Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma

Richard A Moffitt1, Raoud Marayati1, Elizabeth L Flate1, Keith E Volmar2, S Gabriela Herrera Loeza1, Katherine A Hoadley1,3, Naim U Rashid1, Lindsay A Williams1,4, Samuel C Eaton5, Alexander H Chung5, Jadwiga K Smyla1, Judy M Anderson6, Hong Jin Kim1,7, David J Bentrem8,9, Mark S Talamonti10, Christine A Iacobuzio-Donahue11, Michael A Hollingsworth6 & Jen Jen Yeh1,5,7

Pancreatic ductal adenocarcinoma (PDAC) remains a lethal disease with a 5-year survival rate of 4%. A key hallmark of PDAC is extensive stromal involvement, which makes capturing precise tumor-specific molecular information difficult. Here we have overcome this problem by applying blind source separation to a diverse collection of PDAC gene expression microarray data, including data from primary tumor, metastatic and normal samples. By digitally separating tumor, stromal and normal gene expression, we have identified and validated two tumor subtypes, including a ‘basal-like’ subtype that has worse outcome and is molecularly similar to basal tumors in bladder and breast cancers. Furthermore, we define ‘normal’ and ‘activated’ stromal subtypes, which are independently prognostic. Our results provide new insights into the molecular composition of PDAC, which may be used to tailor therapies or provide decision support in a clinical setting where the choice and timing of therapies are critical.

Rigorous sequencing studies have shown that few genetic alterations (occurring in KRAS, CDKN2A, SMAD4 and TPS3) are prevalent in PDAC1–3, but these and other analyses of PDAC tumors are hampered by limited tumor cellularity and the presence of abundant stroma intermixed with normal endocrine and exocrine cells. Additionally, metastatic samples often include cell types from the host organ. Thus, PDAC tumors are complex mixtures in which malignant epithelial cells often represent a minority of the tissue compartment (Fig. 1a). Illustrating this limitation, an important study of the exome copy number in pancreatic cancer removed 43 of 142 patients because of low tumor cellularity affecting the sensitivity of mutation detection1. Although some studies use microdissection to improve tumor content4–7, using microdissection for precision medicine approaches is not yet feasible. When considering the gene expression of bulk tumor samples, normal pancreas and PDAC tissues often cluster together, separate from cell lines, which are assumed to be purely neoplastic8.

Separating the molecular signatures of tissue compartments from measurements of bulk tumor samples belongs to the general class of problems called blind source separation. Previous studies have used chronic pancreatitis samples, which are often fibrotic, to control for the presence of desmoplastic stroma in tumor samples9. In prostate cancer, Stuart et al. have used pathology assessments of cell types to train models of gene expression signatures for tumor, stroma and normal tissue10. In a follow-up study, they used their learned gene lists for the in silico estimation of tissue components in a larger data set11. A similar approach has also been used to quantify stromal content across multiple data sets from The Cancer Genome Atlas (TCGA)12. Among source separation techniques, non-negative matrix factorization (NMF) is especially well suited for biological data because it constrains all sources to be positive in nature, reflecting the goal of identifying positive gene expression exemplars rather than pairwise differences between tissue types. Briefly, we define NMF as modeling the matrix $X$ of expression for $g$ genes and $s$ samples, constituting the product of a matrix $G$ of $g$ gene weights for $k$ factors and a matrix $S$ of $s$ sample weights for $k$ factors. Alexandrov et al. have recently demonstrated that NMF is useful for a similar problem of identifying mutational signatures from the aggregate list of somatic mutations in human cancer samples13,14. Similarly, Biton et al. have applied a related technique, independent component analysis, to examine gene expression in bladder cancer15.

1Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, North Carolina, USA. 2University of North Carolina–Rex Healthcare, Chapel Hill, North Carolina, USA. 3Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, USA. 4Department of Epidemiology, University of North Carolina, Chapel Hill, North Carolina, USA. 5Department of Pharmacology, University of North Carolina, Chapel Hill, North Carolina, USA. 6Eppley Cancer Institute, University of Nebraska, Lincoln, Nebraska, USA. 7Department of Surgery, University of North Carolina, Chapel Hill, North Carolina, USA. 8Department of Surgery, Feinberg School of Medicine Northwestern University, Chicago, Illinois, USA. 9Robert H. Lurie Comprehensive Cancer Center, Feinberg School of Medicine Northwestern University, Chicago, Illinois, USA. 10Department of Surgery, NorthShore University HealthSystem, Evanston, Illinois, USA. 11Department of Pathology, David Rubenstein Center for Pancreatic Cancer Research, Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, New York, USA. Correspondence should be addressed to J.J.Y. (jen_jen_yeh@med.unc.edu).

Received 2 June; accepted 17 August; published online 7 September 2015; doi:10.1038/ng.3398
In this study, we have overcome the challenges of bulk tumor analysis, wherein signal is averaged out between normal, tumor and stromal compartments, by using NMF to perform virtual microdissection of primary and metastatic PDAC samples. This has allowed us to identify tumor-specific and stroma-specific subtypes with prognostic and biological relevance. In addition, by focusing on tumor-autonomous gene expression, we found that within-patient tumor heterogeneity between primary and metastatic sites was unexpectedly low.

**RESULTS**

**Virtual microdissection of PDAC**

We used NMF to analyze gene expression in a collection of microarray data from 145 primary and 61 metastatic PDAC tumors, 17 cell lines, and 46 pancreas and 88 distant site adjacent normal samples generated using Agilent human whole-genome 4x44K DNA microarrays (106 primary tumors were previously used in a separate bulk analysis of gene expression (Gene Expression Omnibus (GEO), GSE21501; ref. 16).

---

**Figure 1** Successful deconvolution of normal tissue with NMF. (a) Schematic depicting the major cell types in primary tumor and liver metastatic samples. (b) Five panels are shown corresponding to different factors. The top of each panel shows the overlap of sample types (solid colors) with factor weights, and the bottom of each panel shows a heat map of gene expression (grayscale) for exemplar genes in all tumors and adjacent normal tissues. Gene expression has been normalized by z score. (c) Box-and-whisker plots (showing the median, quartiles and range) comparing NMF factor weights across tissue types with corresponding t-test results. (d) Percent tumor cellularity plotted against NMF liver factor weight and NMF basal tumor factor weight for metastases to the liver and adjacent liver samples. Linear regression lines are shown in red along with corresponding statistics.
To validate our findings, we performed RNA sequencing (RNA-seq) on 15 primary tumors, 37 pancreatic cancer patient-derived xenografts (PDXs), 3 PDAC cell lines and 6 cancer-associated fibroblast (CAF) lines derived from deidentified samples collected from patients with pancreatic cancer (Supplementary Data Set). The histology of all available samples was reviewed by a single pathologist blinded to sample identity (K.E.V.). The demographic and clinical characteristics of the patients with primary tumors available in our cohorts are summarized in Table 1.

### NMF distinguishes normal and tumor compartments

A key obstacle in the analysis of tumor gene expression data, particularly in PDAC, is the removal of confounding gene expression data for normal tissue and stroma collected at local and distant organ sites. Examples of histology for samples with tumor, stroma and normal tissue are shown in Supplementary Figure 1. We used NMF to identify gene expression patterns that we attributed to normal pancreas, liver, lung, muscle and immune tissues. Expression of exemplar genes from these factors (genes with distinctly large weights in a single column of G) as well as factor weights for the samples (rows of S) showed excellent agreement with known tissue labels (Fig. 1b,c and Supplementary Fig. 2). Investigation of the exemplar genes from these factors further confirmed the role of these factors as confounding normal tissue. For example, using the Kolmogorov-Smirnov test, the top weighted genes from the liver factor showed significant ($P < 1 \times 10^{-10}$) enrichment for the MSigDB term SU_LIVER (Supplementary Table 1), and the gene with the highest weighting, $FGB$ (encoding fibrinogen $\beta$) (Supplementary Table 2), was specifically expressed in normal human liver tissue.

In addition to factors corresponding to normal tissue from distant organs, we found two factors that were exclusive to pancreas tissue but were differentiated from each other by their respective gene lists. One factor described endocrine function, including expression of $GCG$ and $INS$ (encoding glucagon and insulin) (Supplementary Table 2), whereas the other factor described exocrine function, including expression of digestive enzyme genes such as $PNLIP$ (encoding pancreatic lipase) (Supplementary Table 2). This unsupervised discovery of two molecularly distinct yet highly colocalized factors related to normal pancreatic function represents an important proof of concept in the use of NMF to identify new features without predefined knowledge of expression.

To validate our expression signatures for normal tissue, all available samples were reviewed by a single pathologist (K.E.V.) to independently assess the tumor, stromal and normal cellularity. We found that many factor weights were correlated or anticorrelated with tumor cellularity (Supplementary Fig. 3). Among normal and metastatic liver samples, for example, weights for tumor-specific factors were correlated with tumor cellularity, whereas the weight for the factor corresponding to normal liver was inversely related to the tumor content of a sample (Fig. 1d). These findings support our hypothesis that the factor weights obtained from NMF are quantitatively indicative of underlying sample composition.

### Identification of stroma-specific subtypes

Stroma is particularly important in PDAC. According to pathology assessments, the stroma was varied (Supplementary Fig. 1c–e) and comprised on average 48% of our primary tumor samples (s.d. = 30%). Our analysis identified two factors that described gene expression from the stroma (Fig. 2), which were distinctly different from the factors for normal tissue shown in Figure 1. Consensus clustering on exemplar genes from these two stromal factors divided the tumor samples into two stromal subtypes, which we classified as ‘normal’ and ‘activated’ (Fig. 2a). Patients whose samples belonged to the activated stromal subtype had a worse median survival time of 15 months and 1-year survival rate of 60% when compared to patients whose samples belonged to the normal stromal subtype (median survival time of 24 months and 1-year survival rate of 82%; Fig. 2b). The expression signatures for both subtypes were notably absent in PDAC cell lines (Fig. 2c), which exhibited a distinct mitotic expression signature associated with mitotic checkpoints and DNA replication (Supplementary Table 1). The fact that the cell lines did not express the genes from these stromal factors and that many metastatic samples expressed them at low levels (Supplementary Fig. 4) suggests that the genes from these factors are not expressed by the tumor epithelium. To further validate the stromal origin of these gene expression signatures, we isolated six CAF lines from primary tumors (Supplementary Fig. 5) and found that they robustly overexpressed our hypothesized stromal signature genes in comparison to the PDAC tumor cell lines, which had no expression of the stromal signature genes (Fig. 2c).

The vast majority of collagen gene expression was attributable to stromal compartments, with the lone exception being $COL17A1$, which was highly expressed in tumors. Normal stroma was characterized by relatively high expression of known markers for pancreatic stellate cells, $ACTA2$, $VIM$ and $DES$ (encoding smooth muscle actin, vimentin and desmin). Stellate cells

---

**Table 1** Demographics and univariate Cox analysis

|                | All patients | Resected patients with survival data | Univariate Cox $P$ value | Microarray primary samples | RNA-seq primary samples | RNA-seq PDX samples |
|----------------|--------------|-------------------------------------|--------------------------|---------------------------|-------------------------|----------------------|
| **Race**       |              |                                     |                          |                           |                         |                      |
| Caucasian      | 128          | 121                                 | 0.507                    | 99                        | 9                       | 25                   |
| African American | 23          | 18                                  | 0.333                    | 10                        | 3                       | 8                    |
| Other          | 8            | 7                                   | 0.821                    | 5                         | 0                       | 3                    |
| **Sex**        |              |                                     |                          |                           |                         |                      |
| Female         | 90           | 83                                  | 0.348                    | 67                        | 5                       | 23                   |
| Male           | 80           | 68                                  | 0.348                    | 55                        | 8                       | 14                   |
| **T stage**    |              |                                     |                          |                           |                         |                      |
| T1             | 4            | 4                                   | 0.420                    | 2                         | 1                       | 2                    |
| T2             | 22           | 20                                  | 0.530                    | 20                        | 2                       | 5                    |
| T3             | 131          | 122                                 | 0.743                    | 91                        | 9                       | 28                   |
| T4             | 1            | 1                                   | 0.115                    | 1                         | 0                       | 0                    |
| **N stage**    |              |                                     |                          |                           |                         |                      |
| N0             | 49           | 43                                  | 0.068                    | 36                        | 7                       | 10                   |
| N1             | 112          | 106                                 | 0.068                    | 80                        | 5                       | 25                   |
| **M stage**    |              |                                     |                          |                           |                         |                      |
| M0             | 160          | 149                                 | 0.068                    | 129                       | 12                      | 35                   |
| M1             | 15           | 15                                  | 0.068                    | 14                        | 0                       | 1                    |
| **Adjuvant therapy** | | | | | | |
| Yes            | 74           | 70                                  | 0.055                    | 44                        | 5                       | 21                   |
| No             | 30           | 28                                  | 0.055                    | 27                        | 3                       | 7                    |
| **Differentiation** | | | | | | |
| Well           | 16           | 13                                  | 0.940                    | 16                        | 0                       | 1                    |
| Moderate       | 49           | 47                                  | 0.398                    | 49                        | 1                       | 3                    |
| Poor           | 34           | 31                                  | 0.407                    | 34                        | 1                       | 2                    |
| **PDX**        |              |                                     |                          |                           |                         |                      |
| Graft success  | 44           | 37                                  | 0.164                    | 11                        | 8                       | 37                   |
| Graft failure  | 18           | 12                                  | 0.164                    | 9                         | 3                       | 0                    |
| **Margin**     |              |                                     |                          |                           |                         |                      |
| Positive       | 58           | 52                                  | 0.026                    | 34                        | 5                       | 17                   |
| Negative       | 93           | 88                                  | 0.026                    | 75                        | 7                       | 17                   |
| **Total**      | 193          | 163                                 | 0.026                    | 143                       | 15                      | 37                   |
Figure 2 The dual action of stroma is described by distinct gene expression patterns, which are not present in PDAC cell lines. (a) Consensus-clustered heat map of University of North Carolina (UNC) primary tumor samples, metastases and cell lines generated using genes from stromal factors. Samples clustered into three groups, describing samples with activated stroma, samples with normal stroma and samples with low or absent stromal gene expression. (b) Kaplan-Meier survival analysis of patients with resected PDAC from the activated and normal stromal clusters shows that samples in the activated stroma group have worse prognosis, with a hazard ratio of 1.94 (CI = 1.11–3.37, \(P = 0.019\)). (c) Stromal signature genes are overexpressed in CAFs as compared to tumor cell lines. (d) Genes from both stromal signatures are specifically overexpressed by the mouse stroma in PDX tumors and are not expressed by the human tumor cells.

**Activated stroma genes**

| Gene Name |
|----------------|
| SPARC |
| COL1A2 |
| COL1A7 |
| VCAN |
| CD31 |
| CDH11 |
| CTHRC1 |
| COL7A1 |
| COMP |
| FN1 |
| COL1A2 |
| COL3A1 |
| POSTN |
| MMP11 |
| INHBA |
| SYNM |
| GREM1 |
| FNDC1 |
| SULF1 |
| FAP |
| LUM |
| MMP11 |
| POSTN |
| SFRP2 |
| CDH11 |
| COMP |
| FN1 |
| COL11A1 |
| ITGA11 |
| MMP11 |
| INHBA |
| VCAN |
| ADAMTS1 |
| GPM6B |
| ANGPTL7 |
| CDH19 |
| ABCA8 |

**Normal stroma genes**

| Gene Name |
|----------------|
| FABP4 |
| ACTG2 |
| PLP1 |
| RBPMS2 |
| RERGL |
| RERGL |
| CDH19 |
| SCRG1 |
| VIT |
| LPHN3 |
| SYNM |
| GREM1 |
| COMP |
| ADAMTS1 |
| ID4 |
| ID4 |
| CDH11 |
| CTHRC1 |
| FAP |
| LUM |
| MMP11 |
| POSTN |
| SFRP2 |
| SPARC |
| COL1A2 |
| COL3A1 |
| POSTN |
| MMP11 |
| SULF1 |
| THBS2 |
| FN1 |
| COL11A1 |
| ITGA11 |
| MMP11 |
| INHBA |
| VCAN |
| ADAMTS1 |
| GPM6B |
| ANGPTL7 |
| CDH19 |
| SCRG1 |
| VIT |
| LPHN3 |
| SYNM |
| GREM1 |
| COMP |
| ADAMTS1 |
| ID4 |
| ID4 |
| CDH11 |
| CTHRC1 |
| FAP |
| LUM |
| MMP11 |
| POSTN |
| SFRP2 |
| SPARC |

**Mouse expression**

- **Human expression**
have been shown to promote cancer cell survival in vitro but also may restrain PDAC in mouse models. Targeting desmoplastic stroma by Hedgehog pathway inhibition has been shown to both accelerate the development of disease and enhance the delivery of chemotherapy. In patients, the ratio of the area stained for smooth muscle actin to the area stained for collagen has been shown to be predictive of poor outcomes. We found that activated stroma was characterized by a more diverse set of genes associated with macrophages, such as the integrin ITGAM and the chemokine ligands CCL13 and CCL18. Activated stroma also expressed other genes associated with macrophages, such as the integrin ITGAM and the chemokine ligands CCL13 and CCL18.

Figure 3 Tumor-specific gene expression suggests two subtypes of PDAC with similarities to other tumor types. (a) Consensus-clustered heat map of primary tumors, metastatic tumors and cell line models of PDAC generated using correlation, with the underlying distance function showing two subtypes of PDAC. (b) Kaplan-Meier survival analysis of patients with resected primary tumors from each tumor subtype in the ICGC cohort. (c) Consensus-clustered heat map of tumors in the TCGA bladder cancer data set from Prat et al. with tumors split on the basis of basal factor gene expression (n = 72 basal-like and n = 223 not basal) strongly agrees with the previously published division of samples into basal and non-basal subtypes. (h) In survival analysis, basal-like breast cancer, as defined by our labeling, had a hazard ratio of 3.52 (95% CI = 1.94–6.38, P < 0.001).
that point to its role in tumor promotion, including the secreted protein SPARC, the Wnt family members WNT2 and WNT5A, MMP9 (gelatinase B) and MMP11 (stromelysin 3). The presence of FAP (encoding fibroblast activation protein), which has previously been related to worse prognosis, in the gene signature for activated stroma suggests that an activated fibroblast state may be partially responsible for the poor outcomes for patients with this stromal subtype24. This led us to hypothesize that the normal stromal factor might describe a ‘good’ version of stroma, whereas the activated stromal factor might describe an ‘activated’ inflammatory stromal response that has been seen in previous studies to be responsible for disease progression25–27. Our factor analysis supports a complex, multigene model of stroma in PDAC, which may explain why single-gene analysis has yielded mixed results.

**Identification of tumor-specific subtypes**

We found that two tumor-specific factors, independent of normal stromal factors, defined ‘classical’ and ‘basal-like’ subtypes of PDAC. When our samples were split into the two tumor subtypes (Fig. 3a), patients with tumors in the basal-like subtype had an overall significantly (P = 0.007) worse median survival time of 11 months and 1-year survival rate of 44% in comparison to a survival time of 19 months and 1-year survival rate of 70% for patients with tumors in the classical subtype (Fig. 3b). All cell lines assayed in this study

---

**Figure 4 Multivariate survival analysis of tumor and stromal subtypes.** (a) Heat map of tumor samples using 25 genes from each of the tumor and stromal factors, with samples sorted horizontally by classification. Signature scores for selected gene sets appear above for each sample. (b) Combined Kaplan-Meier survival analysis of resected primary tumors from patients with basal-like or classical tumors and normal or activated stroma showing differential survival (P < 0.001, log-rank test). Differential prognosis among the subtypes shows complementarity. Classical tumors with normal stroma (n = 24) had the lowest hazard ratio of 0.39 (95% CI = 0.21–0.73), whereas basal-like tumors with activated stroma (n = 26) had the highest hazard ratio of 2.28 (95% CI = 1.34–3.87). (c,d) Kaplan-Meier survival analysis shows that patients with classical tumors have less response to adjuvant therapy (hazard ratio = 0.76, 95% CI = 0.40–1.43) (c) than patients with basal-like tumors (hazard ratio = 0.38, 95% CI = 0.14–0.91) (d). (e) Kaplan-Meier survival analysis shows that African Americans have worse overall survival for both the basal-like and classical subtypes, with a hazard ratio of 2.28 (95% CI = 1.16–4.5).
Table 2 Summary of associations with clinical covariates and subtypes

| Covariate          | Tumor subtype | Fisher’s exact P value | Stromal subtype | Fisher’s exact P value |
|--------------------|---------------|------------------------|----------------|-----------------------|
|                    | Classical     | Basal-like             | Normal         | Activated             |
| Sex                |               |                        |                |                       |
| Male               | 50            | 16                     | 0.849          | 15                    | 36                     | 1                       |
| Female             | 64            | 12                     | 0.332          | 17                    | 43                     |
| Race               |               |                        |                |                       |
| Caucasian          | 90            | 19                     | 0.521          | 26                    | 65                     | 1                       |
| African American   | 13            | 2                      | 0.026          | 3                     | 7                      |
| T stage            |               |                        |                |                       |
| T2                 | 16            | 6                      | 0.59           | 5                     | 14                     | 1                       |
| T3                 | 85            | 27                     | 0.026          | 25                    | 59                     |
| N stage            |               |                        |                |                       |
| N0                 | 35            | 9                      | 0.532          | 11                    | 22                     | 0.649                   |
| N1                 | 72            | 25                     | 0.385          | 7                     | 22                     | 0.629                   |
| Margin             |               |                        |                |                       |
| Positive           | 38            | 8                      | 0.385          | 7                     | 22                     | 0.629                   |
| Negative           | 65            | 22                     | 0.072          | 10                    | 30                     | 0.769                   |
| Adjuvant therapy   |               |                        |                |                       |
| Yes                | 48            | 13                     | 0.437          | 10                    | 30                     | 0.769                   |
| No                 | 21            | 9                      | 0.037          | 5                     | 19                     |
| Differentiation    |               |                        |                |                       |
| Poor               | 23            | 11                     | 0.042          | 18                    | 43                     | 0.792                   |
| Well               | 49            | 16                     | 0.497          | 13                    | 44                     |
| Mucin              |               |                        |                |                       |
| Low                | 49            | 24                     | 0.042          | 18                    | 43                     | 0.792                   |
| High               | 23            | 3                      | 0.497          | 6                     | 19                     |
| Stroma             |               |                        |                |                       |
| Normal             | 31            | 8                      | 0.144          |                       |
| Activated          | 57            | 31                     |                |                       |

(P < 0.001), as well as a majority of the metastatic samples (P = 0.002), were classified as basal-like, suggesting that cell line models represent only one subset of PDAC. These subtypes as well as their prognostic value were independently validated in the recently published International Cancer Genome Consortium (ICGC) PDAC microarray data set (Fig. 3c,d)28. The expression of genes from the basal-like factor, including ones encoding laminins and keratins, was also consistent with basal subtypes previously defined in bladder29–31 and breast32 cancers (Fig. 3e–h). Interestingly, genes from our basal-like subtype reproduced subtype calls (P < 0.001) in breast cancer, had prognostic value in breast cancer samples (P < 0.001) and reproduced previous subtype calls in bladder cancer (TCGA bladder cancer)29 (P < 0.001). Given these promising results, we developed a single-sample cross-platform classifier of the basal-like subtype that was trained on our microarray data, the TCGA bladder cancer data and the breast cancer data from Prat et al.33, with a cross-validation accuracy of 93%, and was able to classify TCGA breast cancer data with 92% accuracy during external validation (Supplementary Fig. 6, Supplementary Table 3 and Supplementary Note).

Potential subtypes of PDAC have previously been described by Collisson et al.5. We used the published exemplar genes for the ‘exocrine-like’, ‘classical’ and ‘quasimesenchymal’ subtypes to cluster normal pancreas, cell lines and primary PDAC tumors from our cohort (Supplementary Fig. 7a). The three previous classifications were also observed in our data but did not hold prognostic power either by cluster label or by supervised classification with Prediction Analysis for Microarrays (PAM)34 (Supplementary Fig. 7b). Furthermore, inclusion of the subtypes from Collisson et al. in a multivariate Cox regression model with our proposed tumor subtypes did not remove the predictive power of our subtyping (P = 0.014). By cross-referencing the genes from the model of Collisson et al. with those from our NMF model, we observed three key findings. First, genes from the exocrine-like subtype overlapped with genes from our exocrine pancreas factor (17/17). Tumors in this cluster had expression indistinguishable from that of adjacent normal samples from our data set. Second, the genes from the classical subtype of Collisson et al. overlapped with the genes in our classical subtype (20/22), subtypes and clinical variables (sex, ancestry, T stage, N stage, margin status, adjuvant therapy, histological grade and age), both classifications were independently associated with survival (stroma subtypes, P = 0.037; tumor subtypes, P = 0.003).

Although tumors in the basal-like subtype had worse prognosis independent from race and stroma (Fig. 4b,e), patients with basal-like tumors showed a strong trend toward better response to adjuvant therapy (P = 0.072; Fig. 4d). Among patients with the basal-like subtype, adjuvant therapy provided a hazard ratio of 0.38 (95% CI = 0.14–1.09), whereas, in patients with tumors from the classical subtype, adjuvant therapy was associated with a hazard ratio of only 0.76 (95% CI = 0.40–1.43). In our cohort, there was no association of most clinical variables (ancestry, sex, T stage, N stage, differentiation or tumor cellularity) with survival, although positive nodal status trended toward significant association with survival and positive margin status was significantly (P = 0.026) associated with worse survival (Table 1). Two-way associations of all subtype calls with clinical and pathological information from our cohort of patients with PDAC is shown in Table 2. We found no association of tumor or stromal subtype with standard clinical or pathological variables, with the notable exception of mucinous features.

Tumor-specific subtypes in patient-derived xenografts

To assess the tumor or stromal specificity of our signatures, we performed RNA-seq on a group of 37 PDX tumors. PDX tumors are composed of human tumor cells surrounded by mouse stroma (Supplementary Fig. 8)35. Genes from both of our tumor signatures were expressed as human transcripts, whereas genes from both of our stromal signatures were expressed as mouse transcripts (Fig. 2d and Supplementary Fig. 9a). Furthermore, we found that PDX RNA-seq expression divided PDXs into tumors with classical and basal-like groupings (Supplementary Fig. 9b) that both predominantly expressed an activated stromal signature (Fig. 2d). We found that, whereas the tumor-specific subtype was not predictive of graft success (Fig. 5a), patient-derived tumors with the activated stromal subtype had significantly higher graft success rates than those with the normal stromal subtype or low amounts of stroma (Fig. 5b; P = 0.019). Tumors in the basal-like subtype also exhibited faster growth rates than tumors for which we retain the naming convention classical. Third, the gene set associated with the quasimesenchymal subtype appeared to be a mixed collection of genes from our basal-like tumor (6/20) and stromal (6/20) subtypes. Thus, the representation of stromal factors in the list of genes for the quasimesenchymal class in Collisson et al. may explain the apparent mesenchymal-like gene expression that was observed.

Basal-like and classical tumors were found in both the normal and activated stromal subtypes (Fig. 4a). Differential prognosis among tumor and stromal subtypes was cumulative, as tumors from the classical subtype with the normal stromal subtype (n = 24) had the lowest hazard ratio of 0.39 (95% confidence interval (CI) = 0.21–0.73), whereas tumors from the basal-like subtype with the activated stromal subtype (n = 26) had the highest hazard ratio of 2.28 (95% CI = 1.34–3.87) (Fig. 4b). In a multivariate Cox regression model that included tumor subtypes, stromal

© 2015 Nature America, Inc. All rights reserved.
in the classical subtype ($P = 0.032$), as measured by the length of time that tumors took to grow to 200 mm$^3$ ($TT200$; Fig. 5c,d). A previously used metric for PDX growth.36. Retrospective analysis of patients who had matched PDX tumors found that shorter $TT200$ was associated with unfavorable recurrence-free survival ($P = 0.035$; Fig. 5e), suggesting that PDX tumor growth rate may reflect patient biology.

We also measured both mouse- and human-specific expression of the genes from Collisson et al. in our PDX models. We found that, while genes from the classical subtype were expressed by human cells in PDXs, quasimesenchymal genes were expressed by a mixture of human and mouse cells and exocrine-like genes were infrequently expressed (Supplementary Fig. 7c). This finding supports our hypothesis that, whereas the classical subtype is a bona fide group, the quasimesenchymal subtype is partially driven by non-tumor contributions from stroma and the exocrine-like subtype corresponds to normal pancreas.

**KRAS codon mutations, tumor-specific subtypes and ancestry**

Studies of KRAS codon mutations have demonstrated that mutations of different codons may have distinct functions.37,38, and these mutations have been shown to be associated with differential outcome in some clinical studies. Because PDX tumors are enriched for human-specific tumor cells, we evaluated KRAS codon mutations in our PDX cohort using manually curated RNA-seq data. Although the overall frequency of KRAS codon mutations was similar to that in a recent study of PDAC7, we noted that the KRAS mutation encoding p.Gly12Asp was significantly over-represented in our classical subtype, whereas the mutation encoding p.Gly12Val was only found in the classical subtype ($P = 0.030$; Fig. 5f). Furthermore, we found an over-representation of KRAS mutations encoding p.Gly12Val in African Americans ($P < 0.001$; Fig. 5g). In contrast to basal-like breast cancers, which occur most frequently in African-American women and have a worse prognosis9, the tumors from African-American patients in our cohort tended to be mainly of the classical subtype (13 versus 2 tumors). We found that, as in other cancers, the prognosis in African Americans was worse after adjusting for tumor subtype ($P = 0.017$; Fig. 4e). African-American patients with tumors of the classical subtype had a median survival time of 13 months in comparison to Caucasian patients with tumors from the classical subtype who had a median survival time of 19 months.

**Other commonly mutated genes and altered pathways in PDAC**

Previously, loss of SMAD4 has been shown to promote tumor growth.40,41. We found that, as in previous PDX studies of PDAC, loss of SMAD4 was associated with faster growth rates in PDAC models ($P = 0.044$; Fig. 5h and Supplementary Fig. 10). Furthermore, in our PDX cohort, we found that SMAD4 expression was significantly higher in classical than in basal-like PDAC tumors ($P = 0.015$; Fig. 5i), consistent with the observation that SMAD4 loss confers a more aggressive phenotype.

Using mutation, genomic subtype3 and gene expression28 data from the publically available ICGC cohort in which we recapitulated our subtypes and their effects on prognosis, we evaluated significantly mutated genes and pathways in PDAC, including ones recently identified through whole-exome sequencing of microdissected primary PDAC tumors.43-45. We found no significant associations between our expression subtypes and these mutationally altered pathways, including the transforming growth factor (TGF)-β, RB, Notch, CTNNB1, SWI/SNF chromatin-remodeling and DNA repair pathways (Supplementary Fig. 11). Furthermore, we found no overlap between our subtypes and recently identified genomic subtypes or response groups for platinum therapy.46. Consistent with this analysis, a recent comprehensive study of somatic mutations in long-term survivors of PDAC suggested that somatic mutations alone will not be sufficient to explain clinical outcome.47.

Given the overlap of our classical subtype with that of Collisson et al.5, it was not surprising to find that our classical subtype was also enriched for genes associated with GATA6 overexpression44 (Fig. 4a and Supplementary Fig. 12a). GATA6 has been found to promote...
Figure 6 Overcoming limited tumor cellularity shows true heterogeneity among matched primary and metastatic sites. (a) Sample-sample correlations of matched primary and metastatic tumors using the 50 most differentially expressed genes across all samples causes samples to group by organ location. (b) Sample-sample correlations using 25 genes each from the classical and basal-like tumor gene lists caused samples to cluster instead by tumor subtype and patient of origin. (c) The correlation of samples within the same patient is higher when using the tumor-specific genes than when using the most differentially expressed genes. (d) The correlation of samples originating in the same organ is higher when using the most differentially expressed genes than when using the tumor-specific genes. (e) Clustering of multiple samples from two patients using the 50 most differentially expressed genes divides samples by organ. Genes expressed highly in lung and liver tissues are noted with brackets. Clustering of the same samples using the 50 genes specific to tumor subtypes separates samples by patient. Brackets note genes whose expression differentiates the two patients. A diagram of sampled locations for these patients (indicated by concentric circles) illustrates how samples simultaneously exhibit both patient- and organ-specific gene expression.
epithelial cell differentiation\textsuperscript{44,45}. This result prompted us to perform a more detailed analysis of histological markers of differentiation in our samples, and we found that samples with greater than 10% extracellular mucin, a marker of differentiation, comprised mostly tumors of the classical subtype (88.5%; n = 23) in comparison to only 11.5% (n = 3) of tumors from the basal-like subtype (P = 0.042; Table 2 and Supplementary Fig. 13). Consistent with the increased presence of extracellular mucin, our classical subtype was enriched for genes upregulated in mucinous ovarian cancer (‘Wamaunyokoli ovarian cancer grade 1 2 up’ classification\textsuperscript{46}; Supplementary Fig. 12b). Interestingly, our basal-like subtype was enriched for genes related to KRAS activation and STK11 loss in a mouse model of lung cancer where STK11-deficient tumors demonstrated shorter latency and more frequent metastasis\textsuperscript{47} (Supplementary Fig. 12c). We identified one sample with STK11 inactivation in the ICGC data set; this sample was in the basal-like subtype (Supplementary Fig. 11). Notably, our subtypes were not associated with other known signaling pathways in PDAC, including the Fanconi anemia\textsuperscript{48}, DNA repair, chromatin-remodeling\textsuperscript{49}, β-catenin\textsuperscript{50}, RB, ARF and G1 (ref. 51) pathways (Fig. 4a). However, genes from all of these pathways, except for β-catenin signaling, were considerably differentially expressed in cell lines in comparison to patient-derived tumors, suggesting that gene expression in cell lines may be a deceptive representation of the expression in most tumors.

Low within-patient heterogeneity in tumor-specific genes

We expect that only a subset of genes is relevant to the question of within- and between-patient heterogeneity in PDAC. Many methods exist to preselect genes for supervised analysis\textsuperscript{52}, but selection of the most differentially expressed genes is a common preprocessing step during unsupervised analysis\textsuperscript{53}. When clustering matched samples of metastatic and primary lesions using the 50 most differentially expressed genes among all matched samples, the samples separated primarily by organ site instead of by patient (Fig. 6a,c). In contrast, when considering 25 top-ranked exemplar genes each from the basal-like and classical factors, samples from the same patient clustered closer together with the clustering less dependent on organ site (Fig. 6b,d). This trend was further illustrated in a focused analysis of two patients (Fig. 6), whose tumor samples appeared to be patient specific when considering our tumor subtype gene list but clustered by tumor site when considering differentially expressed genes. Overall, we found that our tumor subtype gene list showed higher similarity (mean Pearson’s ρ = 0.53) between all other samples from the same patient than did the differentially expressed gene list (ρ = 0.32; t test, P < 0.001). Furthermore, our tumor subtype gene list produced much lower similarity among all other samples from the same organ site across different patients (ρ = 0.04) than the differentially expressed gene list (ρ = 0.34; P < 0.001). This observed similarity of tumor gene expression among tumors in the same patient suggests overall high between-patient tumor heterogeneity and low heterogeneity between primary and metastatic sites. However, we did observe examples of within-patient heterogeneity between metastatic sites. For example, lung metastases, even those from patients with basal-like tumors in other locations, clustered exclusively with tumors from the classical subtype, suggesting that some within-patient heterogeneity may exist among metastatic sites and supporting the previously reported divergent patterns of metastases in PDAC\textsuperscript{24}.

DISCUSSION

Our study represents the largest investigation of primary and metastatic PDAC gene expression thus far. We have used purity-independent subtyping (PuriSt\textsuperscript{t}) to identify new prognostic subtypes of PDAC that may have been obscured previously by confounding normal and stromal tissues. Our identification of normal-, tumor- and stroma-specific gene expression signatures is supported by both the overlap of these signatures with previously identified gene lists and their expression in appropriate tissue types. Our tumor subtypes are further supported by their relationship to previously identified basal tumor subtypes in breast and bladder cancers and their prognostic relevance in external cohorts. Our findings of two different stromal subtypes may help explain the differential effects of stroma previously seen in preclinical models.

Tumor- and stroma-specific gene expression classified PDAC into four distinct subtypes with prognostic relevance. The orthogonal nature of the tumor- and stroma-specific subtypes suggests an important interplay between tumors in patients that will need to be taken into account as stroma- and immune-modulating therapies are studied. In our cohort, patients with basal-like tumors appeared to derive more benefit from adjuvant therapy. Whether the basal-like and classical subtypes may be associated with response to specific therapies will need to be studied further as more effective therapies become available. One challenge will be defining preclinical model systems that recapitulate these subtypes, as our results suggest that traditional cell lines are lacking in the classical subtype. Although we demonstrate that PDX models recapitulate tumor-specific subtypes, these models alone may not be sufficient because of either the lack of human stroma or over-representation of the activated stromal subtype among the tumors that are successfully grafted. Thus, more detailed characterization of genetically engineered mouse models of PDAC will be needed to determine which models best reflect both our tumor- and stroma-specific subtypes.

Recent exome sequencing studies have confirmed commonly mutated genes in PDAC but have not uncovered mutations that clearly confer differences in survival\textsuperscript{1–3,7,37}. In fact, exome sequencing of a cohort of very long-term survivors of PDAC\textsuperscript{43} found no differences in somatic mutations to explain the improved biology of tumors from these rare patients in comparison with the majority of patients with PDAC, suggesting that examining somatic mutations alone may not be sufficient to understand the biological and clinical differences in PDAC tumors. Furthermore, exome sequencing studies and studies of microdissected samples are limited to the tumor compartment and overlook the stromal compartment, which has been shown to be biologically critical in PDAC, with both tumor-promoting and tumor-inhibiting effects. Our results suggest that RNA-determined subtypes may better capture the molecular landscape of PDAC and better reflect patient outcome. We hypothesize that our RNA subtypes may reflect the broad effect of somatic mutations while also capturing the importance of the neoplastic stroma.

These results provide new insights into the molecular composition of PDAC, which may be used for precision medicine. Furthermore, knowledge of these subtypes and their prognostic value may provide decision support in a clinical setting, where the choice and timing of therapies are critical.

URLs. International Cancer Genome Consortium (ICGC) data portal, https://dcc.icgc.org/; The Cancer Genome Atlas (TCGA) data portal, https://tcga-data.nci.nih.gov/tcga/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Previously unpublished expression data are available through the Gene Expression Omnibus (GEO) under accession GSE71729.
1. Blankin, A.V. et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature 491, 399–405 (2012).
2. Jones, S. et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 321, 1801–1806 (2008).
3. Waddell, N. et al. Whole genomes redefine the mutational landscape of pancreatic cancer. Nature 518, 495–501 (2015).
4. Yachida, S. et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature 467, 1144–1147 (2010).
5. Collisson, E.A. et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. Nat. Med. 17, 500–503 (2011).
6. Iacobuzio-Donahue, C.A. et al. Exploration of global gene expression patterns in pancreatic cancer using cDNA microarrays. Am. J. Pathol. 162, 1151–1162 (2003).
7. Logsdon, C.D. et al. Molecular profiling of pancreatic adenocarcinoma and chronic pancreatitis identifies multiple genes differentially regulated in pancreatic cancer. Cancer Res. 63, 2649–2657 (2003).
8. Stuart, R.O. et al. In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. Proc. Natl. Acad. Sci. USA 101, 615–620 (2004).
9. Wang, Y. et al. In silico estimates of tissue components in surgical samples based on expression profiling data. Cancer Res. 70, 6448–6455 (2010).
10. Yoshihara, K. et al. Inhibiting tumor purity and stromal and immune cell admixture from expression data. Nat. Commun. 4, 2612 (2013).
11. Alexandrov, L.B. et al. Circulating tumor cells in cancer patients and healthy donors. Nature 509, 455–460 (2014).
12. Pritchard, K.P. et al. Inhibition of Hedgehog signaling enhances delivery of chemotherapy in a mouse model of pancreatic cancer. Science 324, 1457–1461 (2009).
13. Erkan, M. et al. The activated stroma index is a novel and independent prognostic marker in pancreatic ductal adenocarcinoma. Clin. Gastroenterol. Hepatol. 6, 1159–1161 (2008).
14. Cohen, S.J. et al. Fibroblast activation protein and its relationship to clinical outcome in pancreatic adenocarcinoma. Pancreas 37, 154–158 (2008).
15. Hwang, R.F. et al. Cancer-associated stromal fibroblasts promote pancreatic tumor progression. Cancer Res. 68, 918–926 (2008).
16. Vonlaufen, A. et al. Pancreatic stellate cells: partners in crime with pancreatic cancer cells. Cancer Res. 68, 2085–2093 (2008).
17. Herrera, M. et al. Functional heterogeneity of cancer-associated fibroblasts from human colon tumors shows specific prognostic gene expression signature. Clin. Cancer Res. 19, 5914–5926 (2013).
18. Nones, K. et al. Genome-wide DNA methylation patterns in pancreatic ductal adenocarcinoma reveal epigenetic deregulation of SLIT-ROBO, ITGA2 and MET (2014).
19. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of uterine bladder carcinoma. Nature 507, 315–322 (2014).
20. Damrauer, J.S. et al. Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology. Proc. Natl. Acad. Sci. USA 111, 3110–3115 (2014).
21. McConkey, D.J., Choi, W. & Dinney, C.P. New insights into subtypes of invasive bladder cancer: considerations of the clinician. Eur. Urol. 67, e76–e78 (2015).
22. Prat, A. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. 27, 1160–1167 (2009).
23. Ihle, N.T. et al. Effect of KRAS oncogene subtypes on protein behavior: implications for signaling and clinical outcome. J. Natl. Cancer Inst. 104, 228–239 (2012).
24. Carey, L., Winer, E., Viale, G., Cameron, D. & Gianni, L. Triple-negative breast cancer: disease entity or title of convenience? Nat. Rev. Clin. Oncol. 7, 683–692 (2010).
25. Bardeesy, N. et al. Smad4 is dispensable for normal pancreas development yet critical in progression and tumor biology of pancreatic cancer. Genes Dev. 20, 3130–3146 (2006).
26. Haege, S.M. et al. Smad4 loss promotes lung cancer formation but increases sensitivity to DNA topoisomerase inhibitors. Oncogene doi:10.1038/onc.2015.112 (20 April 2015).
27. Guardob-Laguna, I. et al. Tumor engraftment in nude mice and enrichment in stroma-related gene pathways predict poor survival and resistance to gemcitabine in patients with pancreatic cancer. Clin. Cancer Res. 17, 5793–5800 (2011).
28. Dal Molin, M. et al. Very long-term survival following resection for pancreatic cancer is not explained by commonly mutated genes: results of whole-exome sequencing analysis. Clin. Cancer Res. 21, 1944–1950 (2015).
29. Zhang, Y. et al. GATA6 activates Wnt signaling in pancreatic cancer by negatively regulating the Wnt antagonist Dickkopf1. PLoS ONE 6, e22199 (2012).
30. Wamunyoki, F.W. et al. Expression profiling of mucinous tumors of the ovary identifies genes of clinicopathologic importance. Clin. Cancer Res. 12, 690–700 (2006).
31. Li, H. et al. LKB1 modulates lung cancer differentiation and metastasis. Nature 448, 807–810 (2007).
32. Croft, D. et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 42, D472–D477 (2014).
33. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29 (2000).
34. Bild, A.H. et al. Oncogenic pathway signatures in human cancers as a guide to drug discovery and the development of targeted therapies. Nature 439, 353–357 (2006).
35. Nishimura, D. BioCarta. Nat. Genet. 35, R68 (2010).
36. Shapira, K. et al. Pathway knowledge graph of colorectal cancer. Bioinformatics 22, 1160–1167 (2006).
37. Cruz, A.L. et al. Comparison of breast cancer intrinsic subtypes. Breast Cancer Res. 12, 106 (2010).
38. Shieh, J. et al. BRCA1/2 mutation screening in breast cancer patients with a history of pancreatic disease. Clin. Gastroenterol. Hepatol. 6, R229–R231 (2008).
39. Iacobuzio-Donahue, C.A. et al. Whole-exome sequencing of pancreatic cancer defines genetic heterogeneity of primary disease. Nature 518, 495–501 (2015).
40. Iacobuzio-Donahue, C.A. et al. Whole-exome sequencing of pancreatic cancer defines genetic heterogeneity of primary disease. Nature 518, 495–501 (2015).
41. Lando, D. et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 42, D472–D477 (2014).
42. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29 (2000).
43. Bild, A.H. et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 439, 353–357 (2006).
ONLINE METHODS

Decomposition by factors and gene ranking. For all analyses in this manuscript, we used \( k = 14 \) as the number of factors. Unsupervised NMF was performed on a gene-by-sample matrix \( X \) first with 20 randomly initialized instances of NMF using the MATLAB (MathWorks, R2013a) multiplicative update NMF solver for 10 steps. The pair with the lowest residual solution from these 20 instances was then used to seed NMF of \( X \) to convergence with the alternating least-squares solver. The result was a matrix of gene loadings, \( G \), and a matrix of sample loadings, \( S \). \( G \) and \( S \) were then scaled such that the mean of each column of \( G \) was 1 to facilitate cross-factor comparisons.

For each of the \( k \) factors, a set of distinct exemplar genes for the \( i \)th factor was established by ranking genes in descending order of the difference between the loading value in the \( i \)th column of matrix \( G \) and the largest loading value not in the \( i \)th column of matrix \( G \).

Two hundred iterations of fivefold resampling, training on a partition of approximately 80% of the samples, were performed to achieve stable NMF results. For each of these 200 data partitions, unsupervised NMF was performed, and the genes that appeared ranked in the top 50 of any factor together were recorded in a gene-by-gene consensus matrix. This gene factor co-occurrence consensus matrix was then used as the basis of a hierarchical clustering operation using correlation as a distance metric and an appropriate cutoff so as to yield \( k \) gene clusters. These \( k \) gene clusters were used to create a seed matrix \( G_0 \) such that the \( i \)th column of \( G_0 \) contained 0.01 for all genes except those in gene cluster \( i \), for which the values were set to 1. \( G_0 \) was then used to seed a final NMF with the multiplicative update solver to completion.

Gene set analysis was performed on the ranked list of genes for each factor with all sets available from MSigDB [1] v3.1. Sets were assessed for significance via the Kolmogorov-Smirnov statistic with Benjamini-Hochberg correction. Because of the positive nature of the ranked gene list, only gene sets with positive enrichment were considered.

Patients and samples. Multiple samples were obtained from 15 patients with metastatic PDAC from the University of Nebraska Medical Center Rapid Autopsy Pancreatic Program and 17 patients from Johns Hopkins Medical Institutions and the Johns Hopkins Gastrointestinal Cancer Rapid Medical Donation Program. Informed consent was obtained from all subjects. To ensure minimal degradation of tissue, organs were collected within 3 h post mortem, and specimens were flash frozen in liquid nitrogen. The cohort further included patients with resected PDAC and/or normal tissue from Johns Hopkins Medical Institutions, Northwestern Memorial Hospital, NorthShore Hospital and UNC hospitals. All samples were collected between 1999 and 2009 and flash frozen in liquid nitrogen at the time of operation, after approval for an epithelial marker, EpCAM (BioLegend, 324209). Fresh tumor was minced into pieces no larger than 1 mm\(^3\) and cultured with DMEM/Ham's F12 (1:1) medium supplemented with 10% FBS. Immunofluorescence was used to confirm the presence of CAFs as defined by staining for smooth muscle actin \( \alpha \) (SMA) (Santa Cruz Biotechnology, sc-32251) and a mesenchymal marker, vimentin (Cell Signaling Technology, 5741), as well as the absence of staining for an epithelial marker, EpCAM (BioLegend, 324209).

Statistical analysis. For all analyses, the sample size was limited to all appropriate cases with full data; no imputation was performed to estimate missing clinical information. Disease-specific survival or recurrence-free survival was analyzed using the Kaplan-Meier product-limit method, and the significance of clinicopathological or subtype variables was measured by Cox proportional hazards regression. Multivariate associations with survival were also performed using the Cox proportional hazards regression method. When more than two survival cohorts were compared, the log-rank test was used to assess global differences in survival. Fisher's exact test was used to analyze associations between two categorical variables. For continuous variables, for example, stain intensity and factor weight, unpaired two-tailed two-sample t tests were performed under the assumption of equal variance. Box-and-whisker plots show the median, quartiles and range of continuous data to demonstrate the variability of data and the degree of normality. Unless otherwise mentioned, sample-to-sample and gene-to-gene similarities were measured by correlation on the basis of log\(^{-}\)transformed gene expression after normalizing the expression for each gene to have a mean of 0 and variance of 1. Unless otherwise noted, clustering was performed via consensus clustering of row-normalized gene expression. Consensus clustering consisted of 1,000 iterations of \( k \)-means clustering, with 50% feature hold-out at each iteration, followed by hierarchical clustering of the consensus matrix with average linkage.

Microarray data. All RNA isolation and hybridization were performed at UNC on Agilent human whole-genome 4x44K microarrays (Agilent Technologies). RNA was extracted from macrodissected snap-frozen tumor samples using AllPrep kits (Qiagen) and quantified using Nanodrop spectrophotometry (Thermo Scientific). RNA quality was assessed with the use of the Bioanalyzer 2100 chip (Agilent Technologies). RNA was selected for hybridization using RNA integrity number and by inspection of 18S and 28S ribosomal RNA. Similar RNA quality was selected across samples. RNA (1 \( \mu \)g) was used as a template for cDNA preparations. cDNA was labeled with Cy5-dUTP and a reference control (Stratagene) was labeled with Cy3-dUTP using the Agilent Low–RNA Input Linear Amplification kit (Agilent Technologies) and hybridized overnight at 65 °C to Agilent 4x44K whole–human genome arrays. Arrays were washed and scanned using an Agilent scanner (Agilent Technologies).
Arrays were annotated using GEO platform GPL4133 and analyzed using log$_2$-transformed background-corrected Cy5 signal to maintain positivity. Multiple probes mapping to the same gene symbol were collapsed by mean probe expression. Samples were normalized to each other via quantile normalization.

**RNA sequencing.** Total RNA (200–1,000 ng) was used to prepare libraries with the TruSeq Stranded mRNA Sample Prep kit (Illumina). Reads (75 bp, paired end) were sequenced on a NextSeq 500 Desktop Sequencer using a high-output flow cell kit (Illumina). Reads were separated by species of origin using Xenome$^{38}$. Human- or mouse-specific reads were then aligned and quantified using TopHat2 (ref. 59), Cufflinks$^{60}$, hg19, mm10, and UCSC transcript and gene definitions. mRNA-based gene expression was analyzed as log$_2$ (1 + FPKM) (where FPKM is fragments per kilobase of transcript per million mapped reads), and KRAS mutation status was determined by manual curation of aligned human reads.

**Validation data sets.** Gene expression array data from resected primary tumor samples from Australian Pancreatic Cancer Genome Initiative and ICGC data were obtained from GSE50827 (ref. 28). Associated open access clinical data were obtained from the ICGC data portal (release_16). Patients with death events before 30 d were assumed to have postoperative complications and were censored. Patients with metastases were excluded from survival analyses. Information on genomic subtypes, mutations and amplifications was obtained from the supplementary information available from Waddell et al.$^3$. Normalized gene expression, survival data and PAM50 (ref. 32) classification from primary breast cancer (Prat et al.) samples ($n = 295$) as part of the UNC337 set were obtained from GSE18229 (ref. 33).

Normalized RNA-seq expression data for 845 primary tumors were obtained as described by Hoadley et al.$^{53}$ from TCGA$^{61}$. Normalized RNA-seq gene expression and partial survival data from 223 urothelial bladder carcinoma (BLCA) samples were obtained from TCGA$^{29}$. Samples were classified as basal or luminal with the BASE47 classifications provided by Damrauer et al.$^{30}$.

55. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 102, 15545–15550 (2005).
56. Neel, N.F. et al. Response to MLN8237 in pancreatic cancer is not dependent on RalA phosphorylation. Mol. Cancer Ther. 13, 122–133 (2014).
57. Bachem, M.G. et al. Pancreatic carcinoma cells induce fibrosis by stimulating proliferation and matrix synthesis of stellate cells. Gastroenterology 128, 907–921 (2005).
58. Conway, T. et al. Xenome—a tool for classifying reads from xenograft samples. Bioinformatics 28, i172–i178 (2012).
59. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36 (2013).
60. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7, 562–578 (2012).
61. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70 (2012).