Supplemental Information

Foa2 and Cdx2 cooperate with Nkx2-1 to inhibit lung adenocarcinoma metastasis

Carman Man-Chung Li, Vasilena Gocheva, Madeleine J. Oudin, Arjun Bhutkar, Shi Yun Wang, Saya R. Date, Sheng Rong Ng, Charles A. Whittaker, Roderick T. Bronson, Eric L. Snyder, Frank B. Gertler, Tyler Jacks
Supplemental Figures

**Li_FigS1**

A. Gene expression levels for Nkx2-1, Foxa2, and Cdx2 under different conditions.

B. Western blot analysis showing protein expression levels for Nkx2-1, Foxa2, Cdx2, and HSP90 across TnonMet, Tmet, and Met conditions.

C. mRNA expression levels for Nkx2-1, Foxa2, Cdx2, Tks5short, and Tks5long under Ctrl shN, Ctrl shF, Ctrl shC conditions.

D. mRNA expression levels for Foxa2 and Cdx2 under Nkx2-1 conditions.

E. Relative FITC-gelatin degradation assay results for shCtrl and shNF1 conditions.

F. Relative FITC-gelatin degradation assay results for Ctrl and NFC conditions.
Supplemental Figure S1. Related to Figure 1
(A) Normalized expression values of Nkx2-1, Foxa2, and Cdx2 in T\textsubscript{nonMet}, T\textsubscript{Met}, and Distant Met cells from microarray data in Winslow et al. (2011). Data are presented as Box-whisker plot (5%–95%). The p-values were calculated by Student’s t test.
(B) Protein level of Nkx2-1, Foxa2, and Cdx2 in representative T\textsubscript{nonMet} (802T4, 394T4, 368T1), T\textsubscript{Met} (482T1, 373T1, 393T5, 393T3), and Met (482M1, 393M1) cells.
(C) Knockdown of Nkx2-1, Foxa2, and Cdx2 in an independent T\textsubscript{nonMet} cell line (368T1) increases Tks5\textsubscript{long} expression, but not Tks5\textsubscript{short}, as measured by qRT-PCR. Data are represented as mean ± SD. The p-value was calculated by Student’s t test.
(D) Combined overexpression of Foxa2 and Nkx2-1 (left) or Cdx2 and Nkx2-1 (right) in an independent T\textsubscript{Met} cell line (393T3) further represses Tks5\textsubscript{long}, but not Tks5\textsubscript{short}, compared to single overexpression, as measured qRT-PCR. Foxa2 and Cdx2 are expressed in a doxycycline-inducible manner, while Nkx2-1 is expressed constitutively. Data are represented as mean ± SD. The p-values were calculated by Student’s t test. **p < 0.01, ***p < 0.001.
(E-F) Effects of Nkx2-1, Foxa2, and Cdx2 knockdown in 394T4 T\textsubscript{nonMet} cells (E) and overexpression in 393T3 T\textsubscript{Met} cells (F) on invadopodia activity as measured by FITC-gelatin degradation assay. Data are represented as mean ± SEM. The p-values were calculated by Student’s t test. ***p < 0.001, ****p < 0.0001.
Supplemental Figure S2. Related to Figure 2
Size of subcutaneous tumors after transplantation of 394T4 $T_{\text{nonMet}}$ cells, $T_{\text{nonMet}}$ with single/double knockdown, $T_{\text{nonMet}}$-shNFC cells, and 373T1 $T_{\text{Met}}$ cells. Each circle represents an individual mouse. Lines (-) indicate control hairpins against firefly or renilla luciferase. Data are represented as mean ± SEM.
Supplemental Figure S3. Related to Figure 3

(A) \( \text{T}_{\text{nonMet}} \) and \( \text{T}_{\text{nonMet}}\)-shNFC cells (394T4) in culture condition have similar morphology, unlike their distinct epithelial and mesenchymal morphology \textit{in vivo}.

(B) \( \text{T}_{\text{nonMet}}\)-shNFC (394T4) subcutaneous tumors lose epithelial marker Krt19, and partially gain mesenchymal markers Twist, Snail, Zeb1, and N-cadherin compared to \( \text{T}_{\text{nonMet}}\)-shCtrl tumors. Each circle represents an individual mouse. Data are represented as mean ± SEM. The p-values were calculated by Student’s t test. *\( p < 0.05 \), **\( p < 0.01 \), ***\( p < 0.001 \).

(C) Immunohistochemical staining of Hmga2 in \( \text{T}_{\text{nonMet}} \), \( \text{T}_{\text{nonMet}}\)-shN, and \( \text{T}_{\text{nonMet}}\)-shNFC subcutaneous tumors. Scale bar represents 100 \( \mu \)m. Insets show nuclear staining of Hmga2.

(D) Measurement of cell migration \textit{in vitro}. \( \text{T}_{\text{nonMet}}\)-shNFC (394T4) cells displayed increased motility compared to \( \text{T}_{\text{nonMet}} \) and \( \text{T}_{\text{nonMet}}\)-shN cells. Knockdown of Tks5 long in \( \text{T}_{\text{nonMet}}\)-shNFC cells using two different shRNAs partially impaired their migration. At least 70 cells pooled from 3-4 independent experiments were analyzed per condition. Data are represented as mean ± SEM. The p-values were calculated by Student’s t test. ****\( p < 0.0001 \).
Supplemental Figure S4. Related to Figure 4

(A) Multidimensional Scaling (MDS) clustering showing sample relationships before and after subtraction of clonal background signature. Removal of cell line-dependent clonal signature unveils underlying clustering of T\textsubscript{nonMet}-shNFC/T\textsubscript{Met} cells away from T\textsubscript{nonMet}/T\textsubscript{nonMet}-shN cells.

(B) qRT-PCR validation of expression changes of pro-metastatic gene TGF\textbeta{} and anti-metastatic gene Mtus1 that were identified in shNFC signature. Data are represented as mean ± SD. The p-values were calculated by Student’s t test.
Supplemental Figure S5. Related to Figure 5
(A) Normalized expression values of Hmga2 and Snail in T_{nonMet}, T_{Met}, and Distant Met cells from microarray data in Winslow et al. (2011). Data are presented as Box-whisker plot (5%–95%). The p-values were calculated by Student’s t test.
(B) ChIP-qPCR analysis of the enrichment of Nkx2-1, Foxa2, and Cdx2 binding at the Hmga2 and Snail genomic loci. Enrichment for Nkx2-1, Foxa2, and Cdx2 binding at Hmga2 as well as Nkx2-1 and Foxa2 binding at Snail suggests that these factors may regulate expression of Hmga2 and Snail at least in part by direct interaction with the indicated enhancers. Data are represented as mean ± SEM of three independent experiments. SftpA, Hnf4a and Wnt5a serve as positive controls for Nkx2-1, Foxa2, and Cdx2 binding, respectively. GD8: negative control mapping to a gene desert region on murine chromosome 8. For each enhancer versus GD8, p < 0.05 by Student’s t test.
(C) qRT-PCR detection of knockdown of Tks5_{long} (shTks5_{long}#1), Hmga2 (shHmga2#1), and Snail (sgSnail#1) in 394T4 T_{nonMet}-shNFC cells. Data are represented as mean ± SD.
(D) Hmga2 and Snail knockdown did not affect the expression of each other, or the expression of Tks5_{long}, as measured by qRT-PCR. Data are represented as mean ± SD.
(E) Size of subcutaneous tumors after transplantation of 394T4 T_{nonMet}-shNFC cells, and T_{nonMet}-shNFC cells with knockdown of Tks5_{long}, Hmga2, or Snail. Each circle represents an individual mouse. Data are represented as mean ± SEM.
Supplemental Figure S6. Related to Figure 6

(A) Representative tumor where moderately-differentiated Cdx2-positive area is adjacent to and not overlapping with poorly-differentiated Hmga2-positive area. Scale bar represents 150 µm.

(B-C) Expression of Cdx2 detected by qRT-PCR (B) and immunoblotting (C) in two independent TnonMet cell lines upon knockdown of Foxa2, Nkx2-1, or both factors. (B) 394T4 and 368T1 cells. Data are represented as mean ± SD. (C) 394T4 cells.

(D) ChIP-qPCR detects binding of Nkx2-1 and Foxa2 to an enhancer of the Cdx2 genomic locus. Data are represented as mean ± SEM of two independent experiments (for Nkx2-1 ChIP) and five independent experiments (for Foxa2 ChIP). SftpA and Hnf4a serve as positive controls for Nkx2-1 and Foxa2 binding, respectively. GD8: negative control mapping to a gene desert region on murine chromosome 8. For each enhancer versus GD8, p < 0.05 by Student’s t test.

(E) A model for regulation of Cdx2 expression by Nkx2-1 and Foxa2.
Supplemental Figure S7. Related to Figure 7
(A) Multivariate Cox proportional hazard regression analysis of overall survival of TCGA lung adenocarcinoma patients (n = 488).
(B) Analysis of the top 10\textsuperscript{th} percentile of lung adenocarcinoma patients with Stage I or II disease (n = 356) in each signature identified by ICA shows that the three signatures correlate with overall survival in these early-stage patients. p < 0.002 by Log-rank test.
Supplemental Figure S8. Expression level of Foxa1 in T_{nonMet}, T_{Met}, and Met cell lines
Foxa1 was not differentially expressed between T_{nonMet} (368T1, 394T4, 802T4), T_{Met} (373T1, 393T3, 393T5, 482T1), and Met (373N1, 393M1, 393N1, 482M1, 482N1) cells as measured by qRT-PCR. Data are represented as mean ± SEM. The p-value was calculated by Student’s t test.
Supplemental Materials and Methods

**cDNA expression and knockdown**

To generate T\textsubscript{Met}-TRE-Nkx2-1, -Foxa2, or -Cdx2 cell lines, the respective coding sequences were cloned into the lentiviral expression vector pCW22tre-optimegaUbcrTA, which contains a TRE promoter for doxycycline-induction of the cDNA and a UBC promoter for constitutive expression of rtTA. In addition, to allow constitutive overexpression of Nkx2-1 in T\textsubscript{Met} cells, the Nkx2-1 coding sequence was cloned into the retroviral expression vector MSCV. These vectors were used to infect T\textsubscript{Met} cells.

To generate T\textsubscript{nonMet}-shNFC cells, shRNAs targeting Nkx2-1, Foxa2, Cdx2, or Firefly/Renilla Luciferases were cloned into lentiviral vectors and used to infect T\textsubscript{nonMet} cells. These cells were subsequently infected with lentiviral vectors expressing shRNAs to knockdown Tks5\textsubscript{long} or Hmga2, or with lentiviral vectors expressing FLAG-Cas9 and sgRNAs targeting Hmga2 or Snail.

### List of shRNAs

| Gene      | shRNA ID         | miR30 shRNA sequence                                                                 |
|-----------|------------------|--------------------------------------------------------------------------------------|
| Nkx2-1    | shNkx2-1         | TGCTGGACAGTGAGCGCCACAGTCGTGCTACCTTTAGTGAGCCACAGATGCAAGAGCTTTTCTTTCTTCTTTGAGGAGCCAACAGATGTAAGGGACAGACATGGCACACTCTAGGCTGCTACTGCTTCGGA |
| Foxa2     | shFoa2           | TGCTGACAGTGAGCGCCACAGTCGTGCTACTTCTTTATAGTGAGCCACAGATGTAAGGGACAGACATGGCACACTCTAGGCTGCTACTGCTTCGGA |
| Cdx2      | shCdx2           | TGCTGGACAGTGAGCGCCACAGTCGTGCTACTTCTTTATAGTGAGCCACAGATGTAAGGGACAGACATGGCACACTCTAGGCTGCTACTGCTTCGGA |
| Tks5\textsubscript{long} | shTks5long#1     | TGCTGGACAGTGAGCGCCACAGTCGTGCTACTTCTTTATAGTGAGCCACAGATGTAAGGGACAGACATGGCACACTCTAGGCTGCTACTGCTTCGGA |
| Tks5\textsubscript{long} | shTks5long#2     | TGCTGGACAGTGAGCGCCACAGTCGTGCTACTTCTTTATAGTGAGCCACAGATGTAAGGGACAGACATGGCACACTCTAGGCTGCTACTGCTTCGGA |
| Hmga2     | shHmga2#1        | TGCTGGACAGTGAGCGCCACAGTCGTGCTACTTCTTTATAGTGAGCCACAGATGTAAGGGACAGACATGGCACACTCTAGGCTGCTACTGCTTCGGA |
| Luciferase| shLuciferase     | TGCTGGACAGTGAGCGCCACAGTCGTGCTACTTCTTTATAGTGAGCCACAGATGTAAGGGACAGACATGGCACACTCTAGGCTGCTACTGCTTCGGA |
| Renilla   | shRenilla        | TGCTGGACAGTGAGCGCCACAGTCGTGCTACTTCTTTATAGTGAGCCACAGATGTAAGGGACAGACATGGCACACTCTAGGCTGCTACTGCTTCGGA |

### List of sgRNAs

| Gene      | sgRNA ID | sgRNA target sequence (5′ to 3′) (PAM sequence in bold) |
|-----------|----------|--------------------------------------------------------|
| Hmga2     | sgHmga2#2| GTCCTGCTTCTTTGGGACCTGGG | |
| Snail     | sgSnail#1| GTCCTGCTTCTTTGGGACCTGGG | |
| Snail     | sgSnail#2| CGCTTATGTTGGGCTTTCGCGG | |
| mRosa26   | sgRosa   | GGAGATGGGCGGAGCTTCTTCTGG | |
**qRT-PCR**

RNA was purified from cultured cells or tissues using the RNAqueous kit (Invitrogen), and was reverse-transcribed using a High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems). Quantitative RT-PCR was performed using KAPA Sybr Fast ABI Prism 2X Mix (Kapa Biosystems), or TaqMan Fast Universal Master Mix and probes (Applied Biosystems). All mRNA levels were normalized to TBP.

| Target gene | qRT-PCR primers | Taqman probe |
|-------------|-----------------|--------------|
|             | Forward | Reverse | probe |
| TBP         | GGGGAGCTGTGATGTAAGT | CCAGGAAATAATTCGCTCA | Mm00446971_m1 |
| Tks5long    | TTATCAACGAGCTTCGTCGT | TTCGGATCTCCTCTGGCCAC |
| Tks5short   | TGGCTCACCCGCTGCCTCTTG | CCTTGCTCTTCAGATGTGCTACAA |
| Nkx2-1      | GCTGTCCTGCTGCATTTG | AGCTGACGCGAGGTTCAAG |
| Foxa1       | GGGCAAGCAGCTACTACGC | TCATTCCAGCGCCACATAG |
| Foxa2       | GACTGGAGCAGCTACTACGC | TCATTCCAGCGCCACATAG |
| Cdx2        | CAGCAGTCCCTAGGAAGC | GCAGCAGCTCAGCTTTTCTC |
| Hnf4a       | AGAGGTTCTGTCACAGCAGATC | CGTCTGTGATGTGCTCGCAATC |
| Hmga2       | GGGCAAGCAGCTCAGCATCAGC | TCAGAGGTTGCTCTTGCTGC |
| Snail       | Mm00441533_g1 |
| Slug        | Mm00441531_m1 |
| Twist1      | Mm00442036_m1 |
| Vimentin    | Mm01333430_m1 |
| Zeb1        | Mm00495564_m1 |
| Cdh1        | Mm01247357_m1 |
| Cdh2        | Mm01162497_m1 |
| Krt17       | Mm00495207_m1 |
| Krt19       | Mm00492980_m1 |
| Mtu1        | Mm00628662_m1 |
| TGFβ        | Mm01178820_m1 |

**Immunoblotting**

Immunoblotting was performed on whole-cell lysates, using antibodies against Nkx2-1 (Epitomics 2044), Foxa2 (Cell Signaling 8186), Cdx2 (Cell Signaling 12306), Tks5 (Santa Cruz sc-30122), Hmga2 (Cell Signaling 5269), Snail (Cell Signaling 3879), HSP90 (Cell Signaling 4877), and β-tubulin (Cell Signaling 2128).
ChIP-qPCR

Adherent T\textsubscript{Met} cells (393T5) overexpressing Nkx2-1, Foxa2, or Cdx2 were washed once in PBS, and cross-linked in 1% formaldehyde diluted in PBS for 10 min at room temperature. The reaction was stopped by 100 mM glycine, followed by 5mg/ml BSA in PBS, and subsequently washed twice in cold PBS. Cells were harvested and re-suspended in lysis buffer [50mM Tris-HCl, pH 8.1, 10mM EDTA, 1% SDS, 1X complete protease inhibitors (Roche)], and sonicated with a Diagenode Bioruptor to obtain 300-500 bp fragment size. Fragmented chromatin was diluted in IP buffer [20mM Tris-HCl pH 8.1, 150mM NaCl, 2mM EDTA, 1% Triton X-100] and incubated overnight at 4°C with Protein G magnetic beads (Dynabeads: Invitrogen) that had been pre-incubated with antibodies against Nkx2-1 (Bethyl A300-8L000), Foxa2 (Santa Cruz sc-6554), Cdx2 (Bethyl A300-691A), or isotype controls (rabbit IgG and goat IgG, Abcam). Immunoprecipitates were washed six times with wash buffer [50mM HEPES pH 7.6, 0.5M LiCl, 1mM EDTA, 0.7% Na deoxycholate, 1% NP-40] and twice with TE buffer. Immunoprecipitated (or no IP input) DNA was recovered in 100ul 1X Elution Buffer [1% SDS, 0.1M NaHCO\textsubscript{3}] over 6 hours at 65°C, and column purified with QiaQuick columns (Qiagen). Quantitative PCR was performed using KAPA Sybr Fast ABI Prism 2X Mix (Kapa Biosystems).

The follow qPCR primers were used for ChIP-qPCR analysis:

| Genetic locus | Forward primer | Reverse primer |
|---------------|----------------|----------------|
| GD8           | GGGCACTGCTAAACTCTTGC | GATGTGGGGAGACTGGAGGA |
| SftpA         | TTGCCTTCTGTTGTTCTTGTG | TACACAGTCAGGTTCCTTCCAG |
| Hnf4a         | TGCCATGACAAAGCAGTAC | GGTGGGGAGATACGTATGACAG |
| Wnt5a         | GTCTGGGTTGTCATTTCTGGA | GCCCAGCCTCTTTATGGC |
| Tks5\textsubscript{long}-enhancer1 | AGCGTGCTACTTCATAGG | TTGTGAGGAGATGGAGGA |
| Tks5\textsubscript{long}-enhancer2 | CAGTGCTCTTCCCGAGTCA | GGTAGCAAGCCCTGGTC |
| Tks5\textsubscript{long}-enhancer3 | TTTGACTGATTCGCGGCTCT | ACTGGCAAGTATGGTAGG |
| Tks5\textsubscript{long}-enhancer4 | TCTCCATCCATGACCTAGCC | GCCAAGTACATGAGGAC |
| Tks5\textsubscript{long}-enhancer5 | TCTGTGTTGCCAACAGGGAGTA | GGTCATAATTTCTGGTCCAGC |
| Tks5\textsubscript{long}-enhancer6 | GCCATCTCCGCTTAAAGGGT | TAGAATTCTCCGAGGGTG |
| Hmga2 - Enhancer 1 | TTTGGGAGACGGGATAG | TCCGTCGTTCGGAGGAG |
| Hmga2 - Enhancer 2 | CAGGCATGCAAACAAACCCG | CGATAAGAGCTGGACAC |
| Hmga2 - Enhancer 3 | GCCAGCAGTGGGGGTTAAAAAG | CCAAGCTTCCTCAGGGGAG |
| Hmga2 - Enhancer 4 | CTATCGGAGAGGGGATAG | GCAGTACCTGGAGGAG |
| Hmga2 - Enhancer 5 | ACGTGCTCTGCTGGCAGG | AGCAATGACATGAGCAG |
| Hmga2 - Enhancer 6 | TTAGTAAGATCTCCTGTGTGCT | GCAGTTCGTCAGGTACAG |
| Hmga2 - Enhancer 7 | TATTTCAAGGCTCCGCTGG | CAGAAGACATGGAAGCTAG |
| Hmga2 - Enhancer 8 | AGAGAGAAGCAAGCCACGC | CCAAGTTCGTTCCACTG |
| Snail - Enhancer 1 | CCAGGGCTACAAAGTGATG | AGAACCTACTGATGAG |
| Snail - Enhancer 2 | TCAGTCATTAGTGGGTACTGTG | CTCTCAAGTCTCCCAGC |
| Snail - Enhancer 3 | CCTGCTTTGAAGAGAAGGTAAC | ATAGAAGAGCTGGAGGAG |
| Cdx2          | CGGTGTGCTATGCGGCTG | GCCGACTTTGGAACCTTACC |
**FITC-gelatin degradation assay**

Glass-bottomed 35-mm plates (MatTek) were coated with FITC-labeled gelatin as described in Bowden et al. (2001) with some modifications. Briefly, MatTek plates were treated with HCl, followed by 50 mg/mL poly-L-lysine, and then coated with a thin layer of FITC-labeled 0.2% gelatin (Sigma) for 1 hour. The gelatin coating was then cross-linked with ice-cold 0.8% glutaraldehyde (Electron Microscopy Sciences)/PBS for 15 min at 4°C and then for 30 min at room temperature. Plates were successively washed in PBS (three times for 5 min each), 5 mg/mL sodium borate in PBS (once for 3 min), and PBS (three times for 5 min each), before being incubated for 30 min with complete tissue culture medium. Cells (8x10⁴) were cultured on the gelatin-coated plates for 72 hours and subsequently processed using standard fluorescence microscopy procedures. Areas of degradation were quantified using ImageJ.

**In vitro cell migration assay**

Glass-bottom MatTek dishes were coated with collagen (100μg/ml in PBS) and Matrigel (0.2% in DMEM) for 1 hour each at 37°C. Cells (5x10⁴) were then plated, and were imaged 5 hours later using a spinning disc confocal microscope (TE2000 microscope, Nikon) at 20X magnification for 16 hours with 5 randomly-selected fields of view per well. Cell motility was tracked using the Manual Tracking Image J plugin, and the migration speed (um/min) was obtained using the Chemotaxis Tool plugin.

**Histology and immunohistochemistry**

Tissues for histology were fixed in 10% formalin for 24 hours and stored in 70% ethanol until paraffin embedding. Histological analysis for tumor grade was performed by a pathologist (R.T.B.) on 4-μm sections stained with haematoxylin and eosin (H&E). Immunohistochemistry (IHC) was performed on a Thermo Scientific Autostainer 360 machine followed by a hematoxylin counterstain, using antibodies against Nkx2-1 (Epitomics), Foxa2 (Cell Signaling), Cdx2 (Cell Signaling), or Hmga2 (Biocheck).

**RNA-sequencing (RNA-seq) analysis**

RNA-seq analysis was performed on 394T4 T_nonMet, T_nonMet-shN, T_nonMet-shNFC, and T_Met 373T1 cells in biological duplicates. RNA was isolated with the RNAqueous Total RNA Isolation Kit (Life Technologies), and cDNA libraries were prepared with the TruSeq RNA Sample Preparation Kit (Illumina). Sequencing was performed on an Illumina HiSeq 2000 instrument to obtain single-end 40-nt reads. All reads that passed quality metrics were mapped to the UCSC mm9 mouse genome build (http://genome.ucsc.edu/) using RSEM (Li and Dewey 2011). Raw estimated expression counts were upper-quartile normalized to a count of 1000 (Bullard et al. 2010). This RNA-seq data set is available through Gene Expression Omnibus (GEO accession GSE71629). Independent Component Analysis (ICA) was performed as described below to identify biologically relevant signatures that characterize the global gene expression profiles of these samples. Targeted differential analysis for overlaps with T_nonMet/T_Met/Met dataset was performed using EBSeq v1.4.0 (Leng et al. 2013). All RNA-seq analyses were conducted in the R Statistical Programming language (http://www.r-project.org/), including signature analysis, hierarchical clustering, and multidimensional scaling (MDS). Gene set enrichment analysis (GSEA) was carried out using the pre-ranked mode with default settings (Subramanian et al. 2005). Heatmaps were generated using the Heatplus package in R.
Clinical analysis

RNA-seq gene expression profiles of the primary tumors and the relevant clinical data of 488 lung adenocarcinoma patients were obtained from the Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov/). Independent Component Analysis (ICA) was performed as described below to identify biologically relevant signatures that characterize the gene expression patterns of NKX2-1, FOXA2, CDX2, and HMGA2 in these human tumors. The intermediate and late signatures were driven by patients that correlated strongly with those expression patterns ($Z_{max} = 4.5$ and $Z_{max} = 2.9$, respectively), thus yielding a strongly correlated patient sub-population within the top 10% of patients. The early signature exhibited lower correlation scores ($Z_{max} = 1.5$) and a larger patient sub-population (top 25%) was chosen for downstream gene expression analysis in order to maximize signal representation. Kaplan-Meier survival analysis was conducted with patients in the top 10th percentile of each signature, and significance was assessed using log-rank test. Multivariate Cox proportional hazard regression analysis with adjustment for gender, age, and stage was performed on the overall survival of patients in the top 10th percentile of each signature.

Independent Component Analysis (ICA)

For the analysis of RNA-Seq samples sequenced in this study and the comparative analysis with TCGA dataset, an unsupervised blind source separation strategy using Independent Component Analysis (ICA) was applied to elucidate statistically independent gene expression signatures within RNA-Seq expression data (Hyvärinen and Oja 2000; Rutledge and Jouan-Rimbaud Bouveresse 2013; A. Bhutkar, manuscript in preparation). ICA is a general-purpose signal processing and multivariate data analysis technique in the category of unsupervised matrix factorization methods. Based on input data consisting of a genes-samples matrix, ICA uses higher order moments to characterize the dataset as a linear combination of statistically independent latent variables. These latent variables represent independent components based on maximizing non-gaussianity, and can be interpreted as independent source signals that have been mixed together to form the dataset under consideration. Each component includes a weight assignment to each gene that quantifies its contribution to that component. Additionally, ICA derives a mixing matrix that describes the contribution of each sample towards the signal embodied in each component. This mixing matrix can be used to select signatures among components with distinct gene expression profiles across the set of samples. All computations were done in the R Statistical Programming Language. The R implementation of the core JADE algorithm (Joint Approximate Diagonalization of Eigenmatrices) (Rutledge and Jouan-Rimbaud Bouveresse 2013; Biton et al. 2013; Nordhausen et al. 2012) was used along with custom R utilities.

Other statistical analyses

All other statistical analyses were performed using Student’s T-test, unless otherwise specified. P-values < 0.05 (two-tailed) were considered statistically significant.
Supplemental References

Biton, A., Zinovyev, A., Barillot, E., and Radvanyi, F. (2013). MineICA: Independent component analysis of transcriptomic data.

Bullard, J.H., Purdom, E., Hansen, K.D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11, 94.

Hyvärinen, A., and Oja, E. (2000). Independent component analysis: algorithms and applications. Neural Netw 13, 411–430.

Biton A, Zinovyev A, Barillot E, Radvanyi F. 2013. MineICA: Independent component analysis of transcriptomic data.

Bowden ET, Coopman PJ,Mueller SC. 2001. Invadopodia: Unique methods for measurement of extracellular matrix degradation in vitro. Methods Cell Biol 63: 613–627.

Bullard JH, Purdom E, Hansen KD, Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11: 94.

Hyvärinen A, Oja E. 2000. Independent component analysis: algorithms and applications. Neural Netw 13: 411–430.

Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BMG, Haag JD, Gould MN, Stewart RM, Kendziorski C. 2013. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics 29: 1035–1043.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323.

Nordhausen K, Cardoso JF, Miettinen J, Oja H, Ollila E. 2012. JADE: JADE and other BSS methods as well as some BSS performance criteria. R package version.

Rutledge DN, Jouan-Rimbaud Bouveresse D. 2013. Independent Components Analysis with the JADE algorithm. TrAC Trends in Analytical Chemistry 50: 22–32.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. P Natl Acad Sci USA 102: 15545–15550.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323.
Nordhausen, K., Cardoso, J.F., Miettinen, J., Oja, H., and Ollila, E. (2012). JADE: JADE and other BSS methods as well as some BSS performance criteria (R package version).

Rutledge, D.N., and Jouan-Rimbaud Bouveresse, D. (2013). Independent Components Analysis with the JADE algorithm. TrAC Trends in Analytical Chemistry 50, 22–32.