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

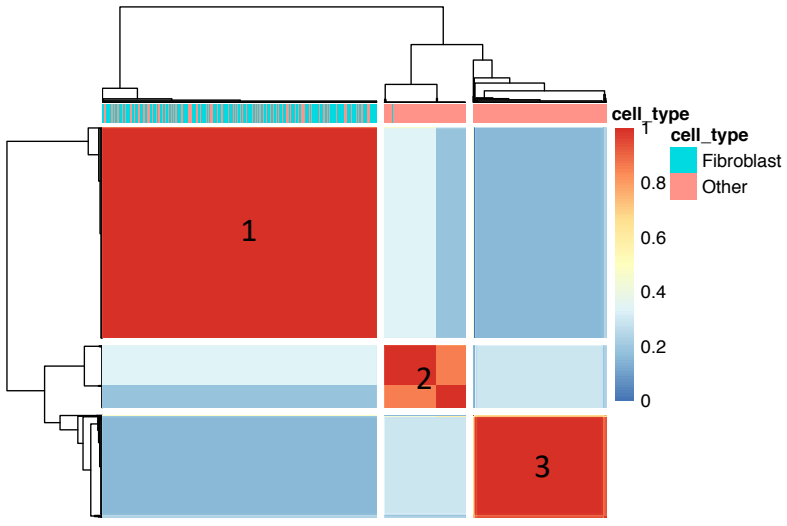

Fig. S1. Consensus matrix for similarity and clustering of cell samples, related to Figures 6 and 9. The symmetric 1034x1034 matrix is colored in element(i,j) by similarity in assigning cells i and j to the same cluster when the clustering parameters are changed. A similarity score of 0 (blue) indicates 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