Machine learning identifies activation of RUNX/AP-1 as drivers of mesenchymal and fibrotic regulatory programs in gastric cancer

Milad Razavi-Mohseni, Weitai Huang, Yu Amanda Guo, et al.

Genome Res. published online May 22, 2024
Access the most recent version at doi:10.1101/gr.278565.123

Machine learning identifies activation of RUNX/AP-1 as drivers of mesenchymal and fibrotic regulatory programs in gastric cancer

Milad Razavi-Mohseni[1], Weitai Huang[2], Yu A Guo[2], Dustin Shigaki[1], Shamaine Wei Ting Ho[3], Patrick Tan[3,4,5,6], Anders J Skanderup*[2], Michael A Beer*[1]

[1] Department of Biomedical Engineering and McKusick-Nathans Department of Genetic Medicine, Johns Hopkins University, Baltimore MD, USA

[2] Laboratory of Computational Cancer Genomics, Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), Singapore

[3] Laboratory of Cancer Epigenetic Regulation, Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), Singapore

[4] Cancer and Stem Cell Biology Program, Duke-NUS Medical School, Singapore

[5] Cancer Science Institute of Singapore, National University of Singapore, Singapore

[6] Department of Physiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[*] AJS and MAB are co-corresponding authors

Abstract

Gastric cancer (GC) is the fifth most common cancer worldwide and is a heterogeneous disease. Among GC subtypes, the mesenchymal phenotype (Mes-like) is more invasive than the epithelial phenotype (Epi-like). While gene expression of the epithelial-to-mesenchymal transition (EMT) has been studied, the regulatory landscape shaping this process is not fully

understood. Here we use ATAC-seq and RNA-seq from a compendium of gastric cancer cell lines and primary tumors to detect drivers of regulatory state changes and their transcriptional responses. Using the ATAC-seq, we developed a machine learning approach to determine the transcription factors (TFs) regulating the subtypes of GC. We identified TFs driving the mesenchymal (RUNX2, ZEB1, SNAI2, AP-1 dimer) as well as the epithelial states (GATA4, GATA6, KLF5, HNF4A, FOXA2, GRHL2) in gastric cancer. We identified DNA copy number alterations associated with dysregulation of these TFs, specifically deletion of *GATA4* and amplification of *MAPK9*. Comparisons with bulk and single-cell RNA-seq datasets identified activation toward fibroblast-like epigenomic and expression signatures in Mes-like GC. The activation of this mesenchymal fibrotic program is associated with differentially accessible DNA *cis*-regulatory elements flanking upregulated mesenchymal genes. These findings establish a map of TF activity in GC and highlight the role of copy number driven alterations in shaping epigenomic regulatory programs as potential drivers of gastric cancer heterogeneity and progression.

Introduction

Gastric cancer (GC) is the fifth most common cancer worldwide and is more prevalent in Asia as well as among Asian Americans in the USA (Ajani et al. 2017; Taylor et al. 2015; Lui et al. 2014). Despite advances in early diagnosis and treatments, the overall 5-year survival rate is still about 30% (Badgwell et al. 2017). Gastric cancer is a heterogeneous disease. A landmark TCGA study proposed four molecular subtypes of GC, including Epstein–Barr Virus (EBV)-associated, Microsatellite Instable (MSI), Genomically Stable (GS), and Chromosomal Instability (CIN) (Sohn et al. 2017). Histological subtyping using the Lauren Classification is also associated with therapeutic recommendations (Berlth et al. 2014).

To determine personalized therapeutic responses to different treatments, previous studies have demonstrated how treatment responses correlate with the gene expression profiles of GC (Tan et al. 2011) as well as the activity and mutation status of frequently-mutated genes such as *CDH1* (E-cadherin), *TP53*, and *ARID1A* (Park et al. 2016; Luo et al. 2018; Xu et al. 2023). Among such classifications are the mesenchymal (Mes) and epithelial (Epi) phenotypes, where the former confers more invasive disease (Oh et al. 2018). The process by which epithelial cells gain a mesenchymal phenotype is called the epithelial to mesenchymal transition (EMT) (Brabletz et al. 2018). Expression profiles of Mes GC and their correlation with response to therapeutic agents such as 5-Fluorouracil and PI3K inhibitors have been investigated (Tan et al. 2011; Oh et al. 2018; Lei et al. 2013). In addition, the activity of transcription factors (TFs) regulating normal stomach and GC (GATA4, GATA6, KLF5, HNF4A, and EBF1) have been implicated in GC invasiveness using TF and histone ChIP-seq (Chia et al. 2015; Sheng et al. 2021; Xing et al. 2020). In mesenchymal (Mes) GC, TEAD1 was identified as a major TF regulator (Ho et al. 2023) using genome-wide chromatin accessibility assays (ATAC-seq). However, there is currently a lack of systematic and unbiased approaches to uncover the complete range of TFs and epigenomic alterations underlying GC heterogeneity.

Here, we expand upon this earlier work by generating the first systematic map of gastric cancer TF activity derived from an unbiased machine learning approach applied to ATAC-seq datasets from a broad compendium of GC cell lines and primary TCGA tumors (Corces et al. 2018; Bass et al. 2014), coupled with RNA-seq analysis for interpretation of the altered regulatory pathways. Our analysis aims to infer the complete set of TFs whose activity explains the heterogeneity of chromatin landscapes across GC and connect TF activity through enhancers to transcriptional activation of specific gene responses. We further investigate mutational and copy number alterations as genomic aberrations underlying the dysregulation of these key TFs. Finally, we

relate GC cell line subtypes to other tissues using models trained on accessibility and single-cell transcriptional profiles.

Results

Sequence-based machine learning model of ATAC-seq profiles identify distinct regulatory programs driving GC heterogeneity

To understand differences in epigenomic state profiles in gastric-cancer derived cell lines, we generated ATAC-seq data from a collection of 25 cell lines (Ho et al. 2023). Principal Component Analysis (PCA) of variation in the activity of distal peaks (Methods) showed significant heterogeneity in chromatin accessibility across the cell lines, forming three clusters (Fig. 1A) with similar genomic accessibility profiles. We used a machine learning approach, gkm-SVM (Ghandi et al. 2014a; Beer et al. 2020; Lee et al. 2015), to identify the sequence features (transcription factor binding sites, TFBS) driving these variations in chromatin accessibility. We chose gkm-SVM for its combination of predictive accuracy and interpretability (Ghandi et al. 2014a; Beer et al. 2020; Lee et al. 2015; Shigaki et al. 2019; Ghandi et al. 2016; Beer 2017; Yan et al. 2021; Kreimer et al. 2017; Lee 2016; Gate et al. 2018; Mo et al. 2016), and because of its advantages relative to DNNs when training on smaller numbers of differentially active peaks (~2000). For each cell line, we trained a gkm-SVM model on the top 10k distal ATAC-seq peaks against random genomic sequences, following our standard pipeline (Beer et al. 2020). On each training, following the default training method in the gkmSVM-R package (Ghandi et al. 2016), the AUROC was evaluated on 5-fold cross validation (CV) and the reported AUROC is the average of 5 CV test sets. The median cross-validation AUROC is

0.925. All cross-fold validation sets produce similar sets of features and have very similar weight vectors. Because this test set AUROC is high, there is little evidence of overfitting. Also, following (Ghandi et al. 2016), for feature interpretation, but not for quantitative assessment, we use the gapped *k*-mer weight vector when training on all the data to reduce variability in the inferred features. Each gkm-SVM model summarizes the TFBS required to classify the ATAC-seq peaks in a vector of gapped *k*-mer weights which can predict the accessibility of the sequences. We clustered the gkm-SVM weight vector from each cell line (Fig. 1B) and found nearly identical clusters to Fig. 1A, thus the similar genomic accessibility profiles are explained by activation of shared TFs.

Based on these differentially active TFs and their resultant differentially expressed target genes, which largely align with EMT and mesenchymal-related pathways (see RNA-seq analysis below), we label these three sets of GC cell lines as Epithelial-like (Epi-like), Intermediate, and Mesenchymal-like (Mes-like). To isolate the TFs driving most of the variability between these sets of cell lines, we next trained gkm-SVM on the most differentially accessible peaks when averaged over the Epi-like and Mes-like cell lines, with Mes-like peaks as the positive set, and Epi-like peaks as the negative set (Supplemental Fig. S1, S2). Again, this generated a highly predictive sequence model and the gkm-SVM weights with the largest difference between the Epi-like and Mes-like cell lines were associated with differential activity of TFs binding the specific motifs: RUNX, AP-1 (FOS/JUN), TEAD, ZEB, HNF4A, FOXA, GRHL, GATA, and KLF (Fig. 1C, Supplemental Fig. S3). Since ZEB1 and ZEB2 are known transcriptional repressors (Aghdassi et al. 2012; Byles et al. 2012), ZEB activity is indicated by a negative motif score. To assess if these TFs also drive variation in tumors, we next trained gkm-SVM on differential peaks in normal stomach vs. primary stomach tumor samples (TCGA-STAD, The Cancer Genome Atlas for Stomach Adenocarcinoma) (Fig. 1C, Supplemental Fig. S4, S5). The TFs detected largely overlap with those in the cell lines, indicating that these TFs also drive variation

in the regulatory state between normal stomach tissue and GC tumors. While there is significant heterogeneity among the tumor samples, and only a fraction of the cells in a tumor are likely in a mesenchymal state, we find significant evidence that the Mes-like regulatory program identified in the cell lines is activated in the STAD samples relative to normal stomach. 70% of the high quality TCGA-STAD ATAC-seq samples detect RUNX, and 70% detect AP-1 (Supplemental Fig. S5, Methods). Finally, we trained gkm-SVM on differentially accessible peaks in each pair of cell lines (300 pairs, see Methods) to detect the full set of TFs explaining the ATAC-seq variation across the cell line collection. RUNX and AP-1 (and lack of ZEB) were strong TFBS signals driving accessibility in Mes-like cell lines, while GRHL, GATA, and KLF TFBS were strongly associated with most Epi-like samples (Fig. 1D, 1E). In summary, this ATAC-seq analysis of open chromatin identifies RUNX, AP-1, ZEB, GRHL, GATA, and KLF as novel regulatory drivers of the Epi-like and Mes-like transcriptional states in gastric cancer. These predictions from gkm-SVM are validated by experiments showing strong RUNX2 and JUN (which binds AP-1) binding at Mes-like enhancers from ChIP-seq in LPS141 (Bevill et al. 2023) and strong KLF5, GATA4, and GATA6 binding at Epi-like enhancers from ChIP-seq generated in AGS (Chia et al. 2015; Liu et al. 2020) (Fig. 1F, Supplemental Fig. S6).

Regulatory and expression signatures of EMT and advanced-vs-early GC explain the differences between GC groups

We used RNA-seq data from the 25 GC cell lines to understand how the differential TF activity inferred from the ATAC-seq data affects gene expression. The three groups of GC cell lines, previously clustered based on their chromatin accessibility, were recapitulated using the expression of tissue-specific genes (Fig. 2A, Supplemental Fig. S7, Methods), demonstrating that the variation in RNA-seq is concordant with that of ATAC-seq. The only exception is SNU484, where the ATAC-seq and RNA-seq profiles are less consistent than they are for every

other cell line. Because we can use gkm-SVM AUROC to assess the quality of the ATAC-seq data, we are more confident of SNU484 ATAC-seq since the AUROC is 0.963. Because we don't have an independent metric to assess the quality of the RNA-seq, we weighted the ATAC-seq more heavily for SNU484 and labelled it Epi-like. For all other samples the ATAC-seq and RNA-seq based cluster labels are consistent.

We next compare gene expression in GC cell lines with that of primary stomach tissue from TCGA-STAD and TCGA normal adjacent tissue (Methods). The combined PCA with cell lines and TCGA shows that there are inherent differences between the stomach-derived cell lines and stomach primary tissue on the first principal component (PC1, Fig. 2B), as noted in previous studies (Aran et al. 2015; Yu et al. 2019). PC2 shows TCGA-Normal samples are more aligned with Epi-like GC cell lines, while TCGA-STAD cancer samples are closer to Mes-like cell lines (Supplemental Fig. S8).

To associate these expression differences with biological processes, we performed gene set enrichment analysis using PCA gene weights and GSEA (Mootha et al. 2003; Subramanian et al. 2005; Liberzon et al. 2011, 2015). The first principal component was strongly associated with mesenchymal and EMT phenotypes, advanced-vs-early gastric cancer and cancer invasiveness (Fig. 2C, hypergeometric FDR < 10^{-40}) (Liberzon et al. 2015; Charafe-Jauffret et al. 2006; Schuetz et al. 2006; Vecchi et al. 2007). EMT is also the top hit when using expression changes between Mes-like and Epi-like clusters (Supplemental Fig. S9, S10) instead of PC1 weights. We discovered a similar pattern of enrichment comparing TCGA-STAD and normal tissues with the cell lines (Fig. 2D, which is gene set enrichment analysis on PC2 of Fig. 2B). Taken together, this demonstrates that the biological differences between the Mes-like and Epi-like clusters of GC cell lines overlap with the EMT and invasiveness phenotypes. Because of these associations, we label the red group of cell lines in Fig. 1A and Fig. 2A "*Mes-like*" for their

mesenchymal expression signature (EMT); the green group "*Epi-like*" for their more normal epithelial expression profile; and the gray group of cell lines is labeled "*Intermediate*". The Intermediate set lies along the continuum between Epi-like and Mes-like, but is distinguishable from the Mes-like set through the stronger activity of ZEB and GRHL motifs, and distinguishable from the Epi-set by the lack of GATA activity, as shown in (Fig. 1E). Our ATAC-seq analysis allows us to identify TF regulators with these EMT and invasiveness gene expression signatures.

TF regulators detected in ATAC-seq are markers of EMT and invasiveness and are differentially expressed in Mes-vs-Epi GC cell lines

To further explore the connection between regulatory states and cancer invasiveness, we generated a set of 1770 differentially expressed (DE) genes between Mes-like and Epi-like cell lines (Methods). These DE genes include the regulator TFs implicated from ATAC-seq, and many genes previously implicated in epithelial GC or EMT (Supplemental Table S1 and Supplemental Table S2). We observed that TFs such as *GATA4, GATA6, KLF5, HNF4A, GRHL2, FOXA1, FOXA2*, and *FOXA3* in the Epi-like cell lines, as well as *SNAI2, ZEB1, FOSL1* (AP-1), and *RUNX2* in Mes-like, not only have a significantly higher expression (Fig. 2E, 2F, Supplemental Table S1) but also have differentially accessible DNA-binding sites in Mes-like vs. Epi-like GC cell line ATAC-seq (Fig. 1C). Further, we assessed the expression of these DE TFs in an independent patient cohort from the Asian Cancer Research Group (ACRG) (Cristescu et al. 2015) (Fig. 2G, Methods), as well as in TCGA-STAD (Supplemental Fig. S11), and found that differential TF expression is associated with overall survival, suggesting that activation of these regulator TFs has both biological and disease prognostic utility.

Many of the differentially expressed genes have been previously implicated in cancer. In the Epi-like upregulated group, *KLF5*, and *HNF4A* along with *GATA4* and *GATA6* have been shown to promote GC (Chia et al. 2015). *GRHL2* plays an anti-EMT role in GC (Sheng et al. 2021; Luo et al. 2019) through TGFB signaling, whereas, *GATA6* suppresses EMT by promoting the expression of *FOXA2* and *CDH1* while decreasing the activity of *SNAI2, TWIST1, ZEB1* and *VIM* in pancreatic cancer (Martinelli et al. 2017). TFs in the Mes-like upregulated group, *ZEB1*, *RUNX2*, and *SNAI2* are known regulators of EMT (Brabletz et al. 2018; Niu et al. 2012; Roche 2018). Additionally, *FOSL1* dimerizes with *JUN* family members to bind the AP-1 motif and drives EMT by upregulating *ZEB1*/*2* and *SNAI2* (Bakiri et al. 2015; Guneri-Sozeri et al. 2023). The combined differential expression of these TF genes and the identification of their TF binding sites as the largest signals in predictive sequence modeling of the chromatin accessibility profiles strongly suggest that the variability in activation of this small set of TFs is responsible for driving the EMT program. Although this finding is in the context of GC, the fact that LPS141 arrived at a similar transcriptional state and has RUNX2 and JUN binding at the Mes-like enhancers suggests that this may be a more general mechanism of activating EMT across other tumor types (Supplemental Fig. S6).

Mutations and copy number alterations are associated with distinct regulatory states in GC

We next sought to identify possible mechanisms driving the variation in regulatory state across the GC cell lines. To understand how the activity of TFs in Mes-like and Epi-like cell lines correlated with known genetic variation in gastric cancer, we examined TF gene expression in TCGA-STAD patients. *CDH1*, a recurrently mutated GC driver gene (Supplemental Table S3) (Luo et al. 2018; Xu et al. 2023; Lei et al. 2013; Fodde 2002), was downregulated in Mes-like

GC cell lines compared to Epi-like. We examined the expression of DE TFs in *CDH1*-mutated vs. *CDH1*-WT TCGA-STAD patients (Fig. 3A, Supplemental Fig. S12, Methods). *CDH1*-mutated TCGA-STAD tumors showed upregulation of the Mes-like TFs. Conversely, the *CDH1*-WT tumors displayed upregulation of Epi-like TFs (Fig. 3A). This result indicates that *CDH1* mutations in GC correlate with upregulation of key EMT-promoting TFs and downregulation of EMT-inhibiting TFs.

In addition, the TFs *GATA4, GATA6, SMAD4, KLF5*, and *HNF4A*, are among the TFs with the highest rates of DNA Copy Number (CN) alterations in TCGA-STAD patients (Fig. 3B, Supplemental Fig. S13). *GATA4, GATA6*, and *SMAD4* are known to be important in endoderm and stomach development (Li et al. 2019), and *SMAD4* has been identified as a stomach cancer driver gene (Martínez-Jiménez et al. 2020). *GATA4* and *GATA6* have lower expression in Mes-like GC cell lines (Fig. 2F), consistent with gkm-SVM analysis (Fig. 1E). ATAC-seq profiles for Mes-like, Intermediate, and Epi-like GC cell lines also identify two distal peaks (*GATA6_E1* and *GATA6_E2*) in the *GATA6* locus (Fig. 3C) whose accessibility correlates with *GATA6* expression (Supplemental Fig. S14, S15). *GATA6_E1* is upstream of *GATA6* and bound by GATA6 in AGS, and *GATA6_E2* is intronic and is also bound by GATA6 in AGS, and by SMAD4 and ZEB1 in HepG2 cells. The low ATAC signal in Mes-like cell lines at *GATA6_E2* is consistent with the upregulation of the repressor *ZEB1* and downregulation of *GATA6*. In further support of their regulatory role, both *GATA6_E1* and *GATA6_E2* are contained in loop extrusion (LE) model predicted CTCF loops containing the *GATA6* promoter (Xi and Beer 2021; Luo et al. 2023). Overall, the up-regulation of the epithelial regulator *GATA6* is consistent with the Mes-like and Epi-like ATAC-seq profiles through *GATA6_E1* and *GATA6_E2*.

Reduced GATA and increased AP-1 regulatory activity in GC cell lines are accompanied by DNA copy number alterations

The gkm-SVM inferred differential GATA activity can be attributed to either GATA6 or GATA4 dysregulation, and both *GATA4* and *GATA6* have lower expression in Mes-like GC (Fig. 2F). To identify potential driving mechanisms, we compared CN changes in Epi-like cell lines vs. non-Epi cell lines (Mes and Intermediate), using DNA sequencing (Methods). On average, *GATA4* has a higher copy number in Epi-like cell lines (Fig. 3D), which correlates with lower expression of *GATA4* in Mes-like cell lines (Fig. 3E, r = 0.56). While less than 1% of TCGA-STAD patients have a mutation in *GATA4* according to cBioPortal (Gao et al. 2013), *GATA4* CN alterations are more frequent in TCGA-STAD (Fig. 3B). Further, an interval of Chr8p containing *GATA4* is frequently deleted in human epithelial cancers (Cai and Sablina 2016). Thus, *GATA* CN alterations are a potential mechanism for *GATA4* expression changes (Fig. 2F) and GATA activity in gastric cancer cell lines and primary tumors (Fig. 1C). We observed an opposite pattern for *MAPK9* kinase which is amplified in the Mes-like cell lines compared to Epi-like (Fig. 3D, 3E). MAPK and JNK family members are known to be regulators of AP-1 (Karin et al. 1997), and *MAPK9* amplification is associated with higher expression of AP-1 complex members (*FOSL1*, Fig. 2F) and higher AP-1 transcriptional activity in Mes-like cell lines and primary stomach cancer (Fig. 1C). Taken together, we see evidence for DNA copy number changes in transcription factors and signaling pathway components as potential drivers of transcriptional dysregulation in gastric cancer cell lines.

Mes-like GC cell lines have expression and regulatory signatures similar to fibroblasts, while Epi-like GC cell lines retain gastrointestinal signatures

Because we detected the activation of non-stomach TFs among many GC cell lines, we next sought to characterize their cell states by comparing to known cell types. We first compared the GC cell line expression profiles to known normal cell types by calculating their correlation with Genotype-Tissue Expression (GTEx) RNA-seq data (Lonsdale et al. 2013) (Methods, Fig.4A). Mes-like GC cell line expression was more similar to fibroblasts (r=0.41) than to stomach (r=0.29), while Epi-like and Intermediate were most similar to stomach and esophagus mucosa, respectively (Fig. 4A). We further confirmed that the gene expression in Mes-like GC cell lines is more similar to fibroblast-derived cells than stomach in two additional independent RNA datasets from ENCODE and TCGA-STAD (Methods, Fig. 4B) (The ENCODE Consortium 2012; Luo et al. 2020). While TCGA-Normal is more similar to normal stomach than TCGA-STAD, the similarity of TCGA-STAD to fibroblast is weaker than the Mes-like cell lines (Fig. 4B).

We extended these comparisons to DNase-seq data and found consistent similarities between the chromatin accessibility of the Mes-like GC cell lines with ENCODE DNase-seq of fibroblast (Fig. 4C). We trained gkm-SVM sequence models on open chromatin regions for each GC cell line, TCGA-STAD, ENCODE fibroblast and stomach tissues, and compared their weight vectors (Methods). The chromatin accessibility of Mes-like GC is highly correlated with the fibroblast-derived cell lines HT1080 and ELR, as well as primary fibroblast tissue from ENCODE, as opposed to primary stomach. In contrast, the Epi-like cell lines and TCGA-STAD ATAC-seq were more correlated with the primary stomach.

In addition to these average correlations of expression and chromatin accessibility, we now show that differential TF activity among GC cell lines positions them along a continuum of cell states between normal stomach and fibroblast. Scatter plots of correlation to stomach and fibroblast (Fig. 4D, 4E, Supplemental Fig. S16, S17) show that TCGA-Normal had the highest correlation with ENCODE normal stomach and the lowest with fibroblast, while TCGA-STAD tumors have a higher correlation with fibroblast and a lower correlation with stomach (Fig. 4D). The Mes-like GC cell lines have a higher similarity to fibroblast than the primary tumors, correlations being as high as 0.8. Consistent with the RNA-seq profiles, gkm-SVM models trained on chromatin accessibility profiles also show that Mes-like cell line accessibility is more similar to fibroblast accessibility than normal stomach tissue (Fig. 4E, Supplemental Fig. S18, S19). While it is clear that GC cell lines are becoming more similar to fibroblasts by activation of fibroblast TFs, the bulk TCGA-STAD profiles may be moving toward a fibroblast cell state either by interaction with stroma through fibrosis or by the presence of cancer-associated fibroblasts (CAF) in the tumor microenvironment (Kalluri and Zeisberg 2006; Piersma et al. 2020).

We next addressed cell line similarity at the single cell level using stomach cancer scRNA-seq data from Kim et al. (Kim et al. 2022) (Methods). We identified 6 distinct populations of cells using UMAP (McInnes et al. 2020) (Fig. 5A, Supplemental Fig. S20), and based on marker gene expression (Methods, Supplemental Table S4), we will identify clusters 2 and 5 as gastrointestinal and cluster 6 as fibroblasts. We compared the expression of GC cell lines to these single-cell clusters and found high similarity between stomach clusters 2 and 5 with Epi-like GC cell lines (Fig. 5B) and high similarity between Mes-like GC cell lines and fibroblast cluster 6 (Fig. 5C). Expression of stomach and fibroblast specific genes from GTEx RNA-seq confirms our association of these clusters with these cell types (Fig. 5D, 5E, Methods).

Fibrotic gene expression in GC cell lines is driven by activation of flanking enhancers

While we observed similarity of ATAC-seq and RNA-seq profiles across Mes-like and Epi-like cell lines (Fig. 1A, 2A), we next show that their differential expression is consistent with direct induction by enhancers flanking Mes-like and Epi-like genes. We calculated the average ATAC-seq signal in all distal ATAC-seq peaks within 50kb of a Transcription Start Site (TSS) of differentially expressed genes between Epi-like or Mes-like GC cell lines. The accessibility of peaks flanking genes more highly expressed in Epi-like cell lines is 1.763 times higher in Epi-like cell lines compared to Mes-like, and the accessibility of peaks flanking genes more highly expressed in Mes-like cell lines is 1.756 times higher in Mes-like cell lines compared to Epi-like (Fig. 6A, 6B), providing a direct connection between altered chromatin state and expression responses.

Among the genes upregulated in Mes-like cell lines are the fibrotic gene *FGF5* and many collagen genes downstream of the TGFB/SMAD pathway (Shin et al. 2019; Xie et al. 2020; Verrecchia et al. 2001). Their role in gastrointestinal cancer (Luo et al. 2019) and cancer-associated fibroblasts (Bordignon et al. 2019) has been well-established. Inhibition of FGF5 decreases proliferation and metastasis in hepatocellular carcinoma (Fang et al. 2015). COL1A1 has a similar effect on tumor behavior and proliferation (Nissen et al. 2019). *FGF5*, *COL1A2*, *COL6A3*, *COL5A1*, *COL12A1*, *COL1A1*, and *COL6A2* are upregulated in Mes-like GC and have 1.5-fold to 6-fold higher accessibility in flanking ATAC peaks. The members of the fibrotic pathways *SMAD3*, *SMAD4*, *TGFB2*, *COL1A1*, *FGF1*, *FGF5*, and *FGF7* are differentially expressed (Fig. 6C, Mann–Whitney *U* test $p < 0.05$) in Mes-like and Epi-like GC.

To understand the regulatory mechanism for the differential expression of *COL1A1* and *FGF5*, we compared the chromatin accessibility of their flanking *cis*-regulatory regions (Fig. 6D, 6E). For *COL1A1*, we found upstream distal peaks *COL1A1_E1* and *COL1A1_E2* differentially active in Mes-like and Epi-like GC cell lines and in primary TCGA-STAD tumors compared to normal stomach (Fig. 6D). They are both located within a loop-extrusion (LE) model predicted CTCF loop (Xi and Beer 2021). Similarly, *FGF5* is flanked by an upstream distal peak (*FGF5_E*) (Fig. 6E), which is active in primary TCGA-STAD cancer and inactive in healthy stomach tissue. Across cell lines, the expression of *COL1A1* is highly correlated with the activity of peaks *COL1A1_E1* and *COL1A1_E2* (r=0.64 and 0.58, Supplemental Fig. S21, S22). Similarly, the relatively higher expression of *FGF5* in 3 of the 5 Mes-like cell lines is correlated with the activity of the distal peak (*FGF5_E*) (r=0.74, Supplemental Fig. S23). *COL1A1_E1, COL1A1_E2*, and *FGF5_E* are all bound by RUNX2 and JUN in LPS141 (Bevill et al. 2023). Therefore, *COL1A1* and *FGF5* are dysregulated in both the Mes-like GC cell lines and in stomach cancer tissue, consistent with increased chromatin accessibility of newly identified flanking distal peaks which are likely acting as enhancers.

Mes-like GC cell lines have a distinct regulatory landscape compared to fibroblasts

Since the Mes-like cell lines are epigenomically and transcriptionally similar to fibroblasts, we next sought to verify that they are not simply normal fibroblasts, or possibly cancer-associated fibroblasts (CAF) (Sahai et al. 2020). Normal fibroblasts would be distinguishable by lower mutation or CN alteration rates. However, we found that the Epi-like, Intermediate, and Mes-like GC cell lines had indistinguishable mutation rates in the coding regions (Fig. 7A, *t*-test $p > 0.12$, Methods). Additionally, we calculated the average rate of CN alterations in each cell line (Fig. 7B, Methods), and the three groups of cell lines had similar rates. While GES1 is derived from

normal stomach epithelium, it was transformed with SV40 virus containing the T antigens, which downregulate TP53 and appear to have mimicked the high mutation burden of the other cell lines, which were not transformed with SV40 and do not have any ATAC-seq reads mapping to SV40. Because of these mutation and copy number profiles, the Mes-like cell lines are clearly distinct from normal fibroblasts.

The Mes-like cell lines are also distinct from normal fibroblasts in their chromatin accessibility profiles. PCA analysis on distal open-chromatin regions in GC cell lines, normal fibroblast and stomach ENCODE DNase-seq, showed differences between each group (Fig. 7C, Methods, analysis similar to Fig. 1A). The variation between Epi-like, Intermediate, and Mes-like GC cell lines is aligned with an axis that also explains most of the variance between stomach and fibroblast accessibility profiles. In addition, gkm-SVM sequence models trained on differentially-accessible ATAC distal peaks between Mes-like GC and normal fibroblast highlight that TFs binding AP-1, RUNX, GATA, RARA, and KLF motifs have higher activity in Mes-like GC, while fibroblasts have higher accessibility explained by TWIST and ZEB DNA-binding sites (Fig. 7D). By aggregating motif activity scores following pairwise comparisons between each Mes-like cell line and each ENCODE fibroblast, we rediscovered AP-1 and RUNX to be active in Mes-like GC as opposed to TWIST and ZEB being more active in normal fibroblasts (Fig. 7E).

Although the expression profiles of Mes-like GC cell lines are similar to fibroblasts, there are also systematic expression differences consistent with the above noted chromatin accessibility changes. TFs with higher expression levels in normal stomach such as *KLF5, FOXA1, HNF4A*, and *POU5F1* were among the TFs having a higher expression in Mes-like GC compared to normal ENCODE fibroblast (Fig. 7F, Methods, FDR < 0.01, $|\log_2 FC| > 2$). The differential activity of KLF, TWIST, and AP-1 discussed in the motif analysis above is consistent with the upregulation of *KLF5* and *FOSL1* in Mes-like GC, compared to the higher expression of *TWIST2*

386 in normal ENCODE fibroblast (Fig. 7F). In addition, fibroblast marker genes such as *FN1*, *DCN*,
387 *COL6A1*, *COL1A1*, and *FGF7* have higher expression in ENCODE fibroblast compared to Mes-
388 like GC (Fig. 7G). Taken together, we see that despite similar chromatin accessibility and gene
389 expression signatures between Mes-like GC and fibroblast tissues, the epigenomic state of
390 Mes-like GC can be clearly distinguished from fibroblasts by the activity of AP-1, RUNX, and the
391 stomach motif KLF, as well as lower activity of TWIST and ZEB compared to fibroblasts.

392 Discussion

393 We provide the first systematic map of gastric cancer TF activity inferred from an unbiased
394 machine learning approach applied to ATAC-seq and RNA-seq data. Our sequence-based
395 machine learning analysis revealed that most of the variation among GC cell lines is driven by a
396 fibrotic vs. epithelial regulatory program differentially activated by a small set of TFs, most of
397 which have not previously been directly associated with EMT in gastric cancer. We used a panel
398 of gastric cancer-derived cell lines in combination with the TCGA stomach-adenocarcinoma
399 (TCGA-STAD) cohort to identify heterogeneous regulatory programs in gastric cancer (GC) and
400 their concomitant transcriptional responses. The inferred cell-line regulatory programs explained
401 much of the variation among TCGA-STAD samples. Our direct analysis of GC cell lines allowed
402 for the isolation of cancer cell regulatory programs without noise derived from intra-tumor
403 heterogeneity and variation in tumor immune infiltration.
404
405 Using ATAC-seq and computational sequence models, along with RNA-seq and scRNA-seq, we
406 identified the largely novel set of TFs involved in mesenchymal (Mes-like) GC (RUNX2, SNAI2,
407 ZEB1, AP-1 dimer) as opposed to the less invasive epithelial state (Epi-like GC) and its
408 regulators (GATA4, GATA6, KLF5, HNF4A, FOXA2, GRHL2) (Fig. 8A). Mutation and DNA copy
409 number analysis identified genetic events associated with the activation of this regulatory

program. The Mes-like transcriptional state was often associated with *GATA4* DNA deletion and *MAPK9* amplification (Fig. 8B). This suggests copy number variation in TFs and signaling components can play a significant role as cancer driver mutations in addition to previously reported enhancer hijacking resulting from genomic structural rearrangement (Wang et al. 2021).

EMT in cancer is a complex process (Lovisa 2021; Yang et al. 2020), but we found that individual GC cell lines activate the gene expression signature associated with EMT to varying degrees. DNase-seq and RNA-seq from ENCODE and GTEx provided evidence for the existence of a fibrotic phenotype in Mes-like GC through EMT (Fig. 8C & 8D). Downstream fibroblast genes such as *FGF5* and *COL1A1* were found to be upregulated in both Mes-like GC and TCGA-STAD, and we identified flanking *cis*-regulatory enhancer elements with increased activity in Mes-like GC vs. Epi-like GC as well as TCGA-STAD compared to normal stomach. Survival analysis shows that activation of Mes-like TFs and their target genes are predictive of disease severity and suggests both prognostic and therapeutic utility.

Taken together, we identified altered regulatory programs in gastric cancer along with their distinct transcriptional responses driven by differential enhancer activity. We identify copy number DNA alterations as genomic aberrations responsible for the dysregulation of these core TFs. Our findings suggest that activation of this small set of TFs driving the Mes-like GC regulatory program plays an important role in cancer progression and highlights new biology and potential therapeutic opportunities in gastric cancer.

Methods

ATAC-seq data processing and gkm-SVM training

The raw GC cell line ATAC-seq data analyzed in this study was reported in (Xu et al. 2023; Sheng et al. 2021; Xing et al. 2020; Ho et al. 2023) and uploaded to GSE264550. ATAC-seq paired reads were mapped to hg38 genome with Bowtie 2 version 2.2.5 (Langmead and Salzberg 2012). For gkm-SVM training, peaks were called by MACS2 (Zhang et al. 2008). Distal peaks were defined 300bp bins centered on the top 10k MACS2 peaks, after removing promoters (<2kb of a TSS) and peaks open in more than 30% of ENCODE Dnase-seq samples (mostly CTCF sites). Negative set genomic regions were selected randomly with matched GC and repeat content as described in (Beer et al. 2020; Shigaki et al. 2019). Training positive peaks vs. the negative sets generated robust models with a median cross-validation AUROC of 0.925 (Supplemental Fig. S24). We used default settings and either the gkmSVM-R or lsgkm packages (Ghandi et al. 2014a, 2016; Lee 2016). This training procedure was shown to generate gkm-SVM models with a strong performance in both MPRA experiments and predictions of variant impact (Yan et al. 2021; Kreimer et al. 2017; Beer 2017). For ENCODE DNase-seq normal stomach we used ENCSR782SSS (Roadmap Epigenomics Consortium et al. 2015) downloaded from encodeproject.org and for TCGA stomach tumor ATAC-seq we used TCGA-BR-A4J6 (Supplemental Fig. S25, S26) (Corces et al. 2018). ENCODE DNase-seq and TCGA STAD ATAC-seq (Corces et al. 2018; Shigaki et al. 2019) were processed in a similar manner, as described in (Shigaki et al. 2019). We chose gkm-SVM for this task because of its robust performance on small (differential accessibility) datasets and because it has been shown that gapped *k*-mers are efficient representations of TFBS (Yan et al. 2021; Ghandi et al. 2014b) and protein motifs (Amanchy et al. 2011). Each gkm-SVM model generates a score function that

can be specified as weights for each gapped *k*-mer (Ghandi et al. 2014b), and the score for each sequence is the sum of weights for all gapped *k*-mers in a sequence. These weights quantify the contribution of each gapped *k*-mer to chromatin accessibility. To map these gapped *k*-mer weights to interpretable TF activity, after training gkm-SVM models, we extracted TF models and their inferred activity using gkm-PWM, which infers the frequency of PWMs (TFBS position weight matrices) by minimizing the error between the observed and generated weight vectors. All of the 10 STAD ATAC samples with >10k distal peaks and AUROC > 0.9 detect some activation of RUNX or AP-1 when trained against normal stomach DNase-seq (ENCODE: ENCSR782SSS): 7 detect AP-1 (TCGA.CD.A48C, TCGA.VQ.A94O, TCGA.VQ.A8PJ, TCGA.BR.A4J6, TCGA.BR.A4CS, TCGA.HF.A5NB, TCGA.BR.A4J4), and 7 detect RUNX (TCGA.VQ.A94O, TCGA.VQ.A91W, TCGA.BR.A4J6, TCGA.BR.A4IY, TCGA.HF.A5NB, TCGA.BR.A4J4, TCGA.CD.A486) (Supplemental Fig. S5). For the GC cell lines, to generate a more compact set of TFBS motifs, we trained each set of positive ATAC distal peaks vs. each other (300 pairs of experiments, n=2000 differentially active sequences in each of the two cell types, median AUROC = 0.922, Supplemental Fig. S27) and used the 19 most commonly detected motifs as PWMs to extract TF activity (AP-1, RUNX, AP2, NFKB, ZEB (AREB6), NFI, ONECUT, EBOX, EHF, TCF/LEF, TEAD, GRHL, HNF4A, HNF1B, HOXB13, FOX, KLF, SOX, GATA) (Fig. 1E, Supplemental Fig. S28). Although many of these motifs are detected in other cell types (McClymont et al. 2018), the combinations are specific to endodermal lineages (Luo et al. 2023). Clustering in the inferred motif activity space generates a PCA plot with the same grouping detected in the ATAC-seq signal (Fig. 1A).

ATAC-seq analysis around differentially expressed genes

We calculated the ATAC-seq signal in peaks within 50k around the TSS of differentially expressed genes in Epi and Mes (higher in Epi, and higher in Mes). The average activity in

flanking peaks averaged over all Epi, Mes, and Intermediate cell lines is shown in Fig. 6A. Similar results were found using peaks within 100k, 150k, 200k, or 250k from TSS.

ChIP-seq data

ChIP-seq samples from ENCODE were used in the genome browser tracks: ZEB1 (ENCODE ENCSR000BVN) and SMAD4 (ENCODE ENCSR826YMT). ChIP-seq data for RUNX2 and JUN in Fig. 1F are from (Bevill et al. 2023). ChIP-seq data for KLF5 in AGS are from (Liu et al. 2020) and ChIP-seq data for GATA4 and GATA6 in AGS are from (Chia et al. 2015) with raw data uploaded to GSE51705.

RNA-seq data processing

GC cell lines RNA-seq

The raw GC cell line RNA-seq data analyzed in this study was reported in refs (Xu et al. 2023; Sheng et al. 2021; Xing et al. 2020; Ho et al. 2023) and uploaded to GSE266159, GSE157750, and GSE85465. To normalize FPKM values, we divided them by the sample's upper-quartile (75th percentile) expression value and multiplied them by the average upper-quartile values across all samples to scale (upper-quartile normalization) and then $\log_2$-transformed.

TCGA-STAD and TCGA-Normal RNA-seq

TCGA-STAD (stomach adenocarcinoma) and TCGA-Normal (normal adjacent stomach tissue) RNA-seq HTSeq counts were downloaded from The Cancer Genome Atlas (TCGA) GDC portal (https://portal.gdc.cancer.gov/) (Anders et al. 2015). 356 tumor (TCGA-STAD) and 32 normal stomach (TCGA-Normal) samples were initially obtained. TCGA-STAD and TCGA-Normal

samples with a higher correlation to GTEx esophagus than to GTEx stomach tissue RNA-seq signature were removed (with a method similar to the GTEx correlation analysis described below). The remaining 322 TCGA-STAD and 17 TCGA-Normal samples were upper-quartile normalized over the set of protein-coding genes as labeled by GENCODE V35 annotation (Frankish et al. 2019).

Chromatin accessibility analysis of Mes-like GC and normal ENCODE fibroblast and stomach

To compare chromatin accessibility profiles, we generated a union set of all distal peaks of ENCODE stomach (n=12) and primary fibroblast (n=30), and 25 GC cell lines to perform PCA. We trained gkm-SVM models on the 2k most differentially accessible peaks in all pairs of 5 Mes-like GC cell lines and 30 ENCODE primary fibroblasts and ranked non-similar motifs by their average gkm-PWM *Z*-score.

RNA-seq analysis

Tissue-specific protein-coding genes

To reduce noise, RNA-seq analyses were performed using 11312 “tissue-specific” genes, which are derived from GTEx RNA-seq profiles of 54 healthy human tissues (GTEx portal: GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz). To define tissue-specific genes, we calculated the number of GTEx tissues, where a gene has a TPM expression above the median expression in that tissue (min = 0, max = 54). If this number is 0, the gene is lowly expressed in all tissues and if this number is 54, the gene is ubiquitously expressed (e.g. ribosomal genes). Tissue-specific genes are then defined as the subset of protein-coding genes

that are neither lowly expressed nor ubiquitously expressed in all GTEx healthy human tissues. Protein-coding genes are found according to GENCODE V35 annotation.

PCA of RNA-seq

Principal Component Analysis (PCA) was performed using prcomp() and drawn with ggplot2 in R (R Core Team 2023). For the joint cell line and primary tissue PCA, GC cell line, and TCGA-STAD FPKM expression values were upper-quartile normalized together over the set of tissue-specific genes. Genes having the highest (positive or negative) weights in the PCA were obtained using R prcomp$rotation[, 1] and prcomp$rotation[, 2] for PC1 and PC2, where the top 100 genes were used for gene set analysis. We used upper-quartile normalized RNA-seq from LPS141 (Chen et al. 2019) to add this cell line to the PCA in Supplemental Fig. S6, which shows that the transcriptional state of LPS141 is very similar to the Mes-like GC cell lines.

Differentially expressed genes in Mes-like vs. Epi-like cell lines

Differentially expressed genes between the Mes-like and Epi-like GC cell lines were found using the edgeR (Robinson et al. 2010) package in R. The list of genes was limited to 11312 tissue-specific protein-coding genes, as described above. 920 genes (including 72 TFs) were found to be upregulated in Mes-like GC cell lines ($|\log_2 FC| > 2$, FDR < 0.05, Supplemental Table S1), while 850 genes (including 51 TFs) have a higher expression in Epi-like cell lines (Supplemental Table S2). The list of all 1639 human TFs was retrieved from http://humantfs.ccbr.utoronto.ca/download.php (Version 1.01) (Lambert et al. 2018).

Survival analysis in ACRG and TCGA-STAD

Gene expression values for 300 ACRG gastric cancer patients were downloaded from NCBI-GEO GSE62254. Patients were stratified into Mes-like (n=150) and Epi-like (n=150) groups

based on the median difference in the mean expression values of Mes and Epi DE TFs. Similarly, TCGA-STAD patients (n=315) were split into two groups using the median expression difference value. Plots and p-values were produced using survminer (https://github.com/kassambara/survminer), ggsurvfit (https://github.com/pharmaverse/ggsurvfit), and survival (https://github.com/therneau/survival) R packages.

Gene set analysis

Gene Set Enrichment Analysis (GSEA) software (Mootha et al. 2003; Subramanian et al. 2005) with hallmark gene sets (Liberzon et al. 2015) was used to identify the biological function of differentially expressed genes in Mes-like vs. Epi-like GC (default GSEA parameters). Additionally, MSigDB (Liberzon et al. 2011) gene sets H (hallmark), C2 (curated gene sets), C5 (ontology gene sets), and C6 (oncogenic gene sets) were used with top 100 PC$rotation (see PCA of RNA-seq in Methods) to find enriched gene sets. Hypergeometric test calculated the p-values, which were then adjusted using p.adjust(method = “fdr”) in R.

Differentially expressed genes in Mes-like GC vs. ENCODE fibroblast

RNA-seq data for normal human fibroblast was downloaded from the ENCODE portal and was upper-quartile normalized with the GC cell line RNA-seq profiles. Normalization and the DE gene analysis were performed similar to what described above (FDR < 0.01, $\log_2$FC > 2) where 982 and 5611 genes are upregulated in Mes-like GC and ENCODE fibroblast samples, respectively. 96 of the upregulated genes in the Mes-like group and 45 in the ENCODE fibroblast group are TFs. To compare the expression of these DE TFs in normal stomach and normal fibroblast, we used GTEx normal stomach and GTEx normal fibroblast to rank the

expression values of these DE TFs (between Mes vs. ENCODE-fibroblast) in the independent GTEx stomach and fibroblast expression profiles.

Copy number and mutation analysis

TCGA-STAD mutation rates were downloaded from cBioPortal (Gao et al. 2013; Cerami et al. 2012) https://www.cbioportal.org/study/summary?id=stad_tcga_pan_can_atlas_2018). Of the TCGA-STAD samples for which the RNA-seq data was available, 255 samples are microsatellite stable (MSS) and were used for further analysis, while the microsatellite instable (MSI) samples were removed because of the possibility of having more passenger mutations. The genes selected in the plot are TFs differentially expressed between Mes-like and Epi-like GC cell lines. The x-axis is the $\log_2$ fold-change of FPKM-UQ values between the 14 samples with a CDH1 nonsynonymous or splice-site mutation and the 241 samples without. Y-axis p-values were calculated using the *t*-test.

Copy number alteration rates for all human TFs in TCGA-STAD (Pan-Cancer) samples were downloaded from cBioPortal, and include copy number deletion and copy number amplification events as calculated by GISTIC (Mermel et al. 2011). GISTIC assigns an integer “score” between -2 and +2 to each gene depending on the number of DNA fragments aligned to it in the sequencing data. GISTIC “scores” “-2” and “-1” represent deep and shallow DNA deletions (possibly homozygous and heterozygous deletions), while “+2” and “+1” indicate deep and shallow DNA amplifications, respectively. A GISTIC “score” “0” for a gene in a sample means that the gene DNA copy number is not changed and is normal (i.e. the copy number is 2 in a healthy diploid human genome).

Whole genome sequencing (WGS) data of GC cell lines were processed as previously described (Xing et al. 2020). Briefly, copy number variations were identified using CNVkit with default parameters (bcbio-nextgen v0.9.3) (Talevich et al. 2016). As GC cell lines have no matched germline samples, CNVs were call against a non-matched normal sample. We filtered high-confidence copy number predictions (n=2061) by removing protein-coding genes whose expression correlation with copy number across cell lines is less than 0.5. We calculated copy number differences between the 15 Epi-like GC cell lines and the combined 5 Mes-like and 5 Intermediate cell lines, and plotted the average copy number differences for each gene in the two groups, i.e. mean(NonEpi CN) - mean(Epi CN)). The y-axis is the p-value for the difference between Epi and NonEpi copy number scores using the Mann–Whitney *U* test. We also analyzed SNVs in the GC cell lines with SmuRF (Huang et al. 2020) and found mutations in *GATA6* and *ZFPM2* ("Friend of GATA family member 2") in 4/5 of the Mes-like cell lines, but their frequency was close to the threshold for common SNPs, and after filtering, the p-values were not significant.

To compare the non-synonymous mutation burden across groups of GC cell lines, performed variant calling on the WGS of GC cell lines using SMuRF (Huang et al. 2020). As no matched germline is available for these cell lines, we filtered out common population variants from the dbSNP database (AF>1%). Then we calculated the number of protein-altering mutations (missense, truncating, start and stop codon gain/loss, and splice region variants) in each cell line (Supplemental Table S5).

To calculate the copy number burden in GC cell lines, we converted the CNVkit output to a GISTIC-like integer score (ranging from -2 to +2, described above). For each cell line, the absolute values of gene CN scores are averaged to report the cell line’s overall CN status.

RNA-seq and ATAC-seq correlations

Correlation analysis using GTEx tissues RNA-seq

GTEx RNA-seq TPM values for 54 healthy human tissues were retrieved as described above. To compare GC cell line RNA-seq with GTEx, the cell line's FPKM values were converted to TPM. To preserve space in the figures, the TPM expression values for all GTEx brain tissues were aggregated (averaged). Mean expression for each tissue-specific gene (described above) for the three cell line groups (Mes, Epi, Intermediate) was calculated and this resulted in a 11312 * 3 TPM expression matrix for the cell lines. Then Pearson's correlation was calculated between the GC cell line groups and GTEx tissues, using TPM expression values.

Correlation heatmap with ENCODE fibroblast and stomach RNA-seq

The following fibroblast and stomach RNA-seq profiles were downloaded from the ENCODE portal: HT1080 (ENCFF754UAP), bronchus fibroblast of lung (ENCFF716LRF), fibroblast skin of abdomen (ENCFF010QUB), fibroblast skin of scalp (ENCFF385POO), stomach 3-yr child (ENCFF299YCQ), stomach 37-yr adult (ENCFF683JSC), stomach 34-yr adult (ENCFF547FBP). Upper-quartile normalized FPKM values of ENCODE samples, GC cell lines, and TCGA-STAD were then compared in a heatmap (Pearson's correlation). The 11312 tissue-specific genes (described above) were used to calculate the correlation. In the scatter plot, correlation values for each individual sample are shown.

Correlation analysis with ENCODE fibroblast and stomach DNase-seq

The following fibroblast and stomach DNase-seq profiles were downloaded from the ENCODE portal and processed as described above: cardiac fibroblast (ENCSR000ENI), lung fibroblast (ENCSR000EPR), ELR fibroblast cell line (ENCSR240TPI), HT1080 fibrosarcoma cell line

(ENCSR000FDI), stomach 34-yr adult (ENCSR782SSS), stomach 54-yr adult (ENCSR163PKT), stomach 3-yr child (ENCSR246PXX). gkm-SVM models were trained on ENCODE DNase-seq, GC cell line ATAC-seq, and TCGA-STAD ATAC-seq profiles, as described above. The gkm-SVM output for each sample is a weight vector corresponding to $4^{11}/2$ *k*-mers (A/T/C/G 11-mers), where the weight value and its sign indicate how overrepresented or underrepresented the *k*-mer (TF binding site) is in the chromatin accessibility data (ATAC or DNase-seq). To find the similarities between the chromatin accessibility profiles, a heatmap (Pearson's correlation) was drawn using the ENCODE, cell lines, and TCGA-STAD samples. To calculate the correlation for each group (i.e Mes-like, Epi-like, TCGA-STAD), their mean gkm-SVM weight vectors were used for each group of samples. In the scatter plot, correlation values for each individual sample are shown. In Fig. 4D and Fig. 4E, lung fibroblast (ENCSR000EPR) and adult stomach (ENCSR782SSS) were used.

scRNA-seq data processing and analysis

scRNA-seq raw counts for 24 patients (gastric tumor and normal adjacent tissue) were downloaded from Kim et al. study (Kim et al. 2022) on NCBI GEO (GSE150290). Cells having 1) less than 500 detected transcripts, 2) more than 20000 transcripts, 3) less than 500 detected genes, or 4) more than 10% mitochondrial or hemoglobin genes were filtered out for being low-quality. This resulted in an expression profile of a total of about 117,000 cells derived from tumor and normal adjacent tissue of 24 GC patients. Log-normalization was performed using Seurat (Hao et al. 2021) and UMAP was used for dimensionality reduction. Clusters were defined by running a Gaussian Mixture Model (GMM) over the UMAP space using the R uwot package (n_components = 2, n_neighbors = 30) (McInnes et al. 2020). To find the correlation and similarities between scRNA-seq populations (clusters) and GC cell line groups, the correlation of scRNA and bulk RNA was calculated for each cell. The average RNA-seq

expression for the Epi-like or Mes-like groups of cell lines over their differentially expressed (DE) genes (1770 genes) was calculated. Then for each cell, the Pearson's correlation of the scRNA-seq expression values (for 1770 DE genes) and the average bulk RNA-seq expression of Mes-like and Epi-like cell lines was calculated.

Kim et al. (Kim et al. 2022) discovered several cell populations including immune cells, fibroblasts, and various gastric cell types. We saw a similar pattern of marker genes in our analysis where clusters 1, 3, and 4 express immune-related markers such as *CD74, CD83*, *PTPRC* (*CD45*), *IL32*, and various MHC/HLA class II variants. Clusters 2 and 5 express gastrointestinal gene markers such as *TFF1, TFF2, TFF3, PGC*, and *REG4* similar to what was discovered in (Kim et al. 2022), whereas cluster 6 expresses fibrotic markers including *COL1A1, COL1A2, COL3A1, COL6A1, COL6A2, PDGFRA, FN1*, and *MMP2*. This indicates that single-cell clusters 2 and 5 include stomach cells, whereas cluster 6 is comprised of fibroblasts. To define a more systematic list of highly-expressed gene markers for fibroblast (or stomach) tissue, we used the same GTEx 54 healthy tissue RNA-seq profiles as described above. We sorted the TPM expression values of 11312 tissue-specific genes (described above) in GTEx fibroblast (or stomach). We used the top 100 highly-expressed genes as the gene markers for fibroblast (or stomach). To find the biological interpretation of each single cell cluster, we used these gene markers. For every single cell, the average expression over the top 100 genes (fibroblast or stomach) was calculated and the boxplot (Fig. 5D, 5E) is drawn based on the single cell clusters, where clusters 2 and 5 were combined as they were both similarly correlated with the same group of GC cell lines. The average expression values across each cluster (in the boxplots) are compared using *t*-test, where $p < 10^{-8}$. Thus, gene expression in Epi-like cell lines is highly correlated to that of stomach cells, unlike Mes-like GC cell lines which have a high correlation with the fibroblast single-cell population.

Loop Extrusion Model Predictions

CTCF loop predictions shown in Fig. 3C, 6D, and 6E use Loop Extrusion model predictions (Xi and Beer 2021) using CTCF ChIP-seq in endoderm (Luo et al. 2023). The loop extrusion model uses orientation and binding strength information from ChIP-seq to predict loop probability, is consistent with CTCF ChIA-PET loop measurements, and generalizes better than more complicated CTCF loop prediction methods (Xi and Beer 2018).

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE264550 and GSE266159. Source files for differential accessibility and expression analysis are included as Supplemental_Code.zip and both code and gkm-SVM models are available at http://beerlab.org/gccl.

Competing interest statement

PT has stock and other ownership interests in Tempus Healthcare, previous research funding from Kyowa Hakko Kirin and Thermo Fisher Scientific, and patents/other intellectual property through the Agency for Science and Technology Research, Singapore (all outside the submitted work). All remaining authors have declared no conflicts of interest.

Acknowledgments

This work was supported by the following grants: MRM, DS, and MAB were supported by National Institutes of Health – National Human Genome Research Institute grant R01 HG012367 to MAB. WH, YAG, SWTH, AJS and PT supported by the National Medical Research Council grant MOH-000967-00 (MOH-STaR21jun-0001) and by the National Research Foundation, Singapore, and Singapore Ministry of Health's National Medical Research Council under its Open Fund-Large Collaborative Grant ("OF-LCG") (MOH-OFLCG18May-0003). The cell line GES1 was a gift from Alfred Cheng from the Chinese University of Hong Kong. We thank Salvador Casaní-Galdón and Brad Bernstein for providing the ChIP-seq bigwigs from (Bevill et al. 2023). We thank Roger Sik Yin Foo and Chukwuemeka George Anene-Nzelu for assisting with ATAC-seq library preparation and sequencing.

Author contributions: Most analysis performed and designed by MRM and MAB, with contributions from AJS, WH, YAG, and DS. Overall study design conceived by MAB and AJS. All authors provided editorial and intellectual input.

References

Aghdassi A, Sendler M, Guenther A, Mayerle J, Behn C-O, Heidecke C-D, Friess H, Büchler M, Evert M, Lerch MM, et al. 2012. Recruitment of histone deacetylases HDAC1 and HDAC2 by the transcriptional repressor ZEB1 downregulates E-cadherin expression in pancreatic cancer. *Gut* **61**: 439–448.

Ajani JA, Lee J, Sano T, Janjigian YY, Fan D, Song S. 2017. Gastric adenocarcinoma. *Nat Rev Dis Primer* **3**: 1–19.

Amanchy R, Kandasamy K, Mathivanan S, Periaswamy B, Reddy R, Yoon W-H, Joore J, Beer MA, Cope L, Pandey A. 2011. Identification of Novel Phosphorylation Motifs Through an Integrative Computational and Experimental Analysis of the Human Phosphoproteome. *J Proteomics Bioinform* **4**: 22–35.

Anders S, Pyl PT, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166–169.

Aran D, Sirota M, Butte AJ. 2015. Systematic pan-cancer analysis of tumour purity. *Nat Commun* **6**: 8971.

Badgwell B, Das P, Ajani J. 2017. Treatment of localized gastric and gastroesophageal adenocarcinoma: the role of accurate staging and preoperative therapy. *J Hematol OncolJ Hematol Oncol* **10**: 149.

Bakiri L, Macho-Maschler S, Custic I, Niemiec J, Guío-Carrión A, Hasenfuss SC, Eger A, Müller M, Beug H, Wagner EF. 2015. Fra-1/AP-1 induces EMT in mammary epithelial cells by modulating Zeb1/2 and TGFβ expression. *Cell Death Differ* **22**: 336–350.

Bass AJ, Thorsson V, Shmulevich I, Reynolds SM, Miller M, Bernard B, Hinoue T, Laird PW, Curtis C, Shen H, et al. 2014. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**: 202–209.

Beer MA. 2017. Predicting enhancer activity and variant impact using gkm-SVM. *Hum Mutat* **38**: 1251–1258.

Beer MA, Shigaki D, Huangfu D. 2020. Enhancer Predictions and Genome-Wide Regulatory Circuits. *Annu Rev Genomics Hum Genet* **21**: 37–54.

Berlth F, Bollschweiler E, Drebber U, Hoelscher AH, Moenig S. 2014. Pathohistological classification systems in gastric cancer: Diagnostic relevance and prognostic value. *World J Gastroenterol WJG* **20**: 5679–5684.

Bevill SM, Casaní-Galdón S, El Farran CA, Cytrynbaum EG, Macias KA, Oldeman SE, Oliveira KJ, Moore MM, Hegazi E, Adriaens C, et al. 2023. Impact of supraphysiologic MDM2 expression on chromatin networks and therapeutic responses in sarcoma. *Cell Genomics* **3**: 100321.

758 Bordignon P, Bottoni G, Xu X, Popescu AS, Truan Z, Guenova E, Kofler L, Jafari P, Ostano P,
759 Röcken M, et al. 2019. Dualism of FGF and TGF-β Signaling in Heterogeneous Cancer-
760 Associated Fibroblast Activation with ETV1 as a Critical Determinant. *Cell Rep* **28**: 2358-
761 2372.e6.

762 Brabletz T, Kalluri R, Nieto MA, Weinberg RA. 2018. EMT in cancer. *Nat Rev Cancer* **18**: 128–
763 134.

764 Byles V, Zhu L, Lovaas JD, Chmilewski LK, Wang J, Faller DV, Dai Y. 2012. SIRT1 induces
765 EMT by cooperating with EMT transcription factors and enhances prostate cancer cell
766 migration and metastasis. *Oncogene* **31**: 4619–4629.

767 Cai Y, Sablina AA. 2016. Cancer-associated chromosomal deletions: Size makes a difference.
768 *Cell Cycle* **15**: 2850–2851.

769 Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer
770 ML, Larsson E, et al. 2012. The cBio Cancer Genomics Portal: An Open Platform for
771 Exploring Multidimensional Cancer Genomics Data. *Cancer Discov* **2**: 401–404.

772 Charafe-Jauffret E, Ginestier C, Monville F, Finetti P, Adélaïde J, Cervera N, Fekairi S, Xerri L,
773 Jacquemier J, Birnbaum D, et al. 2006. Gene expression profiling of breast cell lines
774 identifies potential new basal markers. *Oncogene* **25**: 2273–2284.

775 Chen Y, Xu L, Mayakonda A, Huang M-L, Kanojia D, Tan TZ, Dakle P, Lin RY-T, Ke X-Y, Said
776 JW, et al. 2019. Bromodomain and extraterminal proteins foster the core transcriptional
777 regulatory programs and confer vulnerability in liposarcoma. *Nat Commun* **10**: 1353.

778 Chia N-Y, Deng N, Das K, Huang D, Hu L, Zhu Y, Lim KH, Lee M-H, Wu J, Sam XX, et al. 2015.
779 Regulatory crosstalk between lineage-survival oncogenes KLF5, GATA4 and GATA6
780 cooperatively promotes gastric cancer development. *Gut* **64**: 707–719.

781 Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, Silva TC, Groeneveld C,
782 Wong CK, Cho SW, et al. 2018. The chromatin accessibility landscape of primary human
783 cancers. *Science* **362**. https://science.sciencemag.org/content/362/6413/eaav1898.

784 Cristescu R, Lee J, Nebozhyn M, Kim K-M, Ting JC, Wong SS, Liu J, Yue YG, Wang J, Yu K, et
785 al. 2015. Molecular analysis of gastric cancer identifies subtypes associated with distinct
786 clinical outcomes. *Nat Med* **21**: 449–456.

787 Fang F, Chang R, Yu L, Lei X, Xiao S, Yang H, Yang L-Y. 2015. MicroRNA-188-5p suppresses
788 tumor cell proliferation and metastasis by directly targeting FGF5 in hepatocellular
789 carcinoma. *J Hepatol* **63**: 874–885.

790 Fodde R. 2002. The *APC* gene in colorectal cancer. *Eur J Cancer* **38**: 867–871.

791 Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C,
792 Wright J, Armstrong J, et al. 2019. GENCODE reference annotation for the human and
793 mouse genomes. *Nucleic Acids Res* **47**: D766–D773.

794 Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R,
795 Larsson E, et al. 2013. Integrative Analysis of Complex Cancer Genomics and Clinical
796 Profiles Using the cBioPortal. *Sci Signal* **6**: pl1–pl1.

797 Gate RE, Cheng CS, Aiden AP, Siba A, Tabaka M, Lituiev D, Machol I, Gordon MG,
798 Subramaniam M, Shamim M, et al. 2018. Genetic determinants of co-accessible
799 chromatin regions in activated T cells across humans. *Nat Genet* **50**: 1140.

800 Ghandi M, Lee D, Mohammad-Noori M, Beer MA. 2014a. Enhanced Regulatory Sequence
801 Prediction Using Gapped k-mer Features. *PLoS Comput Biol* **10**: e1003711.

802 Ghandi M, Mohammad-Noori M, Beer MA. 2014b. Robust k-mer frequency estimation using
803 gapped k-mers. *J Math Biol* **69**: 469–500.

804 Ghandi M, Mohammad-Noori M, Ghareghani N, Lee D, Garraway L, Beer MA. 2016. gkmSVM:
805 an R package for gapped-kmer SVM. *Bioinformatics* **32**: 2205–2207.

806 Guneri-Sozeri PY, Özden-Yılmaz G, Kisim A, Cakiroglu E, Eray A, Uzuner H, Karakülah G,
807 Pesen-Okvur D, Senturk S, Erkek-Ozhan S. 2023. FLI1 and FRA1 transcription factors
808 drive the transcriptional regulatory networks characterizing muscle invasive bladder
809 cancer. *Commun Biol* **6**: 1–17.

810 Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C,
811 Zager M, et al. 2021. Integrated analysis of multimodal single-cell data. *Cell* **184**: 3573-
812 3587.e29.

813 Ho SWT, Sheng T, Xing M, Ooi WF, Xu C, Sundar R, Huang KK, Li Z, Kumar V, Ramnarayanan
814 K. 2023. Regulatory enhancer profiling of mesenchymal-type gastric cancer reveals
815 subtype-specific epigenomic landscapes and targetable vulnerabilities. *Gut* **72**: 226–241.

816 Huang W, Guo YA, Chang MM, Skanderup AJ. 2020. Ensemble-Based Somatic Mutation
817 Calling in Cancer Genomes. In *Bioinformatics for Cancer Immunotherapy: Methods and*
818 *Protocols* (ed. S. Boegel), *Methods in Molecular Biology*, pp. 37–46, Springer US, New
819 York, NY https://doi.org/10.1007/978-1-0716-0327-7_3.

820 Kalluri R, Zeisberg M. 2006. Fibroblasts in cancer. *Nat Rev Cancer* **6**: 392–401.

821 Karin M, Liu Z, Zandi E. 1997. AP-1 function and regulation. *Curr Opin Cell Biol* **9**: 240–246.

822 Kim J, Park C, Kim KH, Kim EH, Kim H, Woo JK, Seong JK, Nam KT, Lee YC, Cho SY. 2022.
823 Single-cell analysis of gastric pre-cancerous and cancer lesions reveals cell lineage
824 diversity and intratumoral heterogeneity. *Npj Precis Oncol* **6**: 1–11.

825 Kreimer A, Zeng H, Edwards MD, Guo Y, Tian K, Shin S, Welch R, Wainberg M, Mohan R,
826 Sinnott-Armstrong NA, et al. 2017. Predicting gene expression in massively parallel
827 reporter assays: A comparative study. *Hum Mutat* **38**: 1240–1250.

828 Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR,
829 Weirauch MT. 2018. The Human Transcription Factors. *Cell* **172**: 650–665.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: nmeth.1923.

Lee D. 2016. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **32**: 2196–2198.

Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, Beer MA. 2015. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet* **47**: 955–961.

Lei Z, Tan IB, Das K, Deng N, Zouridis H, Pattison S, Chua C, Feng Z, Guan YK, Ooi CH, et al. 2013. Identification of Molecular Subtypes of Gastric Cancer With Different Responses to PI3-Kinase Inhibitors and 5-Fluorouracil. *Gastroenterology* **145**: 554–565.

Li QV, Dixon G, Verma N, Rosen BP, Gordillo M, Luo R, Xu C, Wang Q, Soh C-L, Yang D, et al. 2019. Genome-scale screens identify JNK–JUN signaling as a barrier for pluripotency exit and endoderm differentiation. *Nat Genet* **51**: 999–1010.

Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. 2015. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst* **1**: 417–425.

Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**: 1739–1740.

Liu Y, Guo B, Aguilera-Jimenez E, Chu VS, Zhou J, Wu Z, Francis JM, Yang X, Choi PS, Bailey SD, et al. 2020. Chromatin Looping Shapes KLF5-Dependent Transcriptional Programs in Human Epithelial Cancers. *Cancer Res* **80**: 5464–5477.

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, et al. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**: 580–585.

Lovisa S. 2021. Epithelial-to-Mesenchymal Transition in Fibrosis: Concepts and Targeting Strategies. *Front Pharmacol* **12**. https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2021.737570/full.

Lui FH, Tuan B, Swenson SL, Wong RJ. 2014. Ethnic Disparities in Gastric Cancer Incidence and Survival in the USA: An Updated Analysis of 1992–2009 SEER Data. *Dig Dis Sci* **59**: 3027–3034.

Luo J, Chen X-Q, Li P. 2019. The Role of TGF-β and Its Receptors in Gastrointestinal Cancers. *Transl Oncol* **12**: 475–484.

Luo R, Yan J, Oh JW, Xi W, Shigaki D, Wong W, Cho HS, Murphy D, Cutler R, Rosen BP, et al. 2023. Dynamic network-guided CRISPRi screen identifies CTCF-loop-constrained nonlinear enhancer gene regulatory activity during cell state transitions. *Nat Genet* **55**: 1336–1346.

Luo W, Fedda F, Lynch P, Tan D. 2018. CDH1 Gene and Hereditary Diffuse Gastric Cancer Syndrome: Molecular and Histological Alterations and Implications for Diagnosis And Treatment. *Front Pharmacol* **9**. https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2018.01421.

Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, Myers Z, Sud P, Jou J, Lin K, et al. 2020. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res* **48**: D882–D889.

Martinelli P, Pau ECS, Cox T, Sainz B, Dusetti N, Greenhalf W, Rinaldi L, Costello E, Ghaneh P, Malats N, et al. 2017. GATA6 regulates EMT and tumour dissemination, and is a marker of response to adjuvant chemotherapy in pancreatic cancer. *Gut* **66**: 1665–1676.

Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, Mularoni L, Pich O, Bonet J, Kranas H, et al. 2020. A compendium of mutational cancer driver genes. *Nat Rev Cancer* **20**: 555–572.

McClymont SA, Hook PW, Soto AI, Reed X, Law WD, Kerans SJ, Waite EL, Briceno NJ, Thole JF, Heckman MG, et al. 2018. Parkinson-Associated SNCA Enhancer Variants Revealed by Open Chromatin in Mouse Dopamine Neurons. *Am J Hum Genet* **103**: 874–892.

McInnes L, Healy J, Melville J. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. http://arxiv.org/abs/1802.03426.

Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, 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.

Mo A, Luo C, Davis FP, Mukamel EA, Henry GL, Nery JR, Urich MA, Picard S, Lister R, Eddy SR, et al. 2016. Epigenomic landscapes of retinal rods and cones. *eLife* **5**: e11613.

Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, et al. 2003. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**: 267–273.

Nissen NI, Karsdal M, Willumsen N. 2019. Collagens and Cancer associated fibroblasts in the reactive stroma and its relation to Cancer biology. *J Exp Clin Cancer Res* **38**: 115.

Niu D-F, Kondo T, Nakazawa T, Oishi N, Kawasaki T, Mochizuki K, Yamane T, Katoh R. 2012. Transcription factor Runx2 is a regulator of epithelial–mesenchymal transition and invasion in thyroid carcinomas. *Lab Invest* **92**: 1181–1190.

Oh SC, Sohn BH, Cheong J-H, Kim S-B, Lee JE, Park KC, Lee SH, Park J-L, Park Y-Y, Lee H-S, et al. 2018. Clinical and genomic landscape of gastric cancer with a mesenchymal phenotype. *Nat Commun* **9**: 1777.

Park S, Lee J, Kim YH, Park J, Shin J-W, Nam S. 2016. Clinical Relevance and Molecular Phenotypes in Gastric Cancer, of TP53 Mutations and Gene Expressions, in Combination With Other Gene Mutations. *Sci Rep* **6**: 34822.

Piersma B, Hayward M-K, Weaver VM. 2020. Fibrosis and cancer: A strained relationship. *Biochim Biophys Acta BBA - Rev Cancer* **1873**: 188356.

R Core Team. 2023. R: A Language and Environment for Statistical Computing. https://www.R-project.org.

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317–330.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140.

Roche J. 2018. The Epithelial-to-Mesenchymal Transition in Cancer. *Cancers* **10**: 52.

Sahai E, Astsaturov I, Cukierman E, DeNardo DG, Egeblad M, Evans RM, Fearon D, Greten FR, Hingorani SR, Hunter T, et al. 2020. A framework for advancing our understanding of cancer-associated fibroblasts. *Nat Rev Cancer* **20**: 174–186.

Schuetz CS, Bonin M, Clare SE, Nieselt K, Sotlar K, Walter M, Fehm T, Solomayer E, Riess O, Wallwiener D, et al. 2006. Progression-Specific Genes Identified by Expression Profiling of Matched Ductal Carcinomas In situ and Invasive Breast Tumors, Combining Laser Capture Microdissection and Oligonucleotide Microarray Analysis. *Cancer Res* **66**: 5278–5286.

Sheng T, Ho SWT, Ooi WF, Xu C, Xing M, Padmanabhan N, Huang KK, Ma L, Ray M, Guo YA, et al. 2021. Integrative epigenomic and high-throughput functional enhancer profiling reveals determinants of enhancer heterogeneity in gastric cancer. *Genome Med* **13**: 158.

Shigaki D, Adato O, Adhikar AN, Dong S, Hawkins-Hooker A, Inoue F, Juven-Gershon T, Kenlay H, Martin B, Patra A, et al. 2019. Integration of Multiple Epigenomic Marks Improves Prediction of Variant Impact in Saturation Mutagenesis Reporter Assay. *Hum Mutat* **40**: 1280–1291.

Shin JY, Beckett JD, Bagirzadeh R, Creamer TJ, Shah AA, McMahan Z, Paik JJ, Sampedro MM, MacFarlane EG, Beer MA, et al. 2019. Epigenetic activation and memory at a TGFB2 enhancer in systemic sclerosis. *Sci Transl Med* **11**: eaaw0790.

Sohn BH, Hwang J-E, Jang H-J, Lee H-S, Oh SC, Shim J-J, Lee K-W, Kim EH, Yim SY, Lee SH, et al. 2017. Clinical Significance of Four Molecular Subtypes of Gastric Cancer Identified by The Cancer Genome Atlas Project. *Clin Cancer Res* **23**: 4441–4449.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* **102**: 15545–15550.

Talevich E, Shain AH, Botton T, Bastian BC. 2016. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLOS Comput Biol* **12**: e1004873.

Tan IB, Ivanova T, Lim KH, Ong CW, Deng N, Lee J, Tan SH, Wu J, Lee MH, Ooi CH, et al. 2011. Intrinsic Subtypes of Gastric Cancer, Based on Gene Expression Pattern, Predict Survival and Respond Differently to Chemotherapy. *Gastroenterology* **141**: 476-485.e11.

Taylor VM, Ko LK, Hwang JH, Sin M-K, Inadomi JM. 2015. Gastric Cancer in Asian American Populations: a Neglected Health Disparity. *Asian Pac J Cancer Prev* **15**: 10565–10571.

The ENCODE Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Vecchi M, Nuciforo P, Romagnoli S, Confalonieri S, Pellegrini C, Serio G, Quarto M, Capra M, Roviaro GC, Contessini Avesani E, et al. 2007. Gene expression analysis of early and advanced gastric cancers. *Oncogene* **26**: 4284–4294.

Verrecchia F, Chu M-L, Mauviel A. 2001. Identification of Novel TGF-β/Smad Gene Targets in Dermal Fibroblasts using a Combined cDNA Microarray/Promoter Transactivation Approach. *J Biol Chem* **276**: 17058–17062.

Wang X, Xu J, Zhang B, Hou Y, Song F, Lyu H, Yue F. 2021. Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. *Nat Methods* **18**: 661–668.

Xi W, Beer MA. 2018. Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy. *PLOS Comput Biol* **14**: e1006625.

Xi W, Beer MA. 2021. Loop competition and extrusion model predicts CTCF interaction specificity. *Nat Commun* **12**: 1046.

Xie Y, Su N, Yang J, Tan Q, Huang S, Jin M, Ni Z, Zhang B, Zhang D, Luo F, et al. 2020. FGF/FGFR signaling in health and disease. *Signal Transduct Target Ther* **5**: 1–38.

Xing M, Ooi WF, Tan J, Qamra A, Lee P-H, Li Z, Xu C, Padmanabhan N, Lim JQ, Guo YA, et al. 2020. Genomic and epigenomic EBF1 alterations modulate TERT expression in gastric cancer. *J Clin Invest* **130**. https://doi.org/10.1172/JCI126726.

Xu C, Huang KK, Law JH, Chua JS, Sheng T, Flores NM, Pizzi MP, Okabe A, Tan ALK, Zhu F, et al. 2023. Comprehensive molecular phenotyping of ARID1A-deficient gastric cancer reveals pervasive epigenomic reprogramming and therapeutic opportunities. *Gut* **72**: 1651–1663.

Yan J, Qiu Y, Ribeiro dos Santos AM, Yin Y, Li YE, Vinckier N, Nariai N, Benaglio P, Raman A, Li X, et al. 2021. Systematic analysis of binding of transcription factors to noncoding variants. *Nature* **591**: 147–151.

Yang J, Antin P, Berx G, Blanpain C, Brabletz T, Bronner M, Campbell K, Cano A, Casanova J, Christofori G, et al. 2020. Guidelines and definitions for research on epithelial–mesenchymal transition. *Nat Rev Mol Cell Biol* **21**: 341–352.

Yu K, Chen B, Aran D, Charalel J, Yau C, Wolf DM, van 't Veer LJ, Butte AJ, Goldstein T, Sirota M. 2019. Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nat Commun* **10**: 3574.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Figure Legends

Figure 1) Variation in chromatin accessibility profiles across GC cell lines is consistent with differential TF activity inferred from sequence-based machine learning. (A) PCA of ATAC-seq detects variable accessibility clustered roughly into three groups of GC cell lines (Mes-like, Intermediate, Epi-like). (B) gkm-SVM training produces similar weight vectors in Mes-like, Intermediate, and Epi-like GC cell lines, resulting from the activation of similar sets of TFs. (C) Training gkm-SVM on differentially active peaks detects RUNX, AP-1, and TEAD activation in Mes-like peaks, and KLF, GATA, GRHL, FOXA, HNF4A activation in Epi-like peaks. Similar TF activation is detected when training on primary GC tumor (TCGA-STAD) vs. normal stomach. ZEB is a transcriptional repressor and the absence of its motif from open-chromatin regions shows its activity and hence, the flipped sign (red color) for ZEB activity score. (D) TFBS PWM logos (E) gkm-SVM inferred activity (dot size) of these TFs across all samples detects common patterns of activation, but additional heterogeneity within each group. (F) ChIP-seq validation experiments show that RUNX2 and AP-1 bind to Mes-high distal peaks and do not bind to Epi-high peaks in LPS141, consistent with the machine learning predictions of RUNX and AP-1 activity. On the other hand, GATA4, GATA6 and KLF5 bind to Epi-high peaks and do not bind to Mes-high peaks in AGS. AGS is one of the Epi-like GC cell lines and LPS141 is a mesenchymal liposarcoma cell line with a very similar transcriptional profile to Mes-like GC cell lines (see Supplemental Fig. S6).

Figure 2) RNA-seq expression of GC cell lines is concordant with EMT and advanced-vs-early GC expression signatures and is consistent with TF activity inferred from chromatin accessibility. (A) PCA of gene expression profiles is consistent with PCA from ATAC-seq. (B) Combined PCA of GC cell lines with TCGA-STAD (tumor) and TCGA-normal shows shared variation between Mes-like vs. Epi-like and tumor vs. normal. (C,D) The genes explaining the variation (top 100 PCA gene weights) in the direction of PC1 in (A) and PC2 in (B) both are most enriched for genes highly expressed in EMT and advanced-vs-early GC. (E) Examples of marker gene expression for EMT in cancer (higher in Mes, red; higher in Epi, green), with likelihood ratio test, FDR < 0.05, and $|\log_2 FC| > 2$. (F) Differentially expressed TFs between Mes-like and Epi-like GC cell lines, with likelihood ratio test FDR < 0.05 and $|\log_2 FC| > 2$. (G) Survival analysis based on the expression of Mes and Epi DE TFs in the ACRG cohort.

Figure 3) Mutations and copy number variants are associated with distinct Mes-like and Epi-like TF activation in GC. (A) Nonsynonymous *CDH1* mutation status separates TFs with differential expression between Mes-like and Epi-like cell lines. Shown is TCGA-STAD expression in tumors with and without a *CDH1* nonsynonymous mutation (*t*-test $p < 0.05$ and $|\log_2 FC| > 0.25$). (B) Human TFs ranked in order of DNA copy number (CN) alteration rate in TCGA-STAD primary tumors. (C) Chromatin accessibility landscape and TF binding of ZEB1 and SMAD4 in the *GATA6* locus. *GATA6_E1* and *GATA6_E2* are putative Epi-like and normal stomach enhancers, where GATA6 binds in GATA6 ChIP-seq in AGS GC cell line. (D) *GATA4* has reduced CN in non-Epi and *MAPK9* has amplified CN in non-Epi cell lines. Average CN differences between Epi GC cell lines and non-Epi (Mes and Intermediate) GC cell lines are shown for protein-coding genes whose expression and copy number correlate. CN values are obtained by DNA sequencing, where “-2” marks deep DNA deletion, “0” shows a normal diploid genome, and “+2” is for deep DNA amplification in the gene locus. Mann–Whitney *U* test $p < 0.05$. (E) CN variation is consistent with altered *GATA4* and *MAPK9* expression across samples. *r* is Pearson’s correlation between CN scores and gene expression.

Figure 4) In both expression and chromatin accessibility, Mes-like GC cell lines are more similar to fibroblasts, and Epi-like GC cell lines are more similar to stomach. (A) Mes-like expression is most similar to GTEx fibroblasts. Correlation of 54 healthy GTEx gene expression profiles for each group of GC cell lines in ~11,300 tissue-specific genes. (B) Correlation of ENCODE stomach and fibroblast gene expression profiles with average gene expression for each group of GC cell lines. HT1080 is a fibrosarcoma cell line. (C) Mes-like chromatin accessibility is also most similar to ENCODE fibroblast, and TCGA-STAD and Epi-like GC cell lines are most similar to ENCODE stomach. We measured the similarity of chromatin accessibility by correlation of gkm-SVM weight vectors trained on distal enhancers of each sample. ELR is a fibroblast-derived cell line. (D) Correlation of GC cell line gene expression profiles with a healthy ENCODE lung fibroblast and a healthy stomach sample. (E) Correlation of GC cell line chromatin accessibility profiles (gkm-SVM weights trained on cell line ATAC-seq) with healthy ENCODE lung fibroblast and stomach (gkm-SVM trained using DNase-seq).

Figure 5) Mes-like GC is most similar in expression profile to the single-cell fibroblast population of both tumor and normal stomach tissue. (A) UMAP of scRNA-seq of gastric tumor and adjacent normal stomach tissue from Kim et al 2022. Six cell populations are detected; 2 and 5 are stomach and 6 is fibroblast. (B) Epi-like cell line gene expression profiles are most highly correlated with the stomach clusters 2 and 5. (C) Mes-like gene expression profiles are most highly correlated with the fibroblast cluster, 6. (D) Highly expressed stomach-specific genes are expressed only in stomach clusters 2 and 5, combined due to similarity. (*t*-test $p < 10^{-8}$ for *cluster 2+5* vs. other clusters). (E) Highly expressed fibroblast-specific genes are expressed only in fibroblast cluster 6 (*t*-test $p < 10^{-9}$ for *cluster 6* vs. other clusters.)

Figure 6) Mes and Epi upregulated genes are flanked by Mes-active and Epi-active distal enhancers. (A) Highly expressed genes in Epi-like GC cell lines, are flanked by distal peaks with increased accessibility in Epi-like cell lines. (B) Highly expressed genes in Mes-like GC cell lines, are flanked by distal peaks with increased accessibility in Mes-like cell lines. All peaks within 50kb of TSS of DE genes were averaged. Many of the upregulated genes in Mes cell lines are members of the TGFB/SMAD and FGF pathways, and their expression is shown in (C) (Mann–Whitney *U* test $p < 0.05$). (D) In the *COL1A1* locus, *COL1A1_E1* and *COL1A1_E2* are putative distal enhancers with increased ATAC-seq signal in Mes-like GC cell lines and TCGA-STAD relative to Epi-like and normal stomach, consistent with higher *COL1A1* expression in Mes-like GC cell lines and TCGA STAD. ChIP-seq experiments in the mesenchymal liposarcoma cell line LPS141 identified AP-1 and RUNX2 binding in *COL1A1_E1* and *COL1A1_E2*. (E) Similarly, in the FGF5 locus, *FGF5_E* is a putative distal enhancer with increased ATAC-seq signal in Mes-like GC cell lines and TCGA-STAD relative to Epi-like and normal stomach, consistent with higher *FGF5* expression in Mes-like GC cell lines and TCGA STAD. *FGF5_E* is a RUNX2 and AP-1 binding site in LPS141. *FGF5_E, COL1A1_E1* and *COL1A1_E2* are contained in CTCF loops enclosing the target gene promoter.

Figure 7) Despite some similarity in gene expression and enhancer activity, Mes-like GC and fibroblast have distinct TF responses. All groups of cell lines have similar levels of (A) somatic mutations (A) and (B) copy number alterations. (C) PCA of chromatin accessibility of GC cell lines, stomach, and fibroblasts shows that while all groups are distinct, there is an axis along which differences in accessibility between stomach and fibroblasts are shared between Epi-like and Mes-like cell lines. Epi-like cell lines are more similar to stomach, and Mes-like cell lines are more similar to but still distinct from, primary fibroblasts. (D) Mes-like vs fibroblast accessibility differences are driven by differential activity of a small set of TFs. (E) Rank plot of the average TF motif weights in all pairwise comparisons between Mes-like cell lines and ENCODE fibroblast samples. (F) Differential expression of TFs between stomach and fibroblast shows that Mes-like cell lines retain high expression of stomach-specific TFs and do not upregulate all fibroblast TFs (especially *TWIST2*) to the same degree. Ranking of DE TFs between Mes-like and ENCODE-fibroblast (FDR < 0.01) in normal GTEx stomach (x-axis) and normal GTEx fibroblast profiles (y-axis) are shown. Some of the DE TFs upregulated in Mes-like GC (*FOXA1, KLF5*, etc.), have a higher expression (lower ranking) in GTEx normal stomach, while *TWIST2* has a higher expression in ENCODE fibroblast and a higher expression (lower ranking) in GTEx fibroblast. (G) Inferred TF motif activities are supported by differential expression of stomach TFs and fibroblast genes in Mes-like GC, normal ENCODE-fibroblast, and ENCODE-stomach (** FDR < 0.01, * FDR < 0.05, likelihood ratio test).

Figure 8) Summary of TFs driving GC epigenomic heterogeneity and EMT. (A) Differentially active TFs in Mes-like and Epi-like GC cell lines. (B) DNA copy number changes in Mes-like and Epi-like GC cell lines. (C) Mes-like GC has a high correlation of expression and distal enhancer activity with fibroblast, whereas Epi-like GC is more similar to stomach. (D) Mes-like GC TF response and chromatin profile are different from that of fibroblast, despite the shared patterns of expression and chromatin accessibility.

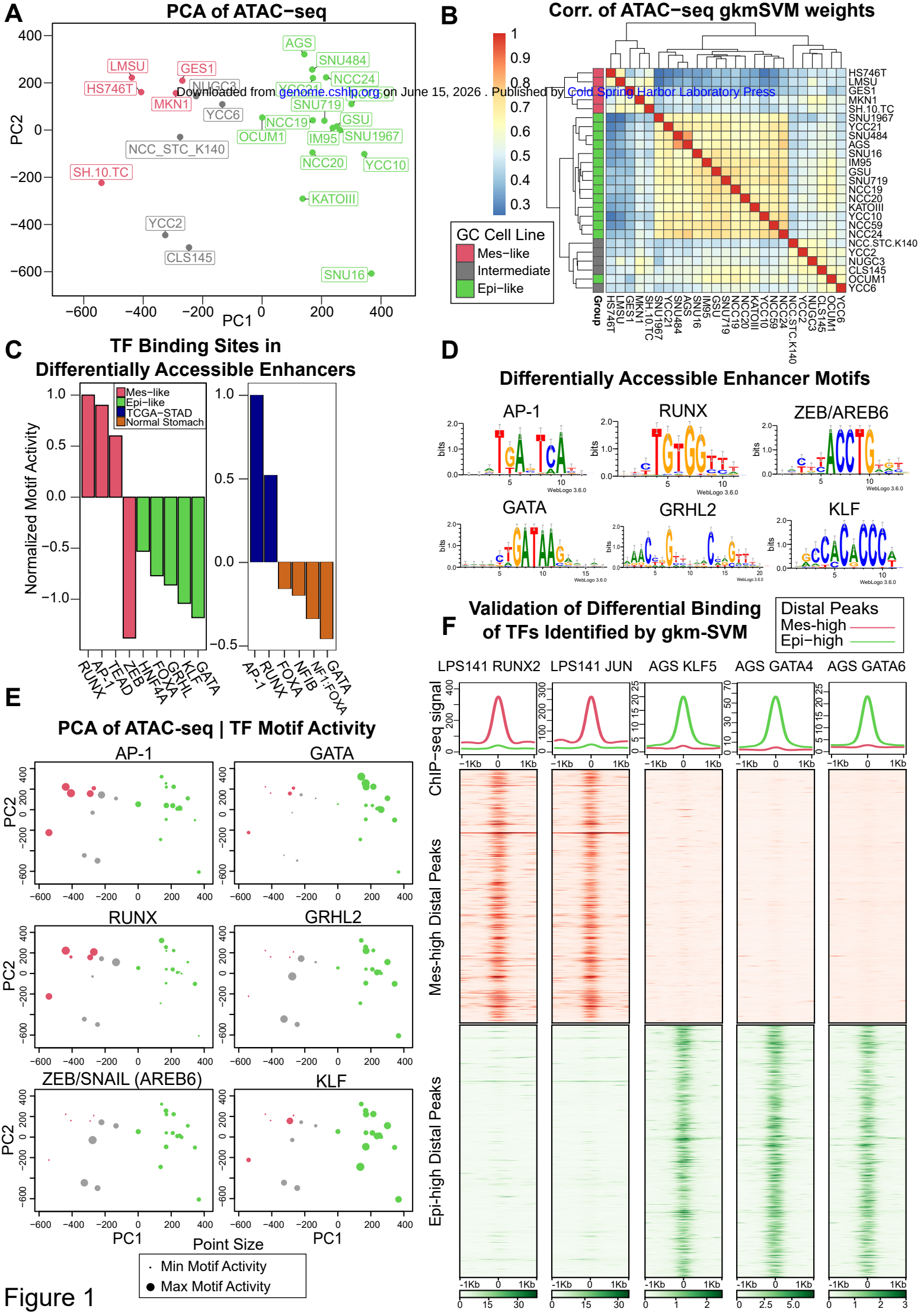
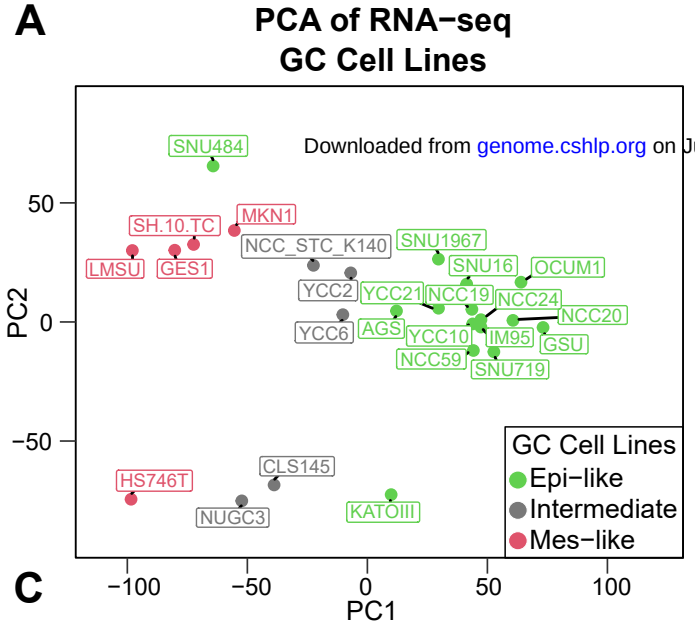
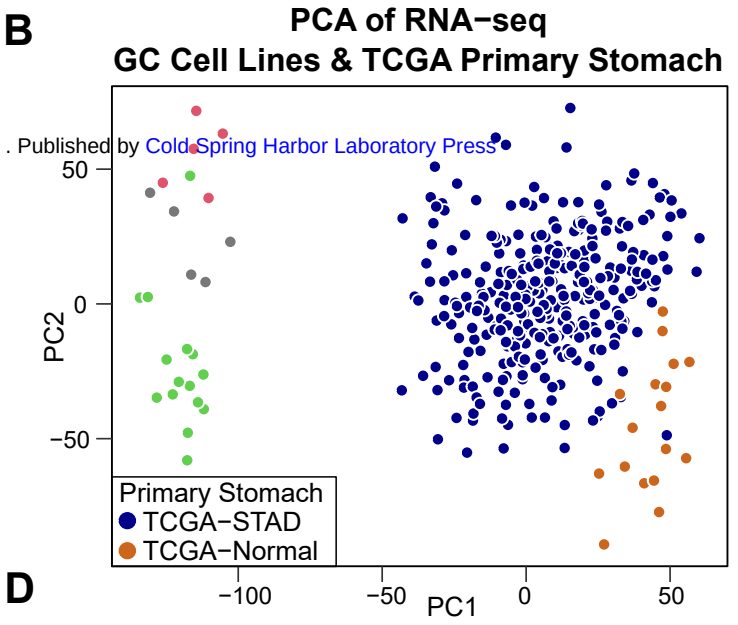


Figure 1





C

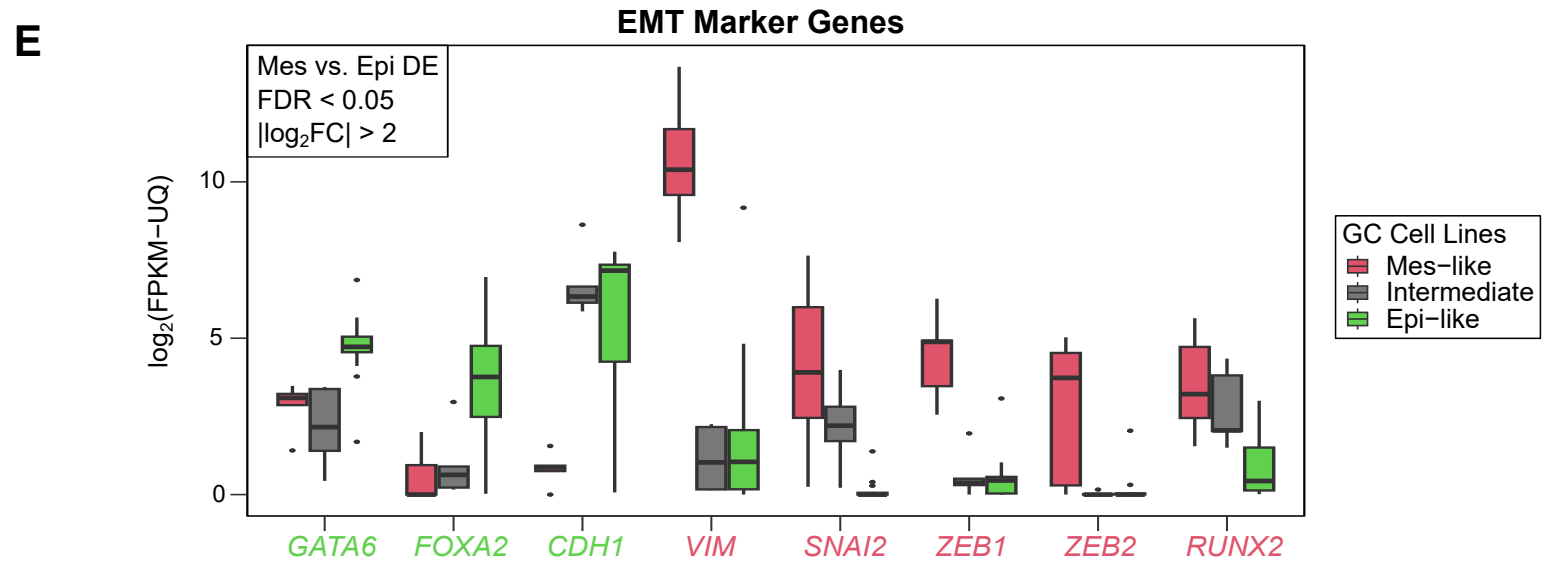
Mes vs. Epi (Top PC1 Genes)

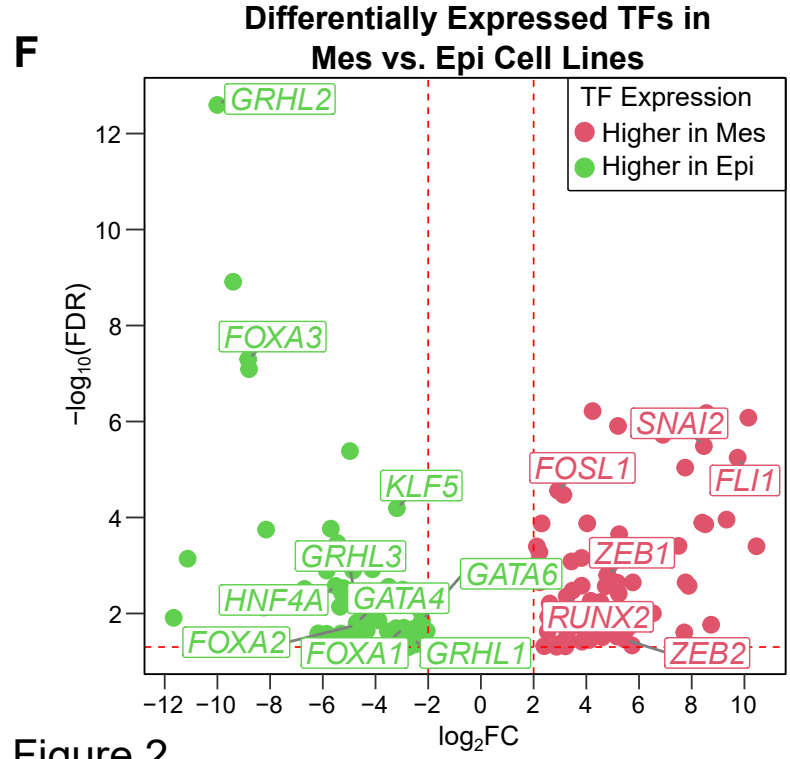
Gene Set	%Overlap	FDR
Charafe breast cancer luminal vs mesenchymal dn	10%	4×10^{-62}
Liu ovarian cancer tumors xenografts dn	3%	2×10^{-49}
Hallmark epithelial mesenchymal transition	16%	7×10^{-44}
Vecchi gastric cancer advanced vs early up	16%	1×10^{-41}
Schuetz breast cancer ductal invasive up	10%	1×10^{-41}

D

STAD & Mes vs. Normal & Epi (Top PC2 Genes)

Gene Set	%Overlap	FDR
Vecchi gastric cancer advanced vs early up	25%	1×10^{-76}
Schuetz breast cancer ductal invasive up	14%	1×10^{-72}
Liu ovarian cancer tumors xenografts dn	4%	2×10^{-72}
Hallmark epithelial mesenchymal transition	22%	3×10^{-72}
Anastassiou multicancer invasive signature up	50%	7×10^{-65}





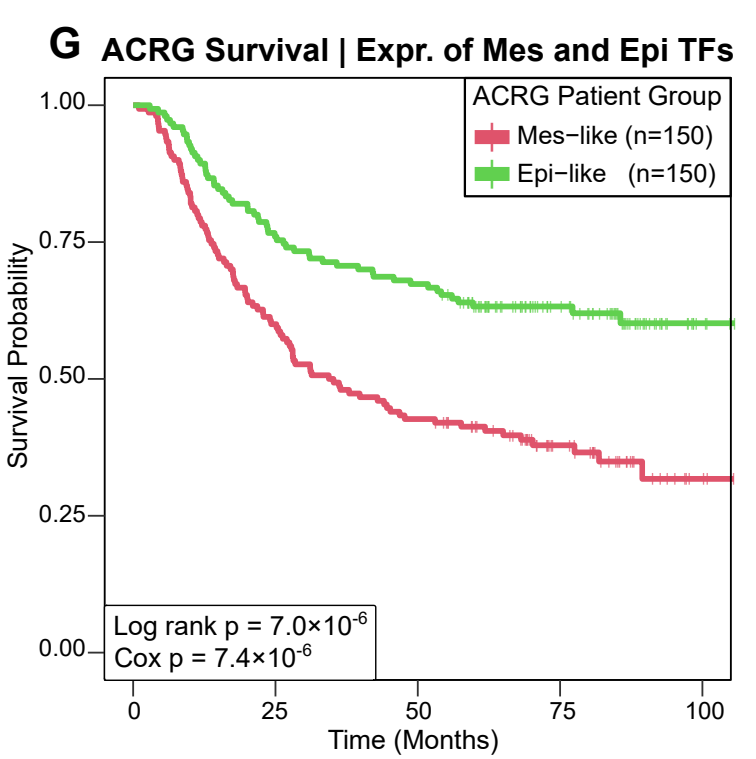


Figure 2

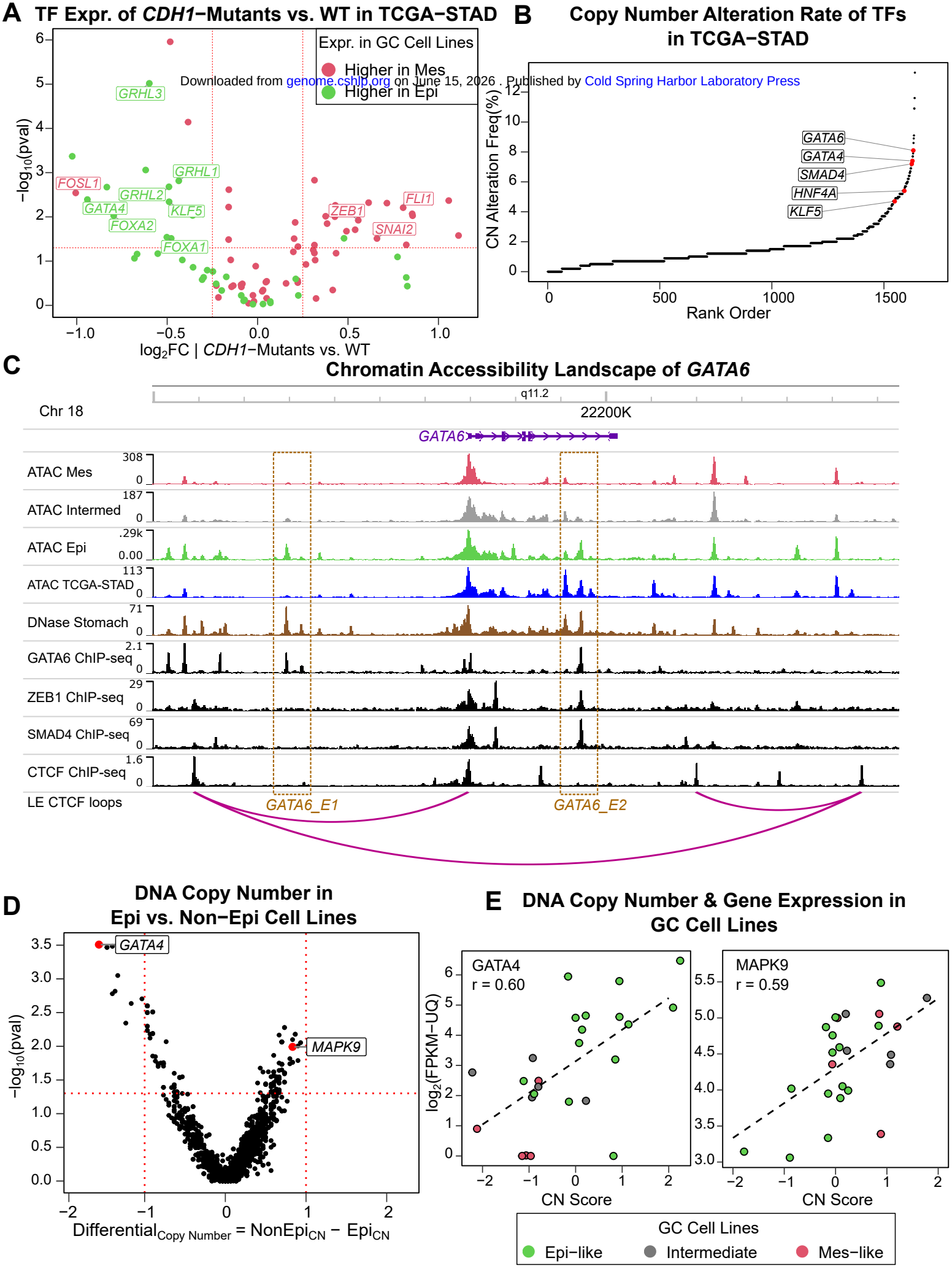


Figure 3

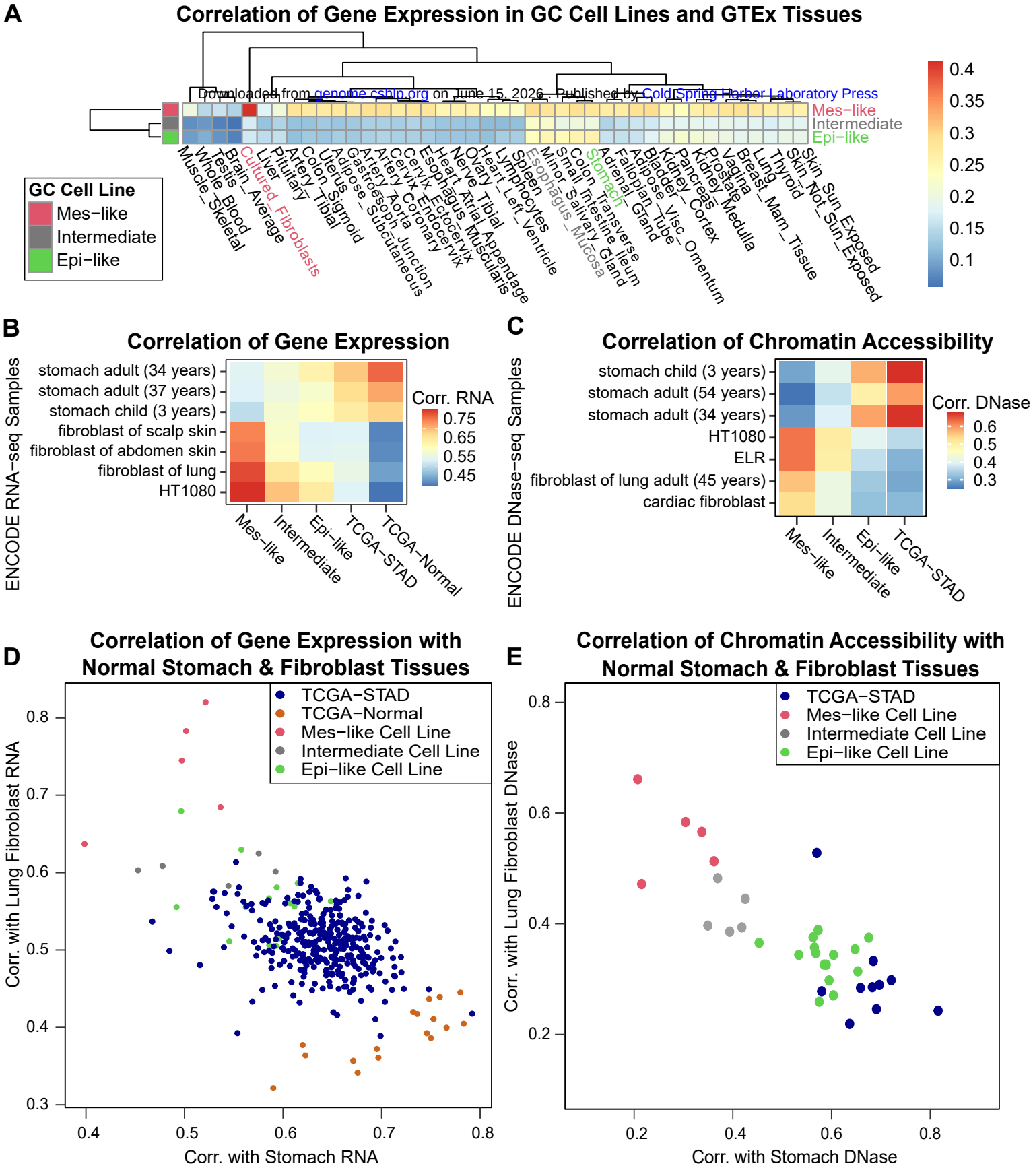


Figure 4

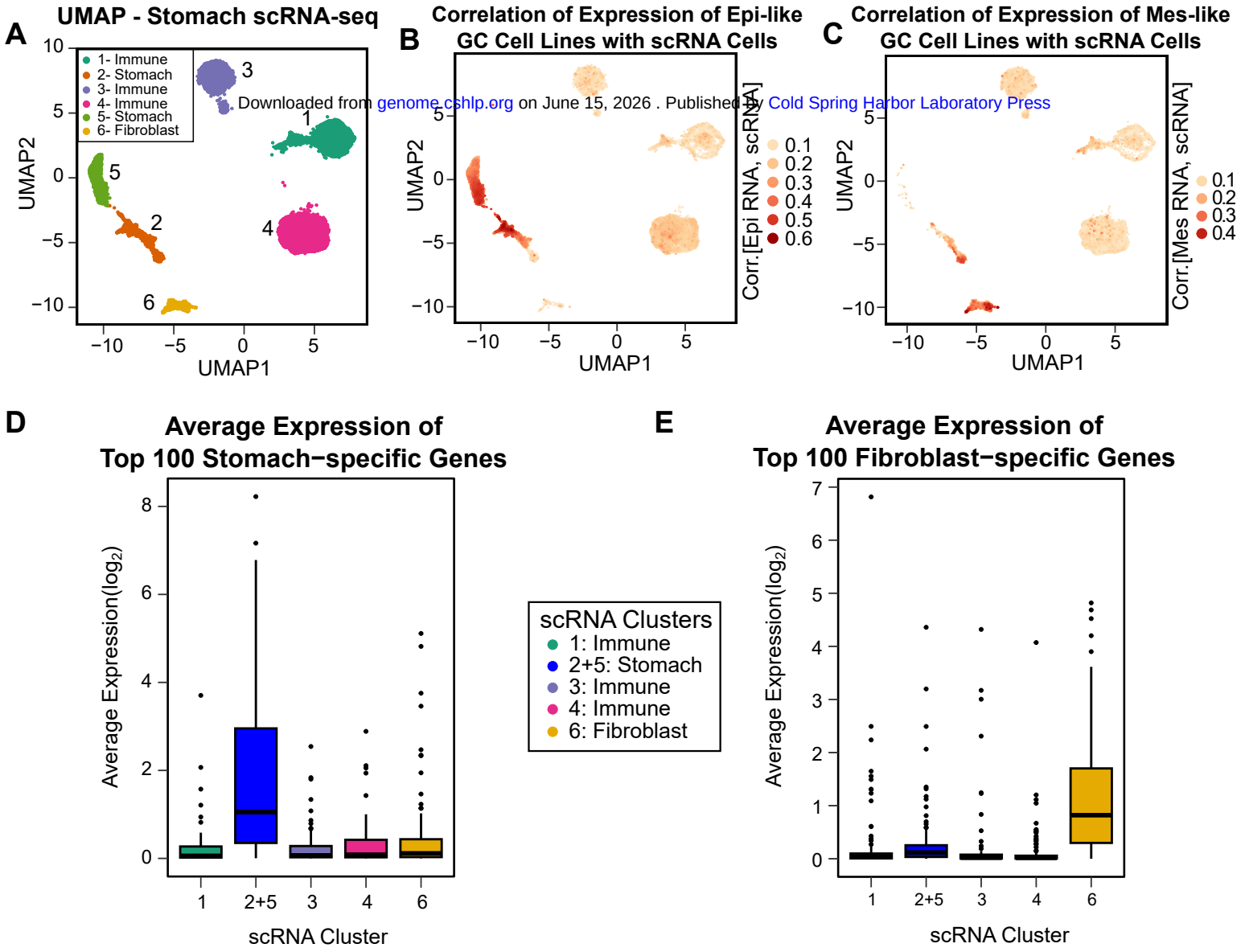


Figure 5

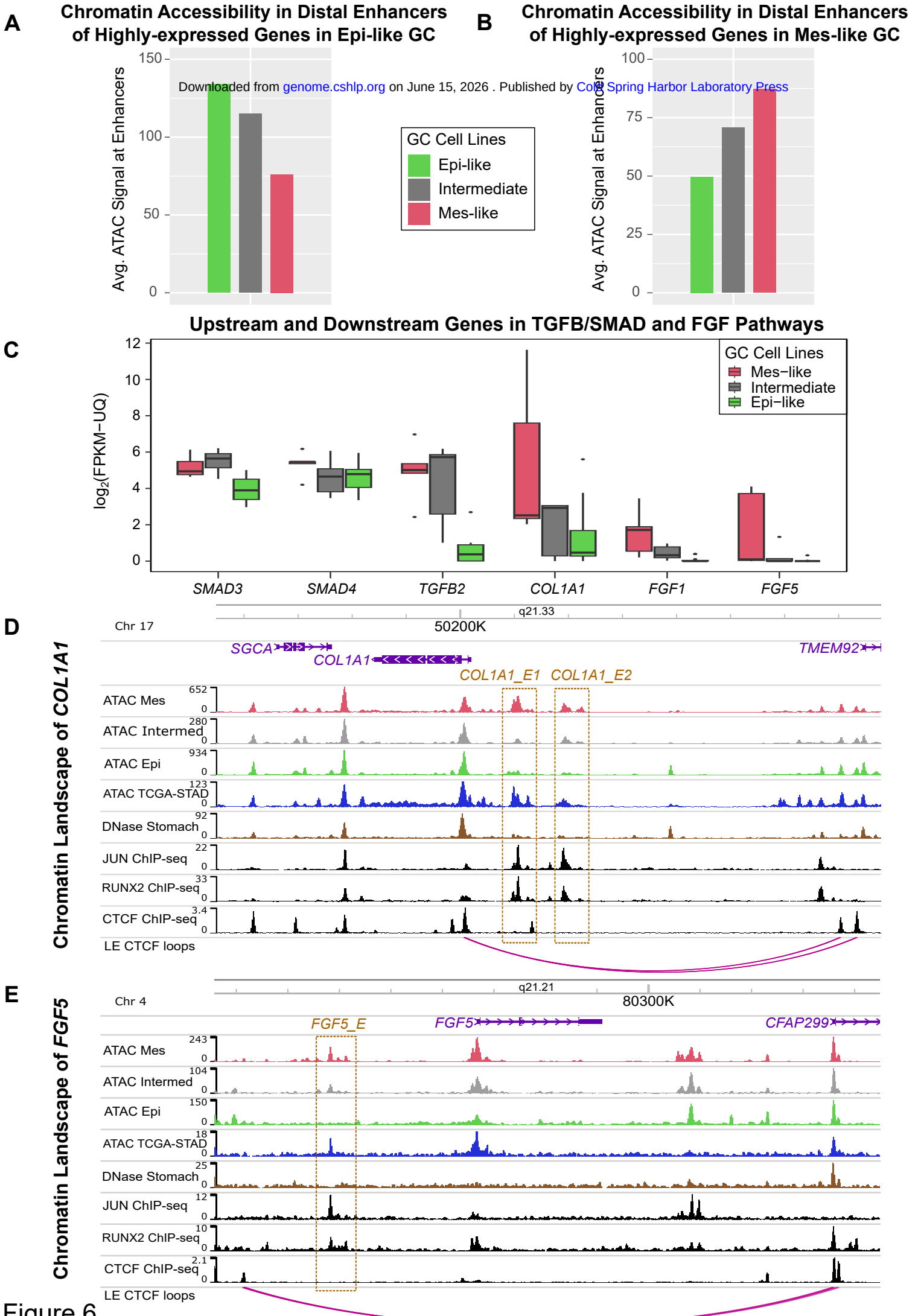


Figure 6

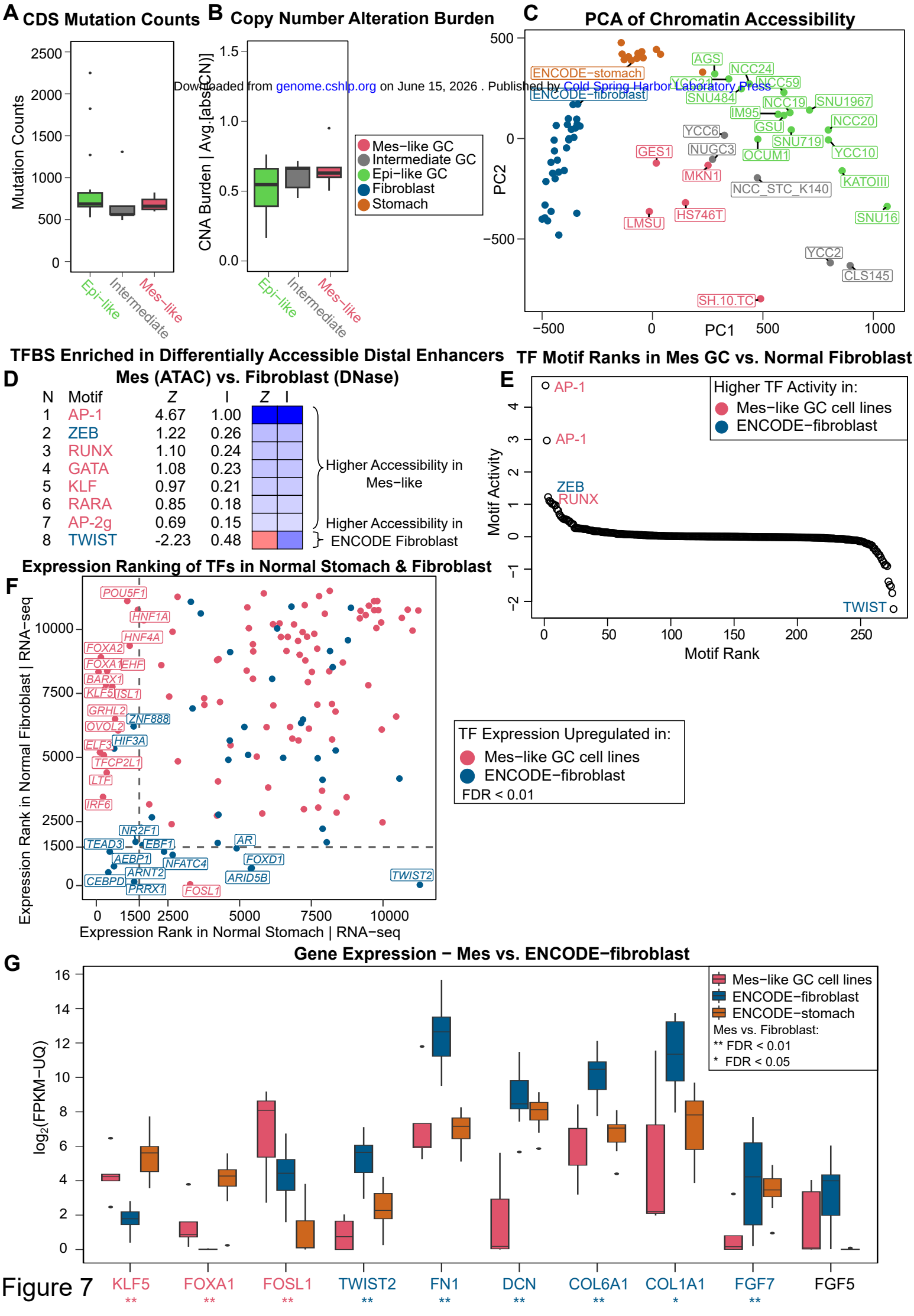


Figure 7

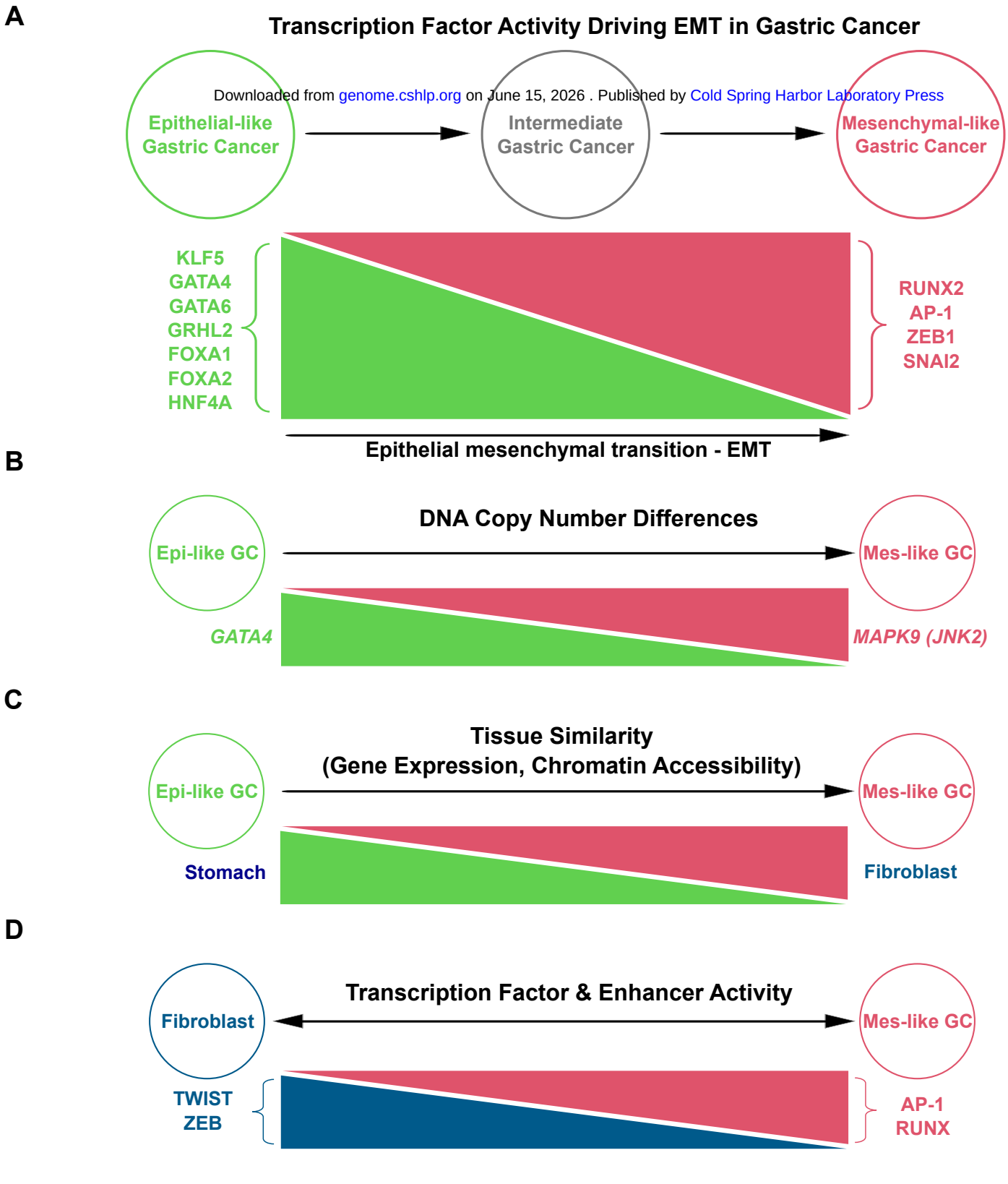


Figure 8