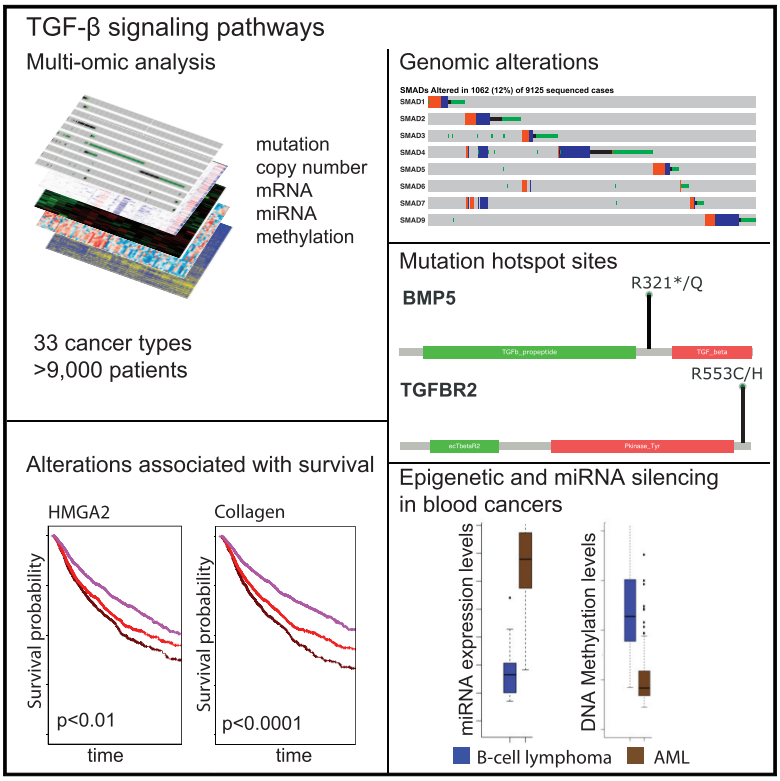
Article

Cell Systems

A Pan-Cancer Analysis Reveals High-Frequency Genetic Alterations in Mediators of Signaling by the TGF-β Superfamily

Graphical Abstract



Highlights

- Genetic alterations in TGF-β pathway members observed in 39% of TCGA cases
- GI cancers enriched with hotspot mutations in TGF-β pathway members
- Gene alterations correlated with expression of metastasis genes and poor prognosis
- TGF-β signaling silenced by miRNAs or DNA methylation in hematologic cancers

Authors

Anil Korkut, Sobia Zaidi, Rupa S. Kanchi, ..., John N. Weinstein, Lopa Mishra, Rehan Akbani

Correspondence

lopamishra2@gmail.com (L.M.), rakbani@mdanderson.org (R.A.)

In Brief

To date, there are no studies of the TGF-β superfamily of signaling pathways across multiple cancers. This study represents a key starting point for unraveling the role of this complex superfamily in 33 divergent cancer types from over 9,000 patients.

Korkut et al., 2018, Cell Systems 7, 422–437
October 24, 2018
https://doi.org/10.1016/j.cels.2018.08.010

CellPress

OPEN ACCESS
CellPress

Cell Systems
Article

A Pan-Cancer Analysis Reveals High-Frequency Genetic Alterations in Mediators of Signaling by the TGF-β Superfamily

Anil Korkut,[1] Sobia Zaidi,[2] Rupa S. Kanchi,[1] Shuyun Rao,[2] Nancy R. Gough,[2] Andre Schultz,[1] Xubin Li,[1] Philip L. Lorenzi,[1] Ashton C. Berger,[3] Gordon Robertson,[4] Lawrence N. Kwong,[5] Mike Datto,[6] Jason Roszik,[7] Shiyun Ling,[1] Visweswaran Ravikumar,[1] Ganiraju Manyam,[1] Arvind Rao,[1] Simon Shelley,[8] Yuexin Liu,[1] Zhenlin Ju,[1] Donna Hansel,[9] Guillermo de Velasco,[10,11] Arjun Pennathur,[12] Jesper B. Andersen,[13] Colm J. O'Rourke,[13] Kazufumi Ohshiro,[2] Wilma Jogunoori,[2,14] Bao-Ngoc Nguyen,[2] Shulin Li,[15] Hatice U. Osmanbeyoglu,[16] Jaffer A. Ajani,[17] Sendurai A. Mani,[5] Andres Houseman,[18] Maciej Wiznerowicz,[19,20,21] Jian Chen,[22] Shoujun Gu,[2] Wencai Ma,[1] Jiexin Zhang,[1] Pan Tong,[1] Andrew D. Cherniack,[3] Chuxia Deng,[2,23] Linda Resar,[24] The Cancer Genome Atlas Research Network, John N. Weinstein,[1,25] Lopa Mishra,[2,14,*] and Rehan Akbani[1,26,*]

[1]Department of Bioinformatics and Computational Biology, MD Anderson Cancer Center, Houston, TX 77030, USA
[2]Center for Translational Medicine, Department of Surgery, George Washington University, Washington, DC 20037, USA
[3]Cancer Program, The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02142, USA
[4]Canada's Michael Smith Genome Sciences Center, BC Cancer Agency, Vancouver, BC V5Z 4S6, Canada
[5]Department of Translational Molecular Pathology, University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[6]Department of Pathology, Duke School of Medicine Durham, Durham, NC 27710, USA
[7]Department of Melanoma Medical Oncology and Genomic Medicine, MD Anderson Cancer Center, Houston, TX 77030, USA
[8]Department of Medicine, University of Wisconsin School of Medicine and Public Health, Madison, WI 53726, USA
[9]Department of Pathology, University of California, San Diego, La Jolla, CA 92093, USA
[10]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[11]Department of Medical Oncology, University Hospital 12 de Octubre, Madrid 28041, Spain
[12]Department of Cardiothoracic Surgery, University of Pittsburgh School of Medicine and University of Pittsburgh Medical Center, Pittsburgh, PA 15213, USA
[13]Department of Health and Medical Sciences, Biotech Research and Innovation Centre, University of Copenhagen, Ole Maaloes Vej 5, Copenhagen 2200, Denmark
[14]Veterans Affairs Medical Center, Institute of Clinical Research, Washington, DC 20422, USA
[15]Department of Pediatrics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[16]Memorial Sloan Kettering Cancer Center, Computational & Systems Biology Program, New York, NY 10065, USA
[17]Department of GI Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[18]College of Public Health and Human Sciences, Oregon State University, Corvallis, OR 9733, USA
[19]Poznań University of Medical Sciences, Poznań 61701, Poland
[20]Greater Poland Cancer Center, Poznań 61866, Poland
[21]International Institute for Molecular Oncology, Poznań 60203, Poland
[22]Department of Gastroenterology, Hepatology & Nutrition, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[23]Faculty of Health Sciences, University of Macau, Macau, Macau SAR, China
[24]Departments of Medicine, Division of Hematology, Oncology and Pathology, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
[25]Department of Systems Biology, MD Anderson Cancer Center, Houston, TX 77030, USA
[26]Lead Contact
*Correspondence: lopamishra2@gmail.com (L.M.), rakbani@mdanderson.org (R.A.)
https://doi.org/10.1016/j.cels.2018.08.010

SUMMARY

We present an integromic analysis of gene alterations that modulate transforming growth factor β (TGF-β)-Smad-mediated signaling in 9,125 tumor samples across 33 cancer types in The Cancer Genome Atlas (TCGA). Focusing on genes that encode mediators and regulators of TGF-β signaling, we found at least one genomic alteration (mutation, homozygous deletion, or amplification) in 39% of samples, with highest frequencies in gastrointestinal cancers. We identified mutation hotspots in genes that encode TGF-β ligands (BMP5), receptors (TGFBR2, AVCR2A, and BMPR2), and Smads (SMAD2 and SMAD4). Alterations in the TGF-β superfamily correlated positively with expression of metastasis-associated genes and with decreased survival. Correlation analyses showed the contributions of mutation, amplification, deletion, DNA methylation, and miRNA expression to transcriptional activity of TGF-β signaling in each cancer type. This study provides a broad molecular perspective relevant for future functional and therapeutic studies of the diverse cancer pathways mediated by the TGF-β superfamily.

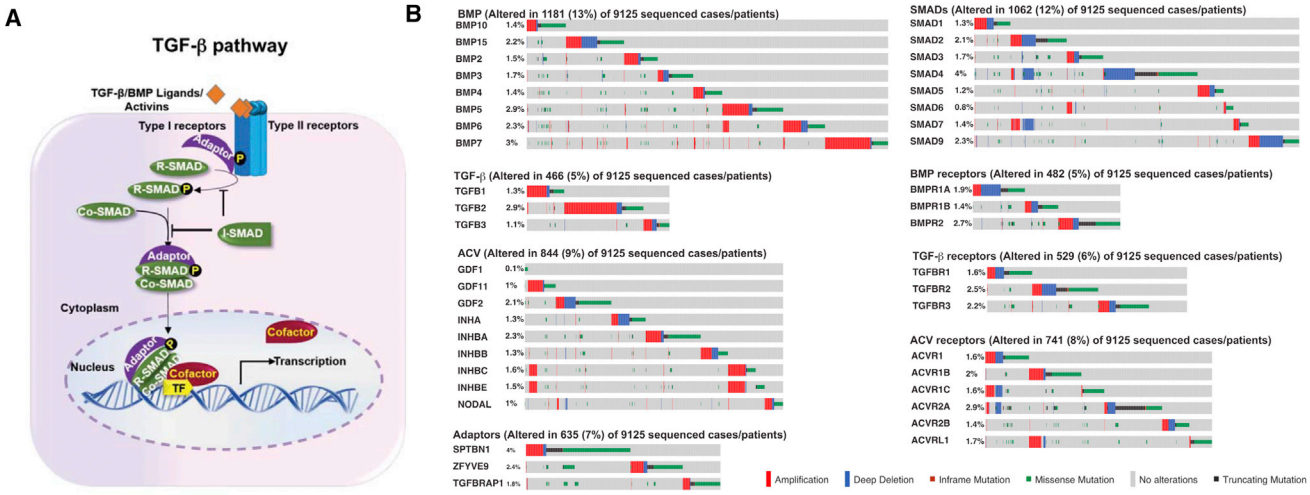


Figure 1. TGF-β Pathway and Its Genomic Alterations in Cancer

(A) The canonical TGF-β pathway. TGF-β superfamily member ligands bind to type II receptors, leading to recruitment and activation of type I receptors through phosphorylation. Subsequently, the activated receptors phosphorylate intracellular R-SMADs, such as SMAD2 and SMAD3, which bind to the receptor through adaptor molecules. The R-SMAD/co-SMAD (SMAD2/3-SMAD4) complex is transported into the nucleus to induce transcriptional programs regulated by the TGF-β superfamily.

(B) Landscape of genomic aberrations in the TGF-β superfamily genes in cancer. The frequency of alterations in TGF-β superfamily ligands, receptors and receptor-associated proteins, intracellular SMADs, and adaptor molecules are presented. Only samples with genomic alterations in the indicated genes are shown in each oncoprint. Alteration rates per gene and gene family are displayed in the left and top labels, respectively.

See also STAR Methods; Figure S1; Tables S1 and S2.

INTRODUCTION

The transforming growth factor β (TGF-β) superfamily of ligands activates Smad proteins to regulate transcription and control cell proliferation and differentiation. The TGF-β pathways are context-dependent signal transduction cascades that can promote seemingly contradictory cell processes, including promotion of differentiation and tumor growth, inhibition of cell proliferation, suppression of immune response, and maintenance of stem cell homeostasis (Akhurst, 2017; Colak and Ten Dijke, 2017; Seoane and Gomis, 2017; Christian and Heldin, 2017; Moustakas and Heldin, 2016; Mishra et al., 2005; Wakefield and Roberts, 2002). Animal models of mammary gland tumorigenesis support a pro-tumorigenic role for signaling by the TGF-β1-Smad2 pathway (Muraoka-Cook et al., 2004), whereas mouse models of gastrointestinal (GI) cancers and hepatocellular cancers indicate a primarily tumor-suppressive role (Chen et al., 2018; Chen et al., 2016b; David et al., 2016; Katz et al., 2016). In pancreatic KRAS-mutant premalignant cells, TGF-β signaling induces expression of metastasis-promoting genes (David et al., 2016) and apoptosis-regulatory genes. Thus, even within a single subfamily of ligands that act through the same downstream Smad complexes, the net outcome can be either tumor-suppressing or tumor-promoting depending on the context. Hence, predicting appropriate TGF-β-based therapeutic interventions is challenging.

To dissect the context-specific roles of the TGF-β pathway across multiple cancer types, we focused on 43 core genes that regulate or mediate TGF-β signaling. We selected the core genes through consensus of The Cancer Genome Atlas (TCGA) TGF-β network members, although we acknowledge that the process of identifying a core subset of genes is inherently subjective to some degree. The "integromic" analysis (Weinstein, 2006) described here reveals potential nodes of crosstalk with other cancer-relevant pathways, and it enables prediction of the activity of TGF-β-Smad pathways in various cancer contexts. The data and analyses provide a rich resource for understanding TGF-β biology, with the potential to identify context-dependent therapeutic targets.

RESULTS

We focus here on the genomic, epigenomic, and transcriptomic landscape of 43 genes that encode proteins that mediate or regulate signaling by the TGF-β superfamily and 50 downstream target genes of Smad-dependent signaling in 9,125 patients across 33 TCGA tumor types (https://tcga-data.nci.nih.gov/docs/publications/tcga/) (Table S1), referred to as the "Pan-Cancer cohort." The analysis is limited to this set of TGF-β pathway-related genes and yet represents a valuable starting point to examine TGF-β signaling across multiple cancers. We analyzed multiple data types: somatic copy number variation (CNV), point mutation, DNA methylation, mRNA expression (from mRNA-seq), miRNA expression (from miRNA-seq), and, for correlative analyses, protein expression (from reverse-phase protein arrays [RPPA]). The data were corrected for batch effects and other systematic biases prior to analysis (see STAR Methods).

Selection of Genes Associated with the TGF-β Superfamily

The list of 43 "core" TGF-β genes includes 2 genes encoding adaptor proteins (*SPTBN1* and *ZFYVE9*) that are important in TGF-β signaling and play roles in other cellular processes. The other 41 genes encode components of each level of the "canonical" TGF-β signaling pathway that activates Smads to regulate gene expression (Figure 1A): 3 TGF-β ligands, 8 bone morphogenetic protein (BMP) ligands, and 9 activin (ACV) ligands; 3 TGF-β

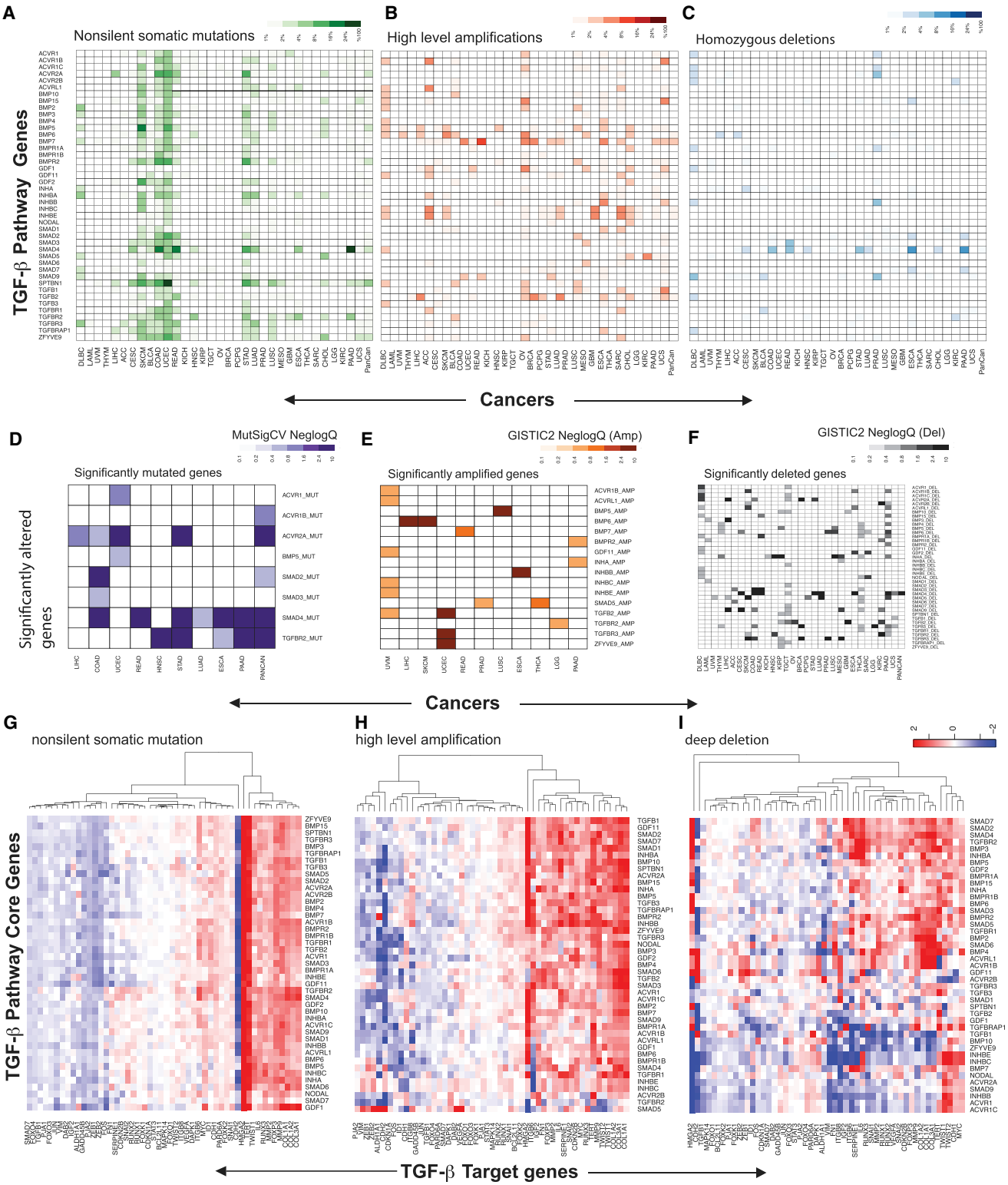


Figure 2. Pan-Cancer Genomic Analysis of the 43 TGF-β Superfamily Pathway Genes in 33 Cancer Types

(A–C) Distribution of genomic alterations over cancer types. (A) Non-silent somatic mutations, (B) copy number amplifications, and (C) homozygous deletion frequencies. SKCM, UCEC, STAD, and COAD show high overall mutation rates.

(legend continued on next page)

receptors and 1 interacting protein (TGFBRAP1), 3 BMP receptors, and 6 ACV receptors; and 8 Smads (Figure 1B). The list of 43 genes is available at cBioPortal (http://www.cbioportal.org) as "General: TGF-β superfamily." Noncanonical signaling (Figure S1A) is excluded from this analysis. Figure S1B shows pairwise correlation coefficients of the 43 genes.

To explore the effect of TGF-β pathway genomic alterations on transcriptional output and to validate pathway activity, we selected a panel of 50 downstream target genes that are regulated by TGF-β-Smad signaling and have important roles in epithelial-to-mesenchymal transition (EMT), metastasis, or tumor suppression (Table S1).

Genomic Alterations in TGF-β Superfamily Genes

We performed mutation and CNV analyses of the 43 genes to identify genomic aberrations across the Pan-Cancer cohort (Figure 1B and Table S2). Using the cBioPortal definitions (Cerami et al., 2012), genomic alterations were classified as gene amplifications, gains (low-level amplifications), deep deletions (equivalent to homozygous deletions for non-aneuploidy cases), shallow deletions (heterozygous loss), truncating mutations, in-frame mutations, or missense mutations. We use the term "alteration" henceforth for mutations (truncating or missense) and CNVs (deep deletion or amplification). Oncoprint representation from cBioPortal revealed the distribution of TGF-β genomic alterations in the Pan-Cancer cohort (Figure 1B). Although alteration frequencies were low, 39% of the tumors contained an alteration in at least one of the 43 genes. *SMAD4* (4%) and *SPTBN1* (4%) were the most frequently altered. Collectively, BMP ligands had an alteration frequency of 13%. Six genes (*GDF1*, *GDF11*, *SMAD6*, *SMAD7*, *INHBE*, and *NODAL*) had mutation frequencies <0.5% (Table S2). When excluding those six, cumulative mutation frequency (23%) in the TGF-β core pathways was significantly higher than expected for a randomly selected set of 37 genes (Figures S1C and S1D). A set of genes in the TGF-β superfamily had recurrent chromosomal deletions of at least one allele (Figure S1E). Heterozygous deletions generally occur with high frequency in tumor suppressor genes and may be accompanied by additional mutations in the remaining allele, leading to complete loss of tumor suppressor function (Haverty et al., 2009). All SMAD-encoding genes had heterozygous deletion frequencies greater than 20%, with several exceeding 30%. Tumors rarely had more than one mutationally altered gene within a category.

Distribution of Gene Alterations across Cancer Types

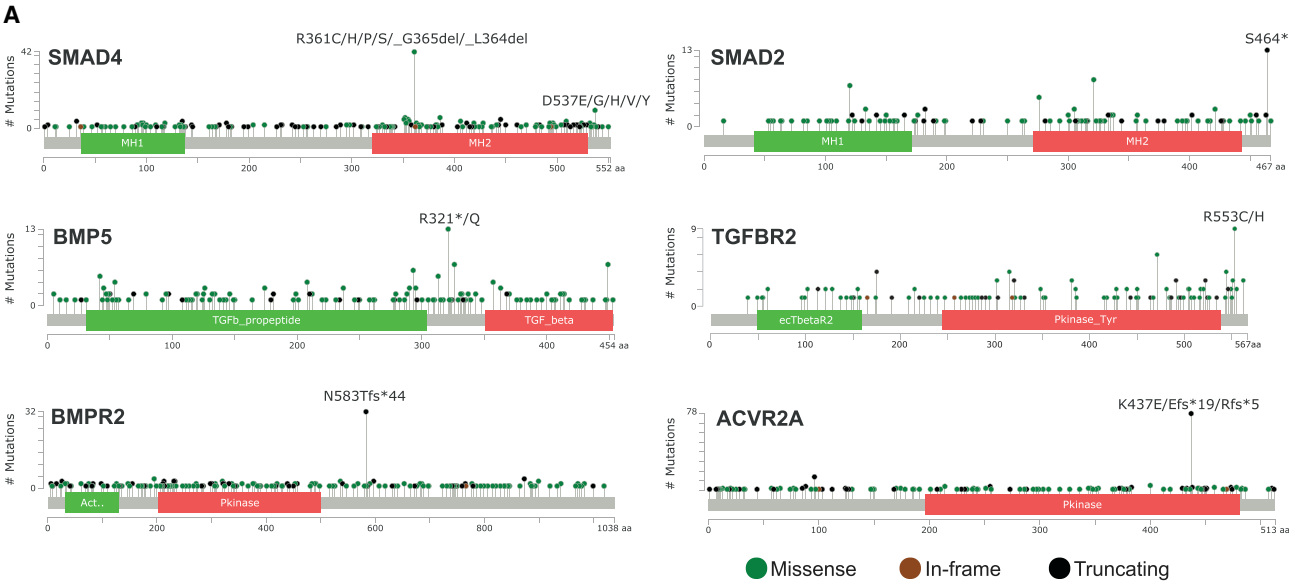
The frequency and type of genomic alteration varied widely across tumor types (Figures 2A and S2A), from no alterations in testicular germ cell tumors (TGCT) to all three types of alterations (mutation, deletion, and amplification) in urothelial bladder cancers (bladder urothelial carcinomas [BLCAs]). There were genomic alterations of TGF-β pathway genes in more than 50% of samples in 12 tumor types (Figure 2A and Tables S2, S3, and S4). Skin cutaneous melanoma (SKCM), colon adenocarcinoma (COAD), esophageal carcinoma (ESCA), stomach adenocarcinoma (STAD), and uterine corpus endometrial carcinoma (UCEC) had high background alteration burdens, including microsatellite instability (MSI) or chromosomal instability (CIN) (Cancer Genome Atlas Network, 2012; Cancer Genome Atlas Network, 2015; Cancer Genome Atlas Research Network et al., 2013). Without adjusting for background alteration burden, among the 39% of TCGA cases that carried TGF-β pathway gene alterations, SKCM (70%), COAD (65%), and ESCA (65%) had the highest percentages of alterations; thyroid carcinoma (THCA) (4%), kidney chromophobe (KICH) (6%), and TGCT (9%) had the lowest (Table S3).

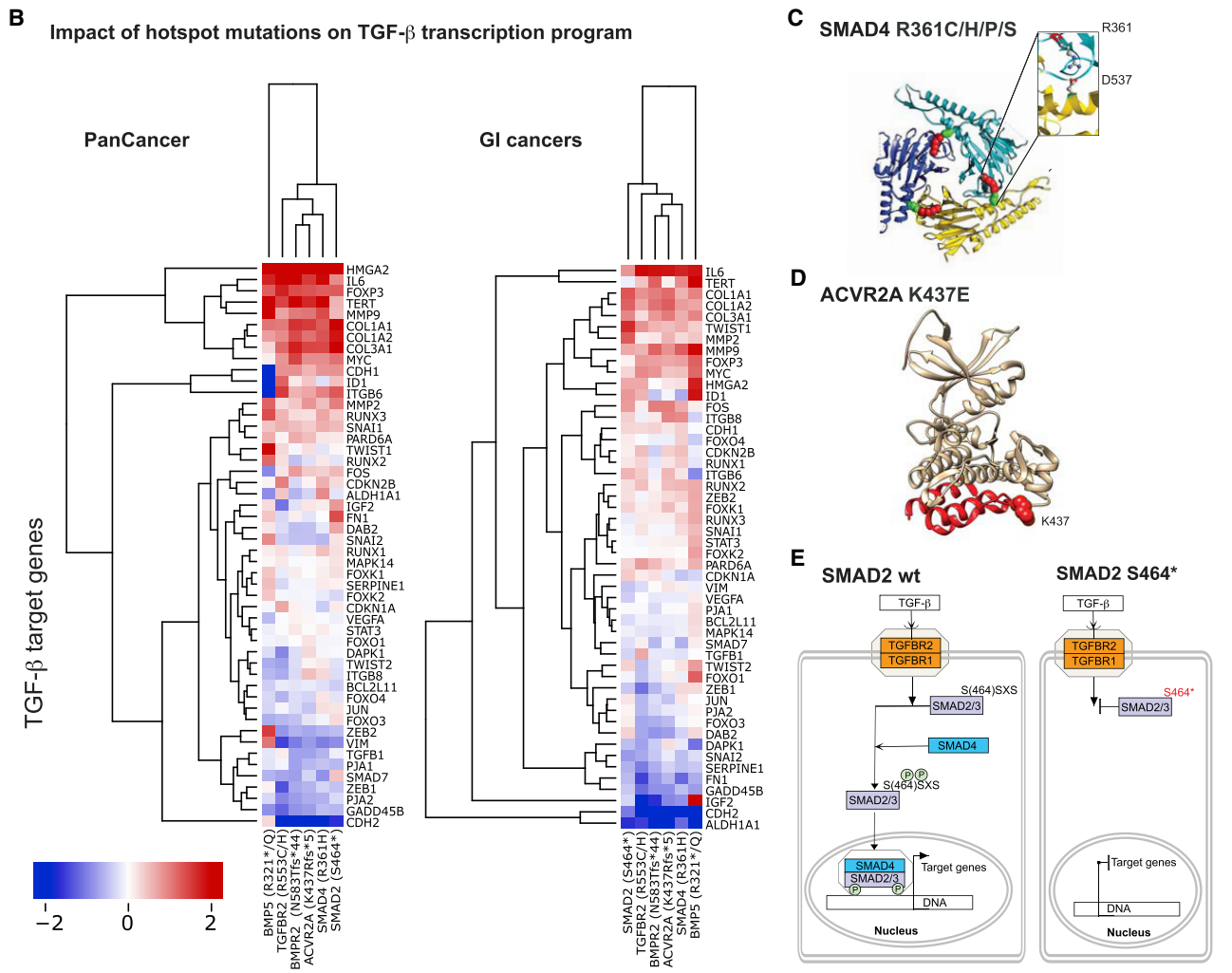
We observed non-silent *SMAD4* mutations in 24% and *SMAD4* deletions in 13% of pancreatic adenocarcinoma (PAAD) samples (Figures 2A and 2C and Table S4). Because *SMAD4* is the Co-Smad required for transducing the Smad signal to downstream effectors, loss of *SMAD4* in PAAD by mutation or deletion suggests a tumor-suppressive role for TGF-β signaling in PAAD, which is consistent with other reports (David et al., 2016).

Among all cancer types, high-grade ovarian cancers (OVs) (Figure 2B) had high amplification frequency, which could be due to genomic instability (Cancer Genome Atlas Research Network, 2011). Prostate adenocarcinoma (PRAD) had the highest deletion frequency, marked by losses in the *SMAD9* (encoding a Receptor-Smad [R-Smad]) and *ACVR2A* (encoding a receptor) (Figures 2B and 2C). Rectal adenocarcinoma (READ) had the greatest frequency of *BMP7* amplification. Diffuse large B-cell lymphoma (DLBC) had a high frequency of deletions spanning different levels of the pathway—ligands (*TGFB2*, *INHBB*, and *GDF1*), receptors or receptor-associated proteins (*BMPR1A*, *ACVR1*, *ACVR1C*, *ACVR2A*, *ACVR2B*, and *TGFBRAP1*), and Smads (*SMAD9*)—indicative of a tumor-suppressive role for TGF-β signaling in these early-stage DLBC cases in the TCGA cohort.

After adjusting for background alteration burden, we analyzed MutSigCV- and GISTIC-precomputed results across all individual cancer types and the Pan-Cancer cohort to identify significantly mutated genes (SMGs) and genes targeted by somatic CNVs (Figures 2D–2F). The analysis revealed *SMAD4*, *ACVR2A*, and *TGFBR2* as the most common SMGs within specific disease types and across the Pan-Cancer cohort. *SMAD4* had a highly overlapping profile with *TGFBR2*; both were SMGs in the GI cancers PAAD, ESCA, and STAD. Among individual disease types, COAD had the highest number of SMGs (*SMAD4*, *SMAD3*, *SMAD2*, and *ACVR2A*). The number of genes targeted by somatic CNVs, particularly deletions, was higher than the number of SMGs (Figures S1C, S2B, and S2C). A common type of CNV was recurrent heterozygous loss (Figure S1E). *SMAD4* was the only statistically significant deletion target in the Pan-Cancer cohort; it was most significantly deleted in GI cancers (PAAD, COAD, READ, STAD, and ESCA). PAAD had deletions

(D–F) Statistical significance of alterations in the TGF-β superfamily pathway genes. Genes that were significantly mutated or targets of copy-number alteration based on MutSigCV results (D) and GISTIC2 (E and F) analyses. Only the genes altered significantly in at least one cancer type are included.
(G–I) Transcriptional output associated with alterations in the TGF-β superfamily pathway genes. Differential mRNA expression of key genes downstream of the TGF-β superfamily pathways, including mutations (G), amplifications (H), and deep deletions (I).
See also Figure S2 and Tables S2, S3, and S4.





(legend on next page)

associated with 14 TGF-β core genes, suggesting synergistic effects from ligands (BMP family), receptors (BMPR, TGFBR), and *SMAD4*. Colorectal cancers (COAD and READ) were marked by *SMAD4* and *SMAD3* deletions. Deletions in genomic regions covering all ACVR genes except *ACVR2B* were identified as significant in DLBC.

Transcriptional Signatures of Genomic Alterations in the TGF-β Pathways

To understand how gene alterations affect transcriptional output of the pathways, we analyzed the mRNA expression of 50 downstream targets of Smad signaling with defined roles as tumor promoters or tumor suppressors (Table S1). Unsupervised hierarchical clustering analysis identified patterns of correlation between target gene expression and each class of genomic alteration (Figures 2G–2I). Point mutations were associated with two predominant patterns of target gene signatures: increased or decreased expression (Figure 2G). Surprisingly, the directionality of target-gene change was consistent for all mutations, even for mutations in the inhibitors *SMAD6/7*. An explanation is that mutations in pathway activators, such as *TGFB1/2/3* and *TGFBR1/2/3*, may result in gain of function, whereas mutations in the inhibitors *SMAD6* and *SMAD7* may result in loss of inhibitory function.

Another explanation is that *SMAD2* was generally co-amplified with *SMAD7* (Figure 1B); both genes are in the same cytogenetic band (18q21.1). Similarly, *SMAD3* was generally co-amplified with *SMAD6*; both are in proximal cytogenetic bands, 15q22.33 and 15q22.31, respectively. Thus, the net effect of those co-amplifications could be an overall increase in pathway activity. In support of that hypothesis, both the amplification and deletion profiles (rows in Figures 2H and 2I) of those gene pairs were similar, and consequently, SMAD2 and SMAD7 are co-clustered, whereas SMAD3 and SMAD6 are clustered close to each other.

The effect of TGF-β pathway amplification events on target gene mRNA expression was similar to that of mutations (Figure 2H), suggesting that most mutations in TGF-β pathway activators are gain-of-function. *HMGA2* (encoding a chromatin remodeling protein with oncogenic properties; Morishita et al., 2013; Thuault et al., 2006) was overexpressed in samples with either mutations or amplifications in the TGF-β pathway genes, with the exception of tumors with amplifications in *TGFB2*, *TGFBR2*, *ACVR2B*, *SMAD4*, *SMAD5*, or *SMAD6*. Those 6 genes may deliver context-specific signals for regulating *HMGA2* expression. Likewise, *CDH2* clustered separately from other genes, and its decreased expression was associated with most point mutations and CNVs. *CDH2* encodes a cadherin important in cell adhesion and migration (Principe et al., 2014; Xu et al., 2009). Another distinct cluster contained over-expressed metastasis-related genes, including collagens (*COL1A1/1A2/3A1*), a metalloprotease (*MMP9*), and a transcription factor (*FOXP3*).

SMAD5 amplification was associated with increased *CDH2* expression; 36 other amplifications were associated with decreased *CDH2* expression. Similarly, *HMGA2* expression was increased with most amplification events but decreased where *SMAD5* was amplified (Figure 2H). Another exception was reduced *HMGA2* expression in samples with amplifications of *SMAD4* or *TGFBR2*, whereas *HMGA2* expression increased in samples with mutations in *SMAD4* or *TGFBR2* (Figure 2G).

Hotspot Mutations in Genes Associated with TGF-β Superfamily Pathways

We focused on sites in the 43 genes that were mutated in at least 9 samples across the 33 tumor types (see Figure S3 for hotspot mutations identified within at least 5 samples). The analysis identified 6 genes with hotspot mutations, representing all levels of the TGF-β pathway (Figures 3A–3E). *BMP5* and *TGFBR2* included previously unreported hotspots.

Hotspot mutations of *BMP5* occurred in 13 cases across 7 cancers. BMP5 is synthesized as a proprotein, and an R321 stop-codon mutation (4 cases) (Figure 3A) results in loss of the functional secreted ligand. An R321 to Q (9 cases) mutation may impact cleavage of the protein to the mature secreted form. Frameshift mutations in *ACVR2A* at the K437 hotspot generate the variants K437Efs*19 (7 cases in 2 cancers) and K437Rfs*5 (69 cases in 5 cancers), resulting in premature stop codons and deletion of 2 C-terminal helices of the 4-helix bundle (Figures 3A and 3D), which likely disrupt ACV signaling (Rossi et al., 2005). Type I receptors ACVR1B and ACVR1C have similar C-terminal frameshift mutation hotspots at R485 (6 cases) and R441 (5 cases), respectively (Figure S3). TGFBR2 R553 to C or H mutations and BMPR2 N583 frameshift might disrupt interaction with other receptor subunits or binding proteins (Chan et al., 2007). Hotspots in SMAD4 at R361 and D537 (two conserved sites in R-Smads) (Shi et al., 1997) normally stabilize homo- or heterotrimer oligomerization (Figure 3C) (Fleming et al., 2013; Shi et al., 1997). Those mutations could have widespread effects because SMAD4 is a binding partner for all Smad-dependent transcriptional regulation. Mutation at either R361 or D537 in SMAD4 correlates with metastasis and decreased survival in colon cancer (Mehrvarz Sarshekeh et al., 2017). SMAD2 exhibited 13 truncating mutations at S464 (Figure 3A). S464 is part of the essential phosphorylation motif SSXS (Ser464-Ser465-X466-S467) of R-SMADs (Fleming et al., 2013) (Figure 3E). S464 is necessary for proper positioning of SMAD2 for phosphorylation at S465 and S467, both of which mediate interaction of SMAD2 with SMAD4 (Macias et al., 2015) and dissociation of SMAD2 from TGFBR1 and the adaptor SARA (encoded

Figure 3. Mutational Hotspots in the TGF-β Superfamily Pathways

(A) Recurrent hotspot sites. Hotspot mutations with >9 incidents are labeled. Mutations that result in substitutions are indicated by single-letter amino acid code separated by slashes. Del, deletions; fs, frameshifts; asterisks, stop codons.

(B) Transcriptional output of pathway hotspot mutations in GI and Pan-Cancer cohorts. Differential mRNA expression of 50 TGF-β pathway target genes quantified in relation to 6 hotspot mutations in the Pan-Cancer cohort (left) and GI cancers (right).

(C) SMAD4 R361C/H/P/S. R361 is located on the SMAD4 homotrimer interaction interface, as shown on the SMAD4 structure (PDB ID: 1DD1).

(D) ACVR2A K437E. K437 is marked on the structure of the ACVR2A C-terminal kinase domain (PDB ID: 4ASX).

(E) SMAD2. Position and putative effect of the C-terminal truncation mutation S464* are shown.

See also Figure S3.

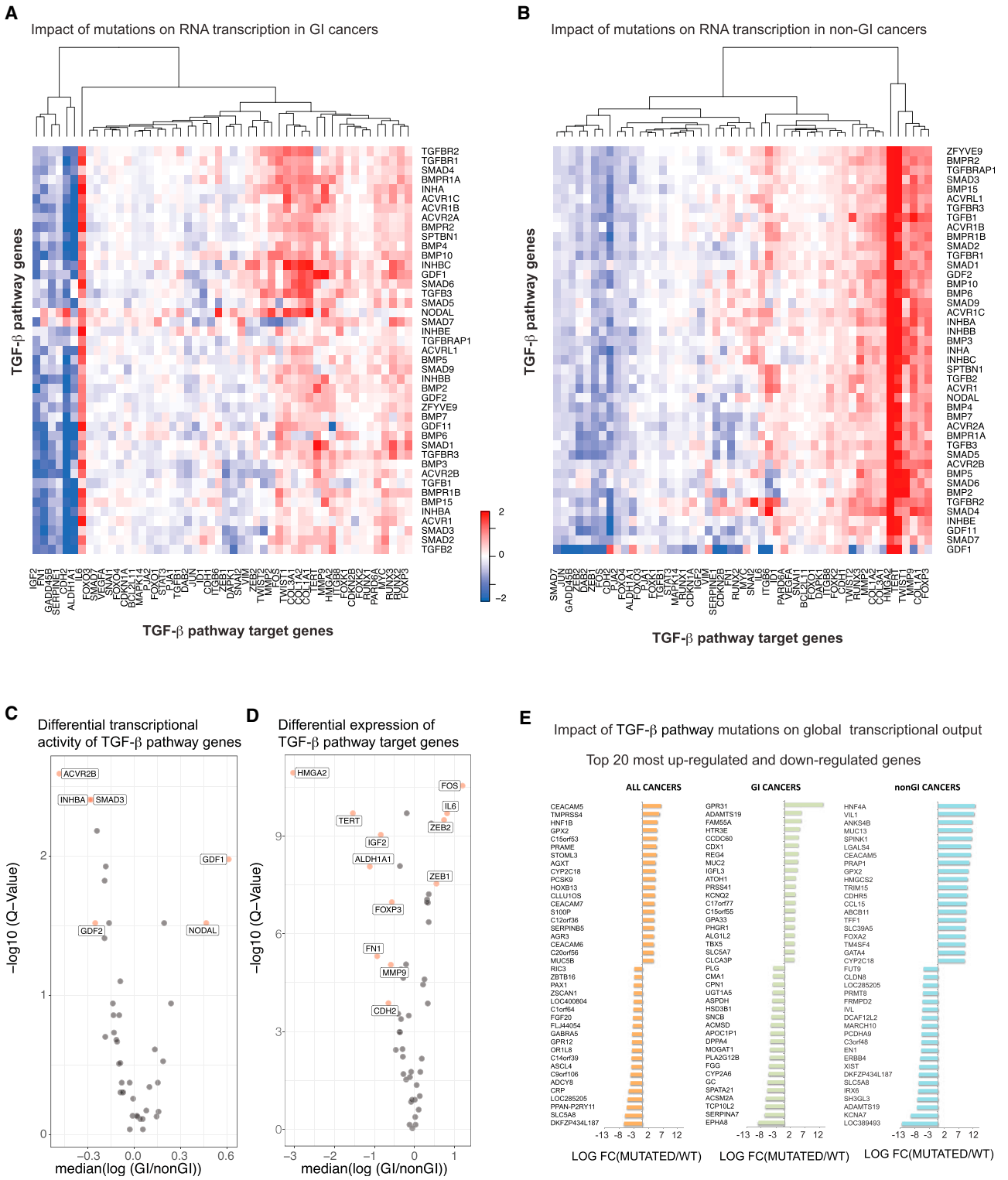


Figure 4. Comparison of TGF-β Superfamily Pathway Activity and Gene Aberrations

(A) The TGF-β superfamily pathway gene expression signature in GI cancers. Heatmap indicating the effects of non-silent somatic mutations in the 43 TGF-β pathway genes on expression of downstream target genes for 1,511 samples of 5 GI cancer types. Color reflects the log ratio of median expression in samples that carry the alteration versus samples that are wild-type (y axis).

(legend continued on next page)

by *ZFYVE9*). Hence, S464 mutations may prevent dissociation of SMAD2 from the receptor-adaptor complex, blocking the downstream signal (Figure 3E).

GI Cancers Are Enriched with TGF-β Pathway Hotspot Mutations

Of 176 mutations at hotspot sites across 6 genes, 115 (65%) were in cancers of the GI system (Figure S3): 60 in ESCA, 51 in COAD, 3 in PAAD, and 1 in liver hepatocellular carcinoma (LIHC). The connection to GI cancers is also supported by other studies (Park et al., 2010; Mehrvarz Sarshekeh et al., 2017). We found the reported *SMAD4* and *BMPR2* hotspots (Park et al., 2010; Mehrvarz Sarshekeh et al., 2017) and identified hotspots in *BMP5* and *TGFBR2*.

To determine if GI cancers possess a unique signature of altered TGF-β pathway activity, we compared changes in the expression of 50 downstream genes related to mutations at hotspot sites (Figure 3B). The expression signatures associated with the *BMP5* hotspot clustered separately from those associated with other hotspots. Notably, *CDH2* exhibited an overall reduction in expression, except in the context of the *BMP5* hotspot mutation. A cluster of genes (*HMGA2, TERT*, *MMP9*, *COL1A1/1A2/3A1*, *MYC*, *FOXP3*, and *IL6*) exhibited increased expression in the GI cancers containing at least one of the 6 hotspot mutations. Unique to the GI tumors was a cluster of genes that included strongly reduced expression of *CDH2, ALDH1A1*, and *IGF2* and a cluster with moderately reduced expression of *SERPINE1*.

When compared with the Pan-Cancer cohort, the GI subset showed an association of hotspot mutations with less expression of downstream genes (Figure 3B). That trend was generally characterized by blunted upregulation of the upregulated genes (*HMGA2*, collagen encoding genes, *FOXP3*, *MMP9*, and *MYC*) and greater downregulation of the downregulated genes (*ALDH1A1* and *CDH2*).

Transcriptional Signatures of TGF-β Pathway Alterations in GI Cancers

Guided by the enrichment of hotspot mutations in GI cancers, we tested for enrichment of TGF-β pathway point mutations in GI cancers. Non-silent mutations were significantly more common in GI cancers (596 of 1,511) than in the non-GI cancers (1,606 of 7,614). Deep deletions and amplifications were also significantly enriched in GI cancers. COAD, READ, and STAD had recurrent aberrations in genes at each level of the pathway (ligands, receptors, and SMADs) and all axes (*TGFBR*, *BMPR*, and *ACVR*), whereas PAAD had frequent mutations in only *SMAD4* and *TGFBR2* (Figure S4A).

To compare the TGF-β pathway transcriptional signatures in GI versus other cancers, we calculated the target gene expression signatures associated with TGF-β pathway mutations in both groups (Figures 4A and 4B). The upregulation of *TERT* and *HMGA2* was less substantial in GI cancers than in the Pan-Cancer cohort. Whereas *IL6* mRNA was increased in most non-GI cancers with TGF-β pathway mutations, *IL6* upregulation was significantly greater in GI cancers than non-GI cancers (Figure S4B), and within GI cancers, *IL6* expression was greater in samples with alterations in the TGF-β pathway genes than those without alterations in the TGF-β pathway genes. Notably, in non-GI cancers associated with *GDF1* mutations, *IL6* mRNA expression was markedly decreased, suggesting that *GDF1* may play different roles in GI and non-GI cancers. A similar analysis revealed a profound difference in *FOS* expression between GI and non-GI cancers (Figure S4C). In GI cancers, most TGF-β pathway gene mutations were associated with increased *FOS* expression; exceptions were *TGFBRAP1*, *SMAD7*, *SMAD5*, *GDF1*, *BMP5*, and *ACVRL1*. In non-GI cancers, only mutations in *TGFBR2* were associated with increased FOS expression; all other TGF-β pathway gene mutations were associated with decreased *FOS* expression.

To compare the transcriptional output resulting from mutations in GI and non-GI cancers, we calculated differences in expression of the 50 target genes associated with mutations in the 43 genes (Figure 4C). The analysis revealed a shift toward repression of transcriptional output in GI cancers, with the most significant shifts occurring with mutations in *ACVR2B*, *INHBA*, *SMAD3*, or *GDF2*. In GI cancers, mutations in *GDF1* were associated with significantly increased target gene transcription. We also analyzed downregulation in each target gene (Figure 4D). Mutations in any of the 43 genes were associated with reduced mRNA expression in GI cancers compared with non-GI cancers for most target genes with the largest reductions found for *HMGA2* and *TERT*. Compared to non-GI cancers, GI cancers had fewer genes with increased expression resulting from pathway mutations. In GI cancers, mutations in any of the 43 genes were associated with a significantly increased expression of *FOS*, *IL6*, *ZEB2*, and *ZEB1* compared to expression changes of the same genes resulting from pathway mutations in non-GI cancers.

Finally, we probed for associations between transcriptional output and TGF-β pathway gene alterations for all cancers and the GI and non-GI subsets (Figure 4E). The top 20 and bottom 20 genes that were up- or downregulated in each case differed. However, all 3 cases included genes associated with metastasis, cell adhesion, and EMT. Members of the CEACAM family, which consists of proteins involved in pathogen sensing, innate immunity, and metastasis (Chen et al., 2016a; Vitenshtein

(B) The TGF-β superfamily pathway gene expression signature in non-GI cancers. Same analysis as (A) for 7,614 samples of 27 non-GI cancer types.
(C) Comparison of disrupted TGF-β superfamily pathway activity in GI and other cancers. Volcano plots for 43 TGF-β pathway genes in GI versus other cancers. Fold changes (x axis) were calculated from the median log ratio of mRNA expression across 50 downstream target genes (normalized to median levels in samples wild-type for the 43 TGF-β pathway genes) associated with mutations in GI versus other cancers. Red Q-values (y axis) identify genes with statistically significant changes in GI versus other cancers. Q-values were calculated by Wilcoxon signed-rank test for each pathway gene, followed by Benjamini-Hochberg (BH) FDR adjustment.
(D) Differential expression of TGF-β superfamily pathway target genes in GI and other cancers. The same as (C) but for TGF-β pathway target genes.
(E) Comparison of global transcriptional output. The ratio of TGF-β target gene expression in samples with and without gene alterations. Genes listed include the highest absolute mRNA expression changes (top 20 increases and top 20 decreases) in the presence of alterations of the 43 TGF-β superfamily genes.
See also Figure S4.

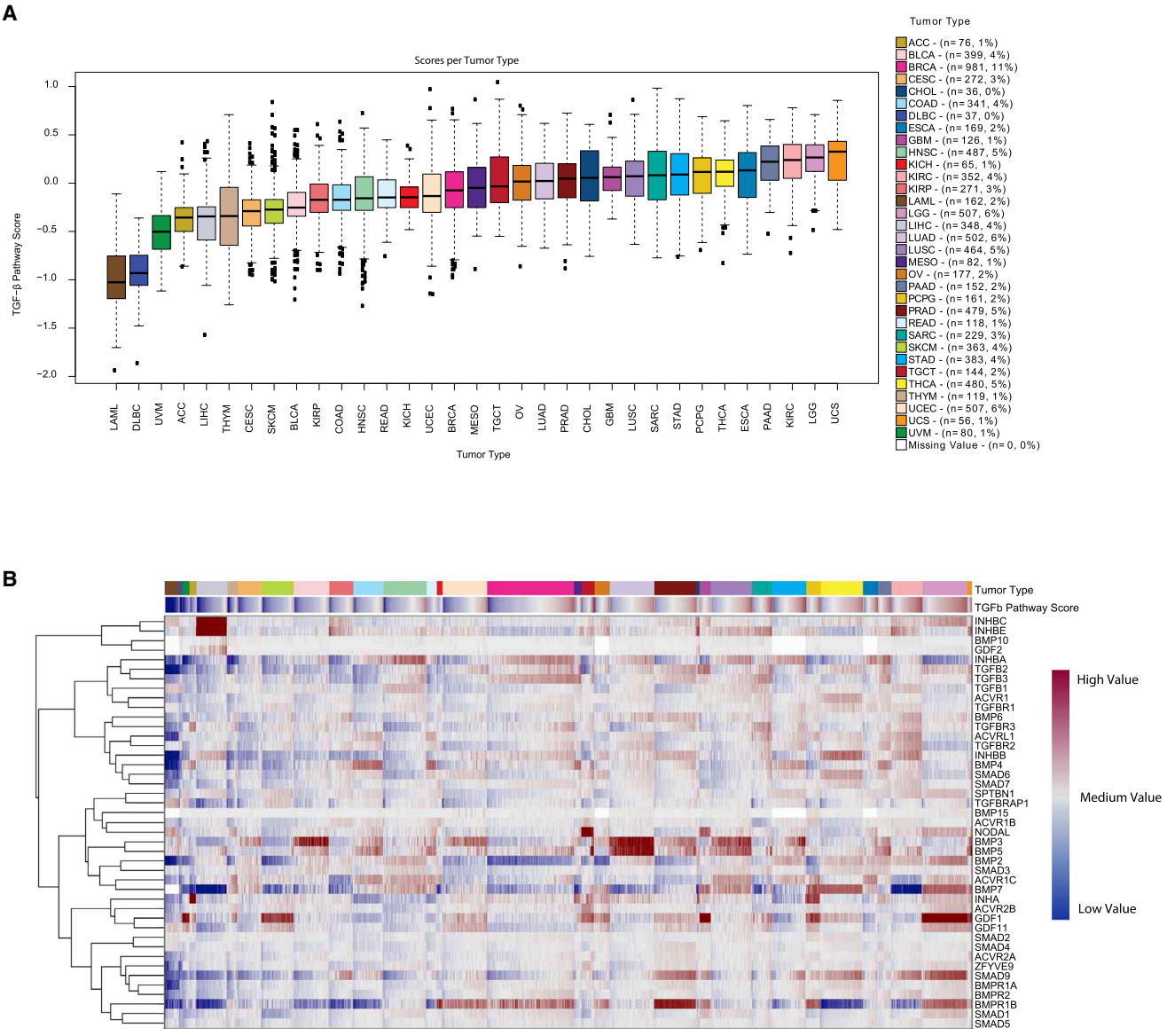


Figure 5. mRNA Analysis of TGF-β Superfamily Pathway Genes

(A) TGF-β superfamily pathway activity across Pan-Cancer tumor types. Boxplot showing the distribution of sample-specific pathway scores across each cancer type. Scores were computed using mRNA transcript levels of genes in the superfamily. The median and interquartile range (boxes) and outliers (whiskers) are shown.

(B) Supervised clustering of mRNA expression. mRNA expression values for the 43 genes, clustered from left to right by tumor type, then by TGF-β superfamily pathway score.

See also STAR Methods.

et al., 2016), were consistently upregulated. *TMPRSS4* and *ADAMTS19*, encoding cell surface proteases, were upregulated in the Pan-Cancer and GI cohorts, respectively. Genes that encode immune-related proteins were also upregulated: *PRAME* in the Pan-Cancer cohort and *GPR31* in GI tumors.

Gene Expression Levels Quantify TGF-β Signaling Pathway Activity

To explore TGF-β signaling pathway variation across the 33 cancers in the Pan-Cancer cohort, we computed a "pathway activity score" based on mRNA expression of the 43 genes. We verified that none of the genes were universally inhibitory in every cancer context. We validated the pathway score by correlating it with the median expression of the 50 TGF-β target genes and, separately, with the median expression of 50 random genes (Figure S5) (see STAR Methods).

Patterns emerged when we grouped activity scores by tumor type (Figure 5A). The two hematologic TCGA cancers, DLBC and acute myeloid leukemia (LAML), had the lowest median pathway activity scores. Uterine carcinosarcoma (UCS) had the highest median pathway activity score (Figure 5A). Five cancers—lung squamous cell carcinoma (LUSC), cervical squamous cell

carcinoma and endocervical adenocarcinoma (CESC), mesothelioma (MESO), TGCT, and kidney renal clear cell carcinoma (KIRC)—had significant differences in overall survival between patients with high and low pathway activity (Figure S6).

Supervised clustering of the 43 genes revealed that *INHBC* and *INHBE* were highly expressed in LIHC, whereas *BMP3* and *BMP5* were highly expressed in lung adenocarcinoma (LUAD) (Figure 5B). *GDF1* expression was high in brain cancers (glioblastoma multiforme [GBM] and brain lower grade glioma [LGG]), rare cancers (UCS and pheochromocytoma and paraganglioma [PCPG]), and in SKCM. *NODAL* expression was high in TGCT. The heatmap indicates the wide range of expression for the 43 genes in different tumor contexts and reveals potential targets for further study.

Unsupervised clustering of the 43 genes produced 11 clusters (Figure S7 and Table S5) that were dominated by cancer type. Cluster C3 was enriched with LAML, LUSC, CESC, squamous ESCA, head and neck squamous cell carcinoma (HNSC), and squamous BLCA. Cluster C3 was characterized by high expression of *BMP3, BMP7, SMAD3*, and *ACVR1C*, coupled with low expression of *BMPR1B*, suggesting that *BMPR1B* signaling may be tumor suppressive, whereas signals involving *BMP3, BMP7, SMAD3*, and *ACVR1C* may be tumor promoting in cancers enriched in that cluster. Cluster C4 was enriched with GI cancers ESCA, STAD, COAD, and READ. Cluster C4 was characterized by high expression of *ACVR1C, BMP4, BMP5*, and *INHBA*, coupled with low expression of *INHA, BMPR1B, GDF1, INHBB, TGFB2*, and *TGFB3*. Those observations suggest tumor-promoting roles for the highly expressed set of genes and tumor-suppressive roles for the set with low expression in cancer types enriched in that cluster.

Cluster C7, which contained most of the breast cancer samples, included two subclusters that did not correspond to clinical breast cancer subtypes (luminal A, luminal B, HER2, basal, or normal-like). Instead, the subclusters separated mainly on the basis of low and high levels of *BMPR1B* expression. Thus, BMPR1B signaling may have a tumor-promoting role and could be a viable therapeutic target for at least a subset of breast cancers.

Figure 6A shows a clustered heatmap of pairwise Pearson's correlations between expression of the 43 TGF-β pathway genes and expression of the 50 downstream target genes. Surprisingly, expression of none of the 43 TGF-β pathway genes was strongly negatively correlated with the activity score, including expression of the pathway inhibitors *SMAD6*/7. We attribute this observation to co-occurring amplifications or deletions of *SMAD7* and *SMAD2* and co-occurring amplifications of *SMAD6* and *SMAD3* (Figure 1B). Expression of ligand-encoding *INHBE* had the strongest negative correlation with pathway activity. Within the downstream targets, expression of *TERT* and *FOXK2* had the strongest negative correlations with activity score, suggesting that their suppression may contribute to the pathway's tumor-suppressor role. By contrast, expression of the EMT genes *ZEB1* and *ZEB2* positively correlated with pathway score, providing a possible mechanism for the tumor-promoting effects of the pathway.

TGF-β Pathway Activity Correlates with Activity of Other Cancer-Related Pathways

With proteomic data and a published method (Akbani et al., 2014), we computed activity scores for 10 other oncogenic pathways: apoptosis, breast reactive, cell cycle, hormone receptor, hormone signaling, PI3K/AKT, RAS/MAPK, RTK, TSC/mTOR, and DNA damage response (DDR). We assigned activity scores for EMT and leukocyte infiltration (an index of immune function) using mRNA and DNA methylation data, respectively (Cancer Genome Atlas Research Network, 2017). A clustered heatmap representation (Figure 6B) shows that the Pan-Cancer cohort exhibited a negative correlation between the TGF-β superfamily pathway score and the activity scores for the cell cycle pathway and apoptosis pathway. In contrast, positive correlations occurred for the EMT pathway, breast reactive pathway, RAS/MAPK, and the RTK pathway. Table S6 shows correlations within individual tumor types and the EMT and cell cycle pathways.

Downstream Target Genes *HMGA2*, *COL1A1*/*COL1A2*/*COL3A1*, and *MMP9* Are Associated with Patient Survival

We analyzed the combined impact of TGF-β target gene expression and the 43 core gene alterations on patient survival across the Pan-Cancer cohort. We compared the survival of patients with 3 different cancer profiles: those with high expression of *HMGA2* and alterations in any one of the 43 TGF-β pathway genes (Figure 6C; high *HMGA2*/TGF-β mutant), those with high *HMGA2* expression and no alterations in any of the 43 genes (Figure 6C; high *HMGA2*/TGF-β wild-type), and those with low expression of *HMGA2* without considering alterations in TGF-β pathway genes (Figure 6C; low HMGA2 expression). Patients with low *HMGA2* expression had the best outcome, followed by patients with high expression of *HMGA2* and no mutations in the 43 genes. A similar trend was observed for genes encoding *MMP9*, collagens, and to a lesser extent for *FOXP3*. *TERT* overexpression had no impact on survival. We saw the opposite for cancers with downregulated *CDH2*; the worst outcome was associated with low *CDH2* expression and mutations in 43 genes (Figure S6B). Thus, the expression profile of specific target genes and alterations in the TGF-β superfamily genes cooperated to increase tumor aggressiveness. The impact on survival was most significant for overexpression of collagen-encoding genes, *HMGA2*, and *MMP9* (Figures 6C–6E). Because of the association of collagen overexpression and alterations in TGF-β pathway genes with poor survival, we hypothesize that altered signaling through the TGF-β superfamily pathways remodels the extracellular matrix to drive metastasis in multiple cancer contexts.

We analyzed survival in GI and non-GI cancers (Figure S6D). In the GI cohort, only *ZEB2* combined with TGF-β pathway gene alteration yielded a significant difference, with low *ZEB2* expression corresponding to a survival benefit. In non-GI patients, high expression of the TGF-β pathway target genes *IL6*, *HMGA2*, *ZEB2*, and *FOS* was associated with reduced survival, particularly when combined with TGF-β pathway mutations. Thus, although TGF-β pathway mutations may not occur as commonly in non-GI cancers, they may be important contributors to mortality.

Epigenetics and miRNAs Modulate TGF-β Pathway Activity

To explore regulation of TGF-β pathway activity, we evaluated DNA methylation (Table S6) and miRNA expression (Table S7);

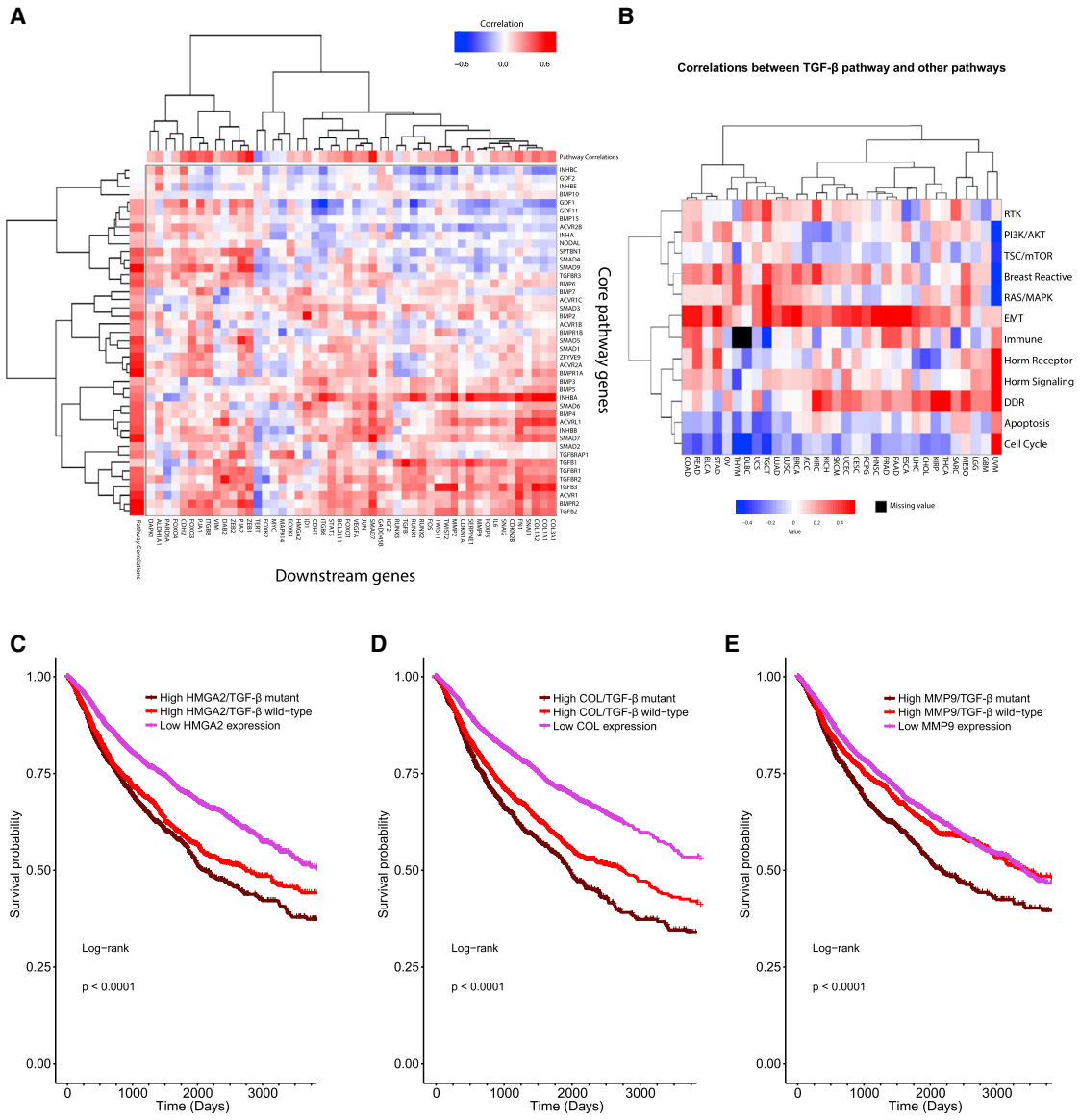


Figure 6. Correlation of TGF-β Superfamily Genes with Other Cancer-Related Pathways and Genes

(A) Clustered heatmap of pairwise correlations between TGF-β pathway gene expression and that of 50 downstream target genes. Unsupervised hierarchical clustering was conducted with 1-Pearson's correlation distance metric and Ward's linkage. The covariate bar on each axis shows median expression values.

(B) Clustered heatmap of correlations between TGF-β pathway activity score and 12 other cancer-associated pathways. Oncogenic pathway activity scores (y axis) were computed from protein data, except for EMT (mRNA) and immune scores (DNA methylation).

(C) Impact of TGF-β pathway-associated *HMGA2* mRNA expression on patient survival. 10-year survival of patients with TGF-β pathway mutations (TGF-β mutant) and high *HMGA2* expression (High HMGA2), no mutations in the TGF-β pathway genes (TGF-β wild-type) and high *HMGA2* expression, and low *HMGA2* expression (regardless of mutation status of 43 genes) was compared in a Kaplan-Meier analysis. Statistical significance was assessed by log-rank test (see STAR Methods and Figure S6 for selection of high and low expression level thresholds).

(D) Impact of collagen-encoding gene (*COL1A1*, *COL1A2*, *COL3A1*) mRNA expression on patient survival. The same analysis as in (C) was performed for aggregated mRNA expression of three collagen genes that showed increased expression in cancers with TGF-β pathway gene mutations.

(E) Impact of *MMP9* mRNA expression on patient survival. The same analysis as in (C) was performed for the impact of *MMP9* expression on patient survival by comparing high *MMP9*/TGF-β pathway mutations, high *MMP9*/wild-type TGF-β pathway, and low *MMP9*.

See also Figures S5 and S6 and Table S6.

both processes are associated with cancer (Dawson and Kouzarides, 2012; Jones and Baylin, 2007; Shen and Laird, 2013). Methylation levels across the 41 genes for each sample grouped by tumor type revealed a high variability (Figure 7A). Despite this variability, when ordered by TGF-β pathway activity, DLBCs with the lowest TGF-β pathway activity score had the highest median and range of DNA methylation scores, and LAML with low pathway activity had a low median DNA methylation score (Figure 7A). Hence, epigenetic silencing appeared to contribute to low pathway activity in DLBC but not LAML. UCSs with the

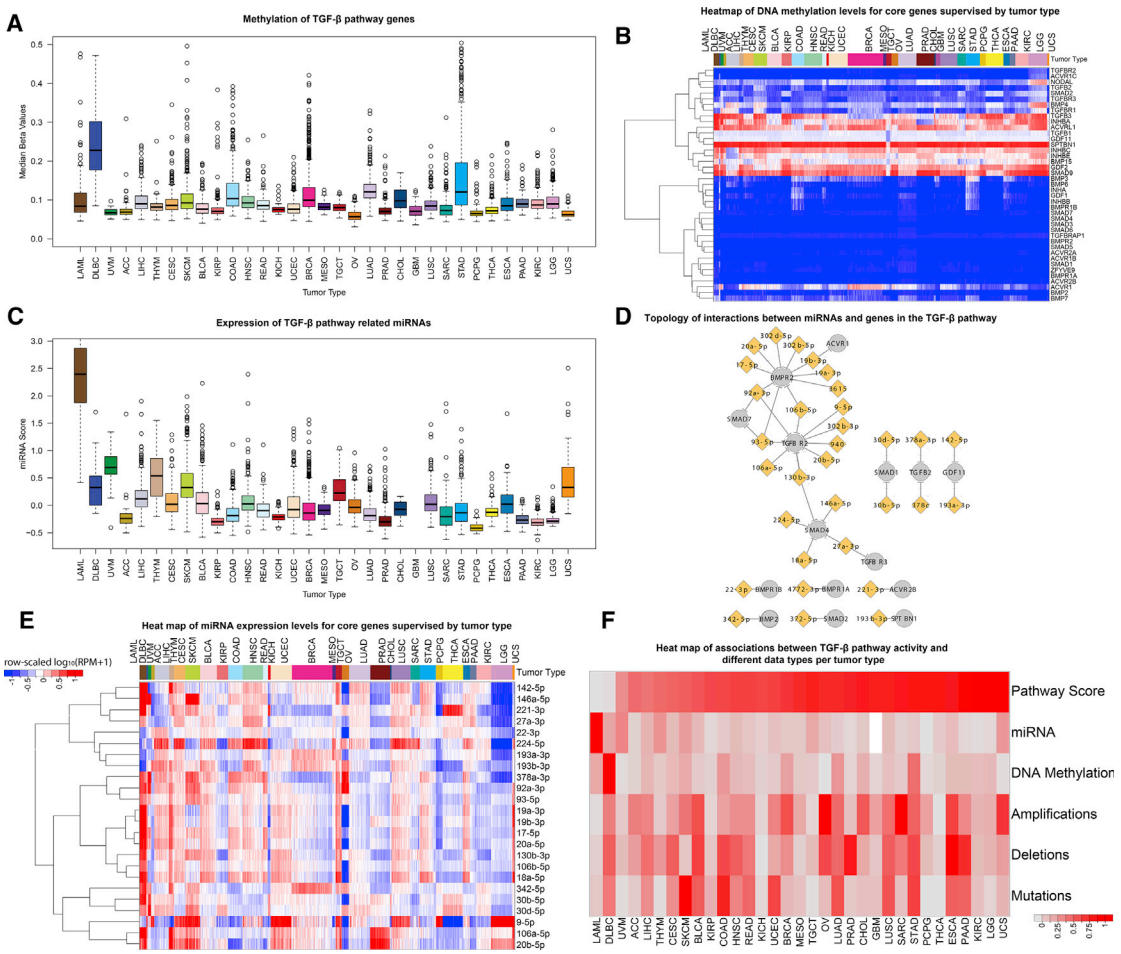


Figure 7. Epigenetic Control of the TGF-β Superfamily Pathways

(A) Methylation levels. Boxes quantify the degree of methylation across the 43 TGF-β genes in a given tumor type. The methylation score is calculated from the median for each gene in a given sample. Scores are grouped by tumor type. The median and interquartile range (boxes) and outliers (whiskers) are shown.

(B) Supervised cluster analysis of methylation patterns. Methylation patterns were clustered as in Figure 6A. Methylation levels were quantified as M-values by first mapping methylation array probes to individual genes. A median beta value for each gene was then calculated as the median beta value across all samples for a given cancer type.

(C) microRNA levels. Boxplot showing the mean miRNA expression levels for the 32 miRNAs that regulate the indicated genes in the TGF-β superfamily pathways. The median and interquartile range (boxes) and outliers (whiskers) are shown.

(D) microRNA regulation. Inferred miR-mRNA targeting for 15 TGF-β superfamily pathway genes by the 32 miRNAs.

(E) Abundance of miRNAs predicted to target the TGF-β superfamily pathway genes. The heatmap illustrates miRNA abundance for 8,930 tumor samples from 32 of the 33 TCGA tumors (GBM excluded, no miRNA data in TCGA).

(F) Contribution of data type to TGF-β superfamily pathways score. Tumor types (columns) ordered from lowest (left) to highest (right) TGF-β superfamily pathway score. Mean miRNA expression levels normalized between 0 and 1 yielded the highest overall correlation with pathway score (R = −0.68). Mean DNA methylation beta values normalized between 0 and 1 had the next highest correlation (R = −0.46). Amplifications (R = 0.24), deletions (R = 0.09), and mutations (R = -0.05) represent proportions of samples with the given type of aberration in at least one of the 43 TGF-β genes.

See also Figure S7 and Tables S5, S6, and S7.

highest TGF-β pathway activity score had a low median methylation score, suggesting that other mechanisms contribute to the differences in activity scores.

We clustered DNA methylation levels (supervised by cancer type) (Figure 7B) and compared the results with supervised clustering of the expression of the 43 TGF-β pathway genes (Figure 5B). The epigenetic cluster analysis divided the genes into two main groups: those with little or no DNA methylation in any cancer and those with DNA methylation in some or all cancers. The cluster with high DNA methylation scores included *SMAD9*, *SPTBN1*, *ACVRL1*, *GDF2*, *INHBC*, *INHBE*, *INHBA*, and *TGFB3*. The presence of ACV ligands suggested that those ligands are tumor suppressive in many cancers. Adaptor *SPTBN1* had a high DNA methylation score in all cancer samples, supporting a tumor-suppressive role.

We focused on miRNAs that, according to miRBase (Kozomara and Griffiths-Jones, 2014), are associated with the 43 TGF-β pathway genes. We selected the top 32 miRNAs anti-correlated with transcript abundance (Table S7). Those miRNAs exhibited variable expression across the 32 tumor types (Figure 7C; GBM had no miRNA data). LAMLs with low TGF-β pathway activity had the highest level of miRNA expression,

suggesting that miRNAs regulate pathway activity in this blood cancer.

We predicted that 15 of the 43 genes were targets of at least 1 miRNA; *BMPR2*, *TGFBR2*, and *SMAD4* were each targeted by 5 or more miRNAs (Figure 7D). An miRNA/mRNA topology map for the GI cancers (COAD, READ, STAD, ESCA, LIHC, and PAAD) (Figure S7B) revealed that *BMP3* was targeted only in GI cancers and that *SMAD4* was targeted only in the Pan-Cancer cohort, suggesting that miRNA/mRNA topologies depend on tumor context.

Cluster analysis (supervised by cancer type) yielded an interesting pattern for miRNA 92a-3p, which is predicted to target the 3 core genes *BMPR2, TGFBR2*, and *SMAD7*. miRNA 92a-3p was overexpressed in breast, ovarian, liver, and head and neck cancers. We also identified *BMPR2* and *TGFBR2* as genes with hotspot sites of mutations that were common in STAD and COAD. The cancers with high frequencies of hotspot mutations in those two genes did not have high expression of miRNA 92a-3p, suggesting that there is little selective pressure for both mutation and downregulation by that miRNA. To examine the contribution of mutations, amplifications, deletions, DNA methylation, and miRNAs to the pathway activity score across tumor types, we computed Pearson's correlations between the pathway activity score and (1) levels of DNA methylation or miRNA expression and (2) percentages of mutations or CNVs in each tumor type and plotted the results in order of increasing pathway activity scores (Figure 7F). The results suggested that miRNAs play a dominant role in LAML, DLBC, uveal melanoma (UVM), and THYM, all of which had low TGF-β pathway activity scores. DNA methylation was dominant in DLBC, STAD, breast invasive carcinoma (BRCA), and COAD. Amplifications positively correlated with the activity score and played a dominant role in UCS, sarcoma (SARC), ESCA, cholangiocarcinoma (CHOL), and ovarian serous cystadenocarcinoma (OV). However, OV has a high background CNV burden, making it difficult to distinguish functionally important effects from passenger alterations. Overall, deletions exhibited a low positive correlation with the pathway activity score, and mutations showed the weakest correlation.

DISCUSSION

Because TGF-β superfamily signaling plays context-dependent roles as both a tumor suppressor and tumor promoter, TGF-β biological function is notably ambiguous. However, given its prominent role in cancer, understanding its function in diverse settings will be necessary to design therapy for tumors with aberrant TGF-β signaling. Hence, this study focused on elucidating salient characteristics of TGF-β-associated genes across a large cohort of different types of cancers. Some of the key findings of the study were that (1) 39% of the cancers carried TGF-β pathway gene alterations; (2) the genomic alterations appeared to affect expression of metastatic and EMT genes; (3) 6 hotspot mutations were identified in 6 genes; (4) the pathway was most frequently aberrant in GI cancers, which exhibited 115 of the 176 hotspot mutations identified; (5) high expression of downstream target genes coupled with mutations in the TGF-β pathway genes was associated with poor outcome, suggesting a net tumor-promoting role of the superfamily across the Pan-Cancer cohort; (6) apparent gene silencing by DNA methylation and deletion of TGF-β pathway genes were observed most frequently in DLBC, whereas miRNA silencing was seen most often in LAML. DLBC and LAML also had the lowest TGF-β pathway activity scores, suggesting a possible tumor-suppressive role of the TGF-β superfamily in hematologic cancers.

Although 39% of the cancers had genomic alterations in at least one of the TGF-β pathway genes, GI cancers were particularly enriched for them. GI cancers were most influenced by recurrent hotspot mutations in 6 genes, *SMAD4, SMAD2, BMPR2, BMP5, TGFBR2*, and *ACVR2A*. The hotspot mutations in *BMP5* and *TGFBR2* had not been identified previously, and their function in GI cancer should be explored.

UCS showed the highest TGF-β superfamily pathway activity. High activity was associated with amplifications or low DNA methylation. In general, epigenetics appeared to play a strong role in regulating the activity of the TGF-β superfamily pathways in DLBC, COAD, BRCA, STAD, and LUAD, whereas miRNAs played a strong role in LAML, UVM, and THYM. Such cancer type-dependent differences in the regulation of the TGF-β pathway could prove important to the development of therapies that target the pathway.

TGF-β signaling pathway activity correlated positively with other cancer-relevant pathways, including EMT, breast reactive, RAS/MAPK, and RTK pathways. Conversely, activity of the TGF-β pathways was anti-correlated with the cell cycle and apoptosis pathways. Overall, this study provides a molecular portrait of genetics, epigenetics, and miRNA-mediated regulation of signaling mediated by the TGF-β superfamily. We expect that this body of organized data and information will be mined by other researchers over time to formulate, test, or validate a variety of additional hypotheses that have not yet come into focus.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
  - Subject Details
- METHOD DETAILS
  - Sample Processing
  - Selection of 43 Core Genes Associated with the TGF-β Superfamily
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Mitigation of Batch Effects and Systematic Biases
  - DNA Sequencing Data
  - Mutation Analysis
  - Copy Number Analysis
  - Pathway Analysis
  - Expression Signatures of Genomic Alterations
  - Gastrointestinal Cancers
  - mRNA Expression Analysis and Pathway Activity Scores
  - microRNA Analysis
  - DNA Methylation Profiles

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and seven tables and can be found with this article online at https://doi.org/10.1016/j.cels.2018.08.010.

ACKNOWLEDGMENTS

This work was supported by the following grants: NCI: U54 HG003273, U54 HG003067, U54 HG003079, U24 CA143799, U24 CA143835, U24 CA143840, U24 CA143843, U24 CA143845, U24 CA143848, U24 CA143858, U24 CA143866, U24 CA143867, U24 CA143882, U24 CA143883, U24 CA144025, U24 CA210949, U24 CA210950, P30 CA016672, P30 CA016672, and P01 CA130821; DoD/CDMRP: W81XWH-16-1-0237, NIAA: R01AA023146, and VA: I01BX003732.

AUTHOR CONTRIBUTIONS

Methodology, R.A., A.K., L.M., N.R.G., and A.C.B.; Genomic Analysis, A.K., X.L., A.C.B., A.D.C., R.S.K., and R.A.; mRNA Analysis, A.K., Y.L., R.A., A.S., X.L., and S. Ling; miRNA Analysis, G.R., A.S., S. Ling, and R.A.; Protein Analysis, A.S., W.M., J.Z., P.T., Z.J., S. Ling, and R.A.; DNA Methylation Analysis, G.M., A.R., V.R., A.S., and R.A.; Integrative Analysis, L.M., C.D., L.R., S. Li, and R.A.; Clinical Analysis, L.M., C.D., B.-N.N, L.R., L.N.K., and S.A.M.; Data Interpretation, R.A., L.M., A.K., S.Z., S.R., S.G., K.O., N.R.G., and J.N.W.; Data Curation, R.A., L.M., A.K., J.C., S.Z., S.R., R.S.K., and S. Ling; Writing, R.A., L.M., A.K., S.Z., N.R.G., L.R., C.D., W.J., J.N.W., J.A.A., V.R., A.R., A.D.C., G.R., and P.L.L.; Visualization, R.A., L.M., A.K., S.Z., N.R.G., A.C.B., A.S., and X.L.; Technical Discussion and Input, R.A., L.M., A.K., S.Z., N.R.G., A.C.B, G.R., L.N.K., M.D., J.R., S.S., Y.L., D.H., G.d.V., A.P., J.B.A., C.J.O., S.Li, H.U.O., S.A.M., A.H., M.W., J.C., A.D.C., and J.N.W.; Overall Concept and Coordination, R.A. and L.M. The Cancer Genome Atlas Research Network contributed collectively to this work.

DECLARATION OF INTERESTS

Michael Seiler, Peter G. Smith, Ping Zhu, Silvia Buonamici, and Lihua Yu are employees of H3 Biomedicine. Parts of this work are the subject of a patent application: WO2017040526 titled "Splice variants associated with neomorphic sf3b1 mutants." Shouyoung Peng, Anant A. Agrawal, James Palacino, and Teng Teng are employees of H3 Biomedicine. Andrew D. Cherniack, Ashton C. Berger, and Galen F. Gao receive research support from Bayer Pharmaceuticals. Gordon B. Mills serves on the External Scientific Review Board of AstraZeneca. Anil Sood is on the Scientific Advisory Board for Kiyatec and is a shareholder in BioPath. Jonathan S. Serody receives funding from Merck. Kyle R. Covington is an employee of Castle Biosciences. Preethi H. Gunaratne is founder, CSO, and shareholder of NextmiRNA Therapeutics. Christina Yau is a part-time employee/consultant at NantOmics. Franz X. Schaub is an employee and shareholder of SEngine Precision Medicine. Carla Grandori is an employee, founder, and shareholder of SEngine Precision Medicine. Robert N. Eisenman is a member of the Scientific Advisory Boards and shareholder of Shenogen Pharma and Kronos Bio. Daniel J. Weisenberger is a consultant for Zymo Research Corporation. Joshua M. Stuart is the founder of Five3 Genomics and shareholder of NantOmics. Marc T. Goodman receives research support from Merck. Andrew J. Gentles is a consultant for Cibermed. Charles M. Perou is an equity stock holder, consultant, and Board of Directors member of BioClassifier and GeneCentric Diagnostics and is also listed as an inventor on patent applications on the Breast PAM50 and Lung Cancer Subtyping assays. Matthew Meyerson receives research support from Bayer Pharmaceuticals; is an equity holder in, consultant for, and Scientific Advisory Board chair for OrigiMed; and is an inventor of a patent for EGFR mutation diagnosis in lung cancer, licensed to LabCorp. Eduard Porta-Pardo is an inventor of a patent for domainXplorer. Han Liang is a shareholder and scientific advisor of Precision Scientific and Eagle Nebula. Da Yang is an inventor on a pending patent application describing the use of antisense oligonucleotides against specific lncRNA sequence as diagnostic and therapeutic tools. Yonghong Xiao was an employee and shareholder of TESARO. Bin Feng is an employee and shareholder of TESARO. Carter Van Waes received research funding for the study of IAP inhibitor ASTX660 through a Cooperative Agreement between NIDCD, NIH, and Astex Pharmaceuticals. Raunaq Malhotra is an employee and shareholder of Seven Bridges. Peter W. Laird serves on the Scientific Advisory Board for AnchorDx. Joel Tepper is a consultant at EMD Serono. Kenneth Wang serves on the Advisory Board for Boston Scientific, Microtech, and Olympus. Andrea Califano is a founder, shareholder, and advisory board member of DarwinHealth and a shareholder and advisory board member of Tempus. Toni K. Choueiri serves as needed on advisory boards for Bristol-Myers Squibb, Merck, and Roche. Lawrence Kwong receives research support from Array BioPharma. Sharon E. Plon is a member of the Scientific Advisory Board for Baylor Genetics Laboratory. Beth Y. Karlan serves on the Advisory Board of Invite.

Received: March 1, 2018
Revised: May 29, 2018
Accepted: August 21, 2018
Published: September 26, 2018

REFERENCES

Akbani, R., Ng, P.K., Werner, H.M., Shahmoradgoli, M., Zhang, F., Ju, Z., Liu, W., Yang, J.Y., Yoshihara, K., Li, J., et al. (2014). A pan-cancer proteomic perspective on the Cancer Genome Atlas. Nat. Commun. *5*, 3887.

Akhurst, R.J. (2017). Targeting TGF-beta signaling for therapeutic gain. Cold Spring Harb. Perspect. Biol. *9*, https://doi.org/10.1101/cshperspect.a022301.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Cancer Genome Atlas Network. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330–337.

Cancer Genome Atlas Network. (2015). Genomic classification of cutaneous melanoma. Cell *161*, 1681–1696.

Cancer Genome Atlas Research Network. (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609–615.

Cancer Genome Atlas Research Network, Kandoth, C., Schultz, N., Cherniack, A.D., Akbani, R., Liu, Y., Shen, H., Robertson, A.G., Pashtan, I., Shen, R., et al. (2013). Integrated genomic characterization of endometrial carcinoma. Nature *497*, 67–73.

Cancer Genome Atlas Research Network. (2017). Comprehensive and integrative genomic characterization of hepatocellular carcinoma. Cell *169*, 1327–1341.e23.

Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A., et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. *30*, 413–421.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404.

Chan, M.C., Nguyen, P.H., Davis, B.N., Ohoka, N., Hayashi, H., Du, K., Lagna, G., and Hata, A. (2007). A novel regulatory mechanism of the bone morphogenetic protein (BMP) signaling pathway involving the carboxyl-terminal tail domain of BMP type II receptor. Mol. Cell Biol. *27*, 5776–5789.

Chen, J., Zaidi, S., Rao, S., Chen, J.S., Phan, L., Farci, P., Su, X., Shetty, K., White, J., Zamboni, F., et al. (2018). Analysis of genomes and transcriptomes of hepatocellular carcinomas identifies mutations and gene expression changes in the transforming growth factor-beta pathway. Gastroenterology *154*, 195–210.

Chen, J., Raju, G.S., Jogunoori, W., Menon, V., Majumdar, A., Chen, J.S., Gi, Y.J., Jeong, Y.S., Phan, L., Belkin, M., et al. (2016a). Mutational profiles reveal

an aberrant TGF-beta-CEA regulated pathway in colon adenomas. PLoS One *11*, e0153933.

Chen, J., Yao, Z.X., Chen, J.S., Gi, Y.J., Muñoz, N.M., Kundra, S., Herlong, H.F., Jeong, Y.S., Goltsov, A., Ohshiro, K., et al. (2016b). TGF- β/ β2-spectrin/CTCF-regulated tumor suppression in human stem cell disorder Beckwith-Wiedemann syndrome. J. Clin. Invest. *126*, 527–542.

Chou, C.H., Chang, N.W., Shrestha, S., Hsu, S.D., Lin, Y.L., Lee, W.H., Yang, C.D., Hong, H.C., Wei, T.Y., Tu, S.J., et al. (2016). miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. Nucleic Acids Res. *44* (D1), D239–D247.

Christian, J.L., and Heldin, C.H. (2017). The TGFbeta superfamily in Lisbon: navigating through development and disease. Development *144*, 4476–4480.

Chu, J., Sadeghi, S., Raymond, A., Jackman, S.D., Nip, K.M., Mar, R., Mohamadi, H., Butterfield, Y.S., Robertson, A.G., and Birol, I. (2014). BioBloom tools: fast, accurate and memory-efficient host species sequence screening using bloom filters. Bioinformatics *30*, 3402–3404.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. *31*, 213–219.

Cibulskis, K., McKenna, A., Fennell, T., Banks, E., DePristo, M., and Getz, G. (2011). ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics *27*, 2601–2602.

Colak, S., and Ten Dijke, P.T. (2017). Targeting TGF-β signaling in cancer. Trends Cancer *3*, 56–71.

David, C.J., Huang, Y.H., Chen, M., Su, J., Zou, Y., Bardeesy, N., Iacobuzio-Donahue, C.A., and Massagué, J. (2016). TGF-beta tumor suppression through a lethal EMT. Cell *164*, 1015–1030.

Dawson, M.A., and Kouzarides, T. (2012). Cancer epigenetics: from mechanism to therapy. Cell *150*, 12–27.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Ellrott, K., Bailey, M.H., Saksena, G., Covington, K.R., Kandoth, C., Stewart, C., Hess, J., Ma, S., Chiotti, K.E., McLellan, M., et al. (2018). Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. Cell Syst. *6*, 271–281.

Fleming, N.I., Jorissen, R.N., Mouradov, D., Christie, M., Sakthianandeswaren, A., Palmieri, M., Day, F., Li, S., Tsui, C., Lipton, L., et al. (2013). SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer. Cancer Res. *73*, 725–735.

Gehring, J.S., Fischer, B., Lawrence, M., and Huber, W. (2015). SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics *31*, 3673–3675.

Gonzalez-Angulo, A.M., Hennessy, B.T., Meric-Bernstam, F., Sahin, A., Liu, W., Ju, Z., Carey, M.S., Myhre, S., Speers, C., Deng, L., et al. (2011). Functional proteomics can define prognosis and predict pathologic complete response in patients with breast cancer. Clin. Proteomics *8*, 11.

Haverty, P.M., Hon, L.S., Kaminker, J.S., Chant, J., and Zhang, Z. (2009). High-resolution analysis of copy number alterations and associated expression changes in ovarian tumors. BMC Med. Genomics *2*, 21.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using Empirical Bayes methods. Biostatistics *8*, 118–127.

Jones, P.A., and Baylin, S.B. (2007). The epigenomics of cancer. Cell *128*, 683–692.

Katz, L.H., Likhter, M., Jogunoori, W., Belkin, M., Ohshiro, K., and Mishra, L. (2016). TGF-beta signaling in liver and gastrointestinal cancers. Cancer Lett. *379*, 166–172.

Korn, J.M., Kuruvilla, F.G., McCarroll, S.A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P.J., Darvishi, K., et al. (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat. Genet. *40*, 1253–1260.

Kostic, A.D., Ojesina, A.I., Pedamallu, C.S., Jung, J., Verhaak, R.G., Getz, G., and Meyerson, M. (2011). PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat. Biotechnol. *29*, 393–396.

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. *42*, D68–D73.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, C., and Hung Wong, W. (2001). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol. *2*, RESEARCH0032.

Macias, M.J., Martin-Malpartida, P., and Massagué, J. (2015). Structural determinants of Smad function in TGF-beta signaling. Trends Biochem. Sci. *40*, 296–308.

McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., de Bakker, P.I., Maller, J.B., Kirby, A., et al. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat. Genet. *40*, 1166–1174.

McPherson, A., Hormozdiari, F., Zayed, A., Giuliany, R., Ha, G., Sun, M.G., Griffith, M., Heravi Moussavi, A., Senz, J., Melnyk, N., et al. (2011). deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput. Biol. *7*, e1001138.

Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. *12*, R41.

Mishra, L., Derynck, R., and Mishra, B. (2005). Transforming growth factor-beta signaling in stem cells and cancer. Science *310*, 68–71.

Morishita, A., Zaidi, M.R., Mitoro, A., Sankarasharma, D., Szabolcs, M., Okada, Y., D'Armiento, J., and Chada, K. (2013). HMGA2 is a driver of tumor metastasis. Cancer Res. *73*, 4289–4299.

Moustakas, A., and Heldin, C.H. (2016). Mechanisms of TGFbeta-induced epithelial-mesenchymal transition. J. Clin. Med. *5*, https://doi.org/10.3390/jcm5070063.

Muraoka-Cook, R.S., Kurokawa, H., Koh, Y., Forbes, J.T., Roebuck, L.R., Barcellos-Hoff, M.H., Moody, S.E., Chodosh, L.A., and Arteaga, C.L. (2004). Conditional overexpression of active transforming growth factor beta1 in vivo accelerates metastases of transgenic mammary tumors. Cancer Res. *64*, 9002–9011.

Olshen, A.B., Venkatraman, E.S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics *5*, 557–572.

Park, S.W., Hur, S.Y., Yoo, N.J., and Lee, S.H. (2010). Somatic frameshift mutations of bone morphogenic protein receptor 2 gene in gastric and colorectal cancers with microsatellite instability. APMIS *118*, 824–829.

Principe, D.R., Doll, J.A., Bauer, J., Jung, B., Munshi, H.G., Bartholin, L., Pasche, B., Lee, C., and Grippo, P.J. (2014). TGF-β: duality of function between tumor prevention and carcinogenesis. J. Natl. Cancer Inst. *106*, djt369.

Radenbaugh, A.J., Ma, S., Ewing, A., Stuart, J.M., Collisson, E.A., Zhu, J., and Haussler, D. (2014). RADIA: RNA and DNA integrated analysis for somatic mutation detection. PLoS One *9*, e111516.

Ramos, A.H., Lichtenstein, L., Gupta, M., Lawrence, M.S., Pugh, T.J., Saksena, G., Meyerson, M., and Getz, G. (2015). Oncotator: cancer variant annotation tool. Hum. Mutat. *36*, E2423–E2429.

Ratan, A., Olson, T.L., Loughran, T.P., Jr., and Miller, W. (2015). Identification of indels in next-generation sequencing data. BMC Bioinformatics *16*, 42.

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., and Mesirov, J.P. (2006). GenePattern 2.0. Nat. Genet. *38*, 500–501.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Mungall, K., Lee, S., Okada, H.M., Qian, J.Q., et al. (2010). De novo assembly and analysis of RNA-seq data. Nat. Methods *7*, 909–912.

Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S., and Swanton, C. (2016). DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. *17*, 31.

Rossi, M.R., Ionov, Y., Bakin, A.V., and Cowell, J.K. (2005). Truncating mutations in the ACVR2 gene attenuates activin signaling in prostate cancer cells. Cancer Genet. Cytogenet. *163*, 123–129.

Mehrvarz Sarshekeh, A.M., Advani, S., Overman, M.J., Manyam, G., Kee, B.K., Fogelman, D.R., Dasari, A., Raghav, K., Vilar, E., Manuel, S., et al. (2017). Association of SMAD4 mutation with patient demographics, tumor characteristics, and clinical outcomes in colorectal cancer. PLoS One *12*, e0173345.

Saunders, C.T., Wong, W.S., Swamy, S., Becq, J., Murray, L.J., and Cheetham, R.K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics *28*, 1811–1817.

Seoane, J., and Gomis, R.R. (2017). TGF-beta family signaling in tumor suppression and cancer progression. Cold Spring Harb. Perspect. Biol. *9*, https://doi.org/10.1101/cshperspect.a022277.

Shabalin, A.A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics *28*, 1353–1358.

Shen, H., and Laird, P.W. (2013). Interplay between the cancer genome and epigenome. Cell *153*, 38–55.

Shi, Y., Hata, A., Lo, R.S., Massagué, J., and Pavletich, N.P. (1997). A structural basis for mutational inactivation of the tumour suppressor Smad4. Nature *388*, 87–93.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. Genome Res. *19*, 1117–1123.

Smith, T.C., and Frank, E. (2016). Introducing Machine Learning Concepts with WEKA. Methods Mol. Biol. *1418*, 353–378.

Thuault, S., Valcourt, U., Petersen, M., Manfioletti, G., Heldin, C.H., and Moustakas, A. (2006). Transforming growth factor-beta employs HMGA2 to elicit epithelial-mesenchymal transition. J. Cell Biol. *174*, 175–183.

Tibes, R., Qiu, Y., Lu, Y., Hennessy, B., Andreeff, M., Mills, G.B., and Kornblau, S.M. (2006). Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. Mol. Cancer Ther. *5*, 2512–2521.

Torres-García, W., Zheng, S., Sivachenko, A., Vegesna, R., Wang, Q., Yao, R., Berger, M.F., Weinstein, J.N., Getz, G., and Verhaak, R.G. (2014). PRADA: pipeline for RNA sequencing data analysis. Bioinformatics *30*, 2224–2226.

Totoki, Y., Tatsuno, K., Covington, K.R., Ueda, H., Creighton, C.J., Kato, M., Tsuji, S., Donehower, L.A., Slagle, B.L., Nakamura, H., et al. (2014). Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. Nat. Genet. *46*, 1267–1273.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J.M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics *26*, i237–i245.

Vitenshtein, A., Weisblum, Y., Hauka, S., Halenius, A., Oiknine-Djian, E., Tsukerman, P., Bauman, Y., Bar-On, Y., Stern-Ginossar, N., Enk, J., et al. (2016). CEACAM1-mediated inhibition of virus production. Cell Rep. *15*, 2331–2339.

Wakefield, L.M., and Roberts, A.B. (2002). TGF-beta signaling: positive and negative effects on tumorigenesis. Curr. Opin. Genet. Dev. *12*, 22–29.

Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., et al. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. *38*, e178.

Weinstein, J.N. (2006). Spotlight on molecular profiling: "Integromic" analysis of the NCI-60 cancer cell lines. Mol. Cancer Ther. *5*, 2601–2605.

Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572–1573.

Xu, J., Lamouille, S., and Derynck, R. (2009). TGF-beta-induced epithelial to mesenchymal transition. Cell Res. *19*, 156–172.

Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhsng, C.Z., Wala, J., Mermel, C.H., et al. (2013). Pan-cancer patterns of somatic copy number alteration. Nat. Genet. *45*, 1134–1140.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
RPPA antibodies	RPPA Core Facility, MD Anderson Cancer Center; Tibes et al. (2006); Gonzalez-Angulo et al. (2011)	https://www.mdanderson.org/research/research-resources/core-facilities/functional-proteomics-rppa-core.html
Biological Samples		
Primary tumor samples	Multiple tissue source sites, processed through the Biospecimen Core Resource	See STAR Methods: Subject Details, Method Details
Critical Commercial Assays		
AmpFISTR Identifier kit	Applied Biosystems	Cat: A30737
DNA/RNA AllPrep kit	Qiagen	Cat: 80204
Genome-Wide Human SNP Array 6.0	ThermoFisher Scientific	Cat: 901153
HumanMethylation450	Illumina	Cat: HM450
Illumina Barcoded Paired-End Library Preparation Kit	Illumina	https://www.illumina.com/techniques/sequencing/ngs-library-prep.html
Infinium HumanMethylation450 BeadChip Kit	Illumina	Cat: WG-314-1002
mirVana miRNA Isolation kit	Ambion	N/A
Phusion High-Fidelity PCR Master Mix with HF Buffer	New England Biolabs	Cat: M0531L
QiaAmp blood midi kit	Qiagen	Cat: 51185
RNA6000 Nano Assay	Agilent	Cat: 5067-1511
SureSelect Human All Exon 50 Mb	Agilent	Cat: G3370J
TruSeq PE Cluster Generation Kit	Illumina	Cat: PE-401-3001
TruSeq RNA Library Prep Kit	Illumina	Cat: RS-122-2001
VECTASTAIN Elite ABC HRP Kit (Peroxidase, Standard)	Vector Lab	Cat: PK-6100
Deposited Data		
Digital pathology images	Genomic Data Commons; Cancer Digital Slide Archive	https://gdc-portal.nci.nih.gov/legacy-archive/; http://cancer.digitalslidearchive.net/
Raw and processed clinical, array, and sequencing data	Genomic Data Commons	https://portal.gdc.cancer.gov/legacy-archive/
Mutations MC3 MAF file	Ellrott et al. (2018)	https://www.cell.com/cell-systems/fulltext/S2405-4712(18)30096-6
Software and Algorithms		
ABSOLUTE	Carter et al. (2012)	Pubmed: 22544022
ABySS v1.3.4	Simpson et al. (2009)	Pubmed: 19251739
ABySS v1.4.8	Robertson et al. (2010)	Pubmed: 20935650
BioBloomTools(v1.2.4.b)	Chu et al. (2014)	Pubmed: 25143290
Birdseed	Korn et al. (2008)	Pubmed: 18776909
Blastn (v2.2.23)	Altschul et al. (1997)	PMID: 9254694
CARNAC	Totoki et al. (2014)	Pubmed: 25362482
Circular Binary Segmentation	Olshen et al. (2004)	Pubmed: 15475419
ConsensusClusterPlus	Wilkerson and Hayes (2010)	Pubmed: 20427518
ContEst	Cibulskis et al. (2011)	Pubmed: 21803805
deconstructSigs	Rosenthal et al. (2016)	Pubmed: 26899170

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
deFuse	McPherson et al. (2011)	Pubmed: 21625565
FireHose	The Broad Institute of MIT & Harvard	https://www.broadinstitute.org/cancer/cga/Firehose
GenePattern	Reich et al. (2006)	Pubmed: 16642009
GISTIC2.0	Mermel et al. (2011)	Pubmed: 21527027
Indelocator	Ratan et al. (2015)	Pubmed: 25879703
MAP-RSeq	N/A	https://bioinformaticstools.mayo.edu/research/maprseq
MapSplice 0.7.4	Wang et al. (2010)	Pubmed: 20802226
MuTect	Cibulskis et al. (2013)	Pubmed: 23396013
MutSigCV v1.4	Lawrence et al. (2013)	Pubmed: 23770567
Next-Generation Clustered Heatmap	MD Anderson Cancer Center	https://bioinformatics.mdanderson.org/TCGA/NGCHMPortal/
Oncotator	Ramos et al. (2015)	Pubmed: 25703262
PARADIGM	Vaske et al. (2010)	Pubmed: 20529912
PathSeq	Kostic et al. (2011)	Pubmed: 21552235
Picard	The Broad Institute of MIT & Harvard	https://picard.sourceforge.net/
PRADA	Torres-García et al. (2014)	Pubmed: 24695405
RADIA	Radenbaugh et al. (2014)	Pubmed: 25405470
RSEM	Li and Dewey, (2011)	Pubmed: 21816040
SNPFileCreator	Li and Hung Wong, (2001)	Pubmed: 11532216
SomaticSignatures	Gehring et al. (2015)	Pubmed: 26163694
STAR	Dobin et al. (2013)	Pubmed: 23104886
Strelka	Saunders et al. (2012)	Pubmed: 22581179
Tophat v2.0.8	Trapnell et al. (2009)	Pubmed: 19289445
WEKA	Smith and Frank (2016)	Pubmed: 27008023

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Rehan Akbani (rakbani@mdanderson.org).

Subject Details

Human Data, Tumor Data and TGF-β Pathway Gene Selection

Molecular data were obtained from patients that had not received prior treatment for their disease (ablation, chemotherapy, or radiation therapy) and had provided informed consent as part of The Cancer Genome Atlas Project (TCGA). Local Institutional Review Boards (IRBs) at the tissue source sites reviewed protocols to approve submission of cases.

We selected samples from 33 TCGA projects to analyze the genomic, epigenomic and transcriptomic alterations in the TGF-β pathway.

TCGA Project Management collected necessary human subjects documentation to ensure the project complies with 45-CFR-46 (the ''Common Rule''). The program has obtained documentation from every contributing clinical site to verify that IRB approval has been obtained to participate in TCGA. Such documented approval may include one or more of the following:

- An IRB-approved protocol with Informed Consent specific to TCGA or a substantially similar program. In the latter case, if the protocol was not TCGA-specific, the clinical site PI provided a further finding from the IRB that the already-approved protocol is sufficient to participate in TCGA.
- A TCGA-specific IRB waiver has been granted.
- A TCGA-specific letter that the IRB considers one of the exemptions in 45-CFR-46 applicable. The two most common exemptions cited were that the research falls under 46.102(f)(2) or 46.101(b)(4). Both exempt requirements for informed consent, because the received data and material do not contain directly identifiable private information.
- A TCGA-specific letter that the IRB does not consider the use of these data and materials to be human subjects research. This was most common for collections in which the donors were deceased.

METHOD DETAILS

Sample Processing

Cases were staged according to the American Joint Committee on Cancer (AJCC). Each frozen primary tumor specimen had a companion normal tissue specimen (blood or blood components, including DNA extracted at the tissue source site). Adjacent tissue was submitted for some cases. Specimens were shipped overnight using a cryoport that maintained an average temperature of less than −180°C.

RNA and DNA were extracted from tumor and adjacent normal tissue specimens using a modification of the DNA/RNA AllPrep kit (Qiagen). The flow-through from the Qiagen DNA column was processed using a mirVana miRNA Isolation Kit (Ambion). This latter step generated RNA preparations that included RNA <200 nt suitable for miRNA analysis. DNA was extracted from blood using the QiaAmp blood midi kit (Qiagen). Each specimen was quantified by measuring Abs260 with a UV spectrophotometer or by PicoGreen assay. DNA specimens were resolved by 1% agarose gel electrophoresis to confirm high molecular weight fragments. A custom Sequenom SNP panel or the AmpFISTR Identifier (Applied Biosystems) was utilized to verify tumor DNA and germline DNA were derived from the same patient. Five hundred nanograms of each tumor and normal DNA were sent to Qiagen for REPLI-g whole genome amplification using a 100 μg reaction scale. Only specimens yielding a minimum of 6.9 μg of tumor DNA, 5.15 μg RNA, and 4.9 μg of germline DNA were included in this study. RNA was analyzed via the RNA6000 nano assay (Agilent) for determination of an RNA Integrity Number (RIN), and only the cases with RIN >7.0 were included in this study. Reasons for rejection are described at https://tcga-data.nci.nih.gov/datareports.

Selection of 43 Core Genes Associated with the TGF-β Superfamily

We selected the list of core TGF-β superfamily genes used in the paper by searching for the keyword "TGF-β" in 4 databases: (i) BIOCARTA_TGFB_PATHWAY from GSEA (http://software.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_TGFB_PATHWAY), (ii) KEGG_TGF_BETA_SIGNALING_PATHWAY from GSEA (http://software.broadinstitute.org/gsea/msigdb/cards/KEGG_TGF_BETA_SIGNALING_PATHWAY), (iii) GO_0007179 full gene set from BioMart, and (iv) subset of GO_0007179 (filtered by "experimental evidence") from AmiGo. The union of the resulting lists comprised 181 genes. We then filtered the list down to 43 genes using the following three criteria. (i) Based on the databases' annotations and prior literature, the genes were divided into two categories: those belonging to the signaling cascades and those that encoded targets of the signaling cascades. We retained genes in the former category. (ii) We then performed extensive literature searches and kept only those genes that satisfied any of the following conditions: (a) the gene had previously been implicated in cancer, or (b) the gene was involved in direct binding to and regulation of Smad function, or (c) the gene was phenotypically associated with the TGF-β superfamily, where mutations or deletions of the gene had resulted in phenotypes similar to those from loss of function of the TGF-β superfamily pathways. (iii) Finally, we discussed the complete list of 181 genes and the results of our literature searches with subject matter experts in the TCGA consortium and, after recommendations, reached a consensus for manual curation.

That selection process resulted in 43 "core" genes, including 2 genes encoding adaptor proteins (*SPTNB1* and *ZFYVE9*) that are important in TGF-β signaling and genetically associated by phenotype (Table S1A). However, those two genes are not exclusive to the TGF-β superfamily and they play roles in other cellular processes as well. The other 41 core genes encode components of each level of the "canonical" TGF-β signaling pathway that activates Smads to regulate gene expression (Figure 1A). Other genes that are not members of the canonical pathway (the "noncanonical" TGF-β signaling pathway) are not included in the set of 43 genes, but noncanonical signaling is represented in Figure S1A for the sake of completeness. The 43 genes used in the study encode 3 ligands in the TGF-β subfamily, 8 ligands in the BMP (bone morphogenetic protein) subfamily, and 9 ligands in the ACV (activin) subfamily; 3 receptors for the TGF-β subfamily and 1 interacting protein (TGFBRAP1), 3 receptors for the BMP family, and 6 receptors for the ACV family; and 8 Smads (receptor-activated R-Smads, inhibitor I-Smads, and the common Co-Smad). The list of 43 genes has been made available at cBioPortal (http://www.cbioportal.org) under the category, "General: TGF-β superfamily," so users can explore them further and/or add their own selected genes to study alongside the gene set we used.

Similarly, 50 downstream genes were selected to study transcriptional output of TGF-β pathway activity. These genes included proteins that function in association with TGF-β pathways (2), proteins that regulate the extracellular matrix (2), extracellular matrix proteins (3), transcription factors (13), apoptosis regulators (9), EMT regulators (10), fibrosis inducers (4), tumor promoters (4), E3 ligases (2), and stemness markers (1) (Table S1B).

QUANTIFICATION AND STATISTICAL ANALYSIS

Mitigation of Batch Effects and Systematic Biases

We investigated batch effects first within individual disease types, and then across tumor types. Specifically, we investigated the effects of multiple confounding factors, including differences in: (i) batches in which the samples were processed, (ii) tissue source sites from where the samples were obtained, (iii) the date on which the samples were shipped to the data generation centers, (iv) the instrument on which the samples were processed, (v) the centers that generated the data. The results from individual tumor type analyses can be found online at: (http://bioinformatics.mdanderson.org/tcgambatch/). We assessed the magnitude of batch effects using the following algorithms, (i) clustered heat maps, (ii) PCA plots, and (iii) box plots. Whenever batch effects were observed, we corrected them using (i) ComBat (Johnson et al., 2007), or an enhanced version of it, (ii) Replicates Based Normalization

(RBN) (Akbani et al., 2014), or (iii) removal of bad gene/probe data. Using those methods, we corrected the mRNA, miRNA, DNA methylation and protein expression data. The mutations and copy number data were already discretized and corrected for background loads.

Differences in tumor purity were adjusted for in genomic and epigenomic data. Tumor purity differences in the expression platforms, however, was completely confounded with tumor type differences. More than 5 normal samples were available for only 15 of the 33 tumor types, so application of deconvolution algorithms to the entire cohort was not possible. We acknowledge that differences in tumor purity is a limitation of TCGA expression data, however, TCGA had ensured that all their samples had high tumor content in the sample acquisition phase. The mutation calls used in all of our analyses were somatic mutations only, not germline, so tumor purity differences had minimal impact on that data type. Copy-number alterations (CNA) were assessed as deviations in the tumor sample from the paired normal tissue sample, so they only reflected somatic changes. However, the amplitude of CNA signals can be suppressed in tumor samples with normal cell contamination. We thus utilized ABSOLUTE-derived tumor purity and ploidy estimates for In Silico Admixture Removal (ISAR) of the segmentation data (Zack et al., 2013) in order to correct for any signal dampening that may have occurred before proceeding to analyze somatic copy number alterations. To minimize the influence of normal tissue contamination and leukocytes infiltration in DNA methylation data, we chose probes not methylated in all relevant normal tissues and blood cells, to get rid of methylation signals from possible confounding factors.

DNA Sequencing Data

Exome capture was performed using Agilent SureSelect Human All Exon 50 Mb according to the manufacturer's instructions. Briefly, 0.5–3 micrograms of DNA from each sample were used to prepare the sequencing library through shearing of the DNA followed by ligation of sequencing adaptors. All whole exome (WES) and whole genome (WGS) sequencing was performed on the Illumina HiSeq platform. Paired-end sequencing (2 x 101 bp for WGS and 2 x 76 bp for WE) was carried out using HiSeq sequencing instruments; the resulting data was analyzed with the current Illumina pipeline. Basic alignment and sequence QC was done on the Picard and Firehose pipelines at the Broad Institute. Sequencing data were processed using two consecutive pipelines:

(1) Sequencing data processing pipeline (''Picard pipeline'')

Picard (http://picard.sourceforge.net/) uses the reads and qualities produced by the Illumina software for all lanes and libraries generated for a single sample (either tumor or normal) and produces a single BAM file (http://samtools.sourceforge.net/SAM1.pdf) representing the sample. The final BAM file stores all reads and calibrated qualities along with their alignments to the genome.

(2) Cancer genome analysis pipeline (''Firehose pipeline'')

Firehose (http://www.broadinstitute.org/cancer/cga/Firehose) takes the BAM files for the tumor and patient- matched normal samples and performs analyses including quality control, local realignment, mutation calling, small insertion and deletion identification, rearrangement detection, coverage calculations and others as described briefly below. The pipeline represents a set of tools for analyzing massively parallel sequencing data for both tumor DNA samples and their patient-matched normal DNA samples. Firehose uses GenePattern (Reich et al., 2006) as its execution engine for pipelines and modules based on input files specified by Firehose. The pipeline contains the following steps:

a. Quality control

This step confirms identity of individual tumor and normal to avoid mix-ups between tumor and normal data for the same individual.

b. Local realignment of reads

This step realigns reads at sites that potentially harbor small insertions or deletions in either the tumor or the matched normal, to decrease the number of false positive single nucleotide variations caused by misaligned reads.

c. Identification of somatic single nucleotide variations (SSNVs)

This step detects candidate SSNVs using a statistical analysis of the bases and qualities in the tumor and normal BAMs, using Mutect (Cibulskis et al., 2013).

d. Identification of somatic small insertions and deletions

In this step, putative somatic events were first identified within the tumor BAM file and then filtered out using the corresponding normal data, using Indellocator (Ratan et al., 2015).

Mutation Analysis

The non-silent mutation frequencies for each gene in the individual cancer and Pan-Cancer settings are determined through mining the MC3 TCGA MAF file (covering n=9125 patients of the Pan-Cancer pathway analysis consortium manuscript freeze sample set) from 33 cancer types. To include only the non-silent mutations, the variant classes, ''Silent,'' "Intron," "3'UTR," "3'Flank," ''5'UTR,''

‘‘5’Flank,’’ "IGR,’’ and "RNA" are excluded from the analyses. The oncoprints are generated using the cBioPortal oncoprinter suite. Each oncoprint visualizes and quantifies the somatic mutation and copy number events in 9,125 patients with 33 cancer types for each gene family in the pathway (Figure 1B). The hotspot mutations are extracted from MC3 MAF file first programmatically for any hotspot site with more than nine counts and validated through a systematic mining in cBioPortal (Figure 3). The hotspots are visualized using the mutationMapper tool in cBioPortal. For ACVR2A and SMAD4 hotspot mutations are mapped onto the respective protein structures (PDB IDs: 4ASX for ACVR2A and 1DD1 for SMAD4) using the UCSF chimera software. The driver mutations in the pathway are detected using MutSigCV for all cancer types in the Pan-Cancer set (Figures 2, 4, and 7F). Although MutSigCV is a well-established method for detecting driver genomic aberrations in cancer, it does have the following limitations. MutSigCV is insensitive to some genomic events, such as the co-occurrence of mutations in genomic proximity and or mutations that are associated with transcription-coupled repair. MutSigCV identifies genomic heterogeneity across patient cohorts. Another challenge that cannot be addressed by MutSigCV is intratumor heterogeneity and detection of driver mutations within subclones of a single tumor. Finally, success of MutSigCV depends on the statistical properties and size of the patient population under study as the algorithm fails to classify rare variants seen within small to mid-sized patient cohorts. Differential mRNA expression of 50 TGF-β pathway target genes is also quantified in relation to 6 hotspot mutations in the Pan-Cancer cohort and GI cancers (Figure 3B). Rows and columns were clustered using the complete-linkage algorithm with Euclidean distance, and dendrogram branches were ordered to minimize the differences between the cube of the mean of adjacent rows and columns.

Copy Number Analysis

Tumor sample DNA was extracted and hybridized to Affymetrix SNP6.0 arrays by the Genome Analysis Platform at the Broad Institute as previously described (McCarroll et al., 2008). The calculated array probe intensities were normalized and combined using SNPFileCreator (Li and Hung Wong, 2001) and then processed with Birdseed (Korn et al., 2008) to yield preliminary copy-number estimates. Segmented relative copy-number profiles were produced by refining and partitioning the preliminary copy-number estimates with tangent normalization and Circular Binary Segmentation (Olshen et al., 2004). The segmented relative copy-number profiles for 9,125 samples were selected for further analysis. For each disease type, GISTIC2.0 (Mermel et al., 2011) was ran on the corresponding copy-number profiles to identify regions undergoing significant focal-level and broad-level somatic copy-number aberrations and to obtain gene-level estimates of copy-number. The significant genomic amplification lesions and genomic deletion lesions identified by GISTIC2.0 were examined to determine if any TGF-β network genes were being targeted as potential oncogenic or anti-oncogenic drivers, and the frequency of amplifications and deletions across the TGF-β network genes were computed from the gene-level thresholded copy-number estimates (-2, -1, 0, +1, +2). Genes assigned positive values of +1 and +2 were considered amplified, with +1 representing low-level amplification events and +2 representing high-level amplification events, and genes with negative values of -1 and -2 were considered deleted, with -1 representing shallow deletion events and -2 representing deep deletion events.

GISTIC2.0 is a tool for detecting independently targeted regions of SCNA, based on data-driven estimation of the background rates of SCNA. GISTIC2.0 used data from SNP arrays, thus the successful application of GISTIC2.0 to detect low frequency differences depends on the resolution of array or sequencing platform and the population size.

GISTIC identifies somatic alterations that occur significantly more frequently than those predicted to occur at random, based on the background rate of copy number changes. The issue with this and all significance methods is that the ability to detect rare but meaningful driver events depends on the frequencies of their occurrence and on the number of the tumors profiled. Tumor types for which few tumors have been profiled and that have infrequently occurring copy number alterations, GISTIC may fail to identify rare but important somatic events. As more copy number profiles become available through large-scale tumor sequencing efforts, the ability to detect these rare but significant events will increase.

Pathway Analysis

A pathway topology is generated to link the 43 core TGF-β pathways based on database searches in KEGG and Pathway Commons, expert curation and literature searches. The pathway diagram is visualized and optimized for layout using the Pathway Mapper program (Figure S4A). The genomic alteration frequencies for copy number gains or losses and mutations are extracted from the cBioPortal and programmatically form the MC3 MAF file. The alterations are mapped to each gene in the pathway diagram. In the GI-focused pathway analysis, only genes with >3% alteration for either copy number or mutation alterations are included in the pathway diagram to capture only those pathways that are substantially altered (Figure S4A).

Expression Signatures of Genomic Alterations

The gene expression signatures of TGF-β pathway alterations are analyzed with a clustering algorithm. The samples with alterations in each core gene and wild type for all TGF-β pathway genes are extracted from the MC3 MAF file. The transcriptional output is quantified using expression of 50 downstream genes (Figures 2G–2I). The median fold change of transcriptional changes are calculated as the ratio of expression of downstream genes among all core pathway gene mutated, amplified and deleted samples to expression levels in TGF-β pathway wild type samples. The transcriptional changes in each downstream gene vs each altered pathway gene is analyzed and visualized with a two-way hierarchical clustered heat map (Figures 4A–4B). The hierarchical clustering is performed using a Euclidean distance and complete linkage. The shift in the transcriptional output shift in different subsets such as Pan-Cancer and GI cancers are visualized with a volcano plot with BH based FDR adjusted P values calculated with a Wilcoxon signed rank test

(null hypothesis is the transcriptional output shift in the two subsets are equal to each other) and log fold change of the fold changes in Pan-Cancer vs. GI cancers (Figure 4C). The global transcriptional output is calculated by comparing fold changes due to TGF-β pathway alterations in all transcripts measured (Figure 4E).

Gastrointestinal Cancers

The cancer types, Colon Adenocarcinoma (COAD, N=341), Esophageal carcinoma (ESCA, N=169), Liver hepatocellular carcinoma (LIHC, N=348), Pancreatic adenocarcinoma (PAAD, N=152), Rectum adenocarcinoma (READ, N=118), Stomach adenocarcinoma (STAD, N=383) are included as the gastrointestinal (GI) samples. The enrichment of TGF-β pathway genomic alterations in the GI cancers was statistically assessed using a one tailed Fisher's exact test, where the null hypothesis was the odds ratio of alterations in GI vs other cancers was not greater than 1. The total number of GI samples was 1511. The transcriptional outcome of GI cancers with TGF-β pathway disruptions were quantified using the same method and downstream target gene list as we did in the analysis of transcriptional output from all cancers (Figures 3B and 4). The pathway analysis was performed as in the case of the Pan-Cancer cohort for each GI cancer type separately (Figure S4). In the pathway analysis, the core genes with that >3% alteration frequency for any of the alteration types (mutations, copy number amplification or deletion) were included into the pathway diagrams while the rest was eliminated.

mRNA Expression Analysis and Pathway Activity Scores

We corrected for batch effects the TCGA mRNA data available from TCGA's web portal (https://portal.gdc.cancer.gov/). The $\log_2$ transformed data were used for all the mRNA analysis in this project. Pathway scores were generated by first Z-normalizing the values for the 43 core genes across all of the samples (Figures 5, 6, and 7). The mean across the 43 genes was then calculated for each sample to yield the pathway activity score per sample. Unsupervised clustering used 1-Pearson's correlation for the distance metric with Ward's linkage. One limitation of the pathway activity score is that it gives equal weight to all the genes in the pathway, meaning that the abundance of each transcript contributes an equal positive value to the score. This is not reflective of the biology, for example, some genes encode inhibitors of pathway activity, and some components interact with multiple partners and thus may be limiting. Another limitation is that the score uses expression of the genes as a surrogate for functional protein abundance, which does not account for loss or gain of function due to mutations. Thus, the pathway activity score represents a relative estimate not an absolute value of pathway function.

microRNA Analysis

We corrected the TCGA miRNA data available from TCGA's web portal (https://portal.gdc.cancer.gov/) for batch effects. For 9310 primary tumor samples, we used MatrixEQTL v2.1.1 or v2.2 (Shabalin, 2012 Pubmed: 22492648) in R 3.4.1 or 3.4.4 to calculate Spearman correlations between batch-corrected, normalized expression data for miRNA mature strands and gene-level mRNA data for 43 pathway genes. We then filtered by records in miRTarBase v6.0 (Chou et al., 2016), retaining both stronger and weaker functional interactions. We further filtered by requiring correlations to have a coefficient <-0.25 and an FDR $<10^{-6}$, which resulted in the retention of 40 miR-mRNA pairs involving 32 miRNA mature strands. For heat maps, we removed eight mature strands, because they were too weakly expressed (<10 RPM) in all or most tumor types, retaining 24 mature strands. For the main heat map of batch-corrected miRNA-seq data, we identified 8930 samples from 32 of 33 tumor types that were from primary tumors, metastatic tumors, or blood cancers. These samples were represented in the ordered heat map for messenger RNAs from the pathway (Figure 7E). We ordered the samples to match the sample order in the messenger RNA heat map (i.e. with cancer types ordered to have increasing mean pathway scores, and samples within a cancer type ordered to have increasing pathway scores). We generated a heat map using the pheatmap v1.0.2 package, in R 3.4.1. We generated a similar heat map for the 1507 primary tumors present in LIHC, COAD, READ, STAD, ESCA, and PAAD data sets. Box plots were generated using the boxplot() function in R (Figure 7C). The data consisted of the mean miRNA value across the 24 miRNAs. A limitation of this approach is that the results are not based on rigorous and objective thresholds for the metrics (like correlations or p values). Rather the thresholds were chosen to yield a reasonably small set of the most statistically significant miRNAs that were easy to evaluate and visualize for human interpretation. Otherwise, the results would appear like the proverbial "hair ball."

DNA Methylation Profiles

We mapped the Illumina methylation array probes to individual genes using the Illumina Human Methylation 27k R annotation data package. Forty-one of forty-three TGF-β pathway genes had at least one probe mapping to their promoter region. For genes with multiple probes, median beta values were used. We then calculated median beta value for these 41 genes in each sample, and plotted them using the boxplot function in R, grouped by cancer type (Figure 7A). For the heat maps, we calculated beta values for each of the 41 genes of TGF-β pathway and the 33 tumor types by taking median across all samples for a given tumor. We then plotted this data as a heat map using the Clustergram function in Matlab (Figure 7B). For the analysis of the GI methylation data, probes were mapped to TGF-β pathway genes for GI cancers (COAD+READ, STAD, ESCA, PAAD and LIHC). Beta values for each gene-sample pair was visualized as a heat map using the ComplexHeatmaps package, with TGF-β pathway genes clustered using Euclidean distances and Ward's linkage. Box plots were generated using the boxplot() function in R. The data consisted of the mean beta value across the 41 genes. This method assumes the mean beta value is reflective of the overall methylation level of the entire pathway, which may not always hold and is a limitation of the approach.

Correlations of Pathway Score Vs. Bootstrapped Genes

Following the calculation of TGF-β pathway scores, the absolute value of the Pearson correlation between gene expression values and pathway scores was calculated for all 20,310 genes where this calculation was possible. Next, 43 correlation values were sampled with replacement from the correlation values of the 43 TGF-β pathway genes a total of 10,000 times, and each time the median sampled correlation was calculated. The same sampling procedure was also performed for the TGF-β target genes, where 50 correlation values were sampled with replacement from the correlation values of the 50 target genes, and for all genes, where 50 correlation values were sampled with replacement from the correlation values of all 20,310 genes. The distribution of the 10,000 median correlations for each of the three gene sets is shown in Figure S5. A p-value was also calculated for each group as the proportion of median correlations for the ''all genes'' group that are greater than or equal to the median of each group.

Survival Analysis

Kaplan-Meier Survival Curves are generated for each patient sub cohort using the Survival and Survminer R packages (Figures 6C–6E and S6B–S6D). The statistical significance of survival differences between multiple subcohorts were determined using the log-rank test to capture relations. In order to segment the cohorts into subgroups characterized by expression levels of the TGF-β target genes, we analyzed the distribution of target gene expression across the Pan-Cancer cohort. We particularly focused on mRNA expression distribution of *HMGA2*, *MMP9*, collagens (*COL1A1*, *COL1A2*, *COL3A1*), *TERT*, *FOXP3*, *CDH2* as the expression of these genes varied significantly between TGF-β pathway mutated vs. wild type samples. For this purpose, we used the batch normalized mRNA expression data. For each gene, the cut-off to separate low and high expressing cohorts was determined empirically based on the distribution. For expression profiles with a unimodal distribution, we used the approximate median values. For bimodal cases, we selected the threshold as the midpoint that separates each peak on the bimodal distribution. The mRNA expression threshold to separate the cohorts with low vs. high target expression groups were HMGA2=5, MMP9=10, mean of collagens (*COL1A1*, *COL1A2*, and *COL1A3*) = 14, TERT=2, FOXP3=6, CDH2=8. The collagen genes are analyzed as a single entity because they showed very strong correlation of mRNA expression with each other. The resulting thresholds divided the cohorts into three groups as TGF-β expression, TGF-β mutant/high target expression, TGF-β wt/high target expression and low target gene expression. We merged the TGF-β mutant/low target expression and TGF-β wt/low expression cohorts as discriminating between these sets do not inform on the combined effect of TGF-β mutations and target expression. The survival differences between each sub cohort are analyzed using the Survival and Survminer R packages.

DATA AND SOFTWARE AVAILABILITY

The raw data, processed data and clinical data can be found at the legacy archive of the GDC (https://portal.gdc.cancer.gov/legacyarchive/search/f) and the PanCanAtlas publication page (https://gdc.cancer.gov/about-data/publications/pancanatlas). The mutation data can be found here (https://gdc.cancer.gov/about-data/publications/mc3-2017). TCGA data can also be explored through the Broad Institute FireBrowse portal (http://gdac.broadinstitute.org) and the Memorial Sloan Kettering Cancer Center cBioPortal (http://www.cbioportal.org). Details for software availability are in the Key Resources Table.

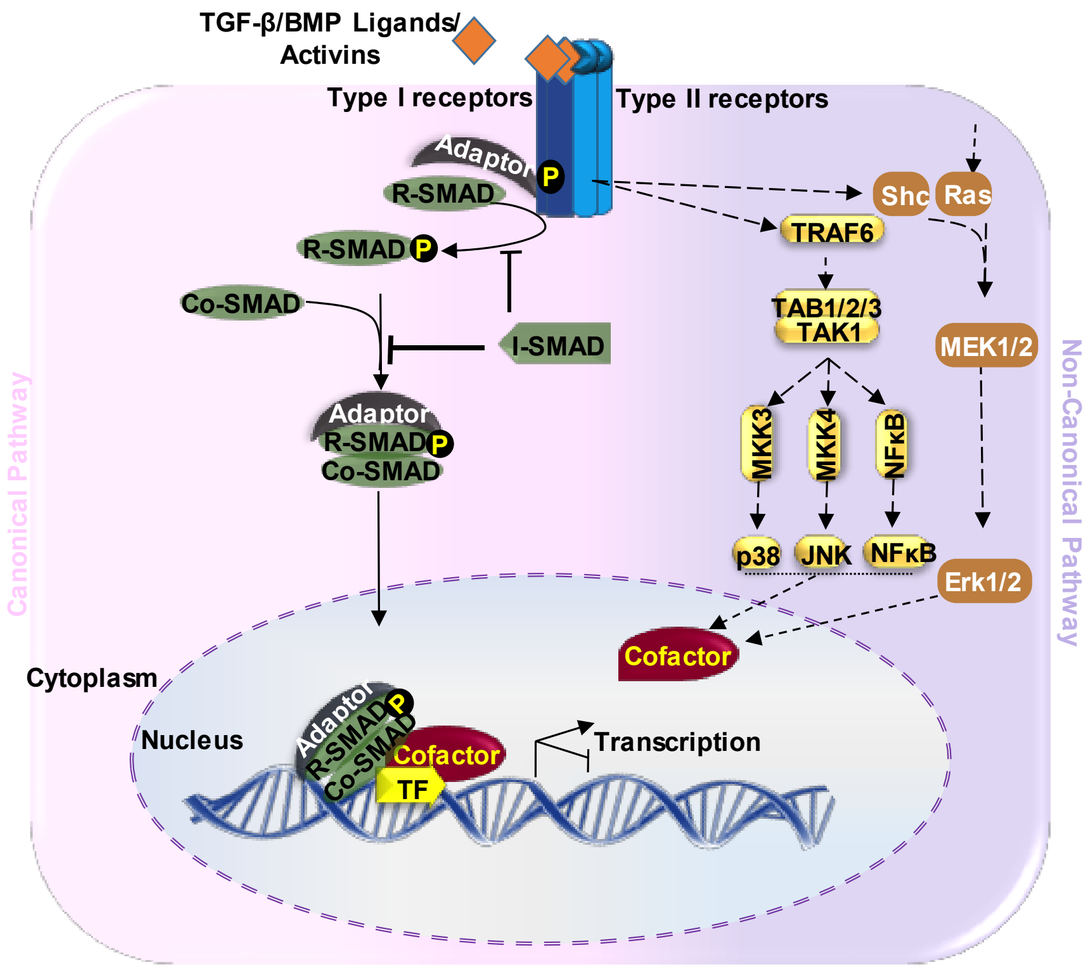
Cell Systems, Volume *7*

Supplemental Information

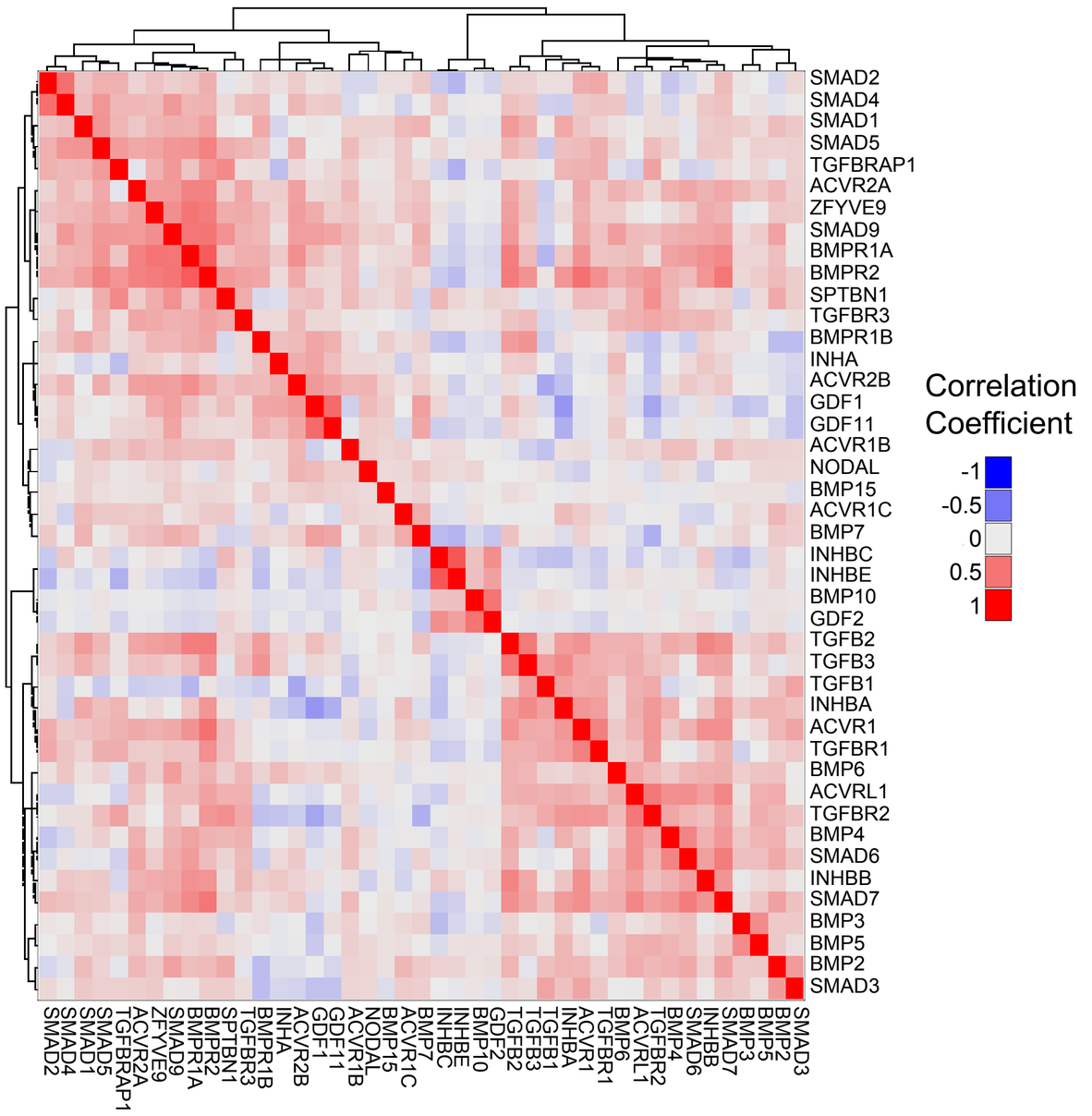
A Pan-Cancer Analysis Reveals High-Frequency Genetic Alterations in Mediators of Signaling by the TGF-β Superfamily

Anil Korkut, Sobia Zaidi, Rupa S. Kanchi, Shuyun Rao, Nancy R. Gough, Andre Schultz, Xubin Li, Philip L. Lorenzi, Ashton C. Berger, Gordon Robertson, Lawrence N. Kwong, Mike Datto, Jason Roszik, Shiyun Ling, Visweswaran Ravikumar, Ganiraju Manyam, Arvind Rao, Simon Shelley, Yuexin Liu, Zhenlin Ju, Donna Hansel, Guillermo de Velasco, Arjun Pennathur, Jesper B. Andersen, Colm J. O'Rourke, Kazufumi Ohshiro, Wilma Jogunoori, Bao-Ngoc Nguyen, Shulin Li, Hatice U. Osmanbeyoglu, Jaffer A. Ajani, Sendurai A. Mani, Andres Houseman, Maciej Wiznerowicz, Jian Chen, Shoujun Gu, Wencai Ma, Jiexin Zhang, Pan Tong, Andrew D. Cherniack, Chuxia Deng, Linda Resar, The Cancer Genome Atlas Research Network, John N. Weinstein, Lopa Mishra, and Rehan Akbani

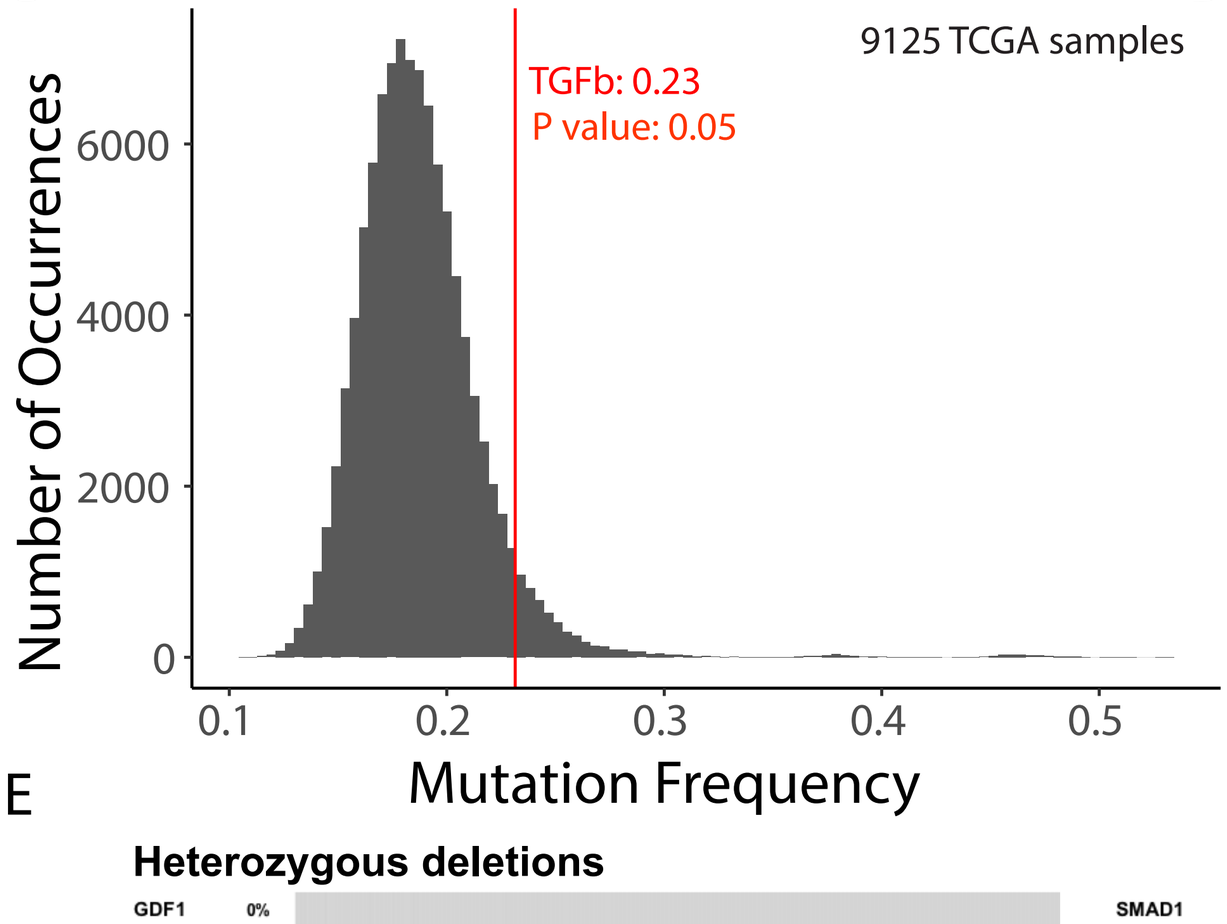
A


TGF-β/BMP Ligands/
Activins
Type I receptors
Type II receptors
Adaptor
P
R-SMAD
R-SMAD P
Co-SMAD
I-SMAD
Adaptor
R-SMAD P
Co-SMAD
Shc
Ras
TRAF6
TAB1/2/3
TAK1
MEK1/2
MKK3
MKK4
NFκB
p38
JNK
NFκB
Erk1/2
Cofactor
Canonical Pathway
Non-Canonical Pathway
Cytoplasm
Nucleus
Adaptor
R-SMAD P
Co-SMAD
Cofactor
TF
Transcription

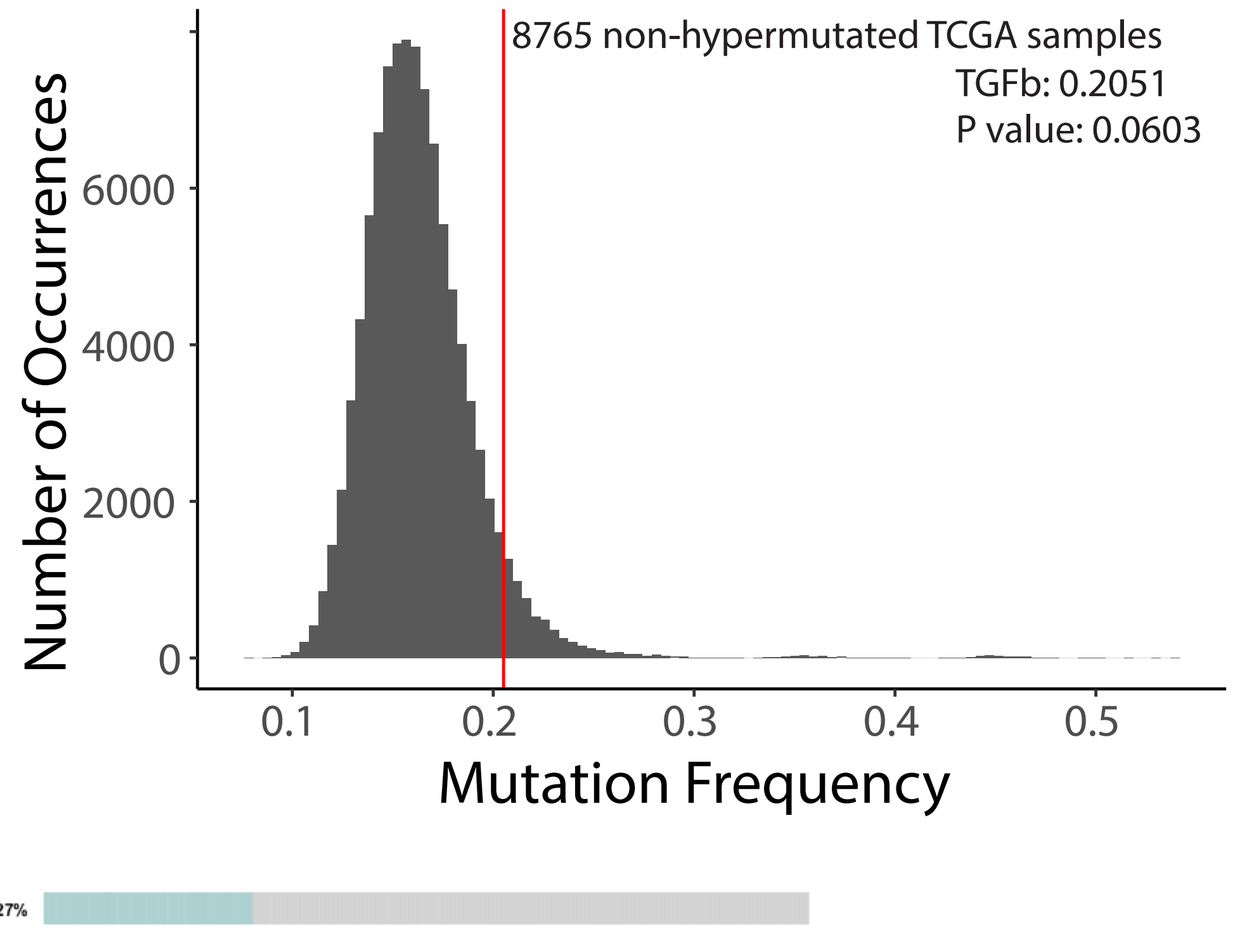
B


SMAD2
SMAD4
SMAD1
SMAD5
TGFBRAP1
ACVR2A
ZFYVE9
SMAD9
BMPR1A
BMPR2
SPTBN1
TGFBR3
BMPR1B
INHA
ACVR2B
GDF1
GDF11
ACVR1B
NODAL
BMP15
ACVR1C
BMP7
INHBC
INHBE
BMP10
GDF2
TGFB2
TGFB3
TGFB1
INHBA
ACVR1
TGFBR1
BMP6
ACVRL1
TGFBR2
BMP4
SMAD6
INHBB
SMAD7
BMP3
BMP5
BMP2
SMAD3
Correlation Coefficient
-1
-0.5
0
0.5
1

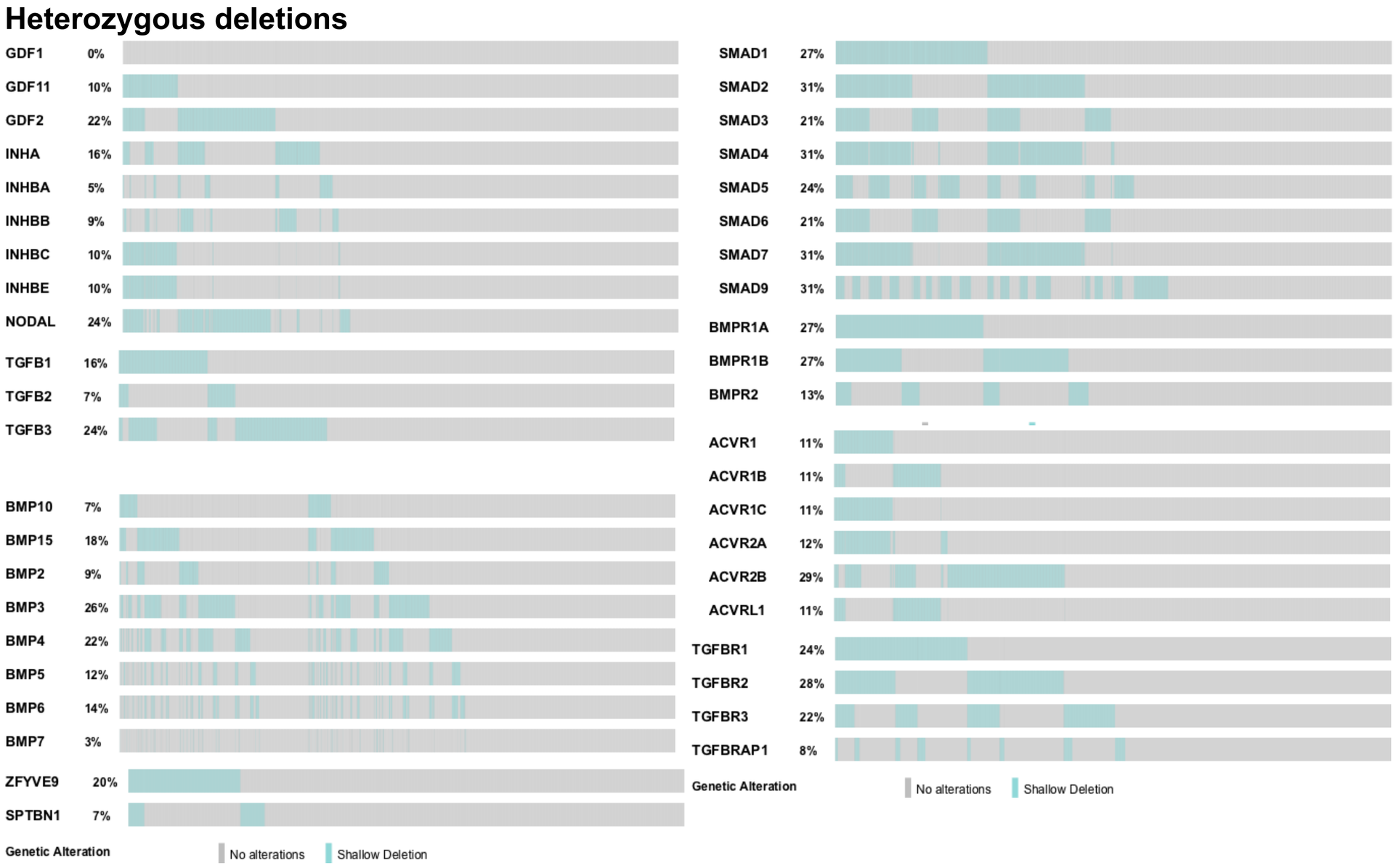
C


9125 TCGA samples
TGFb: 0.23
P value: 0.05
Number of Occurrences
6000
4000
2000
0
0.1
0.2
0.3
0.4
0.5
Mutation Frequency

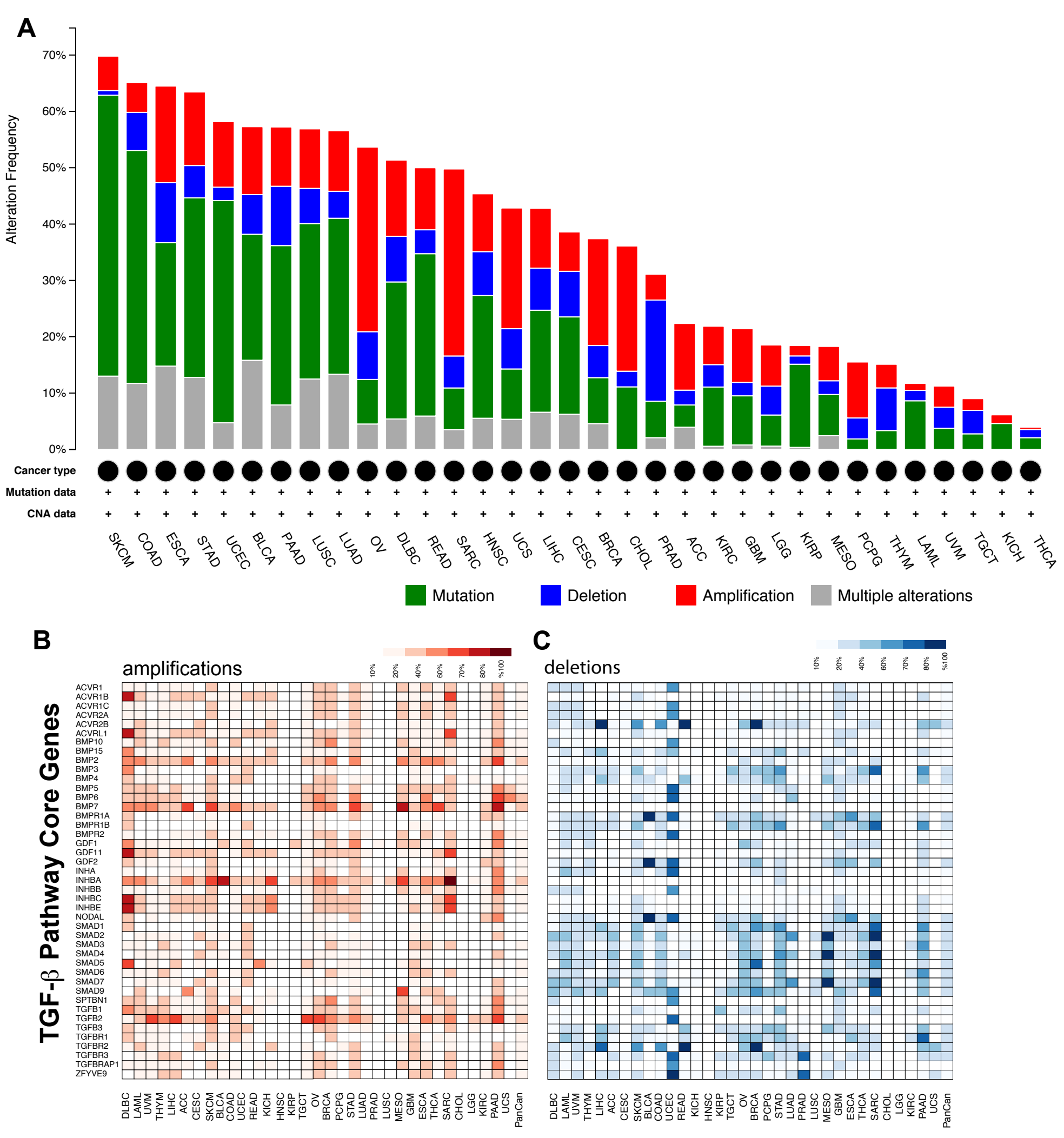
D


8765 non-hypermutated TCGA samples
TGFb: 0.2051
P value: 0.0603
Number of Occurrences
6000
4000
2000
0
0.1
0.2
0.3
0.4
0.5
Mutation Frequency

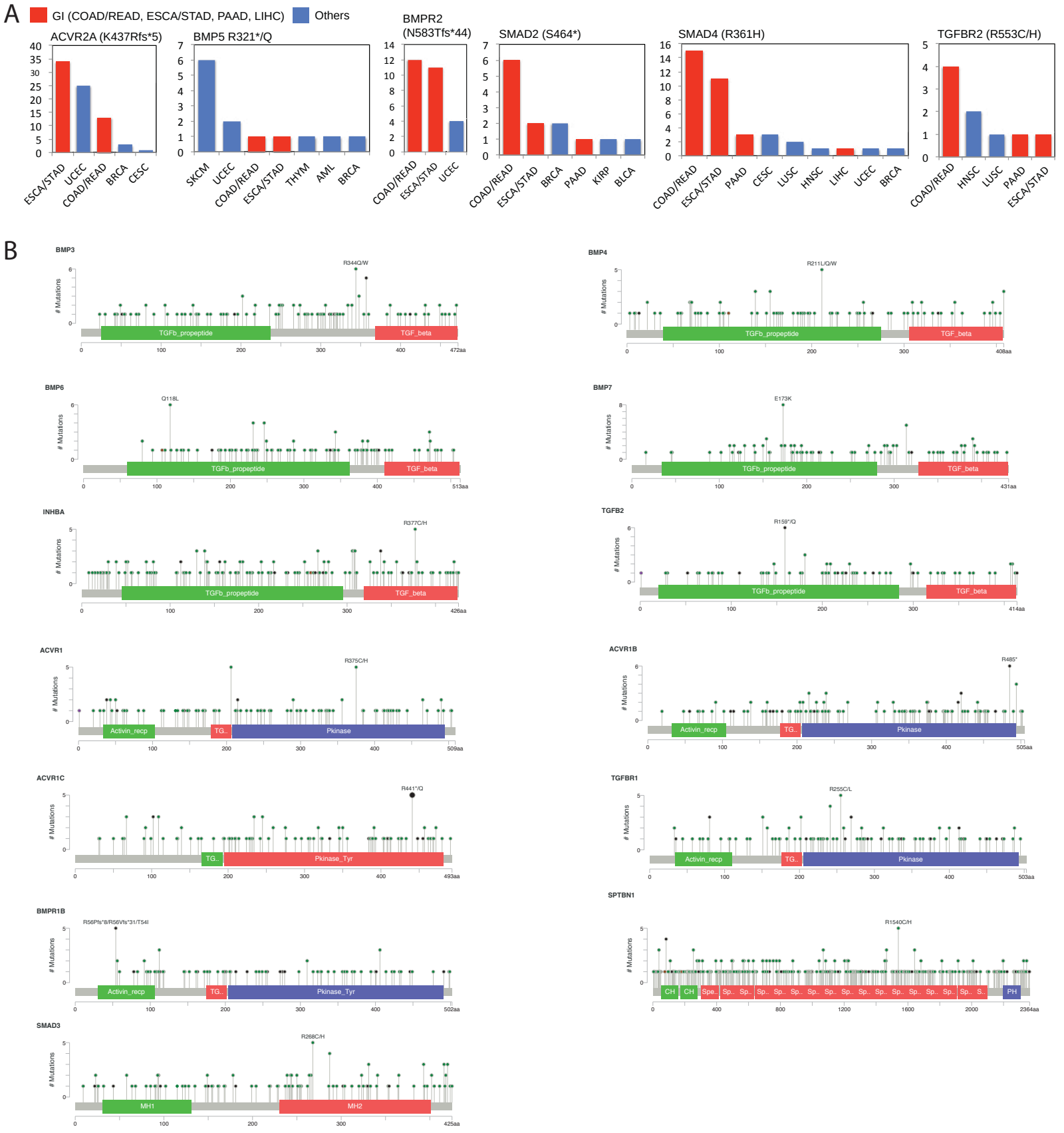
E


Heterozygous deletions
GDF1 0%
GDF11 10%
GDF2 22%
INHA 16%
INHBA 5%
INHBB 9%
INHBC 10%
INHBE 10%
NODAL 24%
TGFB1 16%
TGFB2 7%
TGFB3 24%
BMP10 7%
BMP15 18%
BMP2 9%
BMP3 26%
BMP4 22%
BMP5 12%
BMP6 14%
BMP7 3%
ZFYVE9 20%
SPTBN1 7%
SMAD1 27%
SMAD2 31%
SMAD3 21%
SMAD4 31%
SMAD5 24%
SMAD6 21%
SMAD7 31%
SMAD9 31%
BMPR1A 27%
BMPR1B 27%
BMPR2 13%
ACVR1 11%
ACVR1B 11%
ACVR1C 11%
ACVR2A 12%
ACVR2B 29%
ACVRL1 11%
TGFBR1 24%
TGFBR2 28%
TGFBR3 22%
TGFBRAP1 8%
Genetic Alteration
No alterations
Shallow Deletion

Supplementary Figure S1. Related to Figure 1. A. Signaling by the canonical vs. noncanonical TGF-β pathway. **B.** Clustered heat map of gene-gene correlations of the 43 genes across the entire PanCancer cohort of 9125 samples. **C. Statistical significance of the cumulative mutation frequency in the TGF-β superfamily pathways.** A bootstrapping analysis is performed to assess the significance of the observed mutational frequency in the 43 genes. The null distribution is the mutation frequencies of 100,000 randomly selected sets of 37 genes (the number of genes in the TGF-β superfamily pathway gene set with > 0.5% mutation frequency and hence contribute to the cumulative frequency). The distribution is calculated using the 9125 sample PanCancer cohort (same cohort used for analysis of the TGF-β superfamily pathway alterations). The TGF-β superfamily pathway gene set has significantly more mutations than a randomly selected gene set of identical size. p-value is 0.05 for the cumulative mutation frequency of the 37 TGF-β superfamily genes (each gene has a frequency of >0.5%). **D. Significance of mutations in non-hypermutated samples.** The same bootstrapping analysis is performed after excluding the hypermutator samples. The analysis leads to a lower mutation frequency and significance is not affected. **E. Landscape of heterozygous (shallow) deletions in the TGF-β superfamily pathway genes.** The oncoprints for heterozygous deletions in each gene across all cancer types as identified by GISTIC thresholds (>0.3) in the PanCancer cohort and visualized using the oncoprinter tool. Recurrent heterozygous loss is usually associated with gradual loss of tumor suppressor functions.



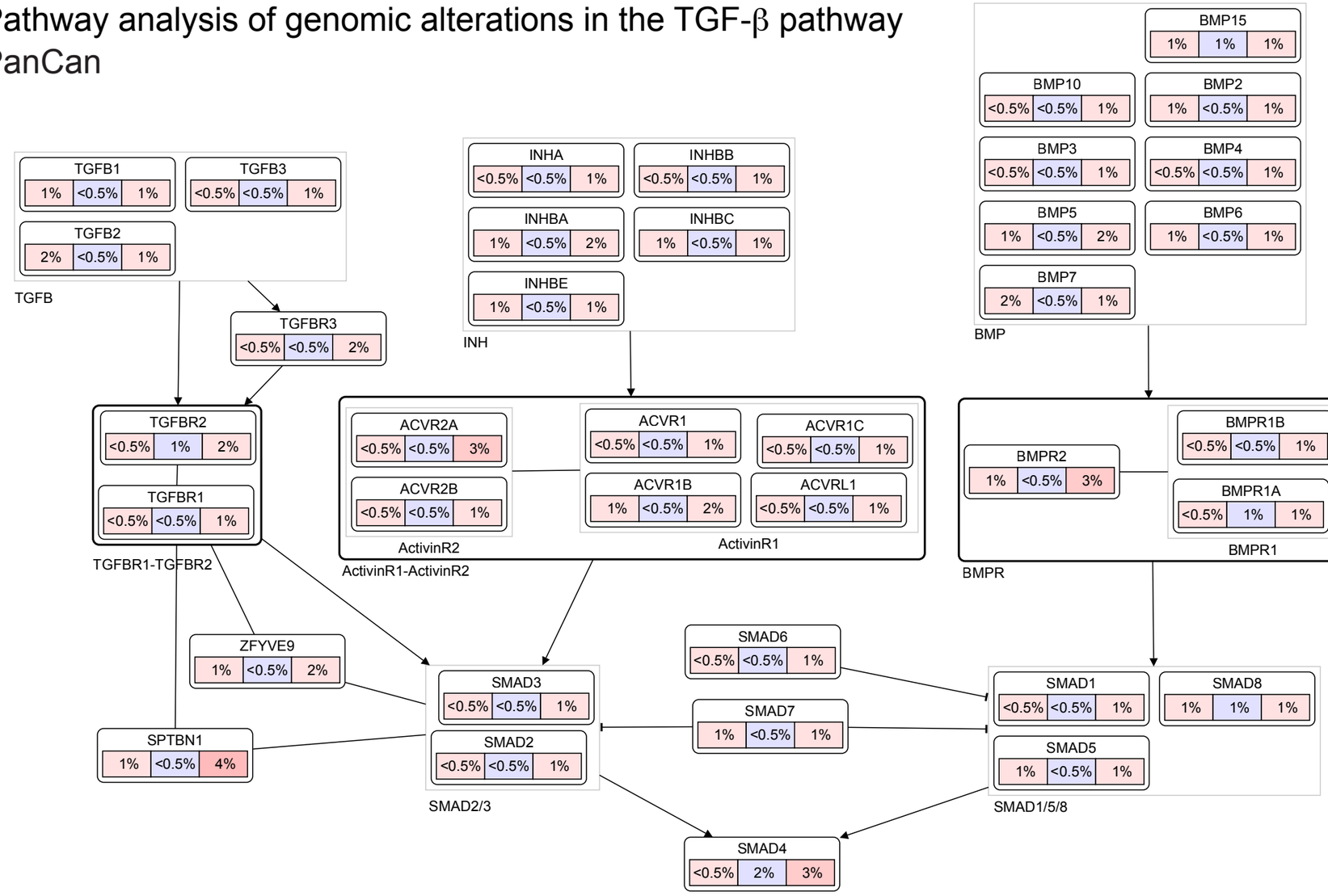
Supplementary Figure S2. Related to Figure 2. A. Distribution of genomic alterations (mutations, high level amplifications and deep deletions) in the 43 TGF-β superfamily pathway genes across the 33 cancer types. **B.** The frequencies of copy number gain and highly amplified TGF-β superfamily genes. **C.** The frequencies of shallow (heterozygous) and deep (homozygous) deletions in TGF-β superfamily genes.



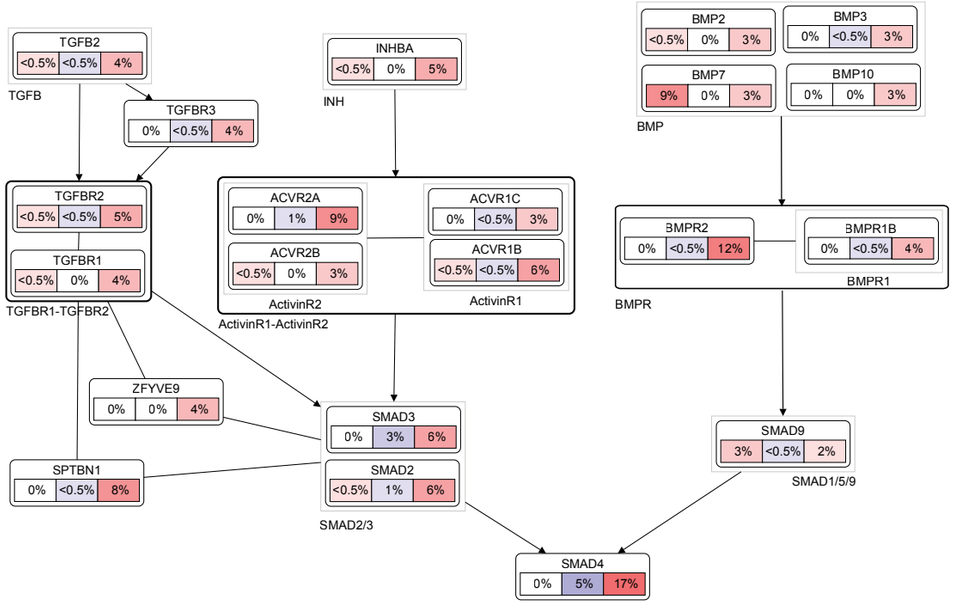
Supplementary Figure S3. Related to Figure 3. A. Distribution of hotspot mutations in cancer types. Hotspots sites are enriched in GI cancers (red) — particularly in colorectal, stomach, and esophageal cancers — compared to other cancer types (blue). **B.** Potential hotspot sites in the TGF-β superfamily pathway genes. The lollipop mutation plots for TGF-β superfamily pathway genes noting individual amino acid residues with codons that are mutated in >5 and <9 incidences across all PanCancer samples.

A

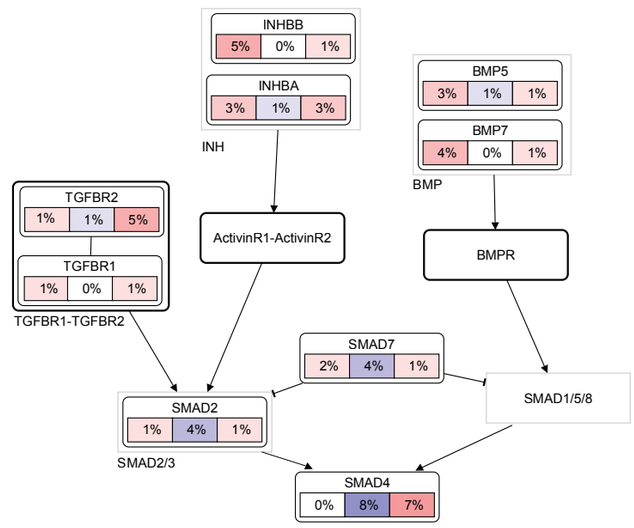
Pathway analysis of genomic alterations in the TGF-β pathway
PanCan



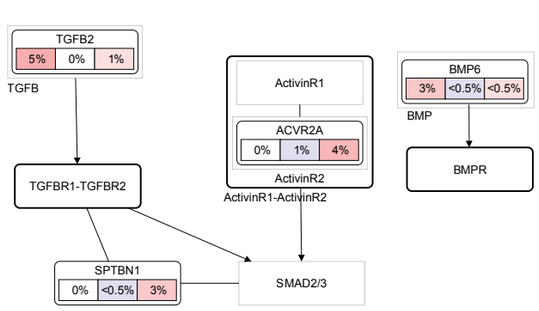
Colorectal carcinoma



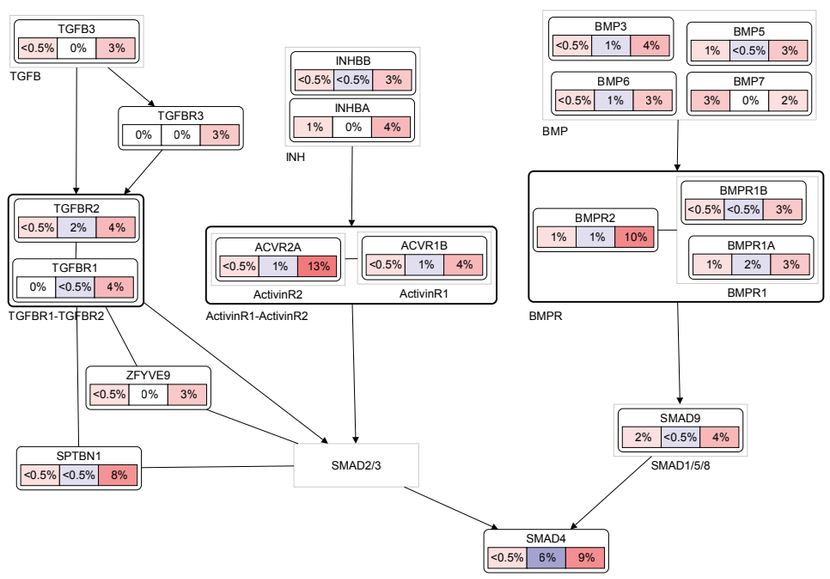
Esophageal carcinoma



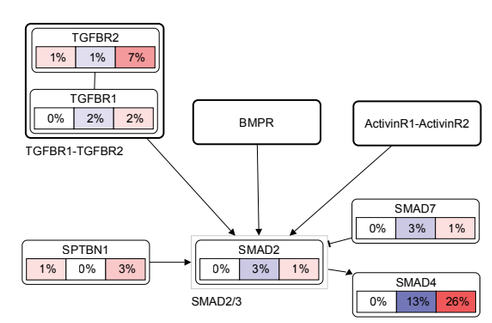
Liver hepatocellular carcinoma



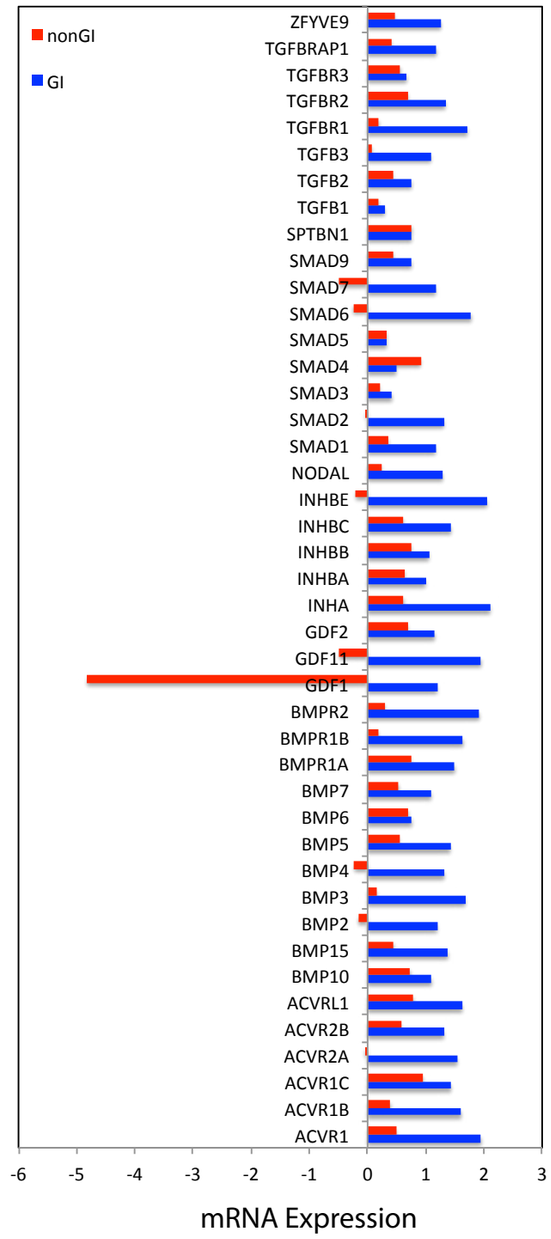
Stomach adenocarcinoma



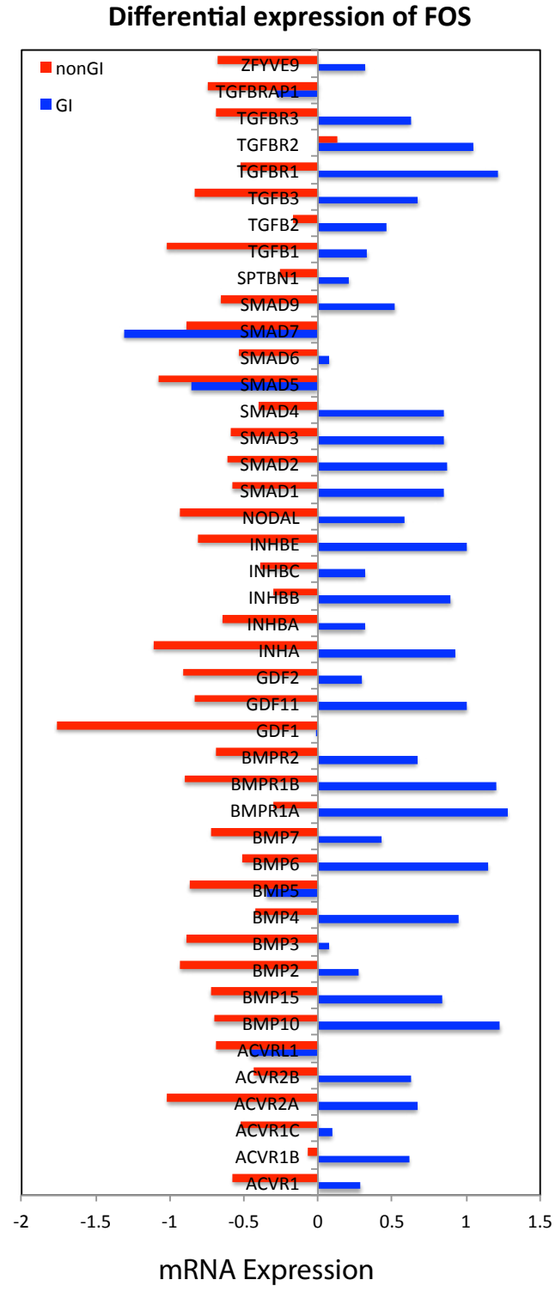
Pancreatic adenocarcinoma



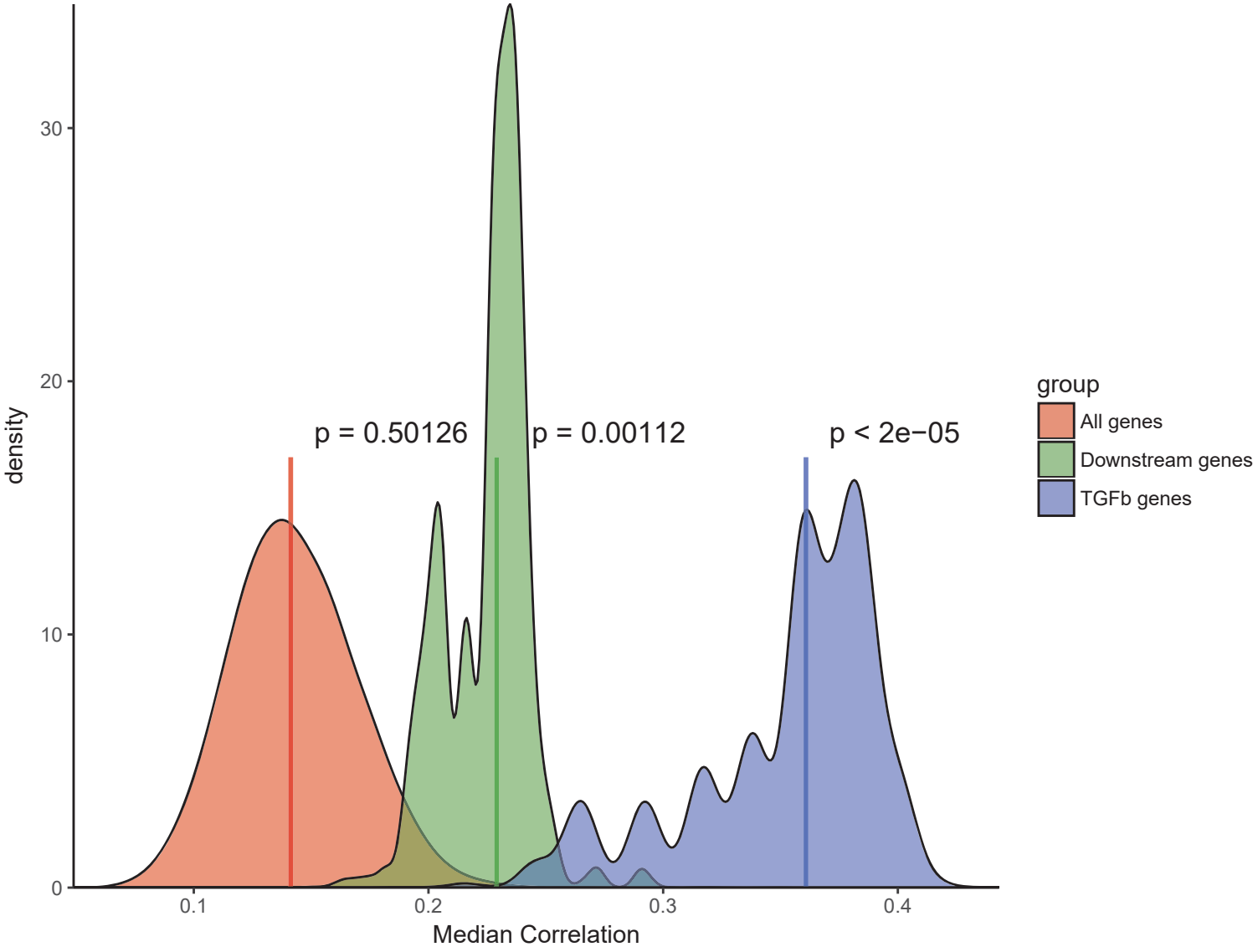
B



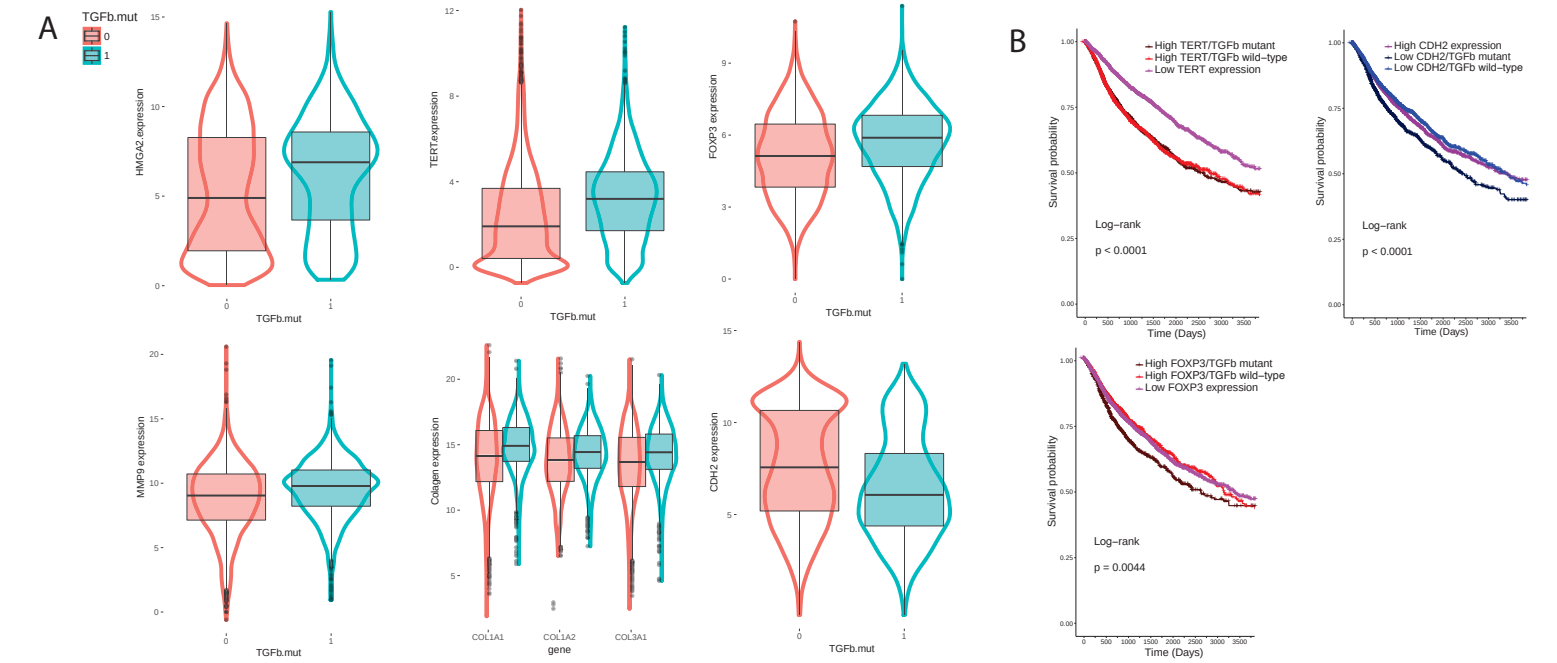
C

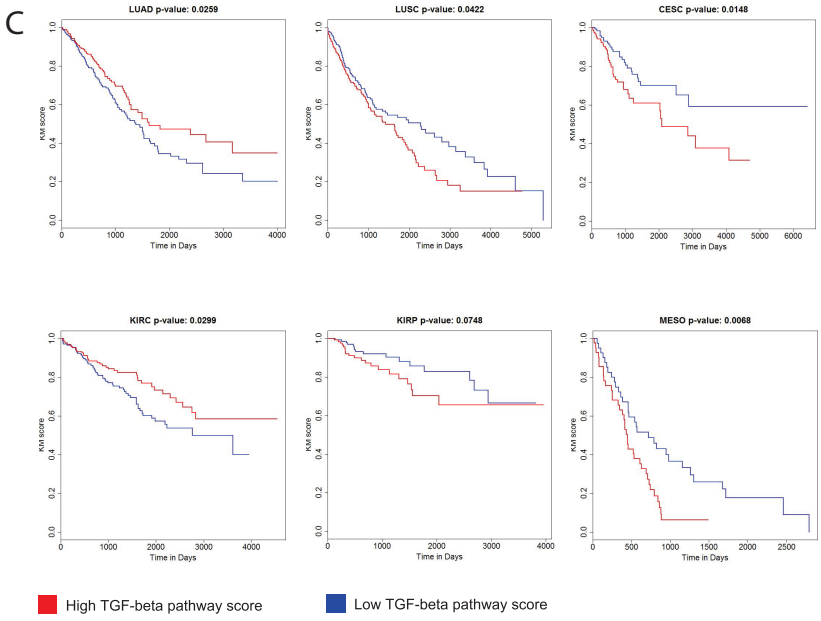


Supplementary Figure S4. Related to Figure 4. A. Pathway analysis of genomic alterations in TGF-β superfamily pathways in PanCancer and GI cancers. The topologies of the pathways are constructed through literature and expert curation. The pathways topologies capture the interactions between ligands, receptors, and intracellular proteins. The pathway diagrams are visualized using the pathway mapper software. Pathway layouts are constructed through a combination of automated optimization in the pathway mapper with default parameters and manual adjustment. The activin and TGF-β receptors activate the R-SMADs, SMAD2/3; the BMP receptors activate SMAD1/5 and SMAD8 molecules. Both parallel pathways converge on the coSMAD, SMAD4. The inhibitory SMADs act on both SMAD2/3 and SMAD1/5/8 to limit the pathway output. The frequencies of CNA and somatic mutations involving components of TGF-β pathways are quantified in boxes. Mutation frequencies are higher than CNV frequencies for virtually all of the subfamilies. Red (left box): high degree copy number amplifications, blue: deep deletions, red (right box): somatic mutations. The mutation and CNA frequencies in each GI tumor type (colorectal carcinoma, esophageal carcinoma, stomach adenocarcinoma, liver hepatocellular carcinoma, pancreatic adenocarcinoma) are quantified and mapped to the literature curated pathway diagrams. The non-altered (> 3%) parts are excluded from the analysis to visualize the sub pathways with substantially recurrent aberrations. Mutations are more common than CNA. Red (left box): high degree copy number amplifications, blue: deep deletions, red (right box): somatic mutations. **B. Differential expression of IL6 in TGF-β pathway mutated vs. wild type samples in GI vs. non-GI cancers. C. Differential expression of FOS for TGF- β pathway mutated vs. wild type samples in GI and non-GI cancers.**

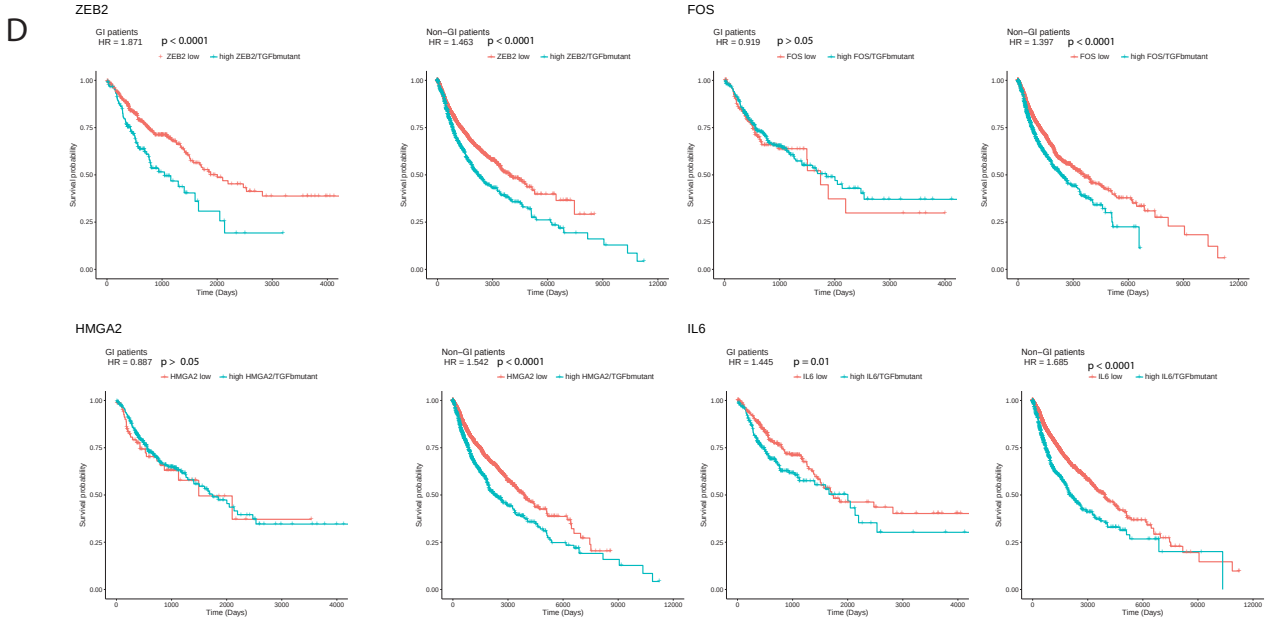


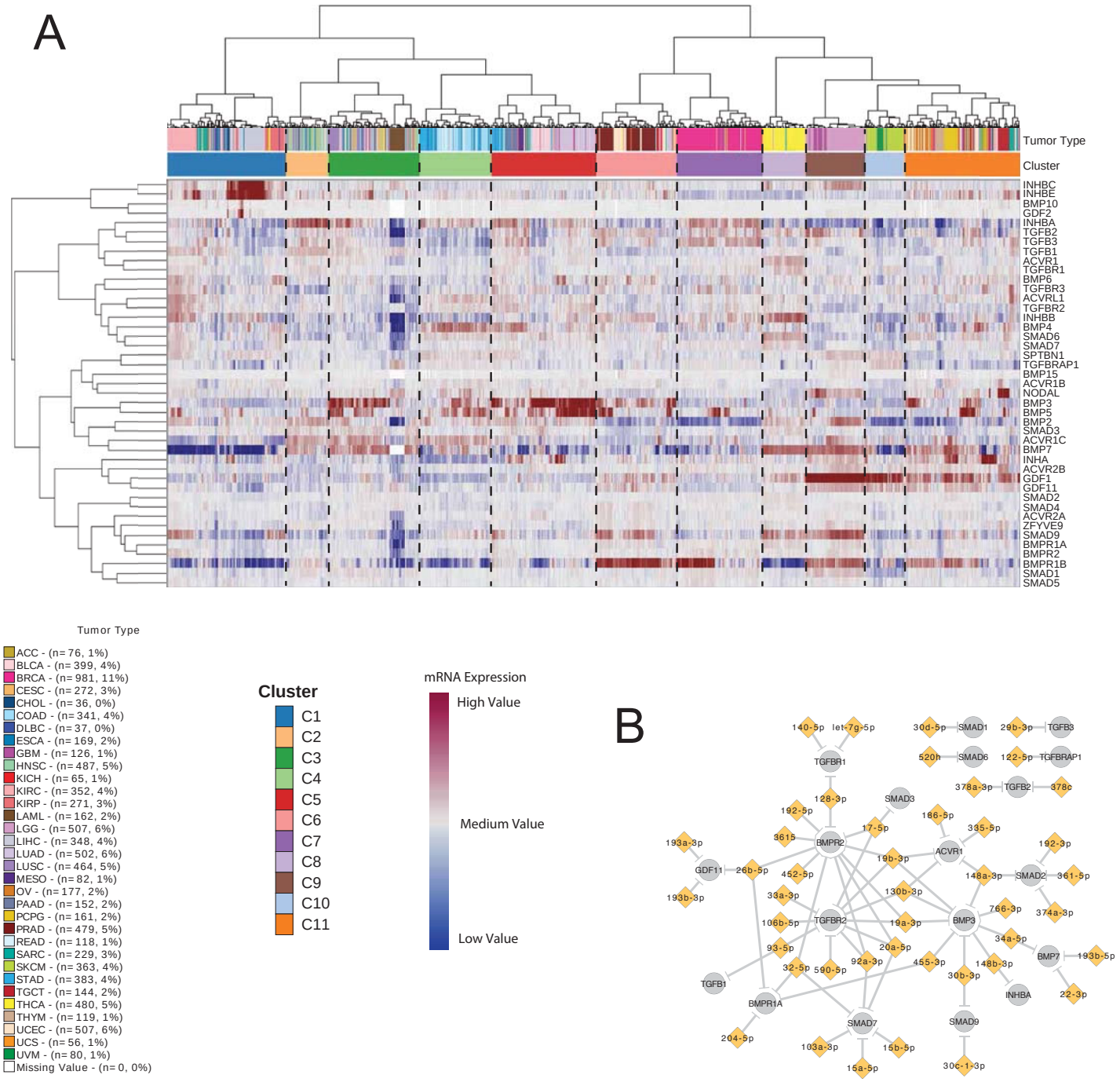
Supplementary Figure S5. Related to Figure 5. Distribution of median absolute value correlations of the TGF-β superfamily pathway score vs. different gene sets using bootstrapping over 10,000 iterations. (Blue) 43 TGF-β superfamily pathway genes, (green) 50 downstream target genes, (red) randomly selected 50 genes from the data set of 20,310 genes. The lines show medians of the densities. Downstream targets have much higher absolute correlations than randomly selected genes, as do the core genes, providing evidence for validation of the TGF-β pathway scores. The p-values are also shown for each peak.





Supplementary Figure S6. Related to Figures 5 and 6. Expression of select TGF-β pathway genes and patient survival. A. The box/violin plots showing the mRNA expression distribution of TGF-β target genes whose expression values are associated with TGF-β pathway mutations. **B.** The impact of TERT, CDH2 and FOXP3 expression in association with TGF-β pathway mutations on patient survival. **C.** Overall survival curves for KIRC, LUSC, MESO, CESC and TGCT showing differences in survival between TGF-β pathway high activity vs. low activity samples. **D.** Overall survival curves for the genes that are differentially regulated by TGF-β pathway mutations in GI vs. non-GI cancers.





Supplementary Figure S7. Related to Figures 5 and 7. A. Unsupervised mRNA clusters. Clustered heat map of the mRNA data using the 43 TGF-β pathway genes across 33 cancer types. The clustered heat maps are generated using a Pearson correlation-based distance metric and ward linkage. Eleven clusters can be seen. The bars above the heat map visualize the distribution over cancer types. **B.** Pan-GI miRNA topology. Inferred miR-mRNA targeting for 43 TGF-β pathway genes for 1507 Pan- GI primary tumor samples (from COAD, READ, STAD, ESCA, PAAD, and LIHC tumor types). Spearman correlations (R < -0.25, FDR < 10-6) shown were supported by miRTarBase v6.0 functional records with stronger and/or weaker evidence types.

Supplementary Table S2. Related to Figures 1 and 2. Genomic alterations in each of the 43 core TGF-β superfamily pathway genes. The columns show numbers or percentages of samples altered using either mutations, CNV events, or both. The numbers for types of mutation alterations (missense, in- frame deletions, truncating) are also shown.

GENE	#CASES (MUT)	%CASES (MUT)	#CASES (CNV)	% CASES (CNV)	#CASES (MUT+ CNV)	% CASES (MUT+ CNV)	#MIS MUT	#TRUN MUT	#INFR MUT
ACVR1	86	0.90%	58	0.60%	143	1.60%	79	7	0
ACVR1B	123	1.30%	56	0.60%	179	2%	105	22	0
ACVR1C	90	1%	55	0.60%	143	1.60%	80	13	0
ACVR2A	205	2.20%	68	0.70%	267	2.90%	82	124	6
ACVR2B	80	0.90%	45	0.50%	125	1.40%	72	8	0
ACVRL1	102	1.10%	53	0.60%	155	1.70%	89	13	0
BMP10	92	1%	36	0.40%	128	1.40%	83	11	0
BMP15	96	1.10%	104	1.10%	200	2.20%	87	11	0
BMP2	81	0.90%	58	0.60%	139	1.50%	66	18	0
BMP3	109	1.20%	42	0.50%	151	1.70%	97	13	0
BMP4	87	1%	40	0.40%	127	1.40%	82	5	1
BMP5	160	1.80%	107	1.20%	264	2.90%	140	22	0
BMP6	92	1%	116	1.30%	208	2.30%	86	6	1
BMP7	97	1.10%	185	2%	281	3%	90	7	0
BMPR1A	81	0.90%	91	1%	170	1.90%	59	25	0
BMPR1B	82	0.90%	45	0.50%	127	1.40%	65	17	0
BMPR2	176	1.90%	72	0.80%	242	2.70%	126	66	0
GDF1	10	0.10%	0	0%	10	0.10%	10	0	0
GDF11	39	0.40%	53	0.60%	92	1%	37	2	0
GDF2	125	1.40%	66	0.70%	190	2.10%	116	11	0
INHA	60	0.70%	62	0.70%	122	1.30%	52	8	0
INHBA	146	1.60%	61	0.70%	206	2.30%	131	20	1
INHBB	53	0.60%	67	0.70%	120	1.30%	51	2	0
INHBC	55	0.60%	89	1%	143	1.60%	48	6	1
INHBE	50	0.50%	87	1%	137	1.50%	46	5	0
NODAL	42	0.50%	51	0.60%	93	1%	40	2	0
SMAD1	57	0.60%	63	0.70%	118	1.30%	47	12	0
SMAD2	107	1.20%	86	0.90%	191	2.10%	74	40	0
SMAD3	109	1.20%	45	0.50%	152	1.70%	98	14	0
SMAD4	256	2.80%	159	1.70%	402	4%	185	79	6
SMAD5	51	0.60%	62	0.70%	112	1.20%	45	8	0
SMAD6	40	0.40%	32	0.40%	72	0.80%	38	3	0
SMAD7	42	0.50%	89	1%	130	1.40%	37	5	0
SMAD9	87	1%	124	1.40%	211	2.30%	78	9	0
SPTBN1	275	3%	66	0.70%	341	4%	235	53	2
TGFB1	49	0.50%	74	0.80%	122	1.30%	39	12	1
TGFB2	75	0.80%	194	2.10%	268	2.90%	61	16	0
TGFB3	52	0.60%	45	0.50%	97	1.10%	39	10	3
TGFBR1	94	1%	55	0.60%	147	1.60%	78	18	0
TGFBR2	145	1.60%	80	0.90%	225	2.50%	106	36	5
TGFBR3	139	1.50%	62	0.70%	198	2.20%	120	18	1
TGFBRAP1	141	1.50%	29	0.30%	168	1.80%	124	16	1
ZFYVE9	167	1.80%	56	0.60%	220	2.40%	147	26	0

Supplementary Table S3. Related to Figure 2. All 33 TCGA tumor types (column 1) with abbreviation (column 2) showing the number (column 3) and percentage (column 4) of samples with aberrations in at least one of the 43 TGF-β superfamily pathway genes. The number (column 3) and percentage (column 4) of wild-type samples without any aberrations in those 43 genes are also shown, along with the total number of samples (column 5). Overall, 39% of the samples from a total of 9,125 samples had an aberration. The lowest alteration percentage among all tumor types is 4.0% as observed in thyroid cancer.

CANCER TYPE	STUDY NAME	#ALTERED CASES	% ALTERATIONS (MUT+CNV)	# SAMPLES
Skin Cutaneous Melanoma	SKCM	253	69.7%	363
Colon adenocarcinoma	COAD	222	65.1%	341
Esophageal carcinoma	ESCA	109	64.5%	169
Stomach adenocarcinoma	STAD	243	63.4%	383
Uterine Corpus Endometrial Carcinoma	UCEC	295	58.2%	507
Bladder Urothelial Carcinoma	BLCA	229	57.4%	399
Pancreatic adenocarcinoma	PAAD	87	57.2%	152
Lung squamous cell carcinoma	LUSC	264	56.9%	464
Lung adenocarcinoma	LUAD	284	56.6%	502
Ovarian serous cystadenocarcinoma	OV	95	53.7%	177
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	19	51.4%	37
Rectum adenocarcinoma	READ	59	50.0%	118
Sarcoma	SARC	114	49.8%	229
Head and Neck squamous cell carcinoma	HNSC	221	45.4%	487
Uterine Carcinosarcoma	UCS	24	42.9%	56
Liver hepatocellular carcinoma	LIHC	149	42.8%	348
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	105	38.6%	272
Breast invasive carcinoma	BRCA	367	37.4%	981
Cholangiocarcinoma	CHOL	13	36.1%	36
Prostate adenocarcinoma	PRAD	149	31.1%	479
Adrenocortical carcinoma	ACC	17	22.4%	76
Kidney renal clear cell carcinoma	KIRC	77	21.9%	352
Glioblastoma multiforme	GBM	27	21.4%	126
Brain Lower Grade Glioma	LGG	94	18.5%	507
Kidney renal papillary cell carcinoma	KIRP	50	18.5%	271
Mesothelioma	MESO	15	18.3%	82
Pheochromocytoma & Paraganglioma	PCPG	25	15.5%	161
Thymoma	THYM	18	15.1%	119
Acute Myeloid Leukemia	LAML	19	11.7%	162
Uveal Melanoma	UVM	9	11.3%	80
Testicular Germ Cell Tumors	TGCT	13	9%	144
Kidney Chromophobe	KICH	4	6.2%	65
Thyroid carcinoma	THCA	19	4%	480
PanCan (Total)		3686	40.4%	9125

Supplementary Table S5. Related to Figures 5 and S7. Contingency table showing percentage of samples in each tumor type vs. mRNA cluster shown in Fig. S7A.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
ACC	6.6	0.0	0.0	0.0	5.3	1.3	0.0	0.0	0.0	0.0	86.8
BLCA	12.0	6.8	13.8	1.0	55.4	3.0	1.3	0.5	0.0	1.0	5.3
BRCA	2.5	0.6	0.6	0.1	1.4	1.2	87.3	0.1	0.0	0.2	5.9
CESC	3.7	15.1	49.6	1.5	13.6	4.4	1.8	0.0	0.0	0.4	9.9
CHOL	86.1	0.0	0.0	0.0	8.3	2.8	0.0	0.0	0.0	0.0	2.8
COAD	0.0	0.0	0.9	95.3	3.5	0.0	0.3	0.0	0.0	0.0	0.0
DLBC	78.4	0.0	13.5	0.0	0.0	0.0	5.4	0.0	0.0	0.0	2.7
ESCA	2.4	18.9	30.8	37.9	7.1	0.6	1.8	0.6	0.0	0.0	0.0
GBM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	96.0	0.0	4.0
HNSC	2.1	61.2	33.1	0.6	0.8	0.0	0.2	0.0	0.0	0.0	2.1
KICH	13.8	0.0	0.0	0.0	1.5	72.3	0.0	0.0	0.0	1.5	10.8
KIRC	90.6	0.0	0.0	0.6	2.0	2.3	0.3	0.0	0.0	0.0	4.3
KIRP	92.3	0.4	1.1	0.4	3.7	0.4	0.0	0.4	0.0	0.4	1.1
LAML	1.2	0.0	98.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LGG	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	99.6	0.2	0.0
LIHC	98.6	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.3	0.9
LUAD	2.2	0.4	4.8	1.2	81.7	4.6	1.0	0.0	0.0	0.2	4.0
LUSC	2.6	8.6	71.6	0.2	5.2	3.7	1.5	0.0	0.0	0.9	5.8
MESO	4.9	1.2	1.2	0.0	75.6	0.0	2.4	0.0	0.0	0.0	14.6
OV	0.0	0.6	1.1	0.0	0.6	10.7	2.3	0.0	0.0	0.0	84.7
PAAD	3.3	0.7	0.0	4.6	87.5	0.0	2.6	0.0	0.0	0.0	1.3
PCPG	0.0	0.0	0.6	0.0	1.2	4.3	0.0	0.0	0.0	1.9	91.9
PRAD	0.0	0.0	0.0	0.0	0.2	98.7	0.0	0.0	0.0	0.0	1.0
READ	0.0	0.0	0.8	95.8	3.4	0.0	0.0	0.0	0.0	0.0	0.0
SARC	42.4	1.7	2.2	0.9	6.1	1.7	0.4	0.0	0.0	2.2	42.4
SKCM	6.1	0.0	0.0	0.0	1.1	0.6	2.5	0.0	0.0	87.9	1.9
STAD	1.0	0.5	1.6	58.7	34.2	0.3	0.8	1.3	0.0	0.8	0.8
TGCT	0.0	0.0	0.0	0.7	0.0	0.7	0.0	0.0	0.7	0.0	97.9
THCA	3.3	0.0	1.0	0.0	0.0	0.0	0.0	95.0	0.0	0.0	0.6
THYM	3.4	0.0	3.4	0.0	0.0	2.5	0.8	4.2	0.0	1.7	84.0
UCEC	1.2	0.0	3.4	0.2	2.6	43.0	0.6	0.2	0.2	0.8	47.9
UCS	0.0	0.0	0.0	0.0	1.8	1.8	1.8	0.0	0.0	0.0	94.6
UVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0

Supplementary Table S6. Related to Figures 6 and 7. The association of TGF-β pathway activity with other key pathways and biological processes and median DNA methylation beta values across the 43 genes of TGF-β pathway and all the samples in each tumor type.

Tumor Type	R values for Cancer types		Median DNA Methylation Beta
	Positively correlated with EMT pathway (t-test p<0.001)	Negatively correlated with Cell Cycle Pathway (t-test p<0.001)	
ACC			0.069
BLCA	0.24	-0.2	0.077
BRCA	0.49	-0.25	0.099
CESC	0.44		0.086
CHOL			0.098
COAD	0.46	-0.29	0.103
DLBC			0.228
ESCA	0.49		0.085
GBM			0.071
HNSC	0.49	-0.25	0.093
KICH			0.075
KIRC	0.28		0.087
KIRP	0.25		0.071
LAML			0.084
LGG			0.09
LIHC	0.38		0.09
LUAD	0.33	-0.19	0.121
LUSC	0.44		0.084
MESO			0.082
OV			0.057
PAAD	0.49		0.09
PCPG	0.37		0.065
PRAD	0.59		0.07
READ	0.46	-0.35	0.086
SARC			0.073
SKCM	0.37	-0.26	0.093
STAD	0.49	-0.37	0.121
TGCT	0.56	-0.42	0.08
THCA	0.28	-0.25	0.072
THYM		-0.48	0.082
UCEC	0.39		0.076
UCS			0.062
UVM			0.067