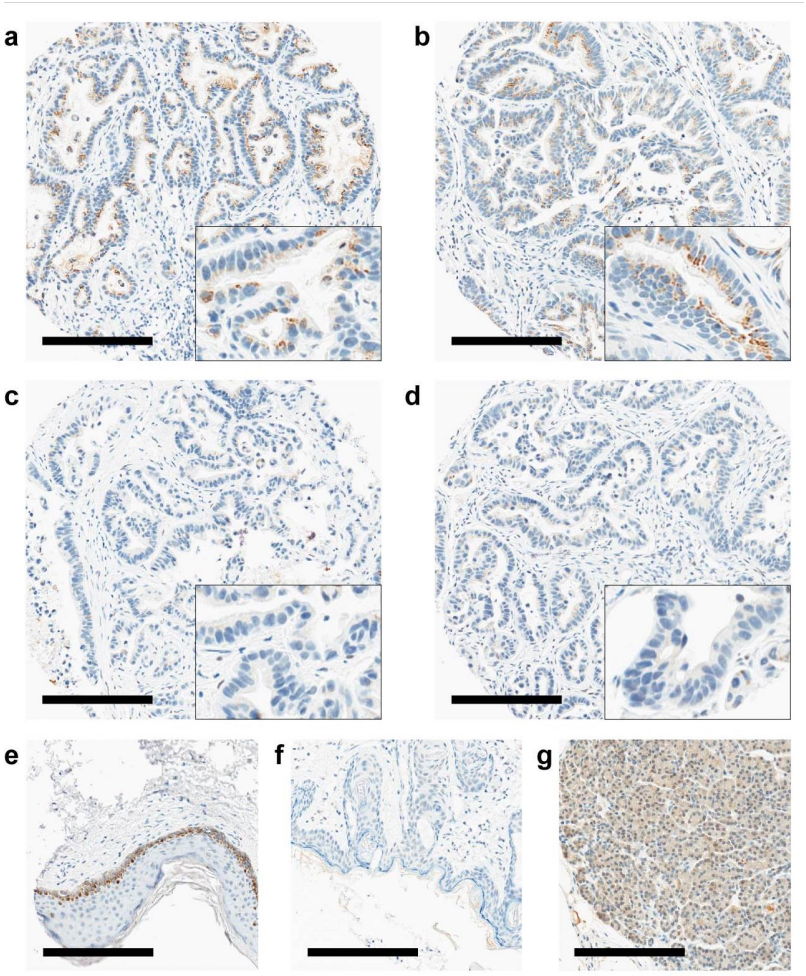


Supplementary Figure 9

Tumor gene expression in PDX models.

(**a**) Mouse- and human-specific gene expression of gene lists from the basal-like and classical subtypes in 37 PDX tumors shown in $\log_2$ (1 + RPKM). Both gene sets are robustly expressed by the human (tumor) but not the mouse (stroma) cells in PDX samples. (**b**) Consensus clustering of these PDX tumors using the gene lists from the basal-like and classical subtypes divides the samples into two groups.

Nature Genetics: doi:10.1038/ng.3398

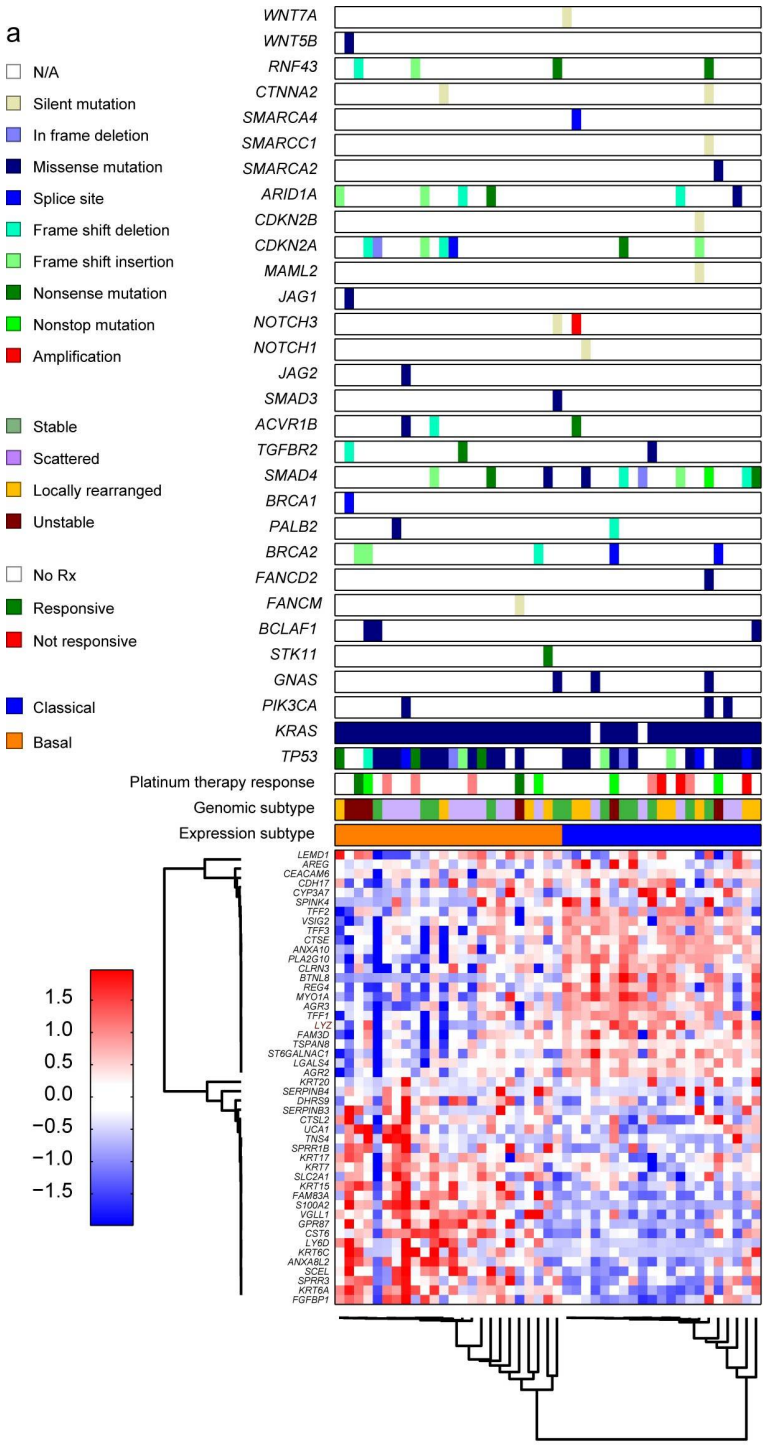


Supplementary Figure 10

SMAD4 staining of representative patient and matched PDX tumors.

Positive SMAD4 staining of a (**a**) patient adenocarcinoma and (**b**) corresponding PDX at passage 4. SMAD4 loss in a (**c**) patient adenocarcinoma and (**d**) corresponding PDX at passage 2. SMAD4 staining of control tissues: (**e**) human skin, (**f**) mouse skin and (**g**) human normal pancreas. Scale bars, 200 μm.

Nature Genetics: doi:10.1038/ng.3398

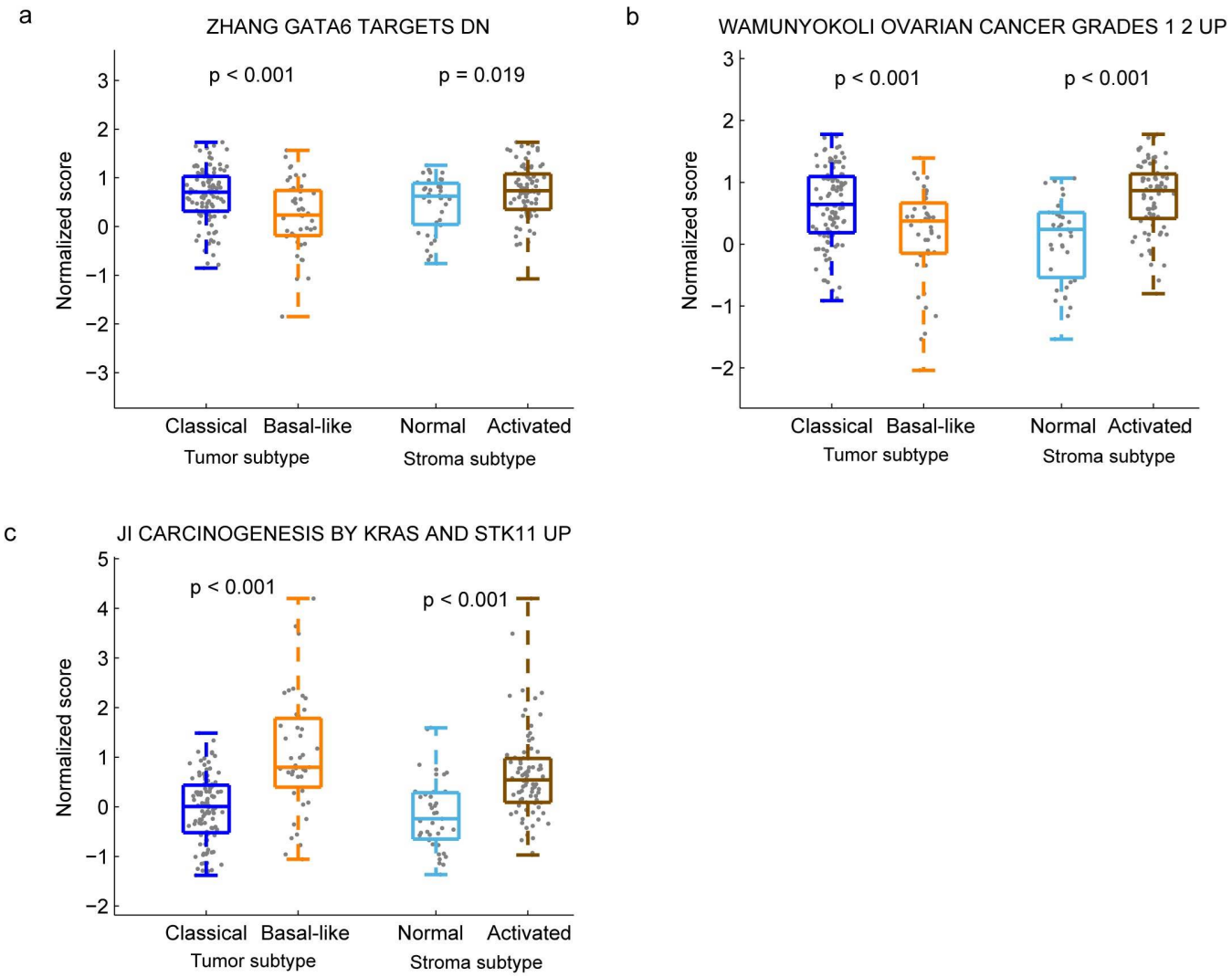


Supplementary Figure 11

Consensus-clustered heat map of ICGC data for which genetic information was available.

The color bars above the heat map show the subtypes and genetic alterations for key genes in PDAC. The heat maps show the *z*-normalized gene expression of genes from the basal-like and classical tumor subtypes.

Nature Genetics: doi:10.1038/ng.3398

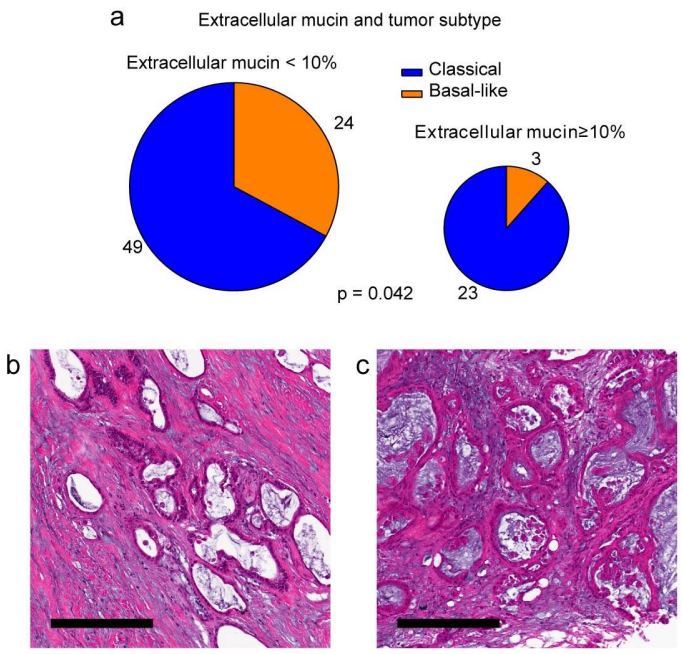


Supplementary Figure 12

Gene signature scores by subtype.

Scores are normalized across the cohort and are calculated as the mean expression across a panel of genes obtained from MSigDB. (**a**) The basal-like subtype shows downregulation of GATA6. (**b**) Tumors from the classical subtype are enriched in genes associated with mucinous ovarian cancer. (**c**) Tumors from the basal-like subtype are enriched in genes related to KRAS activation and STK11 loss.

Nature Genetics: doi:10.1038/ng.3398



Supplementary Figure 13

Differences in extracellular mucin in tumors from the classical and basal-like subtypes.

(**a**) Number of samples with low (<10%) compared to high (≥10%) extracellular mucin content. (**b**) Representative H&E staining of samples with a low degree and (**c**) high degree of extracellular mucin content. Scale bars, 200 μm.

Nature Genetics: doi:10.1038/ng.3398

Supplementary Note

Construction of a cross-platform basal-like classifier

Having established a method for classifying cohorts of PDAC expression data into basal-like and classical samples, we set out to develop a more clinically applicable classification scheme that works on single samples. Such a single-sample classifier would be valuable in a clinical setting, where access to a large cohort of comparative cases is prohibitive. Furthermore, we assessed the ability of such a classifier to work across gene expression platforms and across relevant cancer types. Thus, we developed and tested a platform-independent classifier to discriminate between "basal-like" samples versus others across various cancers, given a sample's individual gene expression profile.

Rank-based classifiers such as the Top Scoring Pair (TSP) [1] and kTSP [2] depend only on the relative ranks of the expression of genes within a sample, allowing such classifiers to be robust against platform-specific effects and study-to-study variations due to data normalization and preprocessing. [3]. Briefly, the kTSP approach selects *k* pairs of genes A and B such that gene A expression > gene B expression implies sample membership to class 1, otherwise implying membership to class 2. The default decision rule in [4] following feature selection weights each TSP equally in their class prediction ("voting"), despite the fact that some TSPs may better discriminate between classes than others. We extend the kTSP approach of [4] by implementing a custom decision rule that inputs the selected *k* gene pairs into a penalized logistic regression classifier to estimate the relative contribution each of the *k* selected TSPs in predicting class membership (defined here as basal-like versus otherwise), similar to [5]. In fitting the model, class membership is the binary outcome variable, and each covariate corresponds to a TSP, consisting of a binary integer vector which takes on the value of 1 for a sample if gene A > gene B in expression for that TSP, and 0 otherwise for each sample. We fit a penalized logistic regression model using the `ncvreg` package [6] to account for potential correlation between TSPs (ridge penalty) and to remove TSPs unhelpful in prediction given the presence of other features in the model (MCP penalty). Given the fitted model and a new sample's expression profile, we can obtain a predicted probability of basal-like class membership.

To build our classifier to predict the basal-like class across various cancers, we train our classifier on a "metadataset" consisting of the the TCGA Bladder (RNA-seq, 20533 genes), UNC Pancreas (Microarray, 19749 genes), and Perou Breast Cancer (Microarray, 17631 genes) datasets, totaling 788 samples. We reduce each dataset to a common set of genes found across each study to the described 50 gene signature in the manuscript. We further filtered the Perou Breast Cancer dataset to remove genes that had missing values for more than 10 samples, leaving 11526 genes. The remaining missing data was imputed using the `impute` package [7] in R using default parameters. Only 29 of the 50 genes from our original gene signature

Nature Genetics: doi:10.1038/ng.3398

remained for feature selection after filtering. Because of this small number, we also utilized a larger 500 gene set encompassing the original 50 gene set, which was derived in a similar fashion. From this larger gene set, 302 genes were found across all three training datasets.

We identified basal-like samples in the TCGA bladder and Perou Breast Cancer datasets from their associated clinical annotation files, and in the UNC Pancreas data we utilized the basal-like clustering calls from our manuscript. Given the known classes (basal-like versus otherwise) and gene expression profiles in each dataset, we perform our feature selection using the `switchBox` package [4] to select the *k* TSPs from the 302 candidate genes, resulting in 16 TSPs being selected. We applied the `ncvreg` function from [6] using the MCP penalty and an alpha parameter of 0.5, allowing for equal contribution of the ridge penalty to account for correlation between TSPs and the MCP penalty for feature selection. The appropriate penalty was chosen via leave-one-out cross validation using the `cv.ncreg` function (788 folds).

We find our final model to contain 14 TSPs when derived from our larger 500 gene signature. The fitted estimates can be found in Supplementary Table 3. Calculating the pair-wise spearman correlation between samples across the classifier's genes, we find that samples from the basal-like state (orange) tend to cluster together in terms of similarity, (Supplementary Fig. S6). We also find that our predictions tended to match the known classes for each sample regardless of platform or tumor type.

To validate our 14 TSP classifier, we applied our model to two independent datasets: TCGA Breast Cancer (RNAseq), and ICGC pancreas cancer data (Microarray). We find that the predictions matched well in the independent TCGA dataset, demonstrating a 92.3% classification accuracy. The only validation dataset that did not have existing subtype calls is the ICGC pancreas dataset. We find that our TSP predictions did not match as well with our clustering results, with a match rate between clustering-based calls and classifier prediction of 85.5%. We also found that spearman correlation of gene expression as a whole was much worse between the ICGC platform and any of our various RNAseq or Agilent Microarray data. In all, we demonstrate excellent within-training set performance of our classifier across multiple platforms in addition to accurate prediction of our classifier in an independent RNAseq data set.

Nature Genetics: doi:10.1038/ng.3398

Supplementary Table 3. Pan-platform basal-like classifier

Gene A	Gene B	Coefficient	Estimated Increase in odds of basal class membership when A > B
CD109	GPR160	0.87	2.38
SLC2A1	AGR2	1.22	3.39
KRT16	SLC44A4	0.52	1.68
CTSL2	TMEM45B	1.43	4.17
KRT6A	BCAS1	0.70	2.01
B3GNT5	VSIG2	0.41	1.51
MET	TFF3	0.72	2.06
CHST6	PLA2G10	0.80	2.24
SERPINB5	HPGD	0.76	2.13
DCBLD2	PLS1	1.40	4.07
IL20RB	FAM3D	1.33	3.79
PPP1R14C	SYTL2	1.58	4.85
NAB1	PLEKHA6	0.41	1.50
MSLN	CAPN9	1.58	4.83
	(Intercept)	-7.16	

References

[1] Leek, Jeffrey T. "The tspair package for finding top scoring pair classifiers in R."*Bioinformatics* 25.9 (2009): 1203-1204.

[2] Afsari, Bahman, Ulisses M. Braga-Neto, and Donald Geman. "Rank discriminants for predicting phenotypes from RNA expression." *The Annals of Applied Statistics* 8.3 (2014): 1469-1491.

[3] Patil, Prasad, et al. "Test set bias affects reproducibility of gene signatures."*Bioinformatics* (2015): btv157.

[4] Afsari, Bahman, et al. "switchBox: an R package for k–Top Scoring Pairs classifier development." *Bioinformatics* (2014): btu622.

[5] Shi, Ping, et al. "Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction." *Bmc Bioinformatics* 12.1 (2011): 375.

Nature Genetics: doi:10.1038/ng.3398

[6] Breheny, P. and Huang, J. (2011) Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Ann. Appl. Statist., 5: 232-253

[7] Trevor Hastie, Robert Tibshirani, Balasubramanian Narasimhan and Gilbert Chu . impute: impute: Imputation for microarray data. R package version 1.40.0

Nature Genetics: doi:10.1038/ng.3398