Supporting Information

Volkmer et al. 10.1073/pnas.1120605109

SI Methods

Data Collection and Processing. We downloaded 1.5 billion data points from 25,237 microarrays in human Affymetrix U133 Plus 2.0; 16,357 microarrays in human Affymetrix U133A; and 3,969 microarrays in human Affymetrix U133A 2.0 platforms from the National Center for Biotechnology Information (NCBI)'s Gene Expression Omnibus (GEO) database (1) and normalized these datasets using the robust multichip average (RMA) algorithm (2). We also downloaded and normalized 10,843 Mouse_430_2.0 microarrays and 3,079 Rat_430_2.0 microarrays from GEO. We computed the thresholds for each gene using the StepMiner algorithm (3) and built a complete database of Boolean relationships between pairs of genes using BooleanNet (4). We identified and manually verified a set of 138 bladder transitional cell carcinoma arrays by automatically searching through the GEO description pages of all downloaded microarrays. These 138 bladder cancer (BC) arrays were distributed in the human Affymetrix platforms described above. We normalized all of the microarrays in all three human Affymetrix platforms (n = 45,563) using a modified chip description file (CDF) with RMA. From this normalized gene expression dataset, we selected our 138 identified BC arrays for further analysis (AffyBC dataset).

Additionally, we downloaded three independent BC datasets with survival data from NCBI GEO public database: 403 samples from Dyrskjøt et al. (European dataset, GSE5479) (5), 256 samples including 165 primary BC samples from Kim et al. (Chungbuk dataset, GSE13507) (6), and 89 samples from Lindgren et al. (Lindgren dataset, GSE19915) (7). We also downloaded another dataset from the Journal of Clinical Oncology (JCO) website by Sanchez-Carbaryo et al. (SanchezC dataset, survival follow-up times were not available) to analyze gene expression from normal bladder tissue (8). These datasets were already normalized and log2 transformed. We renormalized the Lindgren dataset because the KRT14 gene was missing in the original file from GEO. This renormalization was performed using background correction, print tip loess within each microarray, and quantile normalization between microarrays using limma package in R (Bioconductor) (9, 10). The clinical information for the European dataset was downloaded from Oncomine, the Chungbuk dataset was downloaded from GEO, and the Lindgren dataset was downloaded from the Cancer Research website. For the European dataset, overall survival status, progression status, relapse status, age, sex, stage, grade, and treatments were available. For the Chungbuk dataset, overall survival status, cancer specific survival status, age, sex, stage, and grade were available. Similarly, for the Lindgren dataset, cancer-specific survival status, stage, and grade were available. In all datasets, largest follow-up times were at least 5 y.

Statistical Analysis and Software Used. StepMiner software is used to compute thresholds for high and low expression levels (3, 4). For this application, the expression values for each gene were ordered from low to high, and StepMiner was used to fit a rising step function to the data that minimizes the differences (mean squared errors) between the fitted and measured values. This approach places the step at the largest jump from low values to high values (but only if there are sufficiently many expression values on each side of the jump to provide evidence that the jump is not due to noise) and sets the threshold at the point where the step crosses that original data. A noise margin of twofold change (±0.5) was considered around the StepMiner threshold and used as alternative thresholds when a stringent high or low condition is required. BooleanNet statistics was used to infer Boolean relationships between two genes (4). The mining developmentally regulated genes (MiDReG) approach is used to predict developmentally regulated genes in BC (11). We also developed a new web-based software called hierarchical exploration of gene expression microarrays online (Hegemon) that explores gene expression data and their clinical information using a scatterplot of log2 reduced gene expression values of two genes. This software provides a simple framework for automatic selection of patient groups on the basis of gene expression values, as well as other clinical parameters, and performs automated survival analysis. We used this software to test our hypothesis that patients with a basal tumor phenotype have worse survival outcomes compared with a mature tumor phenotype. This algorithm was also used to identify cell-surface markers corresponding to this differentiation state. Kaplan–Meier estimates, univariate and multivariate (using discrete and continuous KRT14) Cox regression analyses, were performed using the R software (R version 2.12.1 2010–12-16 available at http://www.r-project.org). The P values were calculated with the use of the log-rank test (indicated in all Kaplan-Meier plots). The ratio on the right of Kaplan–Meier plots in the main and Supporting Information figures shows the incidence proportions (1 – survival proportions). Boxplot with mean and confidence intervals are prepared with plotCI function of gplots package in the R software.

Identification of Differentiation States in Bladder Cancer Using MiDReG. We used MiDReG to predict genes developmentally regulated in BC as shown in Fig. 2 *A–D* and Fig. S1 *B–E*. We focused specifically on keratins for possible candidates because it is well known that gene expression of keratins changes systematically during epithelial tissue development. The seed genes used for this analysis were KRT5 and KRT20. On the basis of our previous results (12), $CD44^+$ tumor cells are upstream of $CD44^-$ during development and $CD44^+$ cells are mostly $KRT5^+$, and $CD44^-$ cells are mostly $KRT20^+$ (Fig. S1*A*). Therefore, for our analysis we assume that $KRT5^+$ tumor cells are upstream of $KRT20^+$ tumor cells. In this case, we set off to predict genes upstream of $KRT5^+$ tumor cells to further subdivide the original tumor-initiating populations. We formulated this problem in the context of MiDReG as follows. On the basis of previously known biology, KRT5 turns off during development, KRT20 turns on during development, and KRT5 is mutually exclusive with KRT20 ($KRT5^{high} \rightarrow KRT20^{low}$) (Fig. 2*A* and Fig. S1*B*). Our goal is to predict gene X that turns off during development but earlier than KRT5. Therefore, we searched for Boolean relationships $X^{high} \rightarrow KRT5^{high}$ and $X^{high} \rightarrow KRT20^{low}$ in our BooleanNet database (on 25,237 Affymetrix U133 Plus 2.0 microarrays) (Fig. 2 *B–D* and Fig. S1 *B–E*). This analysis identified 137 genes, which were filtered using the AffyBC dataset (n = 138, BC dataset) by good dynamic range of gene expression values (>5), diversity (stddev > 1.5), and keratins. The filtered list of genes contained seven keratins (KRT4, KRT13, KRT14, KRT16, KRT6B, KRT6A, and KRT6C). We examined the heatmaps of these keratins' gene expression in both the AffyBC dataset and the Chungbuk dataset (Fig. S1*F*) and found four good candidate keratins (KRT14, KRT16, KRT6B, and KRT6A) (Fig. S1*G*). Additionally, we used the Chungbuk dataset to check whether the predicted markers are associated with patient survival (by computing hazard ratio (HR) using Cox regression analysis of the high and low expression values as iden-

tified by StepMiner) (Fig. S1*F*). We used the Chungbuk dataset because of the high quality of gene expression values from Illumina platform for many known genes. Two of the identified keratins have strong associations with patient survival (KRT14, HR = 2.75, $P < 0.05$; KRT6B, HR = 3.48, $P < 0.05$) (Fig. S1*F*). We chose KRT14 as a potential candidate for experimental validation because large cohort of patients had high KRT14 gene-expression levels as opposed to other predicted keratins in both the AffyBC and the Chungbuk dataset. KRT14 has strong Boolean relationships: $KRT14^{high} \rightarrow KRT5^{high}$" and "$KRT14^{high} \rightarrow KRT20^{low}$. KRT14 is strongly associated with patient survival in the Chungbuk dataset (HR = 2.75, $P < 0.05$). Therefore, we hypothesize that KRT14 turns off during bladder cancer development earlier than KRT5. On the basis of this hypothesis, BC development is divided into three different states: $KRT14^{+}KRT5^{+}KRT20^{-}$ (basal), $KRT14^{-}KRT5^{+}KRT20^{-}$ (intermediate), and $KRT14^{-}KRT5^{-}KRT20^{+}$ (differentiated) (Fig. 2*F*).

Identification of Corresponding Surface Markers to Keratins Using Hegemon. To experimentally validate the differentiation states in BC defined by keratins, we predicted the corresponding surface markers using two large BC datasets with high quality gene expression values (Fig. S3): the AffyBC dataset, 138 samples in human Affymetrix platforms; and the Chungbuk dataset, with all 256 samples in Illumina Human-6 BeadChip (48K) platform. For the AffyBC dataset, StepMiner threshold for each probeset was computed using all 45,563 publicly available microarray dataset including 25,237 microarrays in human Affymetrix U133 Plus 2.0; 16,357 microarrays in human Affymetrix U133A; and 3,969 microarrays in human Affymetrix U133A 2.0 platforms that were normalized together (Fig. S3*C*). For the Chungbuk dataset, StepMiner threshold was computed using 256 samples (Fig. S3*C*). Within these datasets, we identified BC samples with all three phenotypes as described above: basal ($KRT14^{+}KRT5^{+}KRT20^{-}$), intermediate ($KRT14^{-}KRT5^{+}KRT20^{-}$), and differentiated ($KRT14^{-}KRT5^{-}KRT20^{+}$) as shown in Fig. S3*E*. We searched for two different groups of surface genes. First, we searched for those whose expression values in the basal phenotypes are high in average, but low in differentiated phenotypes, and down-regulated in intermediate phenotypes (Fig. S3*F* and Dataset S1); these genes are strongly down-regulated with differentiation. The high and low gene expression values are computed using StepMiner as described above (Fig. S3*C*). The second group of surface genes is similar to the first except that in differentiated phenotypes the average gene expression values are lower than in the basal phenotypes, but still high (Fig. S3*G* and Dataset S1); these genes are slightly down-regulated with differentiation. Subsequently, we ranked both groups of surface genes on the basis of their association with patient survival in the Chungbuk dataset (by computing hazard ratio using Cox regression analysis of the high and low expression values as identified by StepMiner). Genes from the first group, strongly down-regulated from basal to differentiated tumors contain CD44, the previously identified marker for tumor-initiating cells in BC. However, CD44 was ranked low because of lower hazard ratio and surprisingly it was less than 1. To identify an upstream marker, four top genes based on expression levels and hazard ratios are considered from the first group including CD248, S100A8, COL1A1, and CD90 (THY1). As FACS-compatible antibodies to CD248, S100A8, and COL1A1 are not commercially available, we focused on CD90 first. CD90 and CD248 were highly correlated (Chungbuk dataset, coefficient = 0.67; Affy dataset, coefficient = 0.63) with each other. To identify a downstream marker, we focused on CD49f (ITGA6) from the second group as this marker enriches basal cell populations, is coexpressed with KRT14 in normal and cancer epithelial tissue, and is functionally associated with stem and tumor initiating cells in other epithelial tissues.

Patient Classification for Outcome Analysis According to Bladder Cancer Differentiation Status. For survival analysis, we divided the BC patients using gene expression values of keratins and surface markers separately. We performed all of the following steps using our newly developed software Hegemon.

To determine the clinical relevance of our predicted keratin differentiation states first, we divided the patients into basal, intermediate, and differentiated phenotype whenever possible (all genes present in the datasets). Thus, the European dataset was not used for this analysis because gene expression for KRT5, KRT20 and CD90, CD49f was not measured in that microarray platform. For the Lindgren dataset, we used a stringent low (t – 0.5) for KRT14 and KRT5 and a stringent high (t + 0.5) for KRT20 to collect more patients with basal and intermediate phenotypes (Fig. S4*A*). The basal phenotype in the Lindgren dataset was defined as $KRT14^{+}KRT5^{+}KRT20^{-}$ ($KRT14^{+}$, ID:21409 ≥ 1.2–0.5; $KRT5^{+}$, ID:4006 ≥ 1.92–0.5; $KRT20^{-}$, ID:16873 < -0.05+0.5), the intermediate phenotype was defined as $KRT14^{-}KRT5^{+}KRT20^{-}$ ($KRT14^{-}$, ID:21409 < 1.2–0.5; $KRT5^{+}$, ID:4006 ≥ 1.92–0.5; $KRT20^{-}$, ID:16873 < -0.05+0.5), and the differentiated phenotype was defined as $KRT14^{-}KRT5^{-}KRT20^{+}$ ($KRT14^{-}$, ID:21409 < 1.2–0.5; $KRT5^{-}$, ID:4006 < 1.92–0.5; $KRT20^{+}$, ID:16873 ≥ -0.05+0.5). To evaluate the clinical relevance of our predicted corresponding surface marker differentiation states, we used gene expression values of these surface markers (CD90, CD44, and CD49f) in the Lindgren datasets (Fig. S4*B*). Due to the lack of gene expression values of these surface markers, the European dataset could not be used for analysis. StepMiner threshold (t) was used to determine the high and low gene expression levels of each surface marker. For the Lindgren dataset, the basal phenotype was defined as $CD90^{+}CD44^{+}CD49f^{+}$ ($CD90^{+}$, ID:2627 ≥ −0.59; $CD44^{+}$, ID:6797 ≥ −1.38; $CD49f^{+}$, ID:21822 ≥ 0.47), the intermediate phenotype was as $CD90^{-}CD44^{+}CD49f^{+}$ ($CD90^{-}$, ID:2627 < −0.59; $CD44^{+}$, ID:6797 ≥ −1.38; $CD49f^{+}$, ID:21822 ≥ 0.47), and the differentiated phenotype was defined as $CD90^{-}CD44^{-}CD49f^{+}$ ($CD90^{-}$, ID:2627 < −0.59; $CD44^{-}$, ID:6797 < −1.38; $CD49f^{+}$, ID:21822 ≥ 0.47).

Because a subset of patient samples in our analysis does not fit easily into the three BC subtypes described above (others, gray; Fig. S4 *A* and *B*), we analyzed additionally all possible combinations of keratins and cell surface markers (±) (Fig. S4 *C*–*E*), which reveals additionally heterogeneity and might be correlated with BC dedifferentiation.

Immunofluorescence Staining. Optimal cutting temperature (OCT) compound-embedded frozen tissue was sectioned into 5-μm thick sections and fixed with ethanol. Slides were then blocked with 10% goat serum and probed with anti-CD44 (Calbiochem; 217594), anti-KRT5 (Abcam; ab53121), anti-KRT20 (Abcam; ab962), and anti-KRT14 (Abcam; ab53115) antibodies for 2 h. Samples were stained with goat antimouse, -rabbit, and -rat secondary antibodies conjugated with Alexa 488/594 (Invitrogen) and nuclear counterstained with Hoechst (Invitrogen). Slides were imaged on a Leica fluorescent microscope.

Bladder Tumor Tissue Dissociation. The Stanford University and the Baylor College of Medicine institutional review boards (IRBs) approved the enrollment of human subjects under protocols 1512 and H-26809, respectively. Tumor tissues were mechanically dissociated in Medium 199 containing Liberase TM and TH enzymes (Roche), DNase (Worthington) and Pluronic-F68 (Sigma) at 37 °C until single-cell suspension was achieved (3–6 h). Cells were then washed twice with PBS and filtered through a 70-μm filter.

Flow Cytometry Analysis and Cell Sorting. Tumor cell suspensions were stained with phycoerythrin (PE)-conjugated anti-CD44 (BD

PharMingen; 550989), Alexa 700-conjugated anti-CD90 (Biolegend; 328120), APC-conjugated anti-CD49f (Biolegend; 17–0495), PerCP/Cy5.5-conjugated anti-ESA (Biolegend; 324214), and lineage mixture containing Pacific-blue–conjugated anti-CD45 (Biolegend; human 304022 and mouse 103125), anti-CD31 (Biolegend; human 303114 and mouse 102422), and H-2K^d (Biolegend; 116616) antibodies. Flow cytometry analysis and cell sorting was performed on a BD FACSAria (Becton Dickinson) cell sorting system under 20 psi with a 100-μm nozzle.

Real-Time PCR. Total RNA was isolated from sorted cell populations using the mirVana RNA isolation kit (Ambion) according to the manufacturer's protocol. RNA was then subjected to cDNA synthesis using superscript III (Invitrogen) according to the manufacturer's protocol. The cDNA was then processed through a preamplification step with final Taqman probe PCR reactions run on an ABI 7900 machine (Applied Biosystems) according to the manufacturer's protocol. The qPCR results were normalized using β-actin as endogenous control. (β-actin, Hs00357333_g1; KRT14, Hs00559328_m1; KRT5, Hs00361185_m1; KRT20, Hs00300643_m1; Applied Biosystems).

Xenotransplantation of Patient Cancer Cells into Immunocompromised Mice (Non-obese diabetic scid gamma). The Stanford Administrative Panel on Laboratory Animal Care approved the mouse studies under protocol 10725:4. FACS-purified human patient cancer cells were suspended and mixed with 25% Matrix-Matrigel (Becton Dickinson; 354248) and injected (10^3–10^4 cells), as indicated, intradermal into the dorsal skin of 4- to 8-wk-old NOD.Cg-Prkdc^{scid} Il2rg^{tm1Wjl}/SzJ (NSG) mice, as described previously (12).

KRT14 Immunohistochemistry. A total of 158 (Stanford) and 117 (Baylor) formalin-fixed paraffin-embedded (FFPE) BC tissues (from 1994 to 2006) with at least a 5-y follow-up were collected using Health Insurance Portability and Accountability Act (HIPAA) compliant Stanford (IRB 1512) and Baylor (IRB H-26809) institutional review board approval. FFPE sections (6 μm) were deparaffinized and hydrated with graded ethanol. Antigen retrieval was performed with 10 mM citrate buffer at pH 6.0, followed by primary antibody for KRT14 (Covance; PRB-155P, rabbit polyclonal 1:3,000) incubation. Antibody specificity was determined by Western blot analysis (Fig. S5*C*). Immunoreactivity was visualized using Vector's Vectastain Elite ABC kit (rabbit) following the manufacturer's protocol. A trained pathologist analyzed the staining. Nonepithelial cells (stroma cells) used as internal negative control showed no reactivity. BC specimens with squamous differentiation were used as positive control. Staining was scored according to the extent of KRT14$^+$ epithelial cancer cells, with higher scores indicating a greater proportion of positive cells: 0 (no positive cells), 1 (<5%), 2 (5–50%), and 3 (>50%). Tissue without epithelial cells was excluded. Patients were stratified as follows: negative (score 0–1); positive (score 2–3). Analysis was performed without knowledge of tumor stage, grade, or clinical follow-up. Images and scores are accessible online: http://genepyramid.stanford.edu/microarray/BladderTMA/.

KRT5 and KRT20 Immunohistochemistry. We have approval from Stanford University (IRB 1512) and Baylor College of Medicine (IRB H-26809) to use FFPE bladder cancer tissues from 1994 to 2006 with at least 5-y complete clinical follow-up with survival outcome data. A total of 158 (Stanford) and 117 (Baylor) patients fit with our selection criteria. PPFE sections (6 μm) were deparaffinized and hydrated with graded ethanol. Antigen retrieval was performed with 10 mM citrate buffer at pH 6.0, followed by primary antibodies for KRT5 (Abcam; ab52635, mouse monoclonal 1:100), and KRT20 (Abcam; ab76126, mouse monoclonal 1:100) incubation, washing steps, and secondary antibody incubation. Staining was performed using the Dako EnVision kit (K4006 and K4011) following the manufacturer's protocol. The sections were analyzed and scored by a trained pathologist. The scoring criteria were determined as followed: 0, all negative; 1, <5% of cells positive; 2, 5–50% cells positive; 3, >50% of cells positive. For analysis, 0 and 1 were stratified as negative and 2 and 3 as positive.

KRT14 Western Blot Analysis. Tumor cell pellets were lysed in radio immunoprecipitation assay buffer containing protease and phosphatase inhibitors for 10 min on ice, followed by centrifugation; supernatant was collected for protein quantification. A total of 40 μg of protein was running on a 4–12% gel and probed with KRT14 (Covance; PRB-155P) and GAPDH antibodies as previously described.

Patient Classification for Outcome Analysis for KRT14 as a Single Marker. We tested the hypothesis that KRT14 as a single marker is associated with patient survival using our Hegemon software (Fig. 5). To divide the BC patients according to KRT14 gene expression, we use *K* mean ($k = 3$) in two independent gene expression datasets including the Lindgren and the European datasets. This approach divides the BC patients into three different groups: high, intermediate, and low. For the European dataset of 404 samples, the thresholds to define three different expression levels of KRT14 (ID: 209) are: high (KRT14 ≥ 1.37), intermediate (KRT14 ≥ 0.012; KRT14 < 1.37), and low (KRT14 < 0.012). For the Lindgren dataset of 89 samples, the three different levels of KRT14 (ID: ILMN_1665035) are: high (KRT14 ≥ 0.98), intermediate (KRT14 ≥ 0.1; KRT14 < 0.98), and low (KRT14 < 0.1). Multivariate analysis in the European dataset (largest BC dataset) was performed using both discrete (Fig. 5) and continuous (Table S1) gene-expression values for KRT14.

We performed analysis on all muscle invasive tumors (Clinical stage ≥pT2) in both datasets (Fig. S6 *A* and *B*): the Lindgren and the European datasets. To divide all muscle invasive tumors (≥pT2) into two groups, we computed StepMiner threshold on KRT14 expression in tumors greater than pT2 stage including pT2 stage. In both datasets, we used a stringent low ($t - 0.5$) threshold which is 0.5 lower that the StepMiner threshold. All muscle invasive BC patients (pT2$^+$) were divided into two groups using this threshold: KRT14 high (Lindgren, ID:21409 ≥ 1.24–0.5; European, ID:209 ≥ 1.5–0.5), and KRT14 low (Lindgren, ID:21409 < 1.24–0.5; European, ID:209 < 1.5–0.5). Kaplan–Meier, uni-, and multivariate analyses were performed on the Lindgren and the European datasets as shown in Fig. S6 *A* and *B*.

To evaluate KRT14 association with overall survival of patients that have undergone bladder removal (cystectomy), we used the European dataset (Fig. S6*D*). We divided all cystectomy patients into two groups: KRT14 high (ID:209 ≥ 0.726) and KRT14 low (ID:209 < 0.726). Kaplan–Meier, uni-, and multivariate analyses were performed on this dataset as shown in Fig. S6*D*.

Similar to the above late-stage tumors, we also performed analysis on only pT2 muscle invasive tumors in both datasets (Fig. S6 *E* and *F*): the Lindgren and the European datasets. To divide all muscle invasive tumors (pT2) into two groups, we computed StepMiner threshold on KRT14 expression in only pT2 tumors. For both the Lindgren and the European datasets, we used the threshold computed by StepMiner on only the pT2 patients. The pT2 BC patients were divided into two groups: KRT14 high (Lindgren, ID:21409 ≥ 0.64; European, ID:209 ≥ 1.56), and KRT14 low (Lindgren, ID:21409 < 0.64; European, ID:209 < 1.56).

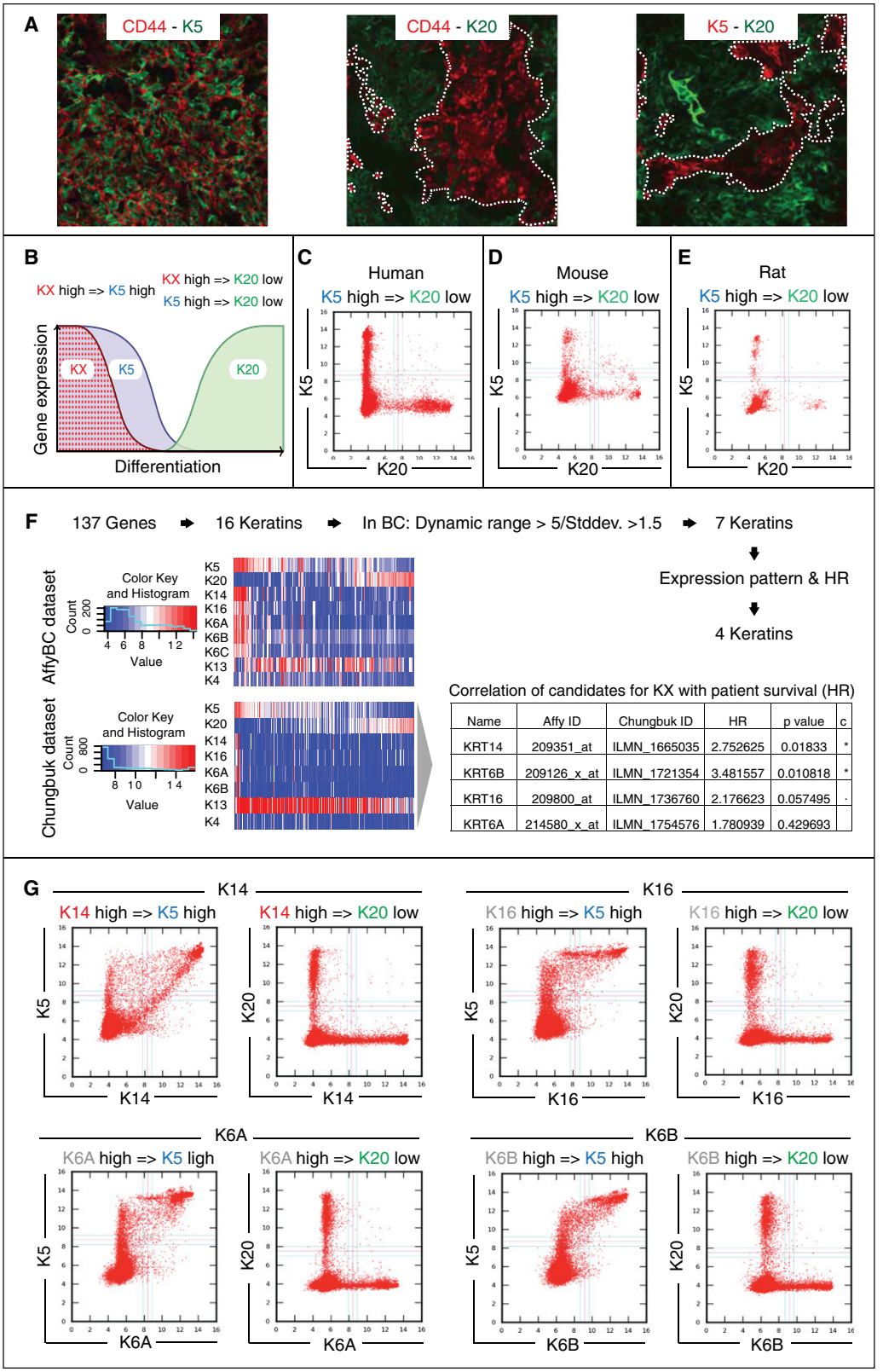
To evaluate KRT14 association with overall survival, progression-free survival, and recurrence-free survival in the early stage tumors (pTa), we used European datasets (Fig. S7). We divided all early stage tumors (pTa) into two groups: KRT14 high (ID:209 ≥ 0.726–0.5) and KRT14 low (ID:209 < 0.726–0.5). In

this case, KRT14 threshold (0.726–0.5) was computed as 0.5 lower than the actual StepMiner threshold 0.726 on all 404 patients to get more highly expressing KRT14 pTa tumors.

To test the association of KRT14 protein expression with BC patient outcome, two independent patient tissue datasets, the Stanford and the Baylor datasets, were used (described above). We evaluated KRT14 association with overall survival of all patients (Fig. 6) and patients with muscle-invasive BC (≥pT2) (Fig. S6*C*). Kaplan–Meier, uni-, and multivariate analyses were performed on this dataset as shown in Fig. 6. In addition, analysis for the combination of KRT14, KRT5, and KRT20 was performed (Fig. S5).

Differentiation States Within Normal Urothelium. Using MiDReG analysis that is based on the Boolean implication relationships as shown in Fig. S2*A*, we hypothesized the expression patterns of the markers of differentiation within normal urothelium as shown in Fig. S2*B*. Heatmaps of the upstream keratins (KRT14, KRT16, and KRT6) are shown in Fig. S2*C* using the normal bladder tissue (n = 48) expression patterns from the Sanchez-Carbayo dataset. KRT14 immunohistochemistry was performed on FFPE normal bladder urothelium (Fig. S2*D*) and immunofluorescence was performed on the fresh frozen OCT-embedded fetal bladder tissue (Fig. S2*E*).

1. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30:207–210.
2. Irizarry RA, et al. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 31:e15.
3. Sahoo D, Dill DL, Tibshirani R, Plevritis SK (2007) Extracting binary signals from microarray time-course data. *Nucleic Acids Res* 35:3705–3712.
4. Sahoo D, Dill DL, Gentles AJ, Tibshirani R, Plevritis SK (2008) Boolean implication networks derived from large scale, whole genome microarray datasets. *Genome Biol* 9:R157.
5. Dyrskjøt L, et al. (2007) Gene expression signatures predict outcome in non-muscle-invasive bladder carcinoma: A multicenter validation study. *Clin Cancer Res* 13: 3545–3551.
6. Kim WJ, et al. (2010) Predictive value of progression-related gene classifier in primary non-muscle invasive bladder cancer. *Mol Cancer* 9:3.
7. Lindgren D, et al. (2010) Combined gene expression and genomic profiling define two intrinsic molecular subtypes of urothelial carcinoma and gene signatures for molecular grading and outcome. *Cancer Res* 70:3463–3472.
8. Sanchez-Carbayo M, Socci ND, Lozano J, Saint F, Cordon-Cardo C (2006) Defining molecular profiles of poor outcome in patients with invasive bladder cancer using oligonucleotide microarrays. *J Clin Oncol* 24:778–789.
9. Smyth GK, Speed T (2003) Normalization of cDNA microarray data. *Methods* 31: 265–273.
10. Ritchie ME, et al. (2007) A comparison of background correction methods for two-colour microarrays. *Bioinformatics* 23:2700–2707.
11. Sahoo D, et al. (2010) MiDReG: A method of mining developmentally regulated genes using Boolean implications. *Proc Natl Acad Sci USA* 107:5732–5737.
12. Chan KS, et al. (2009) Identification, molecular characterization, clinical prognosis, and therapeutic targeting of human bladder tumor-initiating cells. *Proc Natl Acad Sci USA* 106:14016–14021.



Name	Affy ID	Chungbuk ID	HR	p value	c
KRT14	209351_at	ILMN_1665035	2.752625	0.01833	*
KRT6B	209126_x_at	ILMN_1721354	3.481557	0.010818	*
KRT16	209800_at	ILMN_1736760	2.176623	0.057495	.
KRT6A	214580_x_at	ILMN_1754576	1.780939	0.429693	

Fig. S1. Computational prediction of changes in keratin expression and their relationship to differentiation states in bladder cancer. Keratin gene names are abbreviated as K. (*B*) K5 expression is mutually exclusive with K20 in normal urothelium and bladder cancer (BC), as revealed by our previously published data using immunofluorescence staining in tissue sections (*A*). (*C–E*) This relationship is consistent with the Boolean relationship $K5^{high} \rightarrow K20^{low}$ across multiple species (human, mouse, rat), and diverse tissue samples including 75,000 data points. (*A*) Also, on the basis of previously published BC biology, CD44 enriches for tumor-initiating cells in BC and $CD44^{+}$ BC cells (Alexa 594/red) express KRT5 (Alexa 488/green), but not KRT20 (Alexa 488/green) ($CD44^{+}$ cells are highlighted with white

Legend continued on following page

dashed line), whereas downstream CD44^{-} cells express KRT20, but not KRT5 (KRT5^{+} cells are highlighted with white dashed line), we hypothesize that K5^{+} cells are early progenitors upstream of K20^{+} cells. (*B*) Here, we propose to predict upstream progenitor keratin KX using an algorithm, known as MiDReG (mining developmentally regulated genes). This algorithm searches for genes that satisfy KX^{high}→ K5^{high} and KX^{high} → K20^{low} Boolean implication relationships. (*F*) A total of 137 genes (16 keratins) satisfy these Boolean implication criteria that narrowed down to 7 keratins on the basis of dynamic range (>5) and SD (>1.5) in the AffyBC dataset. Heatmaps of these 7 keratins are shown in the AffyBC and 6 keratins in the Chungbuk dataset (keratin 6C is not present in the dataset). BC patient survival association of the predicted upstream keratins (hazard ratio and *P* value) was computed using survival analysis of the KX^{high} and KX^{low} patient groups in the Chungbuk dataset. K13 and K4 have noisy gene expression patterns in the heatmap. (*G*) K14, K16, K6B, and K6A satisfy the required Boolean implication criteria. This analysis revealed that K14 is the best and the most consistent marker of differentiation according to the Boolean implication relationship (*G*), the expression patterns in the heatmaps, and the patient outcome (*F*).

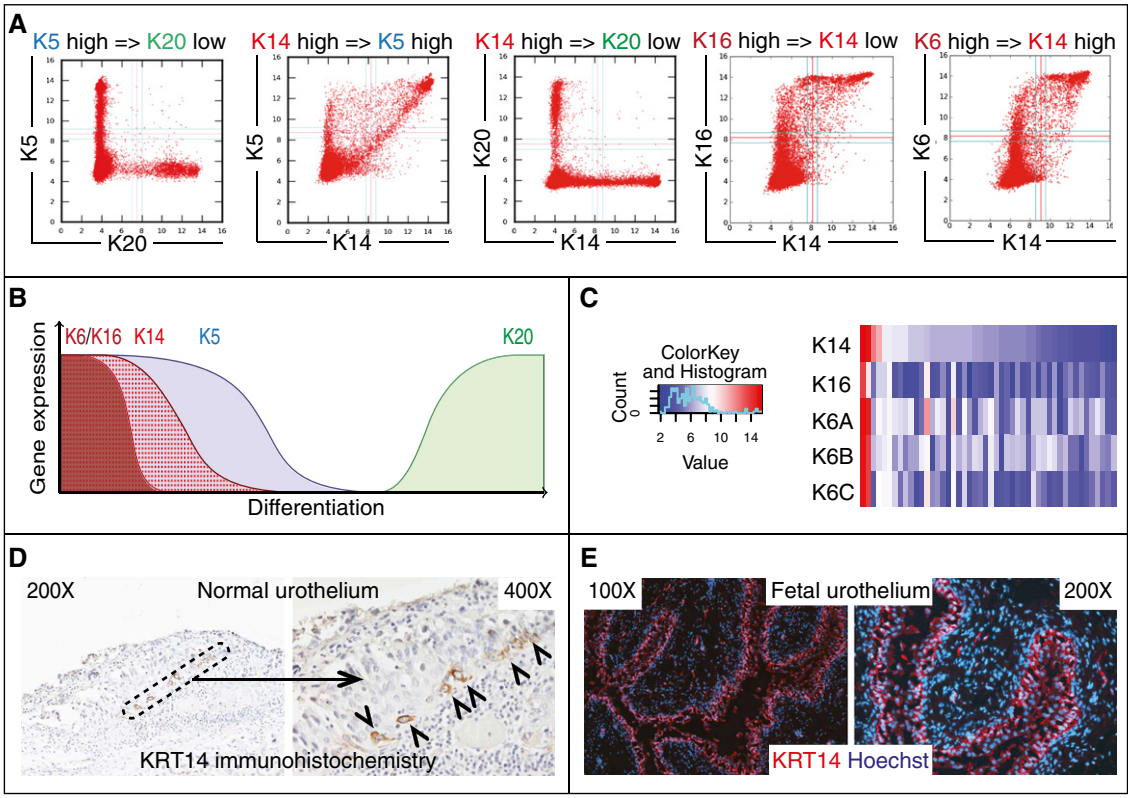


Fig. S2. Differentiation states in normal urothelium. Keratin gene names are abbreviated as K. (*A*) Boolean relationship indicates that K5 expression is mutually exclusive with K20. K14 is coexpressed with K5, whereas K5 can be expressed without K14, and K14 expression is mutually exclusive to K20 expression. These gene expression patterns indicate that K14 is developmentally upstream of K5. K16 and K6 are coexpressed with K14, whereas K14 can be expressed without K16 and K6. This indicates that K16 and K6 are developmentally upstream of K14. (*B*) Schematic illustrating of predicted keratin differentiation states in normal urothelium. (*C*) Heatmap showing the expression patterns of K14, K16, K6A, K6B, and K6C in normal urothelium in the Sanchez-Carbayo dataset ($n = 41$), and the distribution is consistent with predicted differentiation states. (*D*) Immunohistochemistry shows basal localization of KRT14^{+} cells in adult urothelium (paraffin embedded). (*E*) Immunofluorescence staining shows intensive, predominantly basal staining for KRT14 in fetal urothelium (OCT-embedded frozen sections).

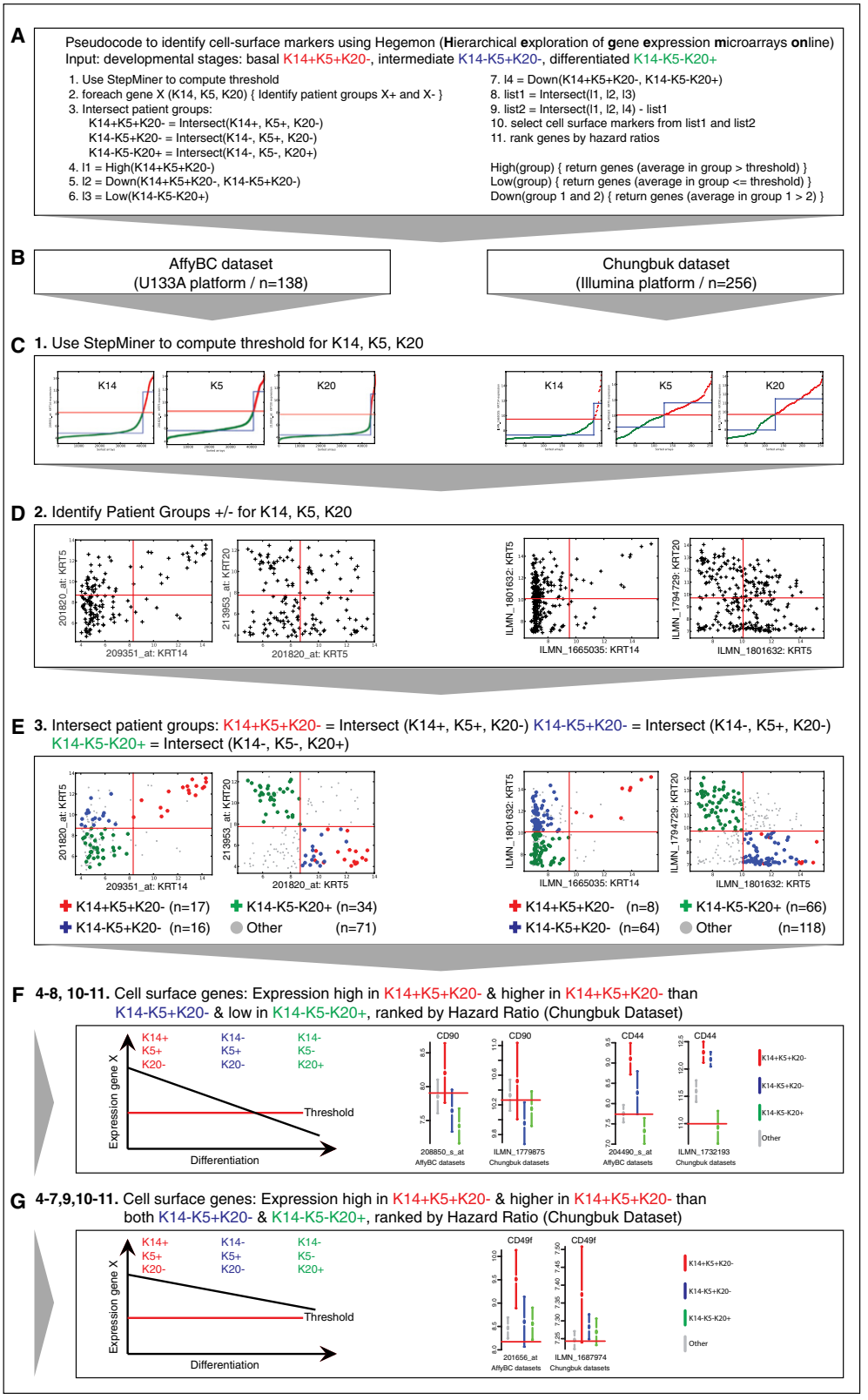


Fig. S3. Identification of surface markers corresponding to the keratin differentiation states using Hegemon. (*A*) Pseudocode of the surface marker identification that takes the input as the keratin differentiation states and return two lists (list 1 and list 2) of genes differentially expressed between the differentiation states. (*B*) The two different datasets used are the AffyBC dataset and the Chungbuk dataset in Affymetrix Human U133A and Illumina microarray platforms, respectively. (*C*) StepMiner is used to compute threshold [above threshold (red): high, below threshold (green): low] for each gene (probeset) for each microarray platform. For Affymetrix Human U133A platform 45,563 publicly available microarrays are used to compute threshold for each probeset. For Illumina platform only Chungbuk dataset (256 microarrays) is used to compute threshold. (*D*) Scatterplots of log2 normalized gene expression between KRT14,

Legend continued on following page

KRT5, and KRT20 with the StepMiner thresholds in the AffyBC (n = 138) and the Chungbuk (n = 256) bladder cancer datasets. (*E*) $K14^{+}K5^{+}K20^{-}$ (red, AffyBC n = 17, Chungbuk n = 8), $K14^{-}K5^{+}K20^{-}$ (blue, AffyBC n = 16, Chungbuk n = 64), $K14^{-}K5^{-}K20^{+}$ (green, AffyBC n = 34, Chungbuk n = 66) patient groups selected and highlighted in respective colors in the scatterplots. (*F*) Schematics of gene expression patterns (corresponds to list 1 of the pseudocode) that are high in $K14^{+}K5^{+}K20^{-}$ (red), down-regulated in $K14^{-}K5^{+}K20^{-}$ (blue), and low in $K14^{-}K5^{-}K20^{+}$ (green) patient groups. (*G*) Schematics of gene expression patterns (corresponds to list 2 of the pseudocode) that are high in $K14^{+}K5^{+}K20^{-}$ (red), down-regulated in both $K14^{-}K5^{+}K20^{-}$ (blue) and $K14^{-}K5^{-}K20^{+}$ (green) patient groups but not all of the way to low in $K14^{-}K5^{-}K20^{+}$ as in *F*. Boxplots with mean and confidence interval are shown for CD90, CD44, and CD49f for different patient groups in their respective colors (gray color indicates "other" patients). The detailed list of cell surface markers (both list 1 and list 2) are presented in Dataset S1.

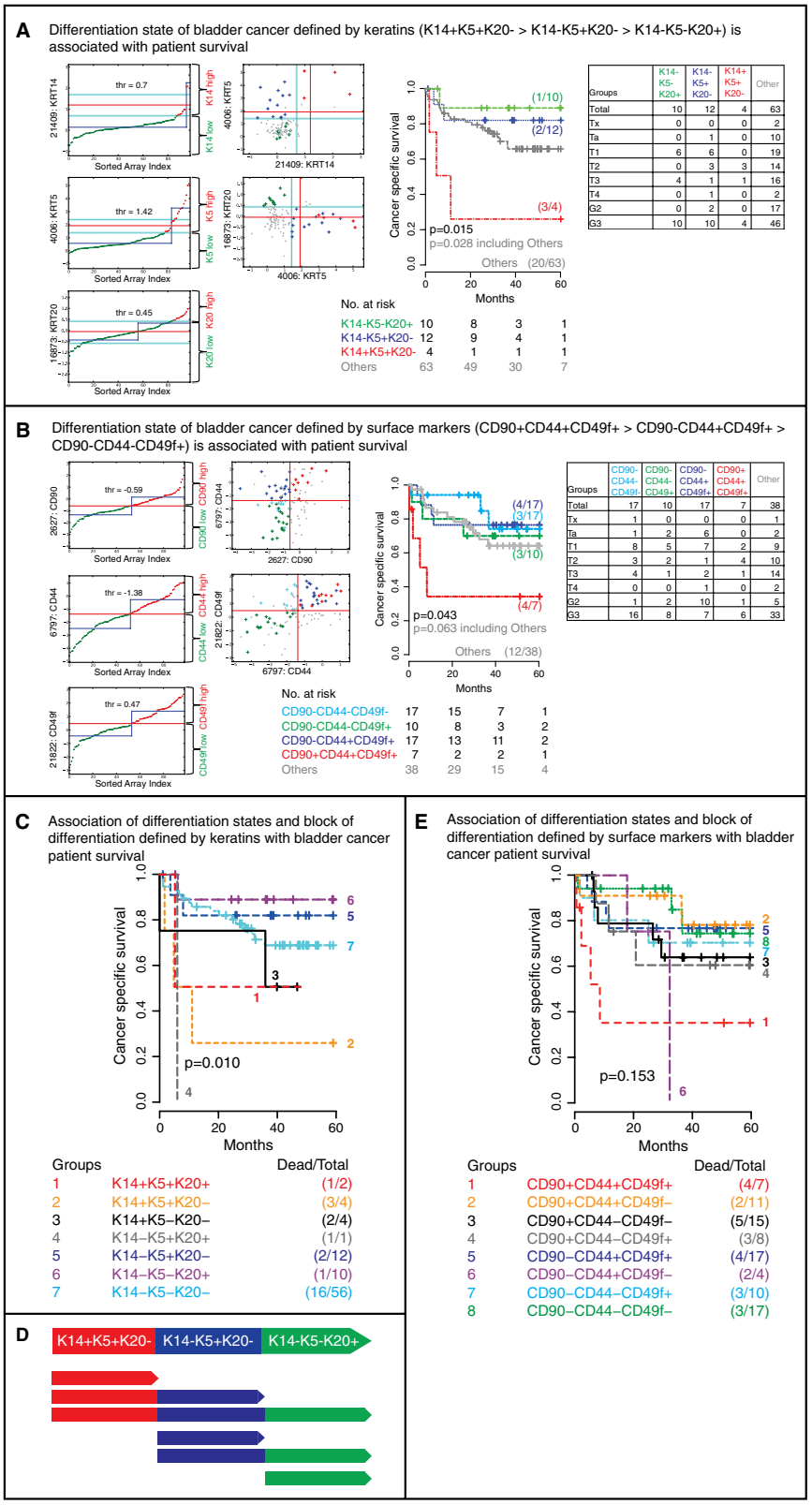
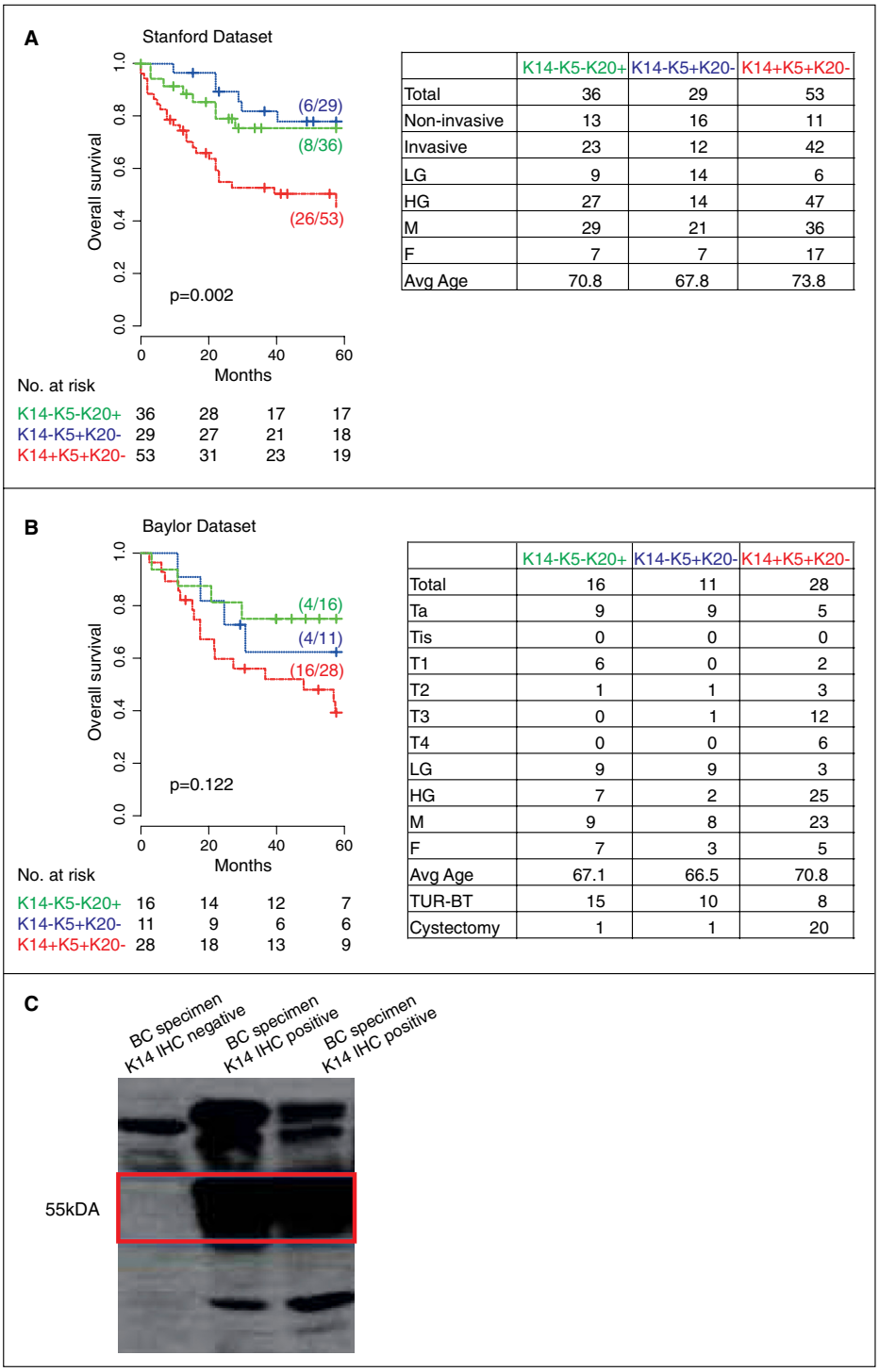


Fig. S4. Lindgren dataset (SWEGENE H_v3.0.1 35K platform). Differentiation state of BC defined by keratins ($K14^+K5^+K20^-$ → $K14^-K5^+K20^-$ → $K14^-K5^-K20^+$) surface markers ($CD90^+CD44^+CD49f^+$ → $CD90^-CD44^+CD49f^+$ → $CD90^-CD44^-CD49f^+$) is associated with patient survival. (*A*, *Left*) Identification of high and low expression values for each gene using StepMiner. Scatterplots of the selected patient groups ($K14^+K5^+K20^-$, red; $K14^-K5^+K20^-$, blue; $K14^-K5^-K20^+$, green; and other, gray) according to the differentiation states. (*Center*) Kaplan–Meier survival curves of all patient groups. The ratio on the *Right* of Kaplan–Meier shows the incidence proportions (1 – survival proportions). *P* values are computed on the basis of the log-rank test both including and excluding the "other" patient group. (*Right*) Clinical and pathological annotations for each patient group. The noise margin is 0.5 above and below (cyan lines) the StepMiner threshold (red

Legend continued on following page

line), which corresponds to a twofold change in gene expression. For KRT14 and KRT5, a stringent low threshold (cyan line, 0.5 below red line) and for KRT20 a stringent high threshold (cyan line, 0.5 above red line) is used to collect more patients in the basal ($K14^+K5^+K20^-$, red) and the intermediate ($K14^-K5^+K20^-$, blue) BC patient groups. (*B*, *Left*) Identification of high and low expression values for each gene using StepMiner. Scatterplots of the selected patient groups ($CD90^+CD44^+CD49f^+$, red; $CD90^-CD44^+CD49f^+$, blue; $CD90^-CD44^-CD49f^+$, green; $CD90^-CD44^-CD49f^-$, cyan; and other, gray) according to the differentiation states. (*Center*) Kaplan–Meier survival curves of all patient groups. The ratio on the right of Kaplan–Meier shows the incidence proportions (1 – survival proportions). *P* values are computed on the basis of the log-rank test both including and excluding the "other" patient group. (*Right*) Clinical and pathological annotations for each patient group. (*C*) Association of differentiation states and block of differentiation defined by keratins with bladder cancer patient survival: Kaplan–Meier survival curves of all patient groups defined by KRT14, KRT5, and KRT20 (including others, Fig. S3*A*). The ratio on the *Right* of Kaplan–Meier shows the incidence proportions (1 – survival proportions). *P* values are computed on the basis of the log-rank test. (*D*) Schematics of differentiation states and block of differentiation in BC. (*E*) Association of differentiation states and block of differentiation defined by surface markers with bladder cancer patient survival: Kaplan–Meier survival curves of all patient groups defined by CD90, CD44, and CD49f. Kaplan–Meier survival curves of all patient groups (including others, Fig. S3*B*). The ratio on the *Right* of Kaplan–Meier shows the incidence proportions (1 – survival proportions). *P* values are computed on the basis of the log-rank test.



	K14-K5-K20+	K14-K5+K20-	K14+K5+K20-
Total	36	29	53
Non-invasive	13	16	11
Invasive	23	12	42
LG	9	14	6
HG	27	14	47
M	29	21	36
F	7	7	17
Avg Age	70.8	67.8	73.8

	K14-K5-K20+	K14-K5+K20-	K14+K5+K20-
Total	16	11	28
Ta	9	9	5
Tis	0	0	0
T1	6	0	2
T2	1	1	3
T3	0	1	12
T4	0	0	6
LG	9	9	3
HG	7	2	25
M	9	8	23
F	7	3	5
Avg Age	67.1	66.5	70.8
TUR-BT	15	10	8
Cystectomy	1	1	20

Fig. S5. Stanford and Baylor BC tissue dataset (KRT14/KRT5/KRT20 immunohistochemistry). Association of differentiation states (basal, intermediate, and differentiated) with patient survival in bladder cancer. In (*A*) Stanford and (*B*) Baylor tissue databases, patient tumors were stratified by immunohistochemical analysis of KRT14, KRT5, and KRT20 protein expression (positive/negative) into three groups: basal ($K14^+K5^+K20^-$, red), intermediate ($K14^-K5^+K20^-$, blue), and differentiated ($K14^-K5^-K20^+$, green). Kaplan–Meier survival curves of the three patient groups and the *P* values based on the log-rank test are shown. The ratio on the *Right* of Kaplan–Meier shows the incidence proportions (1 – survival proportions). Clinical and pathological annotations for each patient group are shown on the *Right*. (*C*) Western blot analysis of keratin 14 (K14) protein expression in BC specimens, which are positive/negative for K14 in IHC analysis.

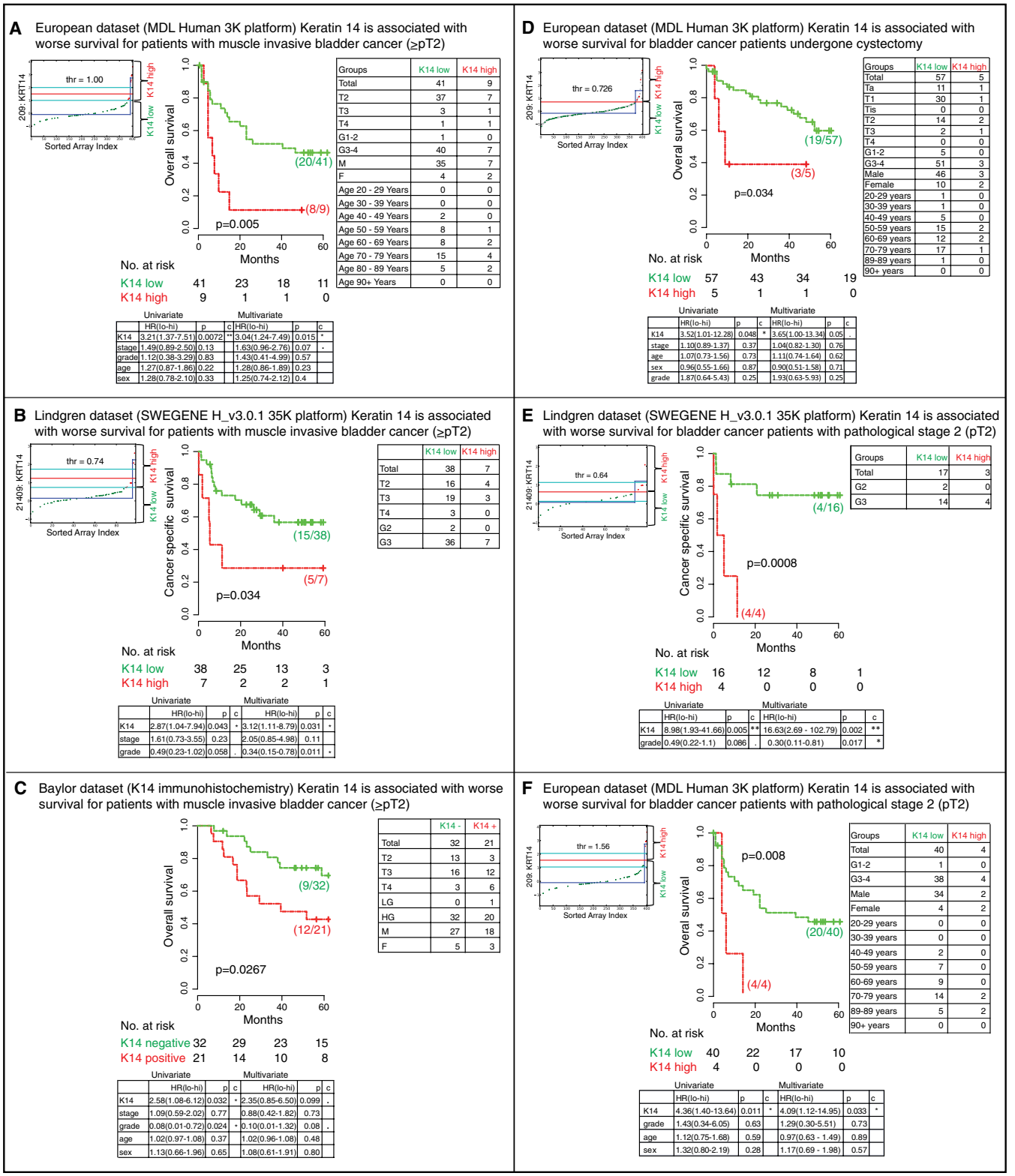
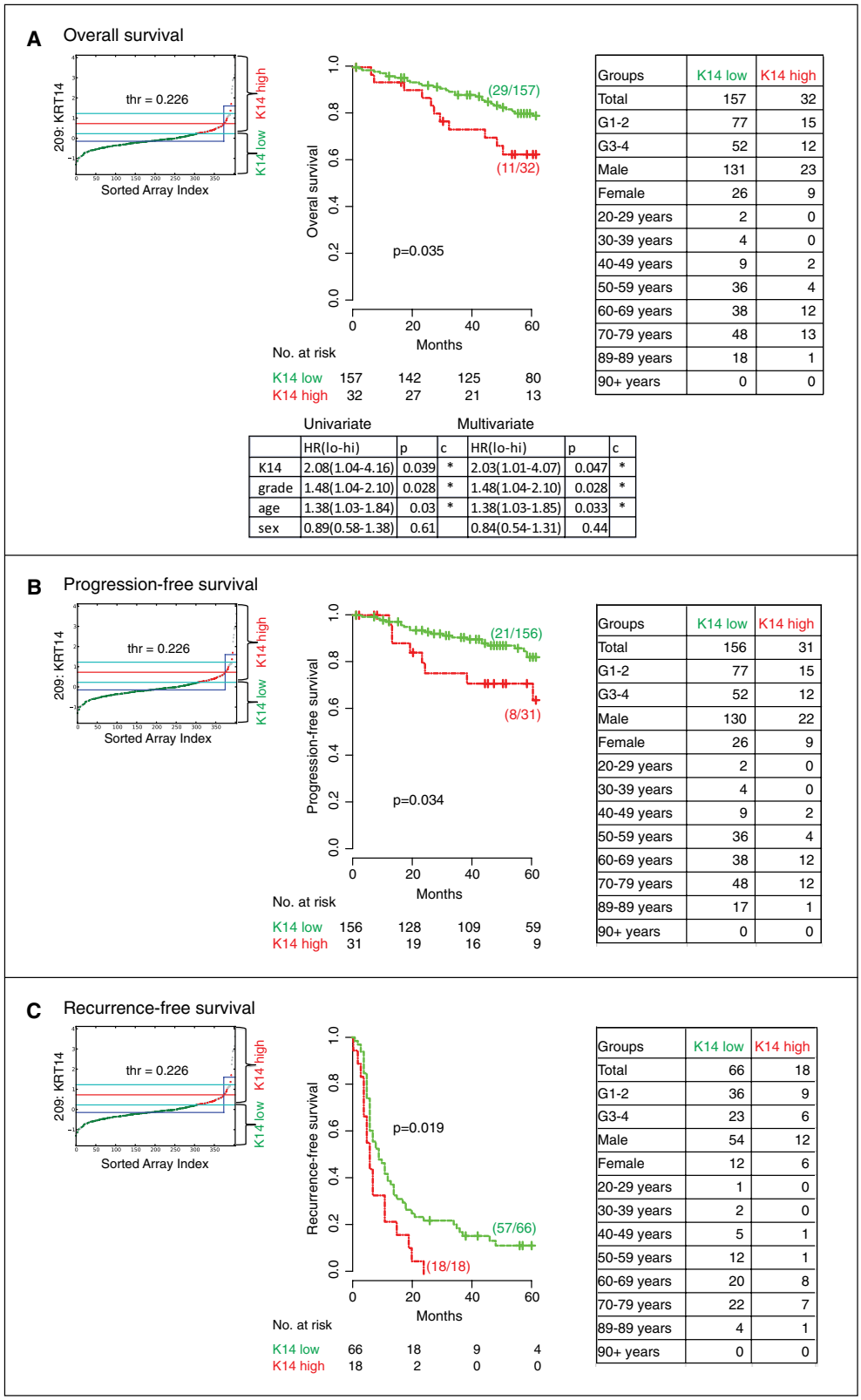


Fig. S6. Keratin 14 is associated with worse survival for bladder cancer patients with muscle-invasive bladder cancer and patients who have undergone cystectomy. (*A*–*C*) Patients with muscle-invasive BC stage ≥pT2: (*A*) KRT14 gene expression levels for bladder cancer patients are divided into two groups (high, red and low, green) in the European dataset. The threshold (red line) was computed by applying StepMiner on all 403 BC data. Kaplan–Meier survival curves of both patient groups and the *P* values based on the log-rank test are shown in the *Center*. Clinical and pathological annotations for each patient group are shown on the *Right*. The table at the *Bottom* shows uni- (*Left*) and multivariate (*Right*) results using Cox regression analysis (HR, hazard ratio; lo-hi, confidence interval; *P*, *P* value; *C*, significance status by number of asterisks). (*B*) KRT14 gene expression levels for bladder cancer patients are divided into two groups (high, red and low, green) in the Lindgren dataset. The threshold (red line) was computed by applying StepMiner on all 89 BC data. Kaplan–Meier survival curves of both patient groups and the *P* values based on the log-rank test are shown in the *Center*. Clinical and pathological annotations for each patient group

Legend continued on following page

are shown on the *Right*. The table at the *Bottom* shows uni- (*Left*) and multivariate (*Right*) results using Cox regression analysis (same abbreviations as above.). The *P* values shown are based on the log-rank test. The ratio on the *Right* of Kaplan–Meier shows the incidence proportions (1 – survival proportions). (*C*) K14 protein expression levels for BC patients are divided into two groups (positive, red and negative, green) in the Baylor tissue dataset. Kaplan–Meier survival curves of all of the patient groups and the *P* values based on the log-rank test are shown in the *Center*. Clinical and pathological annotations for each patient group are shown on the *Right*. The table at the *Bottom* shows uni- (*Left*) and multivariate (*Right*) results using Cox regression analysis (same abbreviations as above). (*D*) Patients who have undergone cystectomy: KRT14 gene expression levels for bladder cancer patients who have undergone bladder removal (cystectomy) are divided into two groups (high, red and low, green) in the European dataset. The threshold (red line) was computed by applying StepMiner on all 403 BC data. Kaplan–Meier survival curves of both patient groups and the *P* values based on the log-rank test are shown in the *Center*. Clinical and pathological annotations for each patient group are shown on the *Right*. The *Bottom Right* table shows uni- (*Left*) and multivariate (*Right*) results using Cox regression analysis (same abbreviations as above). The *P* values shown are based on the log-rank test. The ratio on the *Right* of Kaplan–Meier shows the incidence proportions (1 – survival proportions). (*E* and *F*) Patients with muscle-invasive BC with stage pT2: KRT14 expression levels are divided into two groups (high, red and low, green) in two independent datasets (*E*, Lindgren; *F*, European) in two different microarray platforms (*E*, SWEGENE H_v3.0.1 35K; *F*, MDL Human 3K). The threshold (red line) was computed by applying StepMiner on only the muscle-invasive (pT2) BC patient data. Kaplan–Meier survival curves of the two patient groups and the *P* values based on the log-rank test are shown in the *Center*. The ratio on the *Right* of Kaplan–Meier shows the incidence proportions (1 – survival proportions). Clinical and pathological annotations for each patient group are shown on the *Right*. The table at the *Bottom* shows uni- (*Left*) and multivariate (*Right*) results using Cox regression analysis (same abbreviations as above).



	HR(lo-hi)	p	c	HR(lo-hi)	p	c
K14	2.08(1.04-4.16)	0.039	*	2.03(1.01-4.07)	0.047	*
grade	1.48(1.04-2.10)	0.028	*	1.48(1.04-2.10)	0.028	*
age	1.38(1.03-1.84)	0.03	*	1.38(1.03-1.85)	0.033	*
sex	0.89(0.58-1.38)	0.61		0.84(0.54-1.31)	0.44	

Groups	K14 low	K14 high
Total	157	32
G1-2	77	15
G3-4	52	12
Male	131	23
Female	26	9
20-29 years	2	0
30-39 years	4	0
40-49 years	9	2
50-59 years	36	4
60-69 years	38	12
70-79 years	48	13
89-89 years	18	1
90+ years	0	0

Groups	K14 low	K14 high
Total	156	31
G1-2	77	15
G3-4	52	12
Male	130	22
Female	26	9
20-29 years	2	0
30-39 years	4	0
40-49 years	9	2
50-59 years	36	4
60-69 years	38	12
70-79 years	48	12
89-89 years	17	1
90+ years	0	0

Groups	K14 low	K14 high
Total	66	18
G1-2	36	9
G3-4	23	6
Male	54	12
Female	12	6
20-29 years	1	0
30-39 years	2	0
40-49 years	5	1
50-59 years	12	1
60-69 years	20	8
70-79 years	22	7
89-89 years	4	1
90+ years	0	0

Fig. S7. European dataset (MDL human 3K platform). Keratin 14 is associated with worse overall, progression-free, and recurrence-free survival for patients with nonmuscle-invasive bladder cancer (pTa). KRT14 expression levels for the early stage (pTa) BC patients are divided into two groups in the European dataset: high (red) and low (green). (*A*, overall survival; *B*, progression-free survival; and *C*, recurrence-free survival) Kaplan–Meier survival curves of both patient groups and the *P* values based on the log-rank test are shown in the *Center*. The ratio on the *Right* of Kaplan–Meier shows the incidence proportions (1 – survival proportions). Clinical and pathological annotations for each patient group are shown on the *Right*. (*A*, overall survival) Uni- (*Left*) and multivariate (*Right*) results using Cox regression analysis is shown for the overall survival (see Fig. S4 legend for abbreviations). (*A*–*C*) Stringent low threshold (0.226, cyan line, 0.5 below red line) is used, which is 0.5 below the StepMiner threshold (red line, computed using all 403 microarrays).

Table S1. European dataset (MDL human 3K platform)

	Univariate			Multivariate		
	HR (lo-hi)	*P*	*C*	HR (lo-hi)	*P*	*C*
Analysis exclusive of intravesical treatment (mitomycin/bacillus Calmette–Guérin)						
KRT14	1.52 (1.18–1.98)	0.0015	**	1.37 (1.07–1.76)	0.013	*
Stage	1.37 (1.26–1.48)	<0.0001	***	1.3 (1.18–1.43)	<0.0001	***
Grade	1.84 (1.42–2.39)	<0.0001	***	1.33 (1.01–1.75)	0.043	*
Age	1.47 (1.24–1.73)	<0.0001	***	1.45 (1.23–1.71)	<0.0001	***
Sex	1.01 (0.81–1.25)	0.96		1.01 (0.81–1.26)	0.94	
Analysis inclusive of intravesical treatment (mitomycin/bacillus Calmette–Guérin)						
KRT14	1.64 (0.99–2.71)	0.056	.	1.75 (1.09–2.81)	0.02	*
Stage	1.28 (1.11–1.480	0.00076	***	1.05 (0.87–1.27)	0.59	
Grade	2.69 (1.38–5.24)	0.0037	**	2.27 (1.14–4.50)	0.019	*
Age	1.26 (0.93–1.70)	0.13		1.33 (0.99–1.79)	0.059	.
Sex	0.94 (0.61–1.47)	0.79		0.78 (0.49–1.23)	0.28	
Intrav. mitomycin	1.07 (26–4.47)	0.93		0.78 (0.17–3.49)	0.74	
Intrav. bacillus Calmette–Guérin	3.01 (1.58–5.76)	0.00085	***	2.59 (1.20–5.58)	0.015	*

Keratin 14 gene expression analyzed as continuous variable is associated with significantly worse patient overall survival. Multivariate analysis of KRT14 using continuous KRT14 gene-expression values. Uni- (*Left*) and multivariate (*Right*) results using Cox regression analysis. HR, hazard ratio; lo-hi, confidence interval; *P*, *P* value; *C*, significance status by number of asterisks. (0 if *P* value = 0; *** if $P > 0$ & $P \leq 0.001$; ** if $P > 0.001$ & $P \leq 0.01$; * if $P > 0.01$ & $P \leq 0.05$; . [period] if $P > 0.05$ & $P \leq 0.1$; [blank entry] if $P > 0.1$).

Dataset S1. Computationally predicted cell surface markers

Dataset S1 (XLS)

(*A*) Cell surface genes. Expression high in $K14^+K5^+K20^-$ and higher in $K14^+K5^+K20^-$ than $K14^-K5^+K20^-$ and low in $K14^-K5^-K20^+$, ranked by hazard ratio (Chungbuk dataset). A list of cell surface genes (corresponding to the list 1 of the pseudocode in Fig. S2) whose gene expressions are high in basal BC ($K14^+K5^+K20^-$), higher in the basal BC than the intermediate BC ($K14^-K5^+K20^-$), and low in the differentiated BC ($K14^-K5^-K20^+$), ranked by hazard ratio (Chungbuk dataset). The columns in the excel sheet are: HR, hazard ratio; lo-hi, confidence interval; *P*, *P* value; *C*, significance status by number of asterisks; name, gene name; AffyID, Affymetrix probeset ID; E1, gene expression difference between the basal BC and the differentiated BC in the AffyBC dataset; E2 gene expression difference between the basal BC and the differentiated BC in the Chungbuk dataset; KimID, Illumina probesets in the Chungbuk dataset; Desc, description of the gene name. (*B*) Cell surface genes. Expression high in $K14^+K5^+K20^-$ and higher in $K14^+K5^+K20^-$ than both $K14^-K5^+K20^-$ and $K14^-K5^-K20^+$, ranked by hazard ratio (Chungbuk dataset). A list of cell surface genes (corresponding to the list 2 of the pseudocode in Fig. S2) that are not present in Dataset S1, whose gene expression are high in basal BC ($K14^+K5^+K20^-$), higher in the basal BC than both the intermediate BC ($K14^-K5^+K20^-$), and the differentiated BC ($K14^-K5^-K20^+$), ranked by hazard ratio (Chungbuk dataset). The columns in the excel sheet are: HR, hazard ratio; lo-hi, confidence interval; *P*, *P* value; *C*, significance status by number of asterisks; name, gene name; AffyID, Affymetrix probeset ID; E1, gene expression difference between the basal BC and the differentiated BC in the AffyBC dataset; E2 gene expression difference between the basal BC and the differentiated BC in the Chungbuk dataset; KimID, Illumina probesets in the Chungbuk dataset; Desc, description of the gene name.