A Transcriptional Program for Detecting TGFβ-Induced EMT in Cancer

Momeneh Foroutan1,2, Joseph Cursons2,3,4, Soroos Hediyyeh-Zadeh, Erik W. Thompson1,5,6, and Melissa J. Davis2,7

Abstract

Most cancer deaths are due to metastasis, and epithelial-to-mesenchymal transition (EMT) plays a central role in driving cancer cell metastasis. EMT is induced by different stimuli, leading to different signaling patterns and therapeutic responses. TGFβ is one of the best-studied drivers of EMT, and many drugs are available to target this signaling pathway. A comprehensive bioinformatics approach was employed to derive a signature for TGFβ-induced EMT which can be used to score TGFβ-driven EMT in cells and clinical specimens. Considering this signature in pan-cancer cell and tumor datasets, a number of cell lines (including basal B breast cancer and cancers of the central nervous system) show evidence for TGFβ-driven EMT and carry a low mutational burden across the TGFβ signaling pathway. Furthermore, significant variation is observed in the response of high scoring cell lines to some common cancer drugs. Finally, this signature was applied to pan-cancer data from The Cancer Genome Atlas to identify tumor types with evidence of TGFβ-induced EMT. Tumor types with high scores showed significantly lower survival rates than those with low scores and also carry a lower mutational burden in the TGFβ pathway. The current transcriptomic signature demonstrates reproducible results across independent cell line and cancer datasets and identifies samples with strong mesenchymal phenotypes likely to be driven by TGFβ.

Implications: The TGFβ-induced EMT signature may be useful to identify patients with mesenchymal-like tumors who could benefit from targeted therapeutics to inhibit mesenchymal TGFβ signaling and disrupt the metastatic cascade.


Introduction

Epithelial-to-mesenchymal transition (EMT) is a physiologic process involved in development (1) and wound healing (2) subverted during cancer metastasis, together with the reverse MET process (3). EMT is associated with a migratory, invasive phenotype and cancer stem cell (CSC) characteristics that aid metastasis, promote drug resistance, and repress apoptosis (4). General EMT signatures have been derived using a variety of cell lines and/or stimuli (5–7), with the aim of identifying a general molecular program underlying EMT. Our previous work, however, suggests that EMT is a phenotypic endpoint that can be driven by different stimuli, with important implications for drug responses (8).

TGFβ is a canonical driver of EMT (9). Mutations in the TGFβ pathway can promote cell invasiveness and motility, and reduced TGFβ type II receptor (TGFBR2) expression has been reported in different cancer types (10). In various cancer cell lines, active TGFβ signaling induces the promesenchymal transcription factors SNAI1 and SNAI2, which in turn suppress expression of E-cadherin (CDH1), resulting in loss of cell–cell junctions (11). TGFβ controls important EMT markers through both canonical and noncanonical signaling in different cancers (12). Note that EMT not only is regulated through transcription but also is influenced and maintained through interconnected changes in epigenetics (13), micro-RNAs (14), long, noncoding RNAs (15), and protein synthesis (16, 17).

In a clinical context, it is important to note that TGFβ can play a dual role in human cancer, promoting metastasis or acting as a tumor suppressor (18). To some extent, this parallels contrasting effects of TGFβ on epithelial cells where it inhibits proliferation, and mesenchymal cells where it is stimulatory (19). Although high tumor and circulatory levels of TGFβ are associated with poor prognosis across multiple cancers, TGFβ levels alone are not sufficient to identify patients for TGFβ inhibitor therapies due to these suppressor effects (20). Given the importance of TGFβ-induced EMT in tumor progression (21), we sought to identify a signature which provides evidence of this specific phenotypic program and thus may better identify patients where cancer progression is driven by TGFβ. We
have applied transcript meta-analysis methods, using data from a range of cancer cell lines with TGFβ stimulation and verified EMT. To our knowledge, this is the first time that using comprehensive meta-analysis methods, a transcriptional signature of EMT specifically associated with TGFβ has been obtained. It should be noted that this signature captures the change in gene expression caused by TGFβ-induced EMT and does not necessarily indicate further increases in TGFβ signaling.

We have assessed this signature using a wide range of cancer cell line and clinical tumor data. Specifically, we applied single-sample gene set analysis methods to obtain TGFβ-EMT scores (TES) and identify cancer cell lines and tumor samples with evidence of TGFβ-induced EMT. We illustrate a novel approach of comparing information from multiple signatures, such as more general epithelial and mesenchymal signatures, to examine specific molecular drivers of EMT across a large number of cell lines and clinical samples. Finally, we demonstrate that TES is correlated with a differential response to certain drugs, and we show that higher TES is associated with lower overall survival outcome in pan-cancer data. Our signature may be useful to identify candidate patients for TGFβ signaling inhibitor therapies (22), and the approach of gene set scoring is particularly promising for the strategic design of personalized cancer treatments.

Materials and Methods

All computational and statistical analyses were performed using R (versions 3.1.1, 3.2.2, and 3.2.4) and Bioconductor (version 3.0). Further details are given in the Supplementary Methods, and the digital archive reproducing our results is also available at https://github.com/DavisLaboratory/mforoutan_tgfb_paper_2016.

Microarray data for the gene expression signature of TGFβ-induced EMT

Publicly available microarray data were collected from Gene Expression Omnibus (GEO) using the NCBI portal (http://www.ncbi.nlm.nih.gov/geo/), based upon the following criteria: (i) microarray experiments must have been performed using Affymetrix, Agilent, or illumina platforms; (ii) data must be collected from human cancer cell lines; (iii) there must be evidence of TGFβ-induced EMT through morphologic or phenotypic assays and expression assays for EMT markers (e.g. VIM, CDH1 etc.); and (iv) datasets must include biological replicates.

Data quality was assessed using methods provided within the R/Bioconductor package Simpleaffy (23) for Affymetrix platforms and using unsupervised methods for other platforms to exclude all low quality samples from the analysis. All 10 datasets used in this study were of high quality.

Obtaining our TGFβ-EMT signature

To obtain our TGFβ-EMT signature, we applied 2 meta-analysis techniques. In the first technique, we calculated the “Product of Rank” (PR) metric which has been shown to have better performance characteristics (biologic association, stability, and robustness) than the other available methods for identifying genes that are differentially expressed in all of the datasets (24). In the second method, we integrated all the 10 datasets after RMA normalization and applied SVA (25) and ComBat (26) to estimate and remove batch effects and then performed differential expression analysis using limma.

Public cancer cell line and patient tumor data

Cell line data used in this study were derived from NCI-60 (27), CCLE (Cancer Cell Line Encyclopedia; ref. 28), COSMIC (29) pan-cancer cell lines, as well as the breast cancer cell lines obtained from the study by Neve and colleagues (30) and Heiser and colleagues (31). We also used patients’ data from The Cancer Genome Atlas (TCGA) including breast cancer microarray data as well as the pan-cancer RNASeq data. All of the datasets were scored against the signature, and COSMIC, NCI-60, and Heiser were further analyzed for drug response assessment. Batch analysis of the large datasets is illustrated in Supplementary Results, Supplementary Figures S16–S27.

Single-sample scoring methods applied on the cell line and tumor data

Using the GSVA package in R/Bioconductor, we applied ssGSEA and GSVA methods to obtain the TES for samples in each dataset. We additionally defined a simple single-sample scoring method to examine whether the TES obtained by ssGSEA and GSVA is associated with batches in the large datasets.

Results

Data for TGFβ-induced EMT were identified across the literature and integrated

To identify a gene expression signature associated with TGFβ-induced EMT, we collected microarray data from GEO (32) for 10 studies that examined TGFβ-induced EMT across a range of epithelial cell lines (Table 1). Six differentially expressed genes (DEG) were shared across all studies (COL1A1, FN1, ADAM19, SERPINE1, TAGLN, and PMEP1), all of which are associated with EMT and TGFβ stimulation.

A gene signature for TGFβ-induced EMT was obtained using meta-analysis methods

When using list overlap to derive gene signatures, the results often have low statistical power due to the small number of samples within each study. Given that only 6 DEGs were identified across all datasets, we next applied 2 meta-analysis techniques and took the union of the resulting gene sets for our signature.

First, within each study, genes were ranked by their limma (33, 34) t statistic and the product of ranks (24, 35) was calculated and compared with a permuted null distribution (details in Supplementary Methods). This approach identified 186 up- and 82 downregulated genes implicated in TGFβ-induced EMT (Fig. 1A).

Next, we obtained DEGs by combining the 10 datasets after RMA normalization. Although the merged data had a larger number of samples, providing us with greater statistical power, we needed to minimize batch effects between the different studies. The R package SVA (25) was used with the SVA and ComBat (26) algorithms to estimate and remove batch effects within our integrated data. Using limma (33, 34) with batch corrected data identified 121 up- and 74 down-regulated genes (Supplementary Table S1).

Across both signatures, there were 165 common genes and 301 genes in total. The combined list was used as our signature of TGFβ-induced EMT (108 down- and 193 up-regulated genes, Supplementary Table S1).
The TGFβ-induced EMT gene signature had stimulus-dependent overlap with reported general EMT signatures

A number of transcript abundance signatures have been identified for EMT, including reports by Taube and colleagues ($n_{\text{genes}} = 250$; ref. 5), Groger and colleagues ($n_{\text{genes}} = 131$; ref. 6), Tan and colleagues ($n_{\text{genes}} = 218$; ref. 7), Cursons and colleagues ($n_{\text{genes}} = 206$; Supplementary Table S2; ref. 8), and Du and colleagues ($n_{\text{genes}} = 571$; ref. 36). Comparing our TGFβ-induced EMT signature against these previous studies, only 3 genes were shared across all: vimentin ($VIM$), keratin-15 ($KRT15$), and tetraspanin-1 ($TSPAN1$); however, comparing each pair of signatures, the number of shared genes was higher (Supplementary Results, Supplementary Fig. S1).

Low gene overlap across all signatures may reflect differences in EMT stimuli and experimental and computational methods. Our signature was derived using 10 datasets that specifically examined TGFβ-induced EMT using multiple cancer cell lines (Table 1), and these data were subjected to consistent preprocessing and normalization methods to minimize artifacts. Conversely, Cursons and colleagues derived their signature using EGF- or hypoxia-induced EMT within basal breast cancer cell lines (8). Taube and colleagues used a number of different stimuli (including TGFβ) to induce EMT; however, they only used one cell line and thus their signature may show bias toward the genetic/phenotypic background of HMLE cells (5). Groger and colleagues selected genes differentially expressed in at least 10 (of 18) data sets (6), whereas our hypothesis setting identified genes differentially expressed across all datasets. The signature recently reported by Du and colleagues (36) was obtained using only 2 cancer cell lines, with EMT induced by TGFβ and oncostatin-M (OSM), and their EMT signature was defined simply by taking the intersection of DEGs between the 2 cell lines. Finally, Tan and colleagues (7) applied an alternative approach—rather than studying stimulated cell lines, they used

<table>
<thead>
<tr>
<th>Cell line</th>
<th>Tissue of cell line origin</th>
<th>Platform</th>
<th>Replicates</th>
<th>GEO accession number</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMLE</td>
<td>Breast</td>
<td>Affymetrix- HT_HG-U133A</td>
<td>3</td>
<td>GSE24202</td>
<td>Taube and colleagues (5) PMID: 20713713</td>
</tr>
<tr>
<td>MCF10A</td>
<td>Breast</td>
<td>Agilent-014850 Whole Human Genome</td>
<td>2</td>
<td>GSE28659</td>
<td>Deshiere and colleagues (2013) PMID: 22562247</td>
</tr>
<tr>
<td>HMEC-TR</td>
<td>Breast</td>
<td>Affymetrix- HG-U133_Plus_2</td>
<td>2</td>
<td>GSE28448</td>
<td>Hesling and colleagues (2011) PMID: 21597466</td>
</tr>
<tr>
<td>A549</td>
<td>Lung</td>
<td>Affymetrix- HG-U133_Plus_2</td>
<td>3</td>
<td>GSE17708</td>
<td>Keshamouni and colleagues(2010) PMID: 1918450</td>
</tr>
<tr>
<td>A549</td>
<td>Lung</td>
<td>Affymetrix- HG-U133_Plus_2</td>
<td>3</td>
<td>GSE49644</td>
<td>Sun and colleagues (2014) PMID: 25379179</td>
</tr>
<tr>
<td>HCC827</td>
<td>Lung</td>
<td>Affymetrix- HG-U133_Plus_2</td>
<td>3</td>
<td>GSE49644</td>
<td>Sun and colleagues (2014) PMID: 25379179</td>
</tr>
<tr>
<td>NCI-H358</td>
<td>Lung</td>
<td>Affymetrix- HG-U133_Plus_2</td>
<td>3</td>
<td>GSE49644</td>
<td>Sun and colleagues (2014) PMID: 25379179</td>
</tr>
<tr>
<td>HK-2</td>
<td>Kidney</td>
<td>Affymetrix- HG-U133A_2</td>
<td>3</td>
<td>GSE23338</td>
<td>Walsh and colleagues (2008) PMID: 25379179</td>
</tr>
<tr>
<td>Human proximal tubular cell line</td>
<td>Kidney</td>
<td>Illumina- HumanWG-6 v3.0</td>
<td>3</td>
<td>GSE20247</td>
<td>Hills and colleagues (2010) PMID: 20973508</td>
</tr>
<tr>
<td>Panc-1</td>
<td>Pancreas</td>
<td>Affymetrix- HG-U133_Plus_2</td>
<td>3</td>
<td>GSE23952</td>
<td>Maupin and colleagues (2010) PMID: 20885998</td>
</tr>
</tbody>
</table>

NOTE: To identify a signature for TGFβ-induced EMT that is not specific to a particular type of cancer, we used cell lines that represent a variety of cancers.

Figure 1.

A, Distribution of gene scores by product of ranks. The product of ranks (PR) value for each gene was calculated across all 10 datasets studied and then log2-transformed (histogram). A permuted null distribution was calculated (density line; $n = 119 \times C^{106}$), and a significance threshold was specified at 99.999% of the cumulative density (dashed line) to identify genes with a product of rank lower than expected ($P < 5 \times 10^{-6}$). B, Overlap of our TGFβ-EMT signature (Foroutan) with more general EMT signatures reported by Cursons and colleagues (8) and Du and colleagues (36).
known EMT markers to classify epithelial or mesenchymal cell lines and tumors and then derived their signature.

Supporting our suggestion that stimulus-specific differences underlie the differences in the above signatures, our TGFβ-EMT signature had the greatest overlap with Du and colleagues (using TGFβ and OSM) and the lowest overlap with the EGF/hypoxia-induced EMT signature from Cursons and colleagues (Fig. 1B). In general, the signature obtained in the study by Cursons and colleagues shared the least genes with other studies (Supplemental Results, Supplementary Fig. S1).

There is evidence of active TGFβ signaling in CD44+ breast cancer cells and embryonic stem cells (37, 38) and examining overlap systems.

A wide range of cancer cell lines show graded variations in scoring metrics derived from the signature of TGFβ-induced EMT

To identify cancer cell lines with evidence of TGFβ-induced EMT, we created a relative TES which quantified the degree of concordance with our signature. Using both the ssGSEA (40) and GSVA (41) single-sample scoring methods, scores from the 108 down- and 193 upregulated gene sets in our signature were calculated separately and then summed (Supplementary Methods) to obtain summed-ssGSEA (Figs. 2 and 3) and summed-GSVA (Supplementary Results, Supplementary Figs. S3 and S4) scores. These were calculated for the experimental data used to derive our TES (Fig. 2A), the Neve and colleagues breast cancer cell line data (nCellLines = 54; Fig. 2B; ref. 30), the NCI-60 data (nCellLines = 59; Fig. 2C; refs. 27, 42), and CCLE (nCellLines = 1,036; Fig. 2D; all scores given in Supplementary Tables S7–S11; refs. 43, 44). A comparison between GSVA and ssGSEA scores is provided in Supplementary Fig. S5 (Supplementary Results).

Samples with a high TES show strong concordance with our signature and are thus predicted to have undergone some form of TGFβ-induced EMT. As expected, TGFβ-treated samples used to derive the signature showed higher TES than the control samples (Fig. 2A). All high TES breast cancer cell lines within the Neve data were from the very aggressive basal B subgroup (Fig. 2B), which has CSC-like characteristics and a mesenchymal phenotype (30, 39). Across the NCI-60 (Fig. 2C) and CCLE (Fig. 2D) data, very consistent results were obtained for each cancer type. When cell lines were ranked by their TES and classified as “high TES” (top 10%), “medium TES” (middle 10%), and “low TES” (bottom 10%), however, there appeared to be some enrichment for different tissues of origin. Within the CCLE data, most high TES cell lines are derived from the central nervous system, and there are a number of skin and bone cancer cell lines (Fig. 2D). Conversely, the majority of the low TES cancer cell lines are from the large intestine, followed by lung and breast cancer cell lines (Fig. 2D). Considering the medium TES group, most cell lines are derived from hematopoietic and lymphoid tissues and lung cancers (Fig. 2D). Lists of cell lines with very high, very low, and medium TES are given in Supplementary Table S4.

The Neve, NCI-60, and CCLE data shared some common cell lines, and the TES calculated across these independent studies was very consistent (Fig. 2E and F), suggesting that our TGFβ-EMT signature is stable and robust across different experimental systems.

High TES cell lines also show a high mesenchymal score

To quantify changes across the landscape of epithelial and mesenchymal phenotypes, we used both ssGSEA (Fig. 3A and B) and GSVA (Supplementary Results, Supplementary Fig. S4) to derive epithelial and mesenchymal scores (ES and MS, respectively) from the corresponding Tan and colleagues gene sets (7). For TGFβ-EMT data used to derive our signature, epithelial scores provided reasonable separation of control and TGFβ-treated samples, whereas mesenchymal scores gave good separation (Fig. 3A and B). For CCLE samples grouped by low, medium, and high TES, we extracted the TES, epithelial score, and mesenchymal score (Fig. 3C–E). Low TES cell lines tend to have lower mesenchymal and higher epithelial scores, whereas samples with a high TES have higher mesenchymal and low epithelial scores. Interestingly, samples with higher TES showed a greater range of epithelial and mesenchymal score (Fig. 3D). When all CCLE cell lines considered together, there is a strong negative correlation between TES and epithelial score (Fig. 3F, r = −0.83, P < 2e-16) and positive correlation between TES and mesenchymal score (Fig. 3G, r = 0.80, P < 2e-16). As noted above, a number of stimuli can drive EMT and this raises the possibility that cell lines with a low/moderate TES but relatively high mesenchymal score represent those with a different molecular etiology inducing EMT.

We also scored CCLE using previously identified signatures of general EMT (Supplementary Results, Supplementary Fig. S6). Across the CCLE data, there was a high correlation between TES and scores derived from the Du (r = 0.97), Groset (r = 0.95), or Taube EMT signatures (r = 0.91), all of which used TGFβ as one of their stimuli. It should be noted, however, that there were a number of samples, particularly with lower TES where these metrics diverged (Supplementary Results, Supplementary Fig. S6). Conversely, our TES had almost no association (r = 0.25) with scores from the Cursons and colleagues signature which was derived using EGF and hypoxia stimulation. This appears to support the earlier observation of stimulus-dependent similarities and differences between EMT signature list overlap.

Cell lines with high TES tend to carry a lower mutation load within TGFβ signaling elements

We extracted gene names for all components of the KEGG and Reactome TGFβ signaling pathways (nGenes = 98; Supplementary Table S5), and for CCLE cell lines with a high, medium, and low TES, we looked at the frequency of mutations within these genes (Table 2). The TGFβ mutational load for each group was calculated as the ratio of total mutations in TGFβ pathway genes to the number of cell lines in that group.

The high TES group contains a lower overall relative mutational load within TGFβ signaling pathway genes (Table 2), and although...
mutations were observed for \( TGFB2 \) and \( TGFBR2 \), other important upstream regulators such as \( TGFB1, TGFB3, \) and \( TGFBR1 \) were not mutated within high TES cell lines (Table 2). Furthermore, in high TES cell lines, 69% of mutations within TGF\( \beta \) pathway components are not predicted to change protein primary sequence (Table 3, details in Supplementary Table S6). Conversely, in low TES cell lines, only 32% of TGF\( \beta \) pathway mutations are not predicted to affect protein sequence (Supplementary Table S6).

Figure 2.
Relative TES for cancer cell lines from different datasets. Up- and downregulated sets from our 301-gene signature were scored and combined using a summed-ssGSEA approach. This was applied to (A) the integrated TGF\( \beta \)-EMT dataset (Table 1), (B) Neve and colleagues cell lines as clustered, (C) NCI-60 cell lines, and (D) CCLE cell lines. Where cell lines are present in multiple datasets, we calculated the correlation between TES, for (E) CCLE and NCI-60 data, and (F) CCLE and Neve data, demonstrating that scores are highly correlated in independently measured data.
More than half of the TGFβ component mutations in the high TES group occur within the 3′-untranslated region (UTR; 53%), whereas missense mutations (28%) and 5′-UTR mutations (13%) are also relatively common (Table 3). For the low TES group, however, a large proportion of TGFβ mutations are missense (42%) or 3′-UTR (21%) mutations, whereas frameshift deletions
are more frequent (16%) compared with the high TES group (Table 3). When the mutation frequency in TGFβ signaling genes was normalized against the total number of mutations in each cell line, there was no difference between the high TES and low TES groups, implying that lower TES is associated with higher genomic mutational load.

High TES group responds differently to certain drugs

A wide range of cancer samples show graded variations in scoring metrics derived from the TGFβ-induced EMT signature

A number of drugs targeting TGFβ pathway elements are in development for clinical applications (22); thus, we examined whether our signature could identify clinical samples with evidence of TGFβ-induced EMT. We used ssGSEA to score the TCGA pan-cancer data (Fig. 5A) and breast cancer data (Supplementary Results, Supplementary Fig. S12) against our TGFβ-EMT signature, the Tan and colleagues (2014) tumor epithelial and mesenchymal signatures (7), as well as the Du (36), Groger (6), Taube (5), and Cursons (8) EMT signatures. For breast cancer samples where corresponding microarray data were available ($n_{\text{tumours}}$ = 571), we also calculated the TES, epithelial score, and mesenchymal score and found a high correlation (Supplementary Results, Supplementary Fig. S13), suggesting that our results...
are robust to artifacts associated with different detection platforms.

The TCGA pan-cancer mRNA sequencing data includes 9,756 clinical samples covering 34 cancer types from 29 tissues. The tumor TES of individual cancer types (Fig. 5A) show a very similar distribution to the corresponding cell line data (Fig. 2C and D)—brain, bone, and skin cancer samples tend to display a high TES, hematopoietic cancer samples have an intermediate TES, whereas colorectal cancers have a low TES. Segregating pan-cancer tumors with a high (top 10%; $n_{\text{tumors}} = 976$), moderate (middle 10%), or low (bottom 10%) TES and examining the epithelial and mesenchymal scores (Fig. 5B–D), a high TES was associated with tumors that have high mesenchymal and low epithelial scores. Similar results were observed for the TCGA breast cancer data (Supplementary Results, Supplementary Fig. S12), although there was less agreement between TES and mesenchymal scores for the breast cancer samples, suggesting that TGFβ may not be a common driver of mesenchymal behavior across all breast cancers.

Analysis of the TCGA pan-cancer mutation data showed that samples with high TES have relatively fewer mutations in TGFβ signaling elements compared to samples with low TES, with fewer patients having at least one mutation in the TGFβ signaling elements (35% vs. 47%). These differences, while more modest,
are none-the-less consistent with the results of our cell line analysis, demonstrating that even taking intratumor heterogeneity into account, the genomic observations translate from cell lines to patient samples.

Across all patients, there were significant \((P < 2e-16)\) correlations between TES and mesenchymal scores (Supplementary Results, Supplementary Fig. S8) for breast cancer \((\rho = 0.57)\) and pan-cancer data \((\rho = 0.53)\) and between TES and epithelial scores for breast cancer \((\rho = -0.44)\) and pan-cancer \((\rho = 0.73)\) data. Examining other EMT signatures across the TCGA pan-cancer data, our TES was highly correlated with scores from the Du \((\rho = 0.93)\), Groger \((\rho = 0.80)\), and Taube \((\rho = 0.66)\) signatures but not correlated with the Cursons’ EMT score \((\rho = 0.07);\) Supplementary Results, Supplementary Fig. S7), which intriguingly shows 2 populations with separate TES but a range of overlapping Cursons’ EMT scores. These results again support the notion that signatures derived using similar stimuli are highly associated, whereas EMT signatures derived using alternative stimuli are largely independent, highlighting the need for stimulus-specific EMT signatures.

Given that stromal cells often have a more mesenchymal phenotype and modulate TGFβ signaling (47), it is possible that tumor TES is predominantly derived from stromal cells within the tissue biopsy. To examine this, we calculated the correlation between TES and the consensus measurements of purity estimation (CPE; ref. 48) for TCGA pan-cancer samples (Fig. 5E). There was a weak but significant negative correlation \((\rho = -0.20, P < 2e-16)\) between TES and tumor purity across all cancer types, suggesting that lower purity (i.e., more stromal contamination) is associated with a similar or higher TES. Considering each cancer type separately, weak negative correlations were also observed between TES and tumor purity for all cancer types except glioblastoma \((\rho = 0.003)\) and adrenocortical carcinoma \((\rho = 0.20, P = 0.003, Fig. 5G)\). This is particularly interesting, as colorectal cancers tend to have a very low TES; thus, despite the observation that this tumor type is the most likely to show an inflamed TES due to stromal contamination, we still see low TES overall. Conversely, we see no correlation between tumor purity and TES in glioblastoma \((\rho = 0.003, Fig. 5G)\), despite this tumor type being among those with the highest TES on average. In addition, we calculated the TES for normal and TGFβ-stimulated fibroblasts (Supplementary Results, Supplementary Fig. S15) and found that the TES of TGFβ-stimulated fibroblasts was only slightly higher than untreated samples, perhaps reflecting the specificity

Figure 5.
A, Relative TES for the TCGA pan-cancer data using the ssGSEA method. B-D, Comparing TES, epithelial scores (ES), and mesenchymal scores (MS) in the TCGA pan-cancer data with tumors classified as (B) low TES, (C) medium TES, and (D) high TES. In general, samples with a high TES have a lower ES and higher MS, whereas samples with a low TES tend to have a higher ES and lower MS. E-G, Relationship between TES and tumor purity (CPE) in (E) TCGA pan-cancer data \((n_{\text{samples}} = 7,805)\), (F) colorectal cancer with the strongest negative correlation, and in (G) glioblastoma with the least correlation.
of our signature toward EMT. Collectively, these results demonstrate that even though tumor purity does show some influence on TES, the influence is weak, particularly in cancer types with a very high TES.

The TES effectively separates pan-cancer patient cohort with significant differences in survival rates

Tumors with a high TES ($n_{\text{Tumors}} = 976$) from the TCGA pan-cancer data were associated with a significantly lower survival rate than samples with low TES ($n_{\text{Tumors}} = 976$; Fig. 6A; HR, 0.6; $P = 2e-09$). To investigate whether this effect can be attributed to different survival rates for tumors that could be separated by their relative TES, we compared the overall survival rates of all cancer types. As shown in Fig. 5A, skin and bone tumors have a higher TES; however, these tumor types also have longer median survival time (Fig. 6B), suggesting that tumor-dependent differences in survival may not confound the apparent TES-mediated differences in survival time (Fig. 6A).

Although we observed a highly significant difference in overall survival for patients with low TES versus high TES tumors across the TCGA pan-cancer data, this difference was not statistically significant within all cancer types (Supplementary Results, Supplementary Fig. S14). Highly significant, poor survival rates were associated with high-versus-low TES samples in brain ($P = 2e-5$), kidney ($P = 3e-4$), and lining of body cavity cancers ($P = 0.003$). For cancers of adrenal gland, thyroid gland, eye, and bladder, smaller but significant ($P < 0.05$) differences were seen between survival rates of the low and high TES groups (Supplementary Results, Supplementary Fig. S14). For other cancer types, no significant differences were observed in survival rates of the patients with high versus low TES tumors; however, the trend was similar, as the high TES groups tended to have a worse survival outcome. It should be noted, however, that white blood cell ($P = 0.04$) and head and neck ($P = 0.15$) cancers showed the opposite association, as low TES was associated with poor survival.

Figure 6.

A, Overall survival of pan-cancer patients with high versus low TES ($n_{\text{Tumors}} = 976$ in each group, $P = 2e-9$, HR = 0.6). B, Overall survival rates of different cancer types with 10-year censoring (ordered by median values for overall survival).
Discussion

We have derived a transcriptional signature for TGFB-driven EMT and showed consistent scoring results across large sets of cell line and patient data. This analysis highlights basal breast cancer cell lines and clinical brain, bone, and skin tumors as malignancies with relatively high TGFB signaling activity, whereas colorectal and endometrium cancers tend to show lower scores. The association our TES shows with epithelial and mesenchymal scores suggests that TGFB signaling contributes to EMT in numerous cancer cell lines (Fig. 3C–G) and tumors (Fig. 5B–D); however, there are interesting discrepancies. It is tempting to speculate that samples with a high mesenchymal/general EMT score and low TES represent malignancies where alternative drivers of EMT are active, and conversely, samples with a high TES but low general EMT scores are more likely to display evidence of TGFB-driven EMT (Supplementary Results). This illustrates the advantage of scoring multiple signatures associated with molecular phenotypes to gain a deeper understanding of tumor biology and guide the development and application of targeted therapies. In this context, we note that Tan and colleagues (7) combine their epithelial and mesenchymal scores into a single metric; however, samples with a specific TES can show a range of epithelial or mesenchymal score values (Figs. 3F and G and 5B–D), suggesting that summation into a single metric may hide meaningful variation. The range of correlations between our TES and general EMT signature scores appeared to reflect the relative contribution of TGFB stimulation in the study design (Supplementary Figs. S6 and S7, Supplementary Results), highlighting the need for more specific signatures of EMT, such as the one we present here, that are matched to their underlying molecular etiology.

Examining mutation frequencies in genes of TGFB signaling components, samples with a higher TES tended to contain fewer mutations and these were localized to noncoding regions. None of the 3'-UTR, 5'-UTR, or intron mutations were predicted to cause protein sequence alterations, although some may influence mRNA binding, alternative splicing, or transcript stability (49–51). Given the much greater number of mutations within TGFB signaling pathway genes for the low TES group, it is likely that many of these mutations tend to repress TGFB signaling activity. These data also suggest that for cells within the high TES group, the TGFB signaling pathway tends to be intact, implying that aberrant signaling is not driven by loss-of-function mutations in inhibitory pathway components, and indicating that these cells may respond to targeted inhibitor therapies. In agreement with this, Park and colleagues recently used a novel TGFBR1 inhibitor, IN-1130, to successfully block TGFB-induced EMT in MCF10A and MDA-MB-231 breast cancer cell lines (52) both of which have a relatively high TES (Supplementary Tables S8 and S9).

The effects of EMT on drug efficacy are contentious. Recent reports suggest that EMT can drive general drug resistance (46, 53); however, our analysis shows that high TES and low TES groups respond differentially only for specific drugs, more consistent with the observations of Tan and colleagues which examined response across a spectrum of epithelial and mesenchymal phenotypes (7). We note a number of specific differences, particularly for cancer types such as bone and skin that tend to have a higher TES. For example, Tan and colleagues reported that more mesenchymal melanoma cell lines were resistant to GSK650394 (SGK1/SGK2 inhibitor), whereas those with a more epithelial phenotype were resistant to docetaxel. Our results showed no correlation between melanoma cell line TES and IC50 values for GSK650394 (p = 0.02) and docetaxel (p = 0.05) but instead suggested that high TES cell lines tend to show panherinolide resistance and sensitivity to CGP60474 (CDK1/CDK2 inhibitor). Furthermore, we found that bone cancer cell lines with high TES tended to be epothilone (microtubule inhibitor)-resistant and erlotinib (EGFR inhibitor)-sensitive, whereas Tan and colleagues found little correlation for the efficacy of these drugs. These observations concur with our previous work that demonstrated drug responses were strongly influenced by the nature of the stimulus that generated EMT and highlight the need for stimulus-specific signatures such as the one we present here.

We believe our TES can be used to inform a number of therapeutic options. The correlative associations indicating resistance for several drugs in high TES cell lines raises the possibility that combination therapies with TGFB signaling inhibitors may increase sensitivity for these drugs. Furthermore, a number of studies have shown that chemotherapy-induced TGFB signaling can promote a CSC phenotype and/or drug resistance. In this context, combination therapy using TGFB inhibitors and chemotherapeutic agents has been examined in cell line and xenograft models (54–57) where they show promise (56–58). Finally, for high TES samples, the application of TGFB inhibitors alone may be a potential treatment strategy. In support of this, antisense TGFB1/2 oligonucleotides (e.g., AP12009) are efficacious in the clinical treatment of patients with aggressive gliomas (59), which have a high TES.

When working with complex tissue samples such as those derived from tumor biopsies, there is likely to be at least some degree of stromal contamination. We show that for TCCA pancreas cancer samples, tumor purity (CPE) and TES have a weak negative correlation (Fig. 5E–G), suggesting that higher TES values may be associated with an increased fraction of stromal cells. The strongest negative association between TES and tumor purity was observed for colorectal cancers, suggesting that they may be sensitive to confounding effects from tumor purity. Intriguingly, however, these tumor samples had a relatively low TES, whereas brain cancers (e.g., glioblastoma and low-grade glioma) which tend to show higher TES values often had greater sample purity, suggesting that stromal contamination is not a dominant effect. We note that TGFB mediates stromal–tumor interactions and can drive tumor microenvironment remodeling to promote progression at all stages of carcinogenesis (60, 61), which may also influence the TES; however these effects were not explored in detail.

The relationship between patient survival and EMT-related gene signatures is discordant across the literature. These inconsistencies may reflect the role of phenotypic switching in disease progression, and the need for MET in establishment and growth of metastasis (62, 63); alternatively, these results may have arisen due to difficulties with taking a biopsy from the invasive front of a clinical tumor sample. For example, Taube and colleagues found no correlation between their core EMT signature and breast cancer patient survival (5), whereas Tan and colleagues showed a tumor-type–dependent association between their EMT signature and survival and surprisingly better overall survival outcomes for mesenchymal breast cancer and malignant melanoma (7). Conversely, Shipitsin and colleagues reported an association between poor distant metastasis-free survival and higher expression values for TGFB signaling pathway elements that were upregulated in CD44+ breast cancer samples for one of their datasets (37).
We show that tumors with a high TES are significantly associated with poor survival within the TCGA pan-cancer data, but there were large differences between cancers. The association between TES and survival was significant for brain, kidney, lining of body cavities, adrenal gland, thyroid gland, and eye cancers (P<0.05). For some cancer types, there was no significant association between TES and survival, although a similar trend was observed, such that the high TES group tended to have reduced survival. Surprisingly, for hematopoietic or head and neck cancers, a lower TES was associated with reduced survival rates. This may reflect the fact that hematopoietic cells are not epithelial-derived, thus different molecular mechanisms may be active, whereas a large proportion of the head and neck cancers is squamous cell carcinomas where TGFβ signaling elements such as TGFBR2, SMAD2 and SMAD2 are strongly downregulated (64). Consistent with our observation, it has been shown that patients with head and neck squamous cell carcinoma (HNSCC) who do not have phospho-SMAD2/3 show better survival than patients with active phospho-SMAD2/3 (65).

Given the increasing availability of experimental and clinical drugs targeting the TGFβ signaling pathway (22), our TGFβ-EMT gene signature may be useful to identify clinical subsets of patients who will benefit from targeted treatment to inhibit metastasis. We believe that our results illustrate several advantages for stimulus-specific classification of EMT programs; in particular, the ability to pair signatures may facilitate improved molecular stratification of cancer subtypes for targeted therapies. The future of personalized therapeutics will certainly benefit from further development of driver-specific signatures for clinically relevant phenotypic programs to better target therapies.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

References

Acknowledgments
The authors would like to acknowledge the TCGA Research Network (http://cancergenome.nih.gov/) for generating the data used in this article. We thank Professor Greg Goodall, Associate Professor Frederic Hollande, and Dr. Hong Jian Zhu for the valuable suggestions and feedback and Professor Terry Speed for his advice in data analysis.

Grant Support
M. Foroutan was supported by Melbourne International Fee Remission Scholarship (MIFRS) and Melbourne International Research Scholarship (MIRS). J. Cursons was supported by Australian Research Council Centre of Excellence in Convergent Bio-Nano Science and Technology (project number CE100100036). E.W. Thompson was supported in part by the EmPathy Breast Cancer Network (CG-10-04) a National Collaborative Research Program of the National Breast Cancer Foundation, Australia. The Translational Research Institute is supported by a grant from the Australian Government. M.J. Davis was supported by National Breast Cancer Foundation (NBFC-ECF-043-14).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received September 27, 2016; revised December 19, 2016; accepted January 4, 2017; published online First January 23, 2017.
Characterization of a TGFβ-Induced EMT Signature


