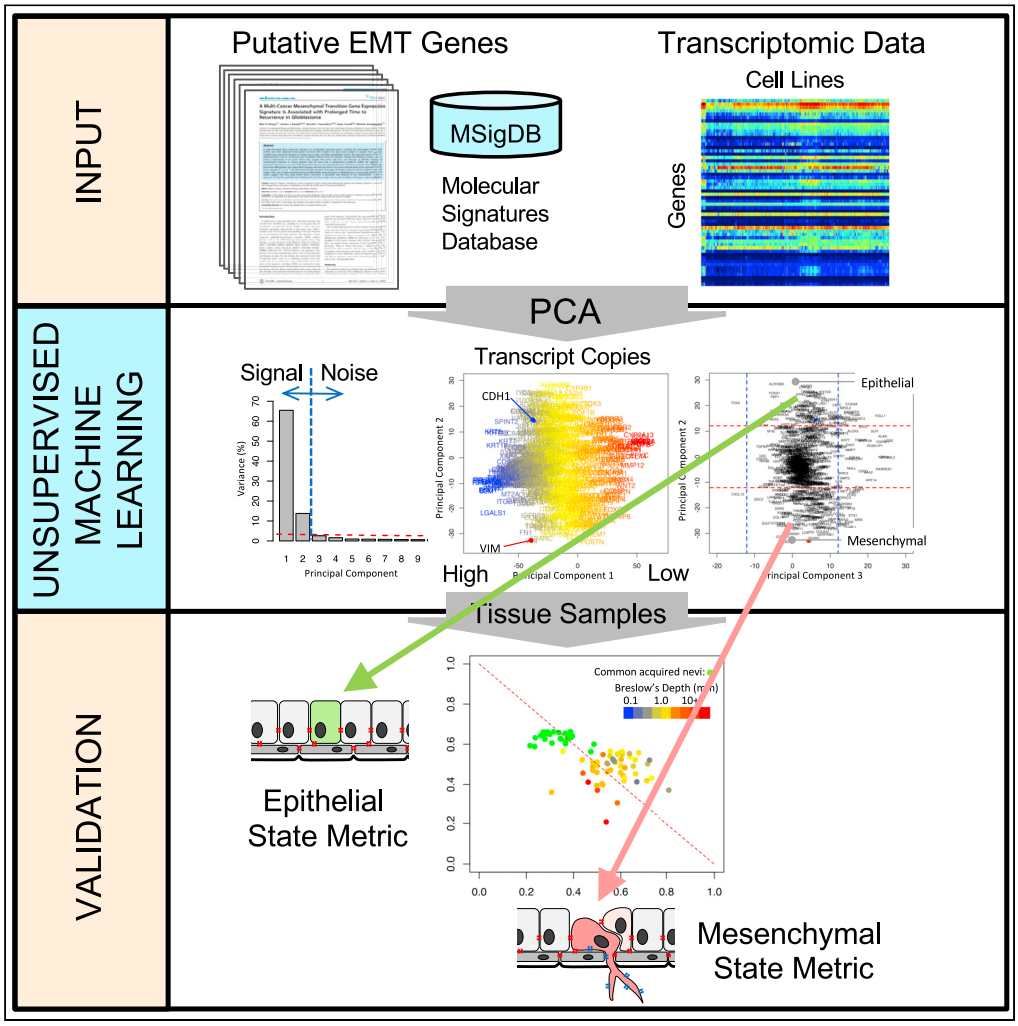
iScience

CellPress
OPEN ACCESS

Article

An Unsupervised Strategy for Identifying Epithelial-Mesenchymal Transition State Metrics in Breast Cancer and Melanoma



David J. Klinke II, Arezo Torang

david.klinke@mail.wvu.edu

HIGHLIGHTS

Unsupervised strategy to generate epithelial and mesenchymal state metrics

Refined metrics for use with bulk RNA-seq data by removing normal fibroblasts genes

Validated state predictions against independent measures of metastatic potential

Breast cancer and melanoma share more common genes in de-differentiated metrics

Klinke & Torang, iScience *23*, 101080
May 22, 2020
https://doi.org/10.1016/j.isci.2020.101080

iScience

CellPress
OPEN ACCESS

Article

An Unsupervised Strategy for Identifying Epithelial-Mesenchymal Transition State Metrics in Breast Cancer and Melanoma

David J. Klinke II[1,2,3,6,*] and Arezo Torang[4,5]

SUMMARY

Digital cytometry aims to identify different cell types in the tumor microenvironment, with the current focus on immune cells. Yet, identifying how changes in tumor cell phenotype, such as the epithelial-mesenchymal transition, influence the immune contexture is emerging as an important question. To extend digital cytometry, we developed an unsupervised feature extraction and selection strategy to capture functional plasticity tailored to breast cancer and melanoma separately. Specifically, principal component analysis coupled with resampling helped develop gene expression-based state metrics that characterize differentiation within an epithelial to mesenchymal-like state space and independently correlate with metastatic potential. First developed using cell lines, the orthogonal state metrics were refined to exclude the contributions of normal fibroblasts and provide tissue-level state estimates using bulk tissue RNA-seq measures. The resulting metrics for differentiation state aim to inform a more holistic view of how the malignant cell phenotype influences the immune contexture within the tumor microenvironment.

INTRODUCTION

Tissues are composed of a diverse set of different cell types that help maintain homeostasis. Oncogenesis is associated with a shift in the cellular composition of a tissue that can be revealed with increasing confidence through direct measurement, such as single-cell RNA sequencing (scRNA-seq), or using digital methods to deconvolute bulk tissue samples (Newman et al., 2019). Given the correlation with response to immunotherapies, the current focus has been on quantifying immune cell types present within the tumor microenvironment (Thorsson et al., 2018; Tirosh et al., 2016). There is also an increasing appreciation for characterizing the heterogeneity among malignant cells that may arise in the same anatomical location (Shannan et al., 2016; Koren and Bentires-Alj, 2015). Given our interest in understanding functional heterogeneity of malignant cells that originate within a particular anatomical organ rather than uncertainty in etiology, we will focus on breast cancer and melanoma as Li et al. show that melanoma and breast cancer cell lines seem to cluster most uniformly, whereas other cell lines defined by anatomical origin seem to have a more heterogeneous composition (Li et al., 2017).

Although the tumor cells that arise in the skin and breast seem to be most similar, patient treatment strategies and outcomes can be diverse. Initial treatment strategies are guided by specific molecular alterations that can be targeted by drugs: aromatase inhibitors for ER-positive breast cancer, anti-HER2 antibodies for HER2-positive breast cancer, or small molecule inhibitors for BRAF V600E-positive or C-KIT-positive melanoma (Taghian et al., 2019; Sosman, 2019). Luckily, these two cancers are diagnosed in greater than 60% of patients while the malignant cells are confined to the organ of origin (Siegel et al., 2019). However, dissemination of primary tumors to vital organs like liver, brain, and lungs is a key limiter for patient survival in breast cancer and melanoma. Specifically, the 5-year survival rate for patients with localized disease versus distant metastases drops from 98% to 23% and from 99% to 27% for melanoma and breast cancer, respectively (American Cancer Society, 2019). In contrast, patient survival for tumors that originate in vital organs is limited by the degree to which malignant cells locally disrupt organ function. Thus, the importance of distal dissemination in determining patient outcomes can vary based on the tissue of origin.

[1]Department of Chemical and Biomedical Engineering, West Virginia University, Morgantown, WV, USA

[2]Department of Microbiology, Immunology and Cell Biology, West Virginia University, Morgantown, WV, USA

[3]WVU Cancer Institute, West Virginia University, Morgantown, WV, USA

[4]Amsterdam UMC, University of Amsterdam, Laboratory for Experimental Oncology and Radiobiology, Center for Experimental and Molecular Medicine, Cancer Center Amsterdam, Amsterdam, the Netherlands

[5]Oncode Institute, UMC, University of Amsterdam, Amsterdam, the Netherlands

[6]Lead Contact

*Correspondence: david.klinke@mail.wvu.edu

https://doi.org/10.1016/j.isci.2020.101080

The distal dissemination and growth of malignant cells—metastasis—is a complex process thought to involve dynamic re-engagement of biological processes used during development that enable migrating cells to form tissues. For carcinomas, initiating metastasis is thought to occur through a process called the epithelial-mesenchymal transition (EMT). EMT is the functional consequence of engaging a genetic regulatory network (GRN) that downregulates the expression of genes associated with an epithelial phenotype and upregulates genes associated with a mesenchymal phenotype. Breast carcinoma primarily originates from either luminal epithelial cells or basal myoepithelial cells within the mammary gland (Zhang et al., 2017). In contrast to breast cancer, melanoma arises from the oncogenic transformation of melanocytes, which follow a different developmental trajectory along the neural crest than epithelial cells and also involves a process similar to EMT (Regad, 2013). Although much of cell specification is imprinted epigenetically via DNA methylation and histone modifications, significant functional changes, such as modifications in cell state due to EMT, may occur within these epigenetic constraints. To characterize cell state based on gene expression, supervised methods have been predominantly used for developing gene signatures that characterize the EMT. Although effective, supervised methods can perform poorly if the strategy is based on misinformation, such as sample misclassification or prior biases as to the number of cell states or defining genes. Although used less frequently, unsupervised methods for feature extraction and selection are advantageous as they can be data driven (Taguchi, 2017). Here, our objective was to use an unsupervised strategy to develop a gene signature capturing this functional plasticity that is tailored to the specific cellular context of breast cancer and melanoma individually, as an illustrative example.

RESULTS

RNA Sequencing Provides an Estimate of Protein Abundance

As extracting metrics to quantify cellular state using bioinformatic approaches depends on the quantity and quality of the underlying information, combining repositories of microarray- and RNA-sequencing (RNA-seq)-based transcriptomic profiling of tissue samples could provide a rich trove of data to mine. We first asked whether assaying the same genes using different transcriptomics profiling platforms provides the same information. To do this, we compared gene expression levels assayed by either Agilent microarray or by Illumina RNA-seq for the same samples (Figure 1). Expression values obtained by RNA-seq are in units of transcripts per million (TPM), whereas the Agilent microarray results are in terms of intensity units. Using samples obtained as part of the breast cancer arm of the TCGA, we focused on genes that have been associated with host immunity, as these genes are likely to span a broad dynamic range within these samples. As the TCGA samples are obtained from homogenized bulk samples of tumor and matched normal breast tissue, expression of these genes could be from the malignant cells, like GATA3 expression by breast cancer cells, or from immune cell infiltrates, like the potential expression of IL4 and IL5 by infiltrating T helper type 2 cells.

Generally, comparing the same row across the two panels illustrates the poor correspondence between transcript abundance assayed using Agilent microarrays and read counts (TPM) obtained by RNA-seq. A subset of genes, like HLA-DRA and HLA-DPA1, exhibit both high microarray intensity units and read counts, whereas other genes, like TBX21 and FASLG, exhibit high microarray intensity units but have low read counts. In addition, the dynamic range observed among these samples is different depending on the platform used, as illustrated in the heatmap by TBX21 and IL17F. Using Illumina RNA-seq, TBX21 is constrained to the low end of the color spectrum (dark to royal blue), whereas the dynamic range spans the middle to upper end of the color spectrum (green to red) when assayed using Agilent microarray. Similarly, IL17F transcripts were not detected by RNA-seq in 87% of the samples but the Agilent microarray shows a rather high average intensity with variation among the samples. The difference in average intensities among genes and in variance among samples assayed by these two platforms suggest that the information provided by these two platforms is not entirely the same. The poor correspondence between Agilent two-channel microarray and RNA-seq data has been attributed to differences in ratio (Agilent two-channel microarray) versus non-ratio (RNA-seq) representations of transcript abundance by the platforms (Guo et al., 2013).

We next asked whether the assayed transcript abundance corresponds to protein abundance. First, we compared RNA-seq counts reported for cell lines associated with the Cancer Cell Line Encyclopedia (CCLE) with protein abundance for the same cell lines measured using Reverse Phase Protein Array (RPPA). We filtered the respective datasets to those cell lines that were reported in both datasets and for genes where there was a positive correlation coefficient greater than 0.36 between read counts in TPM and normalized RPPA values. From the initial datasets, 288 cell lines and 99 genes were retained

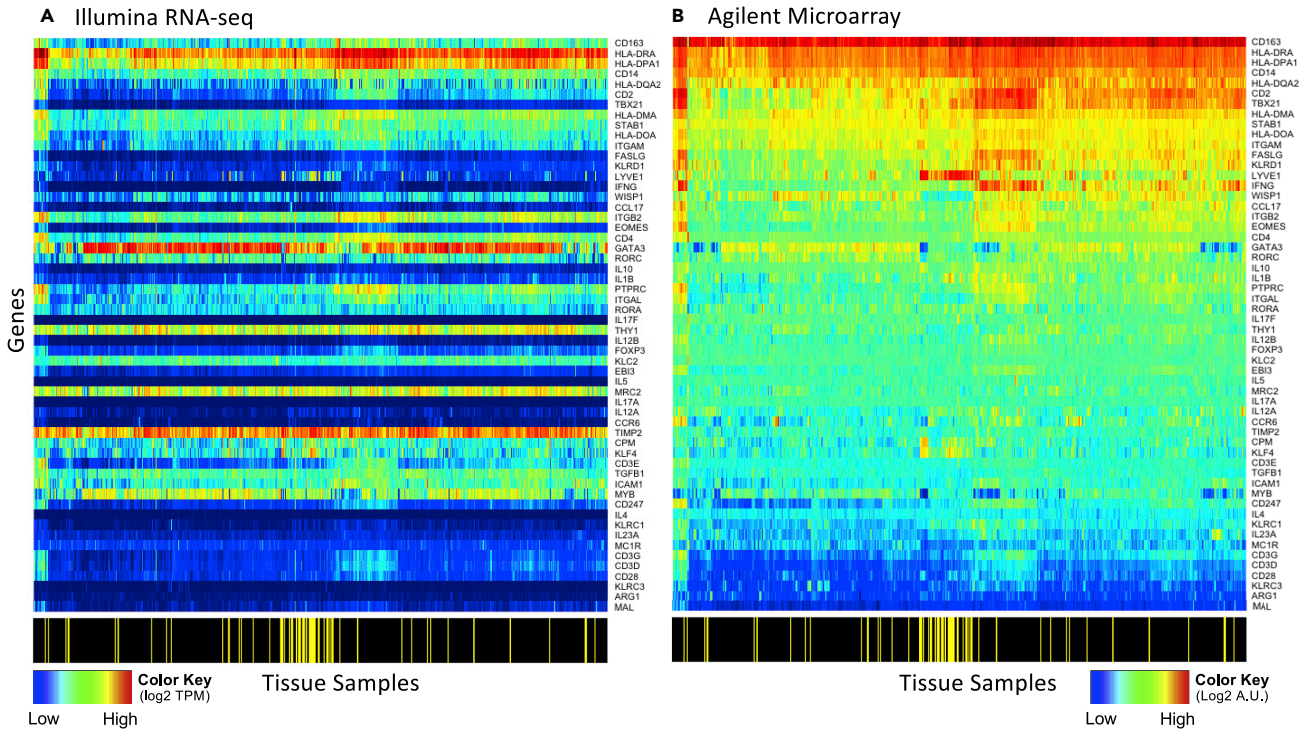


Figure 1. Comparison of Gene Expression within the Same Samples Assayed using RNA Sequencing and Oligonucleotide Microarray

Heatmaps for the expression of a subset of genes in the breast cancer arm of the TCGA study assayed using Illumina RNA-seq (A) and using Agilent microarray (B). Color bar shown at the bottom of the heatmaps indicates samples obtained from tumor tissue (black) versus matched normal tissue (yellow). The genes and samples are similarly ordered in both panels. Values were log2 normalized.

for analysis after filtering. Next, we determined whether the pairs of mRNA and protein measurements share a common value for steady-state transcript abundance that corresponds to steady-state protein abundance measured above background. To do this, we applied a protein expression model to each gene measured across the cell lines where protein abundance was assumed to be a saturable function of transcript abundance (Figure 2A). Using the fitted curve, the threshold of transcript abundance for detecting a change in protein abundance 2.5% above background was back calculated. Example datasets and the corresponding curve fits for the genes CLDN7, AXL, JAG1, and CDH1 are shown in Figure 2B. Interestingly, the median value in the distribution of calculated threshold values was around 1 TPM (1.47, Figure 2C). We repeated this analysis for transcript abundance assayed by Affymetrix U133+2 microarray, a single-channel approach, using robust multi-array average (RMA)-normalized expression values, where 149 genes were retained for analysis after filtering. Qualitatively, data obtained using the single-channel platform (Affymetrix) exhibited better correspondence with RPPA values than data obtained using Agilent's two-channel platform. However, the distribution in calculated threshold values were more broadly distributed compared with the RNA-seq results (see red line in Figure 2C, F-test p value = 6.92×10^{-7}). These results imply that the transcript abundance assayed by RNA-seq provides a higher-quality estimate of protein abundance, that is the signal-to-noise ratio is improved, compared with data obtained using a single-channel microarray platform. Moreover, simple strategies for combining data acquired using different platforms, such as centering across genes, or applying gene signatures cross-platform to interpret new samples warrants caution.

Collectively, the common threshold value observed using RNA-seq data has two implications. First, there are some genes that have a high sensitivity of detection using microarrays such that the observed changes may not be functionally important. From Figure 1, it seems that IL17F, TBX21, FASLG, KLRD1, IFNG, CCL17, and IL10 are but a few examples (i.e., high Agilent microarray intensity but very low read counts) in that dataset. Without knowing the detection sensitivity by microarray, traditional approaches using a *Z* score metric may give equal weight to changes in gene expression driven by a biological signal as to changes dominated by random noise. Second, the threshold value provides a rationale for filtering genes that

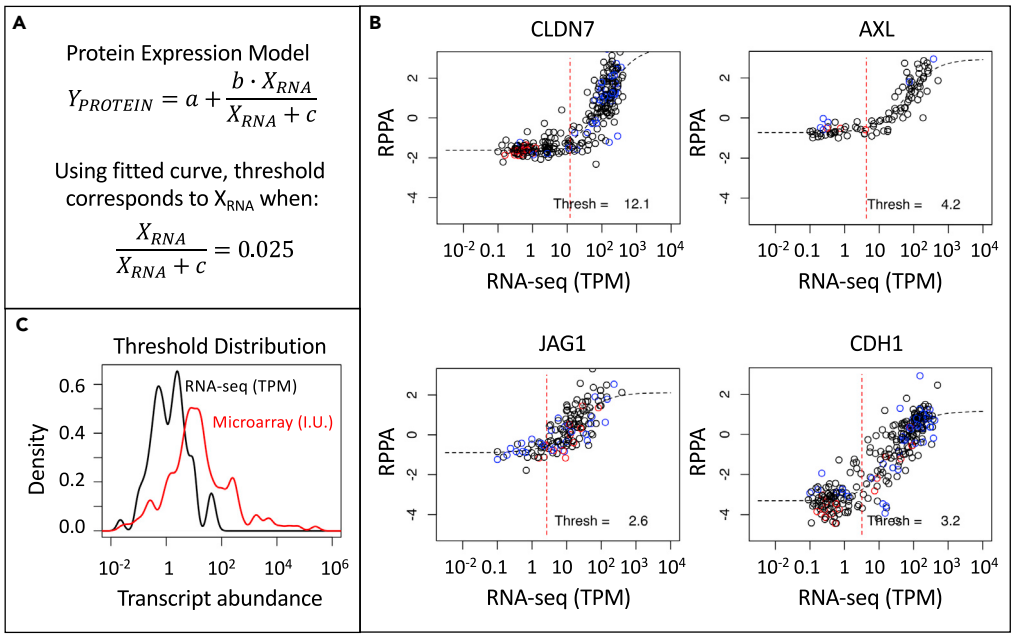


Figure 2. RPPA Measurements Were Used to Determine a Threshold for Biologically Significant Changes in Gene Expression

(A and B) The model for protein dependence on gene expression (A) where representative data (black circles) and model fits (dotted black line) are shown for CLDN7, AXL, JAG1, and CDH1 (B).

(C) The distribution in threshold values calculated for all genes assayed by RNA-seq (black curve, n = 99) and by Affymetrix microarray (red curve, n = 149) meeting the inclusion criteria. Transcript abundance units for RNA-seq corresponds to TPM and intensity units (I.U.) for Affymetrix microarray. In (B), the vertical red dotted line indicates the threshold value and the melanoma and breast cancer cell lines are highlighted by red and blue circles.

are likely to have a low information content when developing gene signatures for phenotypes that are not well defined.

Gene Expression Patterns in Breast Cancer Cells Are Captured by a Single Component

Given the variety of breast cancer subtypes reported in the literature, we next asked how many different GRNs are at work in breast cancer. GRNs associated with development commonly contain transcription factors that interact via positive feedback such that the target genes are either co-expressed or expressed in a mutually exclusive fashion (Alon, 2007). Given the interest in functional responses, we are focusing on patterns of gene expression in response to signal processing by the GRNs rather than trying to identify their topology. In motivating this study, we made four assumptions. First, we assumed that oncogenic mutations alter the peripheral control of GRN but do not alter the core network topology, where signals processed by a GRN change cell phenotype by engaging a unique gene expression pattern. Second, malignant cells derived from a particular anatomically defined cancer represent the diverse ways that hijacking these GRNs can provide a fitness advantage to malignant cells within the tumor microenvironment. Third, culturable tumor cell lines represent a sampling of these ways in which GRNs are hijacked in a particular anatomical location. Fourth, the process of isolating these malignant cells from tumor tissue to generate culturable cell lines does not bias this view. It follows then that the number of different GRNs can be identified by analyzing the transcriptional patterns of genes likely to participate in GRNs among an ensemble of tumor cells lines that share a common tissue of origin. We focused our attention on 780 genes that have been previously associated with the EMT and related gene sets in MSigDB v4.0. (Sarrio et al., 2008; Carretero et al., 2010; Alonso et al., 2007; Cheng et al., 2012; Tan et al., 2014; Kaiser et al., 2016; Deng et al., 2019, 2020) and analyzed the expression of these genes among 57 breast cancer cell lines included in the CCLE database as assayed by RNA-seq using a feature extraction/feature selection workflow summarized in Figure 3. To identify coordinately expressed genes, we used principal component analysis (PCA), a linear statistical approach for unsupervised feature extraction and selection that enables the unbiased discovery of clusters of genes that exhibit coherent patterns of expression (i.e., features) that are independent of other gene clusters (Jolliffe and Cadima, 2016). The relative magnitude of the resulting gene expression patterns can be inferred from the eigenvalues,

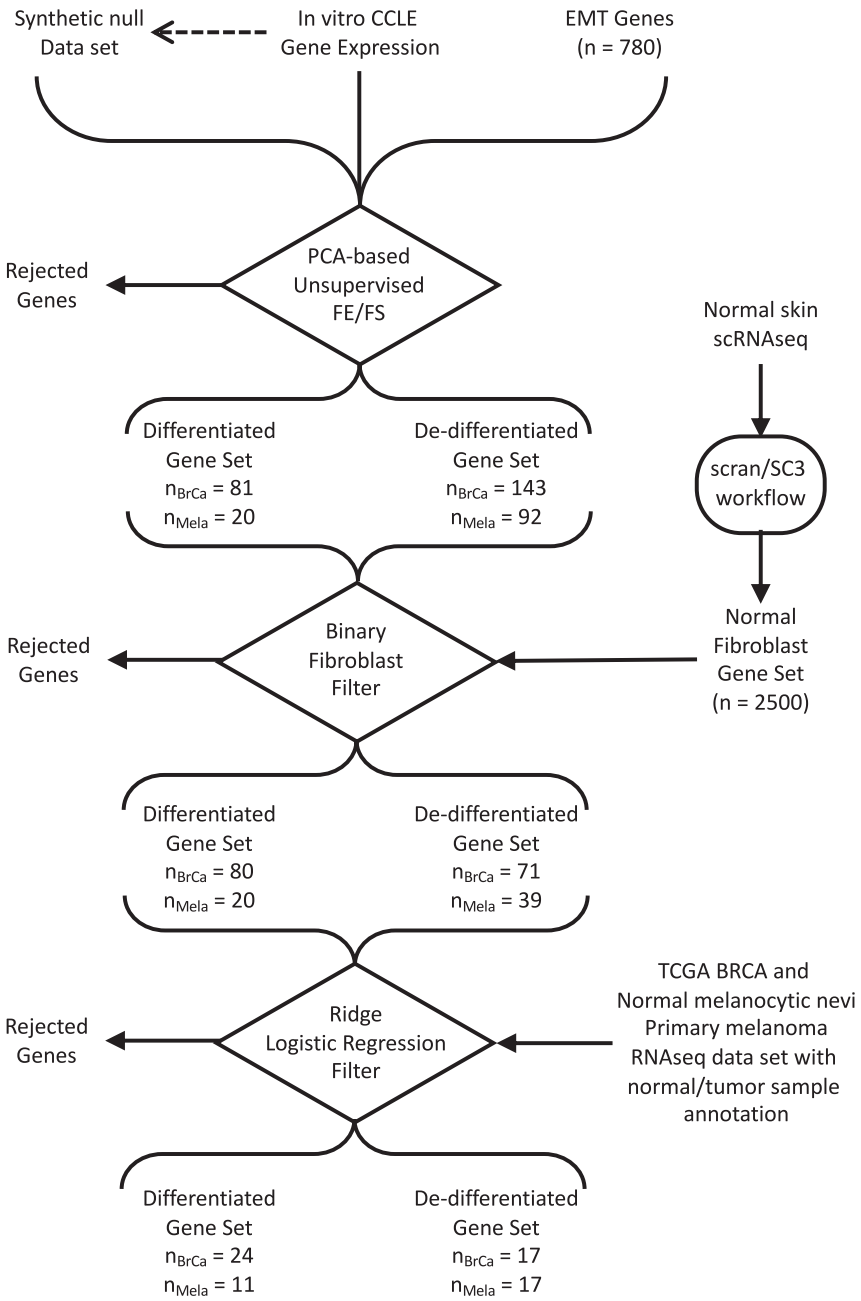


Figure 3. Data Workflow for Identifying Epithelial/Differentiated versus Mesenchymal/De-differentiated State Metrics

Workflow contains three decision points: unsupervised feature extraction (FE)/feature selection (FS) based on PCA, a binary fibroblast filter, and a consistency filter based on Ridge logistic regression of annotated samples.

which represent the extent of the data's covariance captured by a specific principal component. To facilitate comparisons among datasets, we represent the eigenvalues as the percent of total sum over all of the eigenvalues or, simply, percent variance, which is shown in Figure 4. Specifically, PC1 and PC2 captured 66% and 14% of the variance, respectively. Additional principal components each captured less than 3% of the variance.

One of the challenges with PCA is that no clear rules exist to determine how many principal components to consider, such as a gap statistic in clustering (Tibshirani et al., 2001). To select an appropriate number of PCs (i.e., features), we established a threshold for determining significance relative to a null distribution.

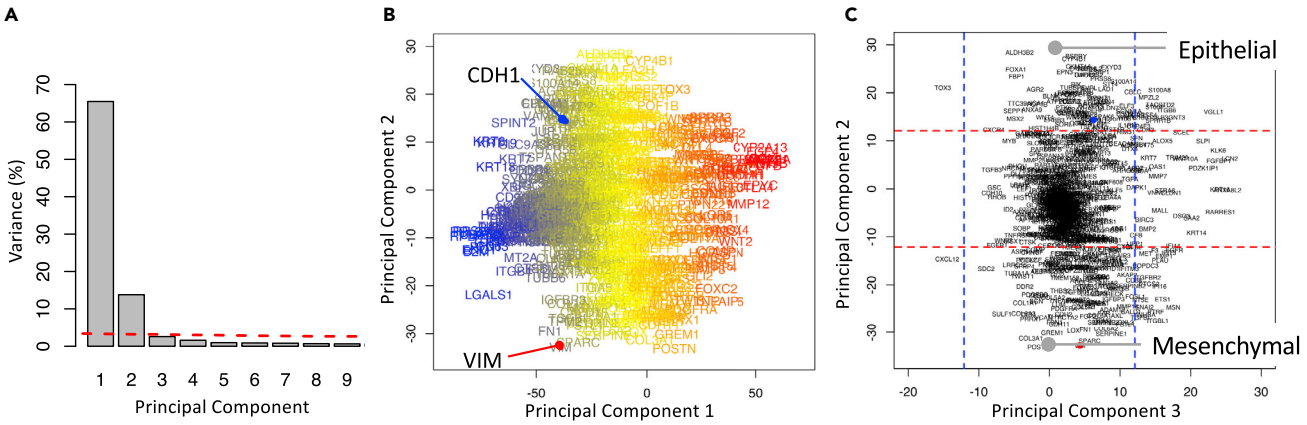


Figure 4. Two Opposing Gene Signatures Were Identified among the Cohort of Breast Cancer Cell Lines

(A) Scree plot of the percentage of variance explained by each principal component, where the dotted line corresponds to variance explained by the null principal components.

(B) Projection of the genes along PC1 and PC2 axes, where the font color corresponds to the mean read counts among cell lines (blue-yellow-red corresponds to high-medium-low read counts).

(C) Projection of the genes along PC2 and PC3 axes, where the dotted lines enclose 95% of the null PCA distribution along the corresponding axis.

Specifically, we applied the same PCA to a synthetic noise dataset generated from the original data by randomly resampling with replacement the collection of gene expression values and assigning the values to particular gene-cell line combinations. The resulting set of eigenvalues represent the values that could be obtained by random chance if the underlying dataset has no information, which are shown as the dotted red line in Figure 4A. In comparison with the null distribution, only the first two PCs were above the threshold. The variance captured by the remaining PCs were below the null PCA distribution suggesting that any potential biological interpretations of these additional PCs could also be explained by random chance. Therefore, we focused on the first two PCs.

As variance in read counts is proportional to abundance, gene projections along the PC1 axis were proportional to the average read counts of the corresponding gene among the samples. Measured transcript abundance is proportional to the basal gene expression associated with cell specification and technical artifacts associated with RNA-seq. Genes that were expressed above the 1-TPM threshold in more than 5% of the cell lines were retained for further analysis. For the breast cancer cell lines, this eliminated 26 genes from potential inclusion in the state metrics. Next, we focused on the projection of retained genes along principal components 2 and 3. The projections were annotated with horizontal and vertical dotted lines that enclose 95% of the projections from the null distribution. Although the majority of the genes were distributed around the origin, a subset of genes were projected along the extreme of the PC2 axis (outside of the dotted horizontal lines) and had no significant projection along the PC3 axis (inside of the dotted vertical lines). The list of genes associated with either the high PC2/null PC3 or the low PC2/null PC3 groups are listed in Table S1 and contained 143 and 81 genes, respectively. Of note, all of the genes excluded based on the 1-TPM threshold were projected within the null PC2 distribution. As the projection of Vimentin (VIM, red dot in Figure 4C) and E-cadherin (CDH1, blue dot in Figure 4C) was prototypical for these two groups of genes, the high PC2/null PC3 genes were annotated as a mesenchymal signature (i.e., a de-differentiated state) and the low PC2/null PC3 group were annotated as an epithelial signature (i.e., a terminally differentiated state). In contrast to supervised approaches that use Vimentin and E-cadherin as the basis to identify associated genes (e.g., Tan et al., 2014; Rokavec et al., 2017), the approach used here is unsupervised whereby the association of Vimentin and E-cadherin with these two opposite groups of genes emerges naturally from the data.

The Epithelial and Mesenchymal State Measures Stratify Intrinsic Subtypes of Breast Cancer and Metastatic Potential

Using these two sets of genes, we developed a state metric to quantify the extent of a gene expression signature associated with epithelial differentiation and mesenchymal de-differentiation using a normalized sum over all of the genes associated with a signature. Although the PCA results suggest that these two sets of genes are inversely related, the metrics were designed to represent each state independently such that

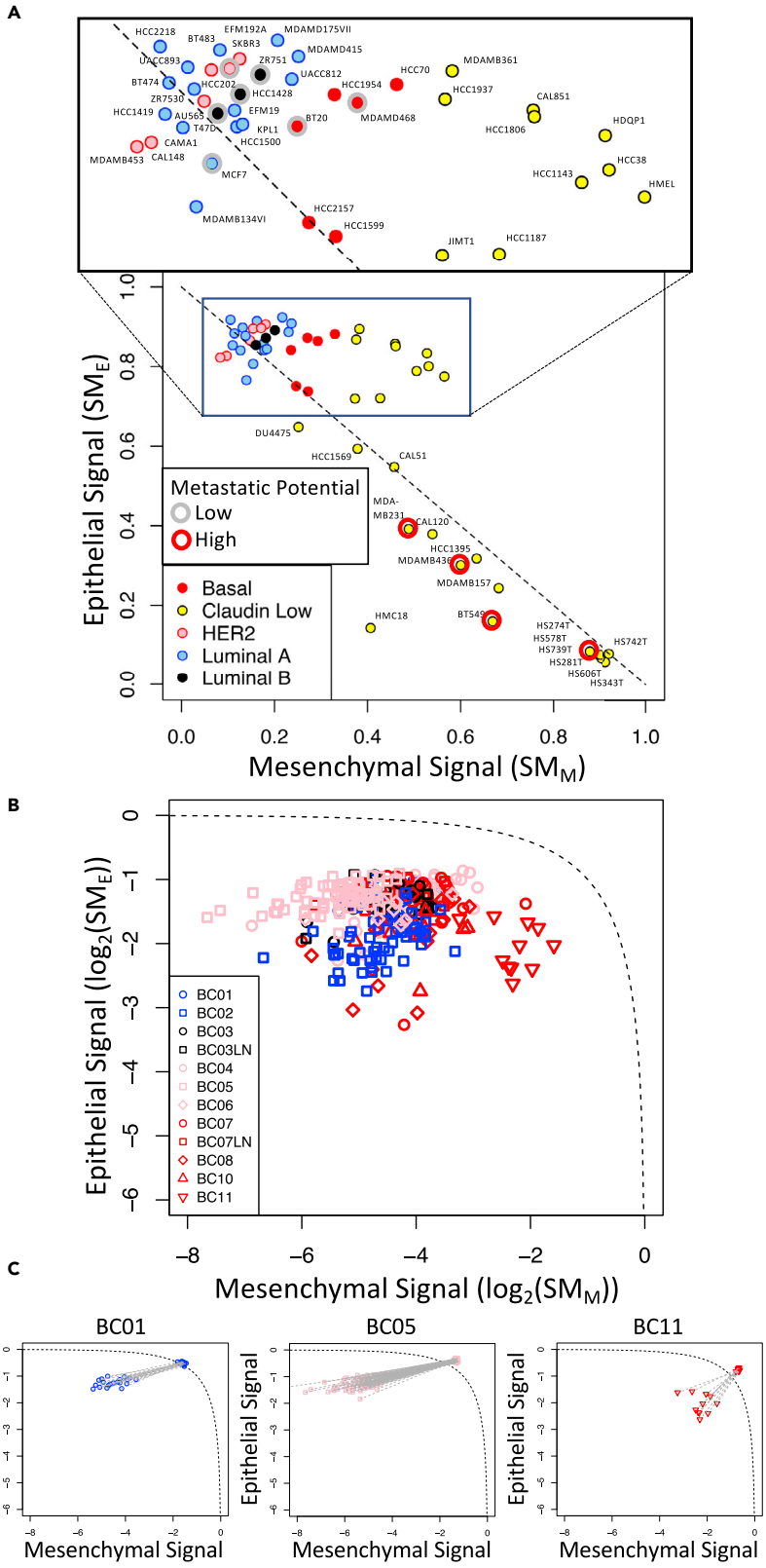

A
HCC2218
EFM192A
MDAMD175VII
BT483
SKBR3
MDAMD415
UACC893
ZR751
UACC812
HCC70
MDAMB361
BT474
HCC202
HCC1954
HCC1937
ZR7530
HCC1428
CAL851
HCC1419
AU565
EFM19
MDAMD468
HDQP1
T47D
BT20
HCC1806
KPL1
CAMA1
HCC1500
HCC38
MDAMB453
CAL148
HCC1143
MCF7
HMEL
HCC2157
MDAMB134VI
HCC1599
JIMT1
HCC1187
DU4475
HCC1569
CAL51
Metastatic Potential
Low
High
MDA-MB231
CAL120
HCC1395
MDAMB436
MDAMB157
HMC18
BT549
HS274T
HS578T
HS739T
HS742T
HS281T
HS606T
HS343T
Basal
Claudin Low
HER2
Luminal A
Luminal B
Epithelial Signal (SME)
Mesenchymal Signal (SMM)
0.0
0.2
0.4
0.6
0.8
1.0
B
BC01
BC02
BC03
BC03LN
BC04
BC05
BC06
BC07
BC07LN
BC08
BC10
BC11
Epithelial Signal (log2(SME))
Mesenchymal Signal (log2(SMM))
0
−1
−2
−3
−4
−5
−6
−8
C
BC01
BC05
BC11
Epithelial Signal
Mesenchymal Signal

Figure 5. The Different Subsets of Breast Cancer Were Clustered Along a Reciprocal Epithelial to Mesenchymal State Axes

(A and B) Log2 projections along the epithelial (SM_E) and mesenchymal (SM_M) state axes for each breast cancer cell line included in the CCLE (A) and primary breast cancer cells (B and C). Values for SM_E and SM_M were estimated by bulk RNA-seq data for cell lines associated with the CCLE and by scRNA-seq data for primary tumor cells (Chung et al., 2017).
(C) Log2 state projections are compared for primary breast cancer cells as originally reported and with dropout values imputed using the values averaged over the rest of the sample population, where gray lines connect the original state values to state values determine after imputation. Symbols were colored based on previously annotated breast cancer PAM50 subtypes: basal, red; claudin low, yellow; HER2, pink; luminal (A), blue; luminal (B), black. In (A), the metastatic potential of a subset of cell lines was annotated based on a recent study (Yankaskas et al., 2019): low metastatic potential, gray circle; high metastatic potential, red circle. The dotted line corresponds to a reciprocal relationship between the SM_E and SM_M state metrics (i.e., $SM_E = 1 - SM_M$).

cells that exhibit a pure phenotype would have values of 1 and 0 associated with their respective state metrics and cells with mixed phenotypes could potentially have values of 1 for both state metrics. Next, we calculated the state metric values for all of the breast cancer cell lines, where their projections in state space are shown in Figure 5. Interestingly, the breast cancer cell lines largely followed a linear reciprocal relationship between epithelial (E) and mesenchymal (M) states (dotted line in Figure 5) and were segregated by intrinsic PAM50 subtype (Parker et al., 2009). Although HER2, Luminal A, Luminal B, and Basal subtypes all have a high E signature, they progressively increased in their M signature. The Claudin Low subset spanned the greatest range with some expressing a high E and moderate M signatures (e.g., HCC1569, MDAMB361, HMEL) and others with a low E and high M signatures (e.g., BT549 and HS578T). Of note, a subset of the Claudin Low cell lines (e.g., HS742T, HS343T, HS281T, HS606T, and HS274T) with high M and very low E signatures have been suggested by the CCLE to be fibroblast-like (see Cell_lines_annotations_20181226.txt). Functionally, cells with low E and high M signatures had a high propensity for metastasis, whereas the propensity for metastasis was low in cell lines with high E and low M signatures (Yankaskas et al., 2019). This functional annotation also provided an external validation of the state metrics for breast cancer.

We next assessed the epithelial and mesenchymal state metrics in breast cancer cells assayed using scRNA-seq (Chung et al., 2017) (see Figure 5B). Similar to the cell lines, the samples were spread across the epithelial to mesenchymal spectrum roughly ordered by their corresponding intrinsic subtype, where HER2 subtype had a high E/low M signature and the basal subtype had the highest M signature without much of a reduction in their E signature. Overall, the state values were farther below the reciprocal trendline than any of the cell lines sampled. As gene-level reads by scRNA-seq are frequently missing (i.e., a dropout read) (Andrews and Hemberg, 2019), we imputed missing values to assess whether the distribution in the E/M state values were a result of read dropouts (see Figure 5C). Although read imputation shifted the cell state metrics toward the reciprocal trendline, the heterogeneity among the cell measurements was lost. Overall, it is unclear whether scRNA-seq measurements can be used to identify biological heterogeneity separately from heterogeneity introduced by technical limits of the assay.

Although single-cell methods are rapidly emerging as tool to assay human tissue samples, bulk transcriptomic assays of tumor tissue samples, like those acquired as part of the Cancer Genome Atlas (TCGA), are more abundant. More samples increase the statistical power for identifying clinical, cellular, and genetic correlates of the EMT. However, applying the epithelial/mesenchymal state metrics to interpret RNA-seq assays of bulk tumor tissue samples requires some additional filtering steps as bulk RNA-seq measurements averages over the heterogeneous normal and malignant cell types present within the tissue. In terms of a gene signature for the EMT, many of the genes commonly associated with acquiring mesenchymal function are also associated with fibroblasts, a relatively common cell type in epithelial tissues. Thus, an enrichment of genes associated with the EMT may be explained solely by a shift in the prevalence of fibroblasts within the tissue sample. Although the functional plasticity of fibroblasts within the tumor microenvironment is of increasing interest (e.g., Vickman et al., 2020), our goal here was to remove the contribution of normal fibroblasts from the state metrics. To deconvolute fibroblast genes from the state metrics, we obtained a list of 2,500 genes that were uniquely associated (Area under receiver operating characteristic curve >0.5) with a cluster annotated as fibroblasts using scRNA-seq data obtained from a digested normal skin sample obtained from a human female. This cluster contained about one-third of the cell samples measured within the CD45-negative population of the digested skin sample (see Figure S1). Although the majority of cells associated with cluster 1 were annotated as fibroblasts, a minor fraction of samples were annotated as "other." The similarity scores and associated hierarchical clustering dendrogram suggest that the gene

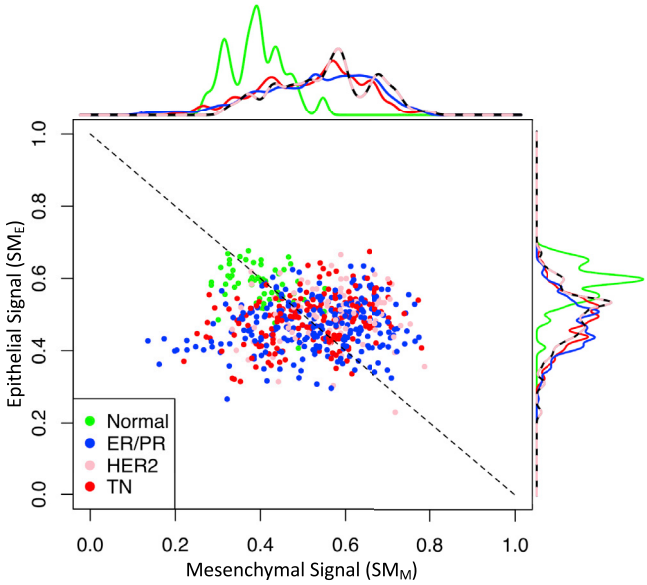


Figure 6. The Samples from Normal Breast Tissue and Breast Cancer Were Clustered Separately Along a Reciprocal Epithelial to Mesenchymal State Axes

Using EMT genes that passed the gene filter workflow, each sample contained within the breast cancer (BrCa) arm of the TCGA was projected along the epithelial (SM_E) versus mesenchymal (SM_M) state axes using the corresponding bulk RNA-seq data. Symbols were colored based on normal breast tissue (green) or clinical breast cancer subtype: ER/PR +, blue; HER2, pink; triple negative (TN), red. The dotted line corresponds to a reciprocal relationship between the SM_E and SM_M state metrics (i.e., $SM_E = 1 - SM_M$).

expression among cells in cluster 1 are very similar and that the cells included in cluster 1 can be considered as a uniform population. The interpretation of this is that, given the variability in scRNA-seq data, it is likely that some true fibroblasts assayed had no reads of COL1A1 or COL1A2. So, some of the cells annotated as "other" in cluster 1 are likely fibroblasts. Using this fibroblast gene list, overlapping genes were removed from the state metrics and highlighted in yellow in Table S1. All but one of the genes removed were contained within the mesenchymal gene list. In developing the overall approach, removing the contribution of normal fibroblasts was critical as projections of tissue samples in the EMT space without removing the contribution of fibroblasts were clustered around values of 1 for both state metrics.

Gene expression assayed from a bulk tissue sample reflects the combined contributions of non-malignant cells plus the changes induced by oncogenic transformation and reciprocal changes due to de-differentiation among malignant cells. Observable changes depend on the relative contributions of each cell source. As the unsupervised PCA analysis of the cell line data suggested that genes associated with EMT can be revealed by identifying a reciprocal pattern of gene expression, we performed Ridge logistic regression using the sample annotation to obtain regression coefficients for the list of EMT genes that passed the fibroblast filter (n = 151). The regression coefficients were used to filter the list of EMT genes for consistency with the reciprocal gene signature identified in the CCLE analysis. Genes that passed the consistency filter were used to define the epithelial and mesenchymal state metrics for bulk tissue samples. Of note, E-cadherin (CDH1) was still associated with the epithelial state metric, whereas N-cadherin (CDH2), Wnt-inducible signaling pathway protein 1 (WISP1/CCN4), and matrix metallopeptidase 3 (MMP3) were also retained in the mesenchymal state metric. The list of genes associated with the corresponding state metrics is given in Table S2.

Next, we projected the tissue samples obtained as part of the breast cancer arm of the TCGA in EMT space using the two tissue-based state metrics. Similar to the CCLE analysis, all samples clustered along the reciprocal SM_E versus SM_M line but exhibited greater dispersion. Samples obtained from normal breast tissue clustered separately from breast cancer samples (Figure 6), with normal breast tissue samples having the highest values for the epithelial state and lower values, on average, for the mesenchymal state. Among the different clinical breast cancer subtypes, the median value for SM_E progressively decreased from

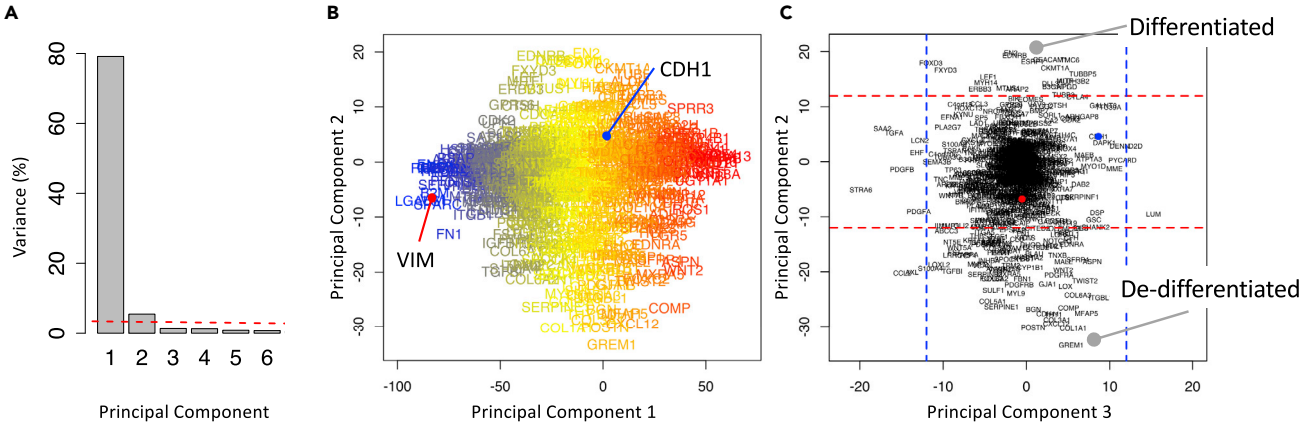


Figure 7. Two Opposing Gene Signatures Were Identified among the Cohort of Melanoma Cell Lines

(A) Scree plot of the percentage of variance explained by each principal component, where the dotted line corresponds to variance explained by the null principal components.

(B) Projection of the genes along PC1 and PC2 axes, where the font color corresponds to the mean read counts among cell lines (blue-yellow-red corresponds to high-medium-low read counts).

(C) Projection of the genes along PC2 and PC3 axes, where the dotted lines enclose 95% of the null PCA distribution along the corresponding axis.

HER2+, triple negative (TN: ER-/PR-/HER2-), and ER/PR+ (luminal) subtypes. The mesenchymal projections were about equal for both TN and ER/PR+ subsets and higher than the HER2+ samples. The higher values for the HER2+ samples along the SM_M axis align with clinical observations. For instance, patients with HER2+ subtype of breast cancer are at increased risk for developing metastatic lesions compared with TN and luminal subtypes (Kennecke et al., 2010). The two different state metrics seem to capture gene signatures that help anchor a cell to its designated location within the tissue and that promote active migration, respectively. In other words, reducing SM_E corresponds to raising the anchor and increasing SM_M corresponds to hoisting the sail. In summary, both cell-level and tissue-level EMT state metrics provide an estimate of metastatic potential and a digital measure of malignant cell differentiation state in the context of breast cancer.

Gene Expression Patterns in Melanoma Cells Are Also Captured by a Single Component

Using the same feature extraction/feature selection workflow as the breast cancer analysis (Figure 3), we applied PCA to the expression of EMT-related genes assayed in an ensemble of 56 melanoma cell lines associated with the CCLE (Figure 7). We focused on the first two principal components, PC1 and PC2, that captured 80% and 6% of the variance, respectively. Additional principal components each captured less than 4% of the variance. PC1 captured the variance associated with read abundance, as gene projections along the PC1 axis were proportional to the average read counts among the samples. Vimentin (VIM) and fibronectin (FN1) were two of the most highly expressed genes, whereas members of the Wnt family were some of the genes with low expression (e.g., WNT1, WNT6, WNT8B, WNT3A, WNT8A, WNT9B). Genes retained for further analysis were expressed above the 1-TPM threshold in more than 5% of the cell lines, which eliminated 78 genes from potential inclusion in the state metrics. These excluded genes were also projected within the null PC2 space.

Similar to the analysis of the breast cancer data, we focused on the projection of retained genes along PC2 and PC3 axes. Specifically, we developed state metrics around a subset of genes that were projected along the extreme of the PC2 axis and had no significant projection along the PC3 axis. The genes associated with either the high PC2/null PC3 or the low PC2/null PC3 groups are listed in Table S1 and contained 26 and 90 genes, respectively. In contrast to the breast cancer results, the projection of Vimentin (VIM, red dot in Figure 7C) and E-cadherin (CDH1, blue dot in Figure 7C) was not associated with either of these two groups of genes. As the high PC2/null PC3 group included MITF, a master regulator of melanocyte differentiation, and the low PC2/null PC3 group included a number of EMT-related genes (e.g., FN1, TCF4, ZEB1, TWIST2, and WISP1), these two gene sets were annotated as a terminally differentiated signature (i.e., an epithelial-like state) and a de-differentiated signature (i.e., a mesenchymal-like state), respectively. We noted that MITF was projected in the null PC2/null PC3 space in the breast cancer analysis.

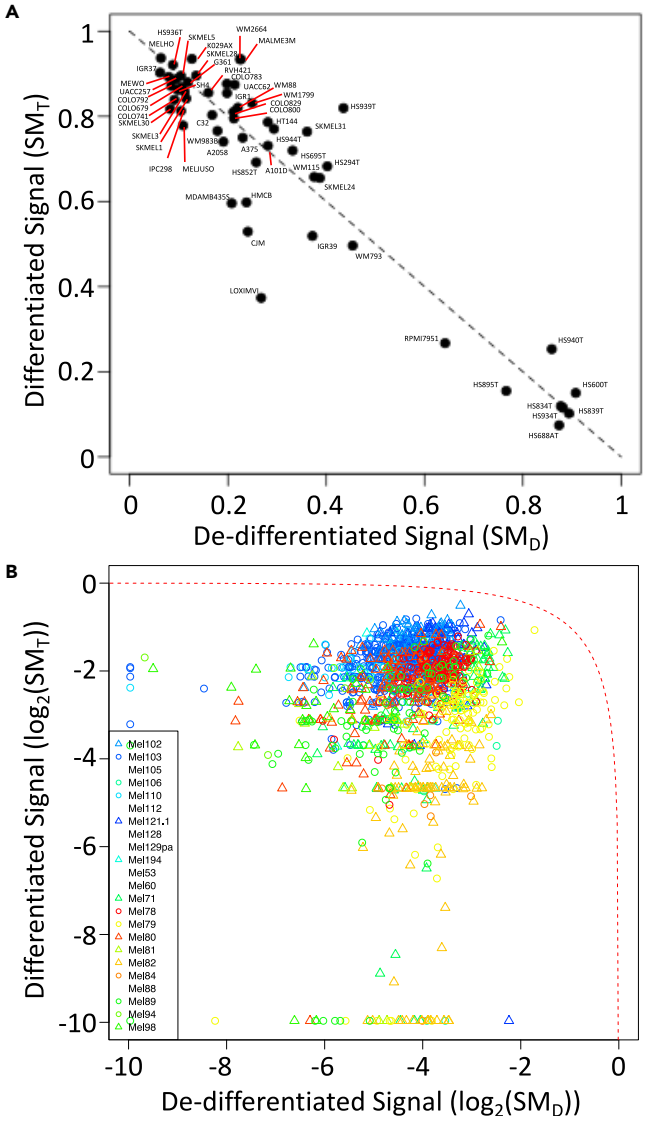


Figure 8. Melanoma Cell Lines and Primary Single Melanoma Cells Are Distributed Along Path between Extremes in Differentiation States

Projections along the terminally differentiated (SM_T) versus de-differentiated (SM_D) state axes for each melanoma cell line included in the CCLE (A) and primary melanoma cells (B). Values for the terminally differentiated and de-differentiated state metrics were estimated by RNA-seq data for cell lines associated with the CCLE and by scRNA-seq data for primary melanoma cells. Symbols for primary melanoma cells were colored differently for each patient sample. The dotted line corresponds to a reciprocal relationship between the SM_T and SM_D state metrics (i.e., SM_T = 1 - SM_D).

Projections of the melanoma cell lines in differentiation state space were calculated using the two state metrics (Figure 8). Similar to the breast cancer cell lines, the melanoma cell lines largely followed a linear reciprocal relationship between terminally differentiated (SM_T) and de-differentiated (SM_D) states (dotted line in Figure 8). The majority of cell lines exhibited primarily a terminally differentiated signature with some expression of de-differentiated genes, whereas only a small subset of the cell lines exhibited primarily a de-differentiated signature. The gene signatures for single melanoma cells were also highly heterogeneous owing to dropout of gene reads.

Using state metrics refined for use with tissue samples (see Table S2), samples acquired from benign melanocytic nevi and untreated primary melanoma tissue were projected onto the state space. Of note, CEACAM1 and MITF were associated with the differentiated state and no genes were shared with the breast

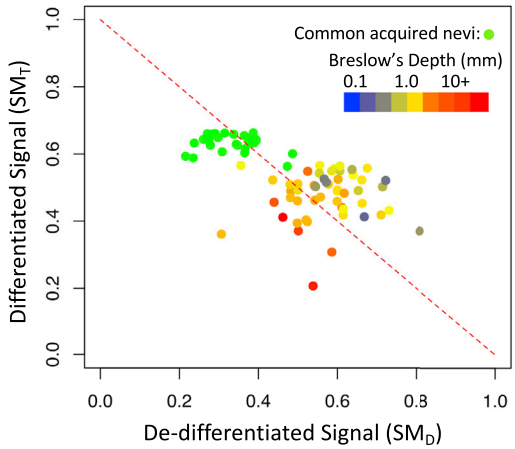


Figure 9. Gene Expression Patterns Associated with Benign Melanocytic Nevi and Primary Melanoma Tissue Samples Are Distributed Along Path between Extremes in Differentiation States

Projections along the terminally differentiated (SM_T) versus de-differentiated (SM_D) state axes for 78 tissue samples obtained from common acquired melanocytic nevi (n = 27, green circles) and primary melanoma (n = 51). The primary melanoma samples are colored based on the Breslow's depth (blue: 0.1 mm to red: 10+ mm). The dotted line corresponds to a reciprocal relationship between the SM_T and SM_D state metrics (i.e., $SM_T = 1 - SM_D$).

cancer epithelial state metric. The de-differentiated state metric had five genes, WISP1/CCN4, FOXC2, ITGA5, SERPINE1, and SPOCK1, that were shared with the breast cancer mesenchymal state metric. Although samples were more narrowly distributed in state space compared with the cell lines (Figure 9), all of the benign nevi exhibited higher terminally differentiated (SM_T) and tended to have lower de-differentiated values (SM_D). The samples from primary melanoma were color-coded based on the annotated Breslow's depth, where higher values were associated with lower terminal differentiation scores. Using Breslow's depth as a surrogate measure of metastatic potential (Balch et al., 2009), tissue-level EMT state metrics provide an estimate of metastatic potential and a digital measure of malignant cell differentiation state in the context of melanoma. This functional annotation also provided an external validation of the state metrics for melanoma.

Terminal Differentiation Is Associated with Distinct Gene Signatures, whereas De-differentiation Seems to Engage Common Gene Regulatory Networks

The separate gene signatures generated for breast cancer cells and melanoma cells using an unsupervised approach provide an opportunity to identify unique and shared aspects of the genetic regulatory mechanisms underpinning cell specification, as summarized in Figure 10. In terms of shared aspects, the overlap in the genes between melanoma and breast cancer metrics were not explained by random chance, as assessed by a Fisher exact test (p value $<2.2 \times 10^{-16}$ for SM_M/SM_D and p value $<4.3 \times 10^{-4}$ for SM_E/SM_T). We also found that the extent of overlap in the mesenchymal state metrics was greater than the overlap in the epithelial state metrics (odds ratio: 3.211 [95% confidence interval, 1.114–9.254], p value $<2.2 \times 10^{-16}$) as assessed by an exact hypergeometric test. Considering just the genes that overlap in the state metrics, the Ki values associated with the SM_M metric, although generally lower, trend similarly to the Ki values associated with the SM_D (see Figure 10B: slope = 0.927 with $R^2 = 0.841$). The Ki values associated with the SM_E, although generally higher, seemed to trend differently than the SM_T metrics (slope = 1.44 with $R^2 = 0.878$), although there are only eight genes in common. In addition, we used GOnet (Pomaznoy et al., 2018) to identify genes with transcription factor activity using the molecular function Gene Ontology term: DNA_binding (GO:0003677). In the breast cancer cell lines, nine transcription factors were upregulated in cells with a terminally differentiated phenotype, including GRHL2 and OVOL2, that have been associated with enforcing epithelial differentiation (Cieply et al., 2012). Correspondingly, four transcription factors were upregulated in melanoma cells, including MITF, which is essential for melanocyte differentiation (Goding and Arnheiter, 2019). Interestingly, there was no overlap in the genes with transcription factor activity in the two differentiated cell signatures. In contrast, melanoma and breast cancer cell lines that exhibited a de-differentiated phenotype shared seven transcription factors, including TWIST2 and ZEB1. De-differentiation in breast cancer cell lines was also associated with an additional eight transcription factors, including TWIST1 (Yang et al., 2004). Overall, the analysis of these transcription factors is consistent with specificity in phenotype as a consequence of engaging gene regulatory networks unique to a specialized cell subset, whereas de-differentiation seemed to engage common gene regulatory networks that facilitate the loss of cell specificity.

DISCUSSION

Here we used an unsupervised feature extraction and selection approach based on PCA and resampling to identify state metrics for the EMT in breast cancer and melanoma individually. Given the importance for

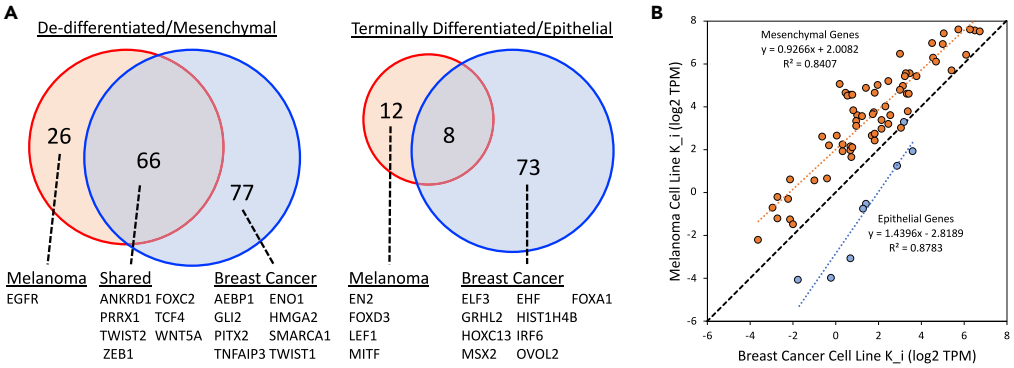


Figure 10. A Comparison of the Genes Included in the Different State Metrics across Cancers

(A) Venn diagram illustrating overlap in genes contained in the opposing state metrics for terminally differentiated/epithelial versus de-differentiated/mesenchymal extracted from breast cancer (blue circle) and melanoma (red circle) cell lines. The subset of the genes listed below the Venn diagram were annotated with transcription factor GO terms.

(B) A biplot of the Ki values for the overlapping genes in the terminally differentiated/epithelial state metrics (blue circles and blue linear trendline) and in the de-differentiated/mesenchymal state metrics (orange circles and orange linear trendline). A 1:1 correspondence is represented by the black dotted line.

identifying patients with tumors likely to metastasize, a number of gene signatures have been developed to predict the prevalence of tumor cells with an EMT signature (Tan et al., 2014; George et al., 2017; Rokavec et al., 2017; Koplev et al., 2018; Malta et al., 2018). Supervised approaches are most common (Tan et al., 2014; George et al., 2017; Rokavec et al., 2017; Koplev et al., 2018; Malta et al., 2018), where samples are classified *a priori*. For instance, Koplev et al. (2018) developed gene signatures that average over all anatomical locations, whereas Levine and coworkers (George et al., 2017; Jia et al., 2019) classify training samples a priori into one of three cell states: epithelial, mesenchymal, or hybrid E/M. Rokavec et al. generate features based on co-expression with E-cadherin and Vimentin (Rokavec et al., 2017). Although effective, supervised methods can perform poorly if the strategy is based on misinformation, such as sample misclassification or prior biases as to the number of cell states or defining genes. Moreover, developing metrics to classify the EMT status of tumors purely based on bulk tumor samples without deconvoluting the contribution of fibroblasts has unclear interpretation (Panchy et al., 2019). We also note that state metrics developed using microarray technology (e.g., Tan et al., 2014; Koplev et al., 2018) are not likely relevant for interpreting data based on RNA-seq, given the unclear relation between transcriptome and protein abundance as assayed using microarray technology. Although these methods rarely used, the data-driven nature of unsupervised methods for feature extraction and selection are attractive (Taguchi, 2017). For instance, Umeyama et al. used an unsupervised approach for feature extraction to identify genes associated with metastasis (Umeyama et al., 2014). To illustrate this data-driven approach, we have focused on breast cancer and melanoma separately, where metastatic dissemination to vital organs is a key limiter of patient survival and the cell-of-origin for these cancers have different developmental trajectories. In summary, we hope that our developed state metrics find use alongside other digital cytometry tools to better understand how oncogenic transformation and associated functional plasticity alters the immune contexture within the tumor microenvironment.

Limitations of the Study

Given the focus on breast cancer and melanoma as illustrative examples, the state metrics developed for these two biological contexts may not apply to cancers that originate in other anatomical locations. In terms of the bioinformatic approach, PCA is a linear approach that is used here for identifying genes that vary in expression together. Given that regulatory networks that underpin gene expression can give rise to non-linear behavior, genes that exhibit non-linear dependence with differentiation state are likely to be excluded from the state metrics. In addition, there may be additional patterns in gene expression that are biologically significant but fall below the null threshold due to a bias sampling of cell lines. In terms of limitations of the underlying data used in the study, next-generation RNA-seq of the cell lines included in the CCLE was performed on RNA isolated from frozen cell pellets using Trizol. We noted that the cells were cultured according to vendors' instructions for preservation, which includes adding the cryoprotectant dimethyl-sulfoxide (DMSO) to the cells prior to freezing. DMSO has also been reported to synchronize cells

by inducing reversible G_1 arrest (Fiore et al., 2002), which might impact transcriptional profiles. Although some work suggests that DMSO has no impact on the transcriptome (Guillaumet-Adkins et al., 2017), differences in whether samples were fresh or frozen when processed could be convoluted with differences in the state metrics compared between cell line versus tissue samples. Ultimately, the resulting state metrics should be revisited for consistency as additional transcriptomic datasets are reported and additional EMT-related genes are identified.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

DATA AND CODE AVAILABILITY

The code used in the analysis can be obtained from the following GitHub repository:

- https://github.com/KlinkeLab/DigitalCytometry_EMT_2020

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101080.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation (NSF CBET-1644932 to D.J.K.) and National Cancer Institute (NCI R01CA193473 to D.J.K.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF or NCI.

AUTHOR CONTRIBUTIONS

Conceptualization: D.J.K.; Methodology: D.J.K. and A.T.; Software: D.J.K. and A.T.; Formal Analysis: D.J.K. and A.T.; Visualization: D.J.K.; Supervision: D.J.K.; Funding acquisition: D.J.K.; Project administration: D.J.K.; Writing – original draft: D.J.K.; Writing – review & editing: D.J.K. and A.T.

DECLARATION OF INTERESTS

The authors declare no competing financial interests.

Received: December 19, 2019
Revised: March 24, 2020
Accepted: April 14, 2020
Published: May 22, 2020

REFERENCES

Alon, U. (2007). An introduction to systems biology: design principles of biological circuits. In Chapman & Hall/CRC Mathematical and Computational Biology Series, *volume 10*Chapman & Hall/CRC Mathematical and Computational Biology Series (Chapman & Hall/CRC), pp. 97–104.

Alonso, S.R., Tracey, L., Ortiz, P., Perez-Gomez, B., Palacios, J., Pollan, M., Linares, J., Serrano, S., Saez-Castillo, A.I., Sanchez, L., et al. (2007). A high-throughput study in melanoma identifies epithelial-mesenchymal transition as a major determinant of metastasis. Cancer Res. *67*, 3450–3460.

American Cancer Society (2019). Cancer Facts & Figures 2019 (American Cancer Society).

Andrews, T.S., and Hemberg, M. (2019). False signals induced by single-cell imputation [version 2; peer review: 4 approved]. F1000Res. *7*, 1740.

Balch, C.M., Gershenwald, J.E., Soong, S.J., Thompson, J.F., Atkins, M.B., Byrd, D.R., Buzaid, A.C., Cochran, A.J., Coit, D.G., Ding, S., et al. (2009). Final version of 2009 AJCC melanoma staging and classification. J. Clin. Oncol. *27*, 6199–6206.

Carretero, J., Shimamura, T., Rikova, K., Jackson, A.L., Wilkerson, M.D., Borgman, C.L., Buttarazzi, M.S., Sanofsky, B.A., McNamara, K.L., Brandstetter, K.A., et al. (2010). Integrative genomic and proteomic analyses identify targets for Lkb1-deficient metastatic lung tumors. Cancer Cell *17*, 547–559.

Cheng, W.Y., Kandel, J.J., Yamashiro, D.J., Canoll, P., and Anastassiou, D. (2012). A multi-cancer mesenchymal transition gene expression signature is associated with prolonged time to recurrence in glioblastoma. PLoS One *7*, e34705.

Chung, W., Eum, H.H., Lee, H.O., Lee, K.M., Lee, H.B., Kim, K.T., Ryu, H.S., Kim, S., Lee, J.E., Park, Y.H., et al. (2017). Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat. Commun. *8*, 15081.

Cieply, B., Riley, P., Pifer, P.M., Widmeyer, J., Addison, J.B., Ivanov, A.V., Denvir, J., and Frisch, S.M. (2012). Suppression of the epithelial-mesenchymal transition by Grainyhead-like-2. Cancer Res. *72*, 2440–2453.

Deng, W., Fernandez, A., McLaughlin, S.L., and Klinke, D.J. (2019). WNT1-inducible signaling pathway protein 1 (WISP1/CCN4) stimulates melanoma invasion and metastasis by promoting the epithelial-mesenchymal transition. J. Biol. Chem. *294*, 5261–5280.

Deng, W., Fernandez, A., McLaughlin, S.L., and Klinke, D.J. (2020). Cell communication network factor 4 (CCN4/WISP1) shifts melanoma cells from a fragile proliferative state to a resilient metastatic state. Cell. Mol Bioeng. *13*, 45–60.

Fiore, M., Zanier, R., and Degrassi, F. (2002). Reversible G(1) arrest by dimethyl sulfoxide as a new method to synchronize Chinese hamster cells. Mutagenesis *17*, 419–424.

George, J.T., Jolly, M.K., Xu, S., Somarelli, J.A., and Levine, H. (2017). Survival outcomes in cancer patients predicted by a partial EMT gene expression scoring metric. Cancer Res. *77*, 6415–6428.

Goding, C.R., and Arnheiter, H. (2019). MITF-the first 25 years. Genes Dev. *33*, 983–1007.

Guillaumet-Adkins, A., Rodríguez-Esteban, G., Mereu, E., Mendez-Lago, M., Jaitin, D.A., Villanueva, A., Vidal, A., Martinez-Marti, A., Felip, E., Vivancos, A., et al. (2017). Single-cell transcriptome conservation in cryopreserved cells and tissues. Genome Biol. *18*, 45.

Guo, Y., Sheng, Q., Li, J., Ye, F., Samuels, D.C., and Shyr, Y. (2013). Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. PLoS One *8*, e71462.

Jia, D., George, J.T., Tripathi, S.C., Kundnani, D.L., Lu, M., Hanash, S.M., Onuchic, J.N., Jolly, M.K., and Levine, H. (2019). Testing the gene expression classification of the EMT spectrum. Phys. Biol. *16*, 025002.

Jolliffe, I.T., and Cadima, J. (2016). Principal component analysis: a review and recent developments. Philos. Trans. A Math. Phys. Eng. Sci. *374*, 20150202.

Kaiser, J.L., Bland, C.L., and Klinke, D.J. (2016). Identifying causal networks linking cancer processes and anti-tumor immunity using Bayesian network inference and metagene constructs. Biotechnol. Prog. *32*, 470–479.

Kennecke, H., Yerushalmi, R., Woods, R., Cheang, M.C., Voduc, D., Speers, C.H., Nielsen, T.O., and Gelmon, K. (2010). Metastatic behavior of breast cancer subtypes. J. Clin. Oncol. *28*, 3271–3277.

Koplev, S., Lin, K., Dohlman, A.B., and Ma'ayan, A. (2018). Integration of pan-cancer transcriptomics with RPPA proteomics reveals mechanisms of epithelial-mesenchymal transition. PLoS Comput. Biol. *14*, e1005911.

Koren, S., and Bentires-Alj, M. (2015). Breast tumor heterogeneity: source of fitness, hurdle for therapy. Mol. Cell *60*, 537–546.

Li, J., Zhao, W., Akbani, R., Liu, W., Ju, Z., Ling, S., Vellano, C.P., Roebuck, P., Yu, Q., Eterovic, A.K., et al. (2017). Characterization of human cancer cell lines by reverse-phase protein arrays. Cancer Cell *31*, 225–239.

Malta, T.M., Sokolov, A., Gentles, A.J., Burzykowski, T., Poisson, L., Weinstein, J.N., Kamińska, B., Huelsken, J., Omberg, L., Gevaert, O., et al. (2018). Machine learning identifies stemness features associated with oncogenic dedifferentiation. Cell *173*, 338–354.

Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat. Biotechnol. *37*, 773–782.

Panchy, N., Azeredo-Tseng, C., Luo, M., Randall, N., and Hong, T. (2019). Integrative transcriptomic analysis reveals a multiphasic epithelial-mesenchymal spectrum in cancer and non-tumorigenic cells. Front. Oncol. *9*, 1479.

Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. *27*, 1160–1167.

Pomaznoy, M., Ha, B., and Peters, B. (2018). GOnet: a tool for interactive Gene Ontology analysis. BMC Bioinformatics *19*, 470.

Regad, T. (2013). Molecular and cellular pathogenesis of melanoma initiation and progression. Cell. Mol. Life Sci. *70*, 4055–4065.

Rokavec, M., Kaller, M., Horst, D., and Hermeking, H. (2017). Pan-cancer EMT-signature identifies RBM47 down-regulation during colorectal cancer progression. Sci. Rep. *7*, 4687.

Sarrio, D., Rodriguez-Pinilla, S.M., Hardisson, D., Cano, A., Moreno-Bueno, G., and Palacios, J. (2008). Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype. Cancer Res. *68*, 989–997.

Shannan, B., Perego, M., Somasundaram, R., and Herlyn, M. (2016). Heterogeneity in melanoma. Cancer Treat. Res. *167*, 1–15.

Siegel, R.L., Miller, K.D., and Jemal, A. (2019). Cancer statistics, 2019. CA Cancer J. Clin. *69*, 7–34.

Sosman, J. (2019). Overview of the Management of Advanced Cutaneous Melanoma (UpToDate).

Taghian, A., El-Ghamry, M.D., and Merajver, S.D. (2019). Overview of the Treatment of Newly Diagnosed, Non-metastatic Breast Cancer (UpToDate).

Taguchi, Y.H. (2017). Principal components analysis based unsupervised feature extraction applied to gene expression analysis of blood from dengue haemorrhagic fever patients. Sci. Rep. *7*, 44016.

Tan, T.Z., Miow, Q.H., Miki, Y., Noda, T., Mori, S., Huang, R.Y., and Thiery, J.P. (2014). Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. EMBO Mol. Med. *6*, 1279–1293.

Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.H., Porta-Pardo, E., Gao, G.F., Plaisier, C.L., Eddy, J.A., et al. (2018). The immune landscape of cancer. Immunity *48*, 812–830.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. B *63*, 411–423.

Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189–196.

Umeyama, H., Iwadate, M., and Taguchi, Y.H. (2014). TINAGL1 and B3GALNT1 are potential therapy target genes to suppress metastasis in non-small cell lung cancer. BMC Genomics *15* (*Suppl 9*), S2.

Vickman, R.E., Broman, M.M., Lanman, N.A., Franco, O.E., Sudyanti, P.A.G., Ni, Y., Ji, Y., Helfand, B.T., Petkewicz, J., Paterakos, M.C., et al. (2020). Heterogeneity of human prostate carcinoma-associated fibroblasts implicates a role for subpopulations in myeloid cell recruitment. Prostate *80*, 173–185.

Yang, J., Mani, S.A., Donaher, J.L., Ramaswamy, S., Itzykson, R.A., Come, C., Savagner, P., Gitelman, I., Richardson, A., and Weinberg, R.A. (2004). Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. Cell *117*, 927–939.

Yankaskas, C.L., Thompson, K.N., Paul, C.D., Vitolo, M.I., Mistriotis, P., Mahendra, A., Bajpai, V.K., Shea, D.J., Manto, K.M., Chai, A.C., et al. (2019). A microfluidic assay for the quantification of the metastatic propensity of breast cancer specimens. Nat. Biomed. Eng. *3*, 452–465.

Zhang, M., Lee, A.V., and Rosen, J.M. (2017). The cellular origin and evolution of breast cancer. Cold Spring Harb. Perspect. Med. *7*, a027128.

iScience, Volume *23*

Supplemental Information

An Unsupervised Strategy for Identifying Epithelial-Mesenchymal Transition State Metrics in Breast Cancer and Melanoma

David J. Klinke II and Arezo Torang

Table S1. List of genes and corresponding K_i values for state metrics developed separately for breast cancer and melanoma cell lines based on CCLE gene expression, related to Figures 5, 8, and 10. Genes that overlap with the fibroblast gene list are highlighted in yellow.

Breast Cancer Cell Lines

Epithelial Signature

GENE_SYMBOL	*Ki (log2 TPM)*	GENE_SYMBOL	*Ki (log2 TPM)*
AGR2	3.411	SORL1	0.575
ALDH3B2	-0.162	SPINT1	3.224
ANXA9	0.842	SPINT2	6.114
AP1M2	3.229	SPRR3	-4.907
ARHGAP8	1.512	ST14	1.847
ATP2C2	0.834	TMC6	2.909
BIK	-0.165	TMPRSS2	-1.064
BLNK	-1.478	TSPAN1	2.861
BSPRY	-0.331	TSPAN15	3.200
C1orf106	0.586	TTC39A	1.869
C4orf19	-0.034	TUBBP5	-1.711
CBLC	-1.002	VAMP8	4.388
CDH1	3.017	VAV3	1.434
CDS1	1.307	WNT3A	-4.361
CEACAM6	-0.326	WNT4	-0.875
CGN	1.816	WNT6	-3.454
CKMT1A	0.752		
CLDN4	4.194		
CLDN7	3.465		
CXCR4	-0.294		
CYP4B1	-2.556		
DSC2	0.489		
EDN2	-0.558		
EFNA1	3.246		
EHF	1.722		
ELF3	3.661		
EPCAM	3.937		
EPN3	1.121		
ERBB3	3.212		
ESRP1	1.473		
ESRP2	2.192		
EVPL	1.066		
F11R	3.831		
FA2H	-1.695		
FBP1	-0.025		
FOXA1	1.528		
FXYD3	3.654		
GRB7	2.017		
GRHL2	0.375		
HIST1H4B	-3.644		
HOXC13	0.820		
ICA1	1.830		
IL1RN	-1.914		
IRF6	1.130		
JUP	4.619		
LAD1	1.392		
LLGL2	3.403		
MAP7	1.918		
MST1R	1.586		
MSX2	0.294		
MYH14	1.305		
MYO5C	0.668		
OR7E14P	-1.719		
OVOL2	-0.791		
PAK6	-0.971		
PDGFB	0.847		
POF1B	-2.139		
PPL	1.697		
PRSS8	1.768		
PTK6	0.323		
RAB25	2.307		
S100A14	3.154		
SCNN1A	2.039		
SEPP1	1.217		
SLC37A1	1.599		

Mesenchymal Signature

GENE_SYMBOL	*Ki (log2 TPM)*	GENE_SYMBOL	*Ki (log2 TPM)*
ACTA2	3.826	LOX	3.049
ADAM12	1.053	LOXL2	5.029
AEBP1	0.789	LRRC15	-2.078
AKAP12	1.603	LUM	0.844
AKAP2	3.466	MAP1B	2.602
AKT3	1.980	MFAP5	0.349
ANK2	0.019	MME	1.240
ANKRD1	0.750	MMP14	4.173
ASPN	-3.605	MMP2	3.048
AXL	2.980	MMP3	-1.962
B2M	10.551	MT2A	8.204
BAG2	3.408	MVP	5.543
BGN	2.601	MXRA7	5.324
C1S	3.387	MYL9	5.467
CALD1	5.718	NID2	2.203
CCL2	1.674	OLFML2B	0.722
CD68	3.956	PAPPA	-0.374
CDH11	1.735	PCOLCE	5.155
CDH2	2.608	PDGFC	3.104
CFH	0.357	PDGFRA	-0.591
CHN1	2.675	PDGFRB	0.074
CLIC4	6.317	PHLDA1	4.025
COL1A1	6.150	PITX2	-0.011
COL3A1	2.372	PLAUR	4.586
COL5A1	3.435	PMP22	3.951
COL5A2	3.181	POSTN	1.271
COL6A1	5.100	PRKCA	2.585
COL6A2	4.555	PROCR	2.850
COL6A3	1.839	PRRX1	0.603
COMP	-2.917	RCN3	3.333
COPZ2	2.300	RECK	1.623
CTSB	7.845	S100A4	6.338
CXCL3	-0.458	SACS	2.270
CYBRD1	3.311	SDC2	4.282
DAB2	2.936	SERPINB2	0.838
DCN	0.752	SERPINE1	4.732
DDR2	1.732	SERPINE2	5.020
EDNRA	-2.082	SFRP4	-2.690
EIF5A2	2.134	SH3KBP1	3.958
EMP3	4.799	SMARCA1	2.936
ENG	3.451	SPARC	6.312
ENO1	10.469	SPOCK1	3.297
FABP5	5.250	SRPX	1.971
FAP	0.521	SULF1	1.789
FBN1	3.493	TCF4	0.814
FERMT2	4.687	TFPI	3.398
FGF1	-0.276	TGFB1	3.796
FGF2	1.013	TGFB1I1	2.768
FHL1	2.509	TGFB2	2.218
FHL2	5.727	THBS2	0.240
FN1	7.767	THY1	1.438
FOXC2	-2.163	TIMP3	4.142
FST	1.858	TMEFF1	0.345
FSTL1	5.764	TMEM158	1.339
FZD7	2.928	TNC	3.700
GAS1	-0.404	TNFAIP3	2.728
GEM	1.912	TNFAIP6	-2.10206
GFPT2	1.747	TPM2	6.788168
GJA1	1.859	TRPC1	1.200131
GLI2	-1.515	TUBA1A	6.360364
GLT8D2	0.756	TUBB3	-4.36173
GREM1	0.866	TUBB6	7.167798
HGF	-1.971	TWIST1	1.57171
HMGA2	1.475	TWIST2	-0.95998
HTRA1	3.909	VCAN	1.996064
IFITM3	7.043	VEGFC	2.485532
IGFBP3	6.549	VIM	7.664345
ITGA5	4.594	WISP1	-2.69714
ITGB1	8.822	WNT5A	2.181108
LEPRE1	4.883	WNT5B	1.845579
LGALS1	10.647	ZEB1	0.997605
LHFP	2.476		

Melanoma Cell Lines

Differentiated Signature

GENE_SYMBOL	*Ki (log2 TPM)*
ALDH3B2	-4.011
ARAP2	-2.582
B3GAT1	-3.133
CEACAM1	-0.123
CKMT1A	-3.107
DLL3	0.188
EDNRB	1.235
EN2	-1.807
ERBB3	3.265
ESRP1	-0.554
FOXD3	-1.932
FXYD3	1.928
HPGD	-1.439
LEF1	2.589
MITF	3.481
MTUS1	1.819
MYH14	-0.758
TMC6	1.244
TUBB3	-3.873
TUBBP5	-4.099

Dedifferentiated Signature

GENE_SYMBOL	*Ki (log2 TPM)*	GENE_SYMBOL	*Ki (log2 TPM)*
ABCC3	1.406	SERPINB2	4.536
ACTA2	5.421	SERPINE1	6.088
ADAM12	3.608	SFRP4	-1.224
ANKRD1	2.131	SPOCK1	5.385
ASPN	-2.238	SULF1	3.621
BGN	3.567	TCF4	2.083
C1S	4.607	TFPI	3.789
CDH11	2.617	TGFBI	8.522
CFH	1.919	THBS2	5.056
CITED2	5.845	THY1	4.842
CLU	5.813	TNXB	1.628
COL1A1	6.420	TPM2	7.498
COL3A1	3.977	TWIST2	0.560
COL5A1	4.609	VCAN	5.018
COL5A2	4.965	VEGFC	3.187
COL6A1	7.388	WISP1	-0.241
COL6A2	6.933	WNT2	-2.599
COL6A3	3.714	WNT5A	3.375
COMP	-0.745	WNT5B	2.735
CXCL12	1.898	ZEB1	3.080
CYP1B1	1.646		
DCN	4.524		
DES	-1.976		
EDNRA	-1.273		
EGFR	2.254		
EPS8L2	2.871		
FAP	4.634		
FBN1	5.531		
FGF1	2.181		
FGF2	3.328		
FHL1	5.185		
FN1	11.332		
FOXC2	-0.334		
FST	4.619		
FSTL1	7.571		
GJA1	2.395		
GLT8D2	1.638		
GREM1	3.809		
HGF	-1.492		
IFITM2	6.038		
IGFBP3	7.526		
IL1R1	2.317		
INHBA	3.419		
ITGA5	6.247		
ITGBL1	3.675		
KRT14	2.253		
KRT7	3.666		
LGR5	-2.537		
LOX	4.766		
LOXL2	6.880		
LRRC15	0.581		
MALL	0.788		
MFAP5	2.224		
MMP2	6.450		
MXRA5	-1.191		
MYL9	5.656		
NID2	2.973		
NOTCH3	2.482		
NT5E	6.282		
OLFML2B	1.962		
PAPPA	0.616		
PDGFC	2.985		
PDGFRA	2.568		
PDGFRB	2.617		
PLAU	2.660		
POSTN	3.522		
PRRX1	4.508		
PTGS1	0.887		
PTRF	7.111		
RCN3	5.553		
RHOD	0.830		
S100A4	7.575		

Table S2. List of genes and associated Ki values for refined state metrics based on TCGA breast cancer tissue samples and tissue samples of common acquired melanocytic nevi and primary melanoma, related to Figures 6 and 9. Genes that overlap in the state metrics between breast cancer and melanoma are highlighted in green.

TCGA Breast Cancer Tissue Samples

Epithelial Signature		Mesenchymal Signature	
GENE_SYMBOL	*Ki (log2 TPM)*	GENE_SYMBOL	*Ki (log2 TPM)*
ALDH3B2	5.860	ASPN	5.774
C1orf106	1.334	B2M	10.641
C4orf19	1.790	CDH2	1.251
CDH1	7.929	CLIC4	7.324
CLDN4	7.355	CTSB	8.506
CLDN7	6.914	EDNRA	4.265
CYP4B1	2.776	FOXC2	0.656
DSC2	4.018	IFITM3	9.645
EHF	5.528	ITGA5	5.290
FA2H	1.907	MMP3	2.950
GRB7	4.897	POSTN	8.570
ICA1	5.021	SERPINE1	5.388
IRF6	6.778	SPOCK1	3.846
JUP	8.167	SULF1	6.026
MSX2	3.530	TGFB1	5.376
OR7E14P	2.594	TUBB3	0.189
POF1B	1.498	WISP1	3.040
PPL	4.865		
SPRR3	-2.907		
TMPRSS2	2.745		
TUBBP5	1.073		
WNT3A	-2.816		
WNT4	2.170		
WNT6	-0.415		

Melanocytic Nevi and Melanoma Tissue Samples

Differentiated Signature		De-differentiated Signature	
GENE_SYMBOL	*Ki (log2 TPM)*	GENE_SYMBOL	*Ki (log2 TPM)*
ARAP2	5.322	ACTA2	6.008
CEACAM1	3.142	DES	1.865
CKMT1A	0.335	FGF1	2.198
EDNRB	8.655	FOXC2	-4.064
ERBB3	6.930	HGF	2.130
ESRP1	5.179	INHBA	2.599
FXYD3	7.487	ITGA5	4.301
HPGD	6.732	KRT7	1.376
MITF	7.547	NID2	3.532
MTUS1	5.913	NOTCH3	4.783
MYH14	3.382	PDGFRB	5.790
		SERPINE1	2.665
		SPOCK1	2.186
		TPM2	5.225
		VEGFC	2.026
		WISP1	1.830
		WNT5A	3.416

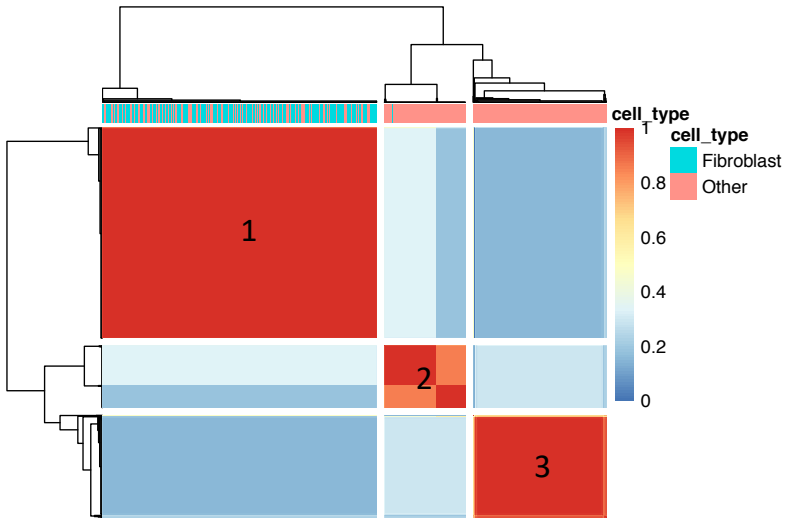


Fig. S1. Consensus matrix for similarity and clustering of cell samples, related to Figures 6 and 9. The symmetric 1034x1034 matrix is colored in element(i,j) by similarity in assigning cells i and j to the same cluster when the clustering parameters are changed. A similarity score of 0 (blue) indicates that the two cells are always assigned to different clusters while a score of 1 (red) indicates that the two cells are always assigned to the same cluster. The similarity of the samples are also illustrated by the dendrograms shown on the top and side. The top bar indicates whether the cell was annotated as a fibroblast based on COL1A1 and COL1A2 co-expression (aqua - fibroblast, pink - other).

Transparent Methods

'Omics Data. Transcriptomics profiling of the same samples using both Agilent microarray and Illumina RNA sequencing for the breast cancer arm (BRCA) of the Cancer Genome Atlas was downloaded from TCGA data commons. Values for gene expression, expressed in TPM for RNA-seq and gene-centric RMA-normalized data for Affymetrix U133+2 microarray, for the cell lines contained within the Cancer Cell Line Encyclopedia were downloaded from the Broad data commons (Website: https://portals.broadinstitute.org/ccle Files: CCLE_RNAseq_rsem_genes_tpm_20180929.txt accessed 04/04/2019 and CCLE_Expression_Entrez_2012-10-18.res accessed 6/15/2018). Reverse phase protein array (RPPA) results for the cancer cell lines were obtained from the M.D. Anderson proteomics website (Website: https://tcpaportal.org/mclp/ File: MCLP-v1.1-Level4.txt accessed 6/15/2018) (Li et al., 2017). Single-cell gene expression (scRNA-seq) for breast cancer and melanoma cells expressed in TPM were downloaded from the Gene Expression Omnibus (GEO) entries GSE75688 and GSE72056, respectively. 10X Genomics scRNA-seq data for CD45-negative cells digested from a normal human female skin sample and expressed in counts of gene-level features was downloaded from European Bioinformatics Institute (EMBL-EBI) ArrayExpress entry E-MTAB-6831. RNA-seq data expressed in counts assayed in samples acquired from benign melanocytic nevi and untreated primary melanoma tissue and associated annotation were downloaded from GEO entry GSE98394.

Non-linear regression of protein abundance to mRNA expression. All data was analyzed in R (V3.5.1) using the 'stats' package (V3.5.1). For each gene where complementary CCLE transcriptomic and RPPA data exist and for which their correlation coefficient was above 0.36, the non-linear function,

$$Y_{protein} = a + \frac{b \cdot X_{mRNA}}{X_{mRNA} + c}, \tag{S1}$$

was regressed using the *nls* function to the corresponding protein ($Y_{protein}$) and transcript (X_{mRNA}) abundance data. As the RPPA values are normalized, the parameters a and b represent the background value and maximum detectable increase above background, respectively, while the parameter c represents the midpoint in transcript abundance within the dynamic range of the assay. A minimum in the summed squared errors between model-predicted and observed RPPA values were used to determine the optimal values of the model parameters. Using the optimal values, a threshold was estimated independently for each gene based on the transcript abundance that yields a 2.5% increase in protein abundance above background. The regression was repeated using both RNA-seq and Affymetrix transcriptomics data.

Statistical analysis for cell-level signatures. Principal component analysis (PCA) was performed on log base 2 transformed TPM values using the *prcomp* function in R on the CCLE RNA-seq data, which was filtered to 780 genes previously associated with epithelial-mesenchymal transition. The collective list of genes were assembled from prior studies (Sarrio et al., 2008; Carretero et al., 2010; Alonso et al., 2007; Cheng et al., 2012; Tan et al., 2014; Kaiser et al., 2016; Deng et al., 2019, 2020) and additional gene sets from MSigDB V4.0 including: "EPITHELIAL TO MESENCHYMAL TRANSITION" and "REACTOME TGF BETA RECEPTOR SIGNALING IN EMT EPITHELIAL TO MESENCHYMAL TRANSITION". PCA was applied to the genes to extract the features, where the resulting eigenvectors capture the relative influence of a gene's expression on a specific principal component and the eigenvalues represent how much information contained within the dataset is captured by a specific principal component. Drawing upon conventional hypothesis testing where significance is established by rejecting the null hypothesis that experimental observations could be explained by random chance, we used a resampling approach to establish a null hypothesis related to the eigenvalues, that is to determine the true rank of the noisy expression matrix. The resampling approach involved repetitively applying PCA (n = 1000) to a synthetic noise dataset with the same dimensions that was generated from the original data by randomly resampling with replacement from the collection of gene expression values and assigning the values to particular gene-cell line combinations. The resulting distribution of eigenvalues and eigenvectors represent the values that could be obtained by random chance if the underlying dataset has no information (i.e., the null PCA distribution). Principal components with eigenvalues greater than the null PCA distribution were used to define the principal subspace for subsequent analysis, that is the selection of features. Similarly, the distribution in the projection of genes within the null PCA space were used to determine whether the projection of a gene along a particular PC axis was explained by random chance or not by setting thresholds along the PC2 and PC3 axes that enclosed 95% of the null PCA space. The PC projection of genes relative to the null PCA space was used to refine the extracted features.

A metric was developed to estimate the extent that a cell exhibits a gene signature corresponding to a "Epithelial/Terminally Differentiated" versus "Mesenchymal/De-differentiated" state. The state metrics (SM) quantify the cellular state by averaging over a normalized expression level of each gene in the signature ($reads_i$, expressed in TPM) according to the formula:

$$SM = \frac{1}{n_{gs}} \sum_{i=1}^{n_{gs}} \frac{reads_i}{reads_i + 2^{K_i}}. \tag{S2}$$

The genes included in a signature with their corresponding K_i values are listed in Table S1 and n_{gs} corresponds to the number of genes within a signature. The K_i values were estimated by clustering the log2 expression of each gene into two groups using the k-means method and the value was set as the mid-point in expression between the two groups.

Statistical analysis for tissue-level signatures. Genes differentially expressed in normal epidermal fibroblasts were obtained by analyzing single-cell RNA-seq data of normal skin obtained using a Genomics 10x platform and a bioinformatics workflow based on the scater (V1.12.2) and SC3 (V1.12.0) packages in R. Briefly, scRNA-seq data were filtered to retain samples that had less than 50% of the reads in the top 50 genes and to remove outlier samples based on PCA analysis. Gene-level features were limited to those that were expressed at greater than 1 count in more than 10 cell samples. Read depth was normalized using a variant of CPM contained within the *scran* (V1.12.1) package, which develops a sample-specific normalization factor repetitive sample pooling followed by deconvoluting a sample-specific factor by linear algebra. Following from Davidson et al. (bioRxiv 467225), fibroblasts were annotated based on co-expression of COL1A1 and COL1A2.

Samples were clustered and genes differentially associated with each cluster were identified using the *SC3* workflow (V1.14.0) using default parameters (see Figure S1).

Prior to logistic regression analysis, TCGA BRCA data and the benign nevi and melanoma data were filtered to remove sample outliers and normalized based on housekeeping gene expression (Eisenberg and Levanon, 2013). Using normal versus tumor annotation associated with the data, ridge logistic regression was performed on log base 2 transformed TPM and median-centered values using the *glmnet* package (V2.0-18), which was limited to EMT-related genes identified in the CCLE analysis and not associated with normal fibroblasts. To minimize overfitting, ridge logistic regression was repeated 500 times using a subsample of the original data set using the genes associated with each signature separately. In each iteration, the samples were randomly assigned in an 80:20 ratio between training and testing samples. Regression coefficients were captured for each iteration using a lambda value that minimized the misclassification error of a binomial prediction model estimated by cross-validation. Accuracy was assessed using the testing samples. Genes were determined to have a consistent expression pattern if greater than 95% of the distribution in regression coefficients had the correct sign. Similarly to the cell-level analysis, state metrics were developed for bulk tissue-level RNA-seq measurements to estimate the extent that a tissue sample exhibits a gene signature corresponding to a “Epithelial/Terminally Differentiated" versus “Mesenchymal/De-differentiated" state. The genes included in a signature and their corresponding K_i values are listed in Table S2.

Data and Code Availability. The code used in the analysis can be obtained from the following GitHub repository:

- https://github.com/KlinkeLab/DigitalCytometry_EMT_2020