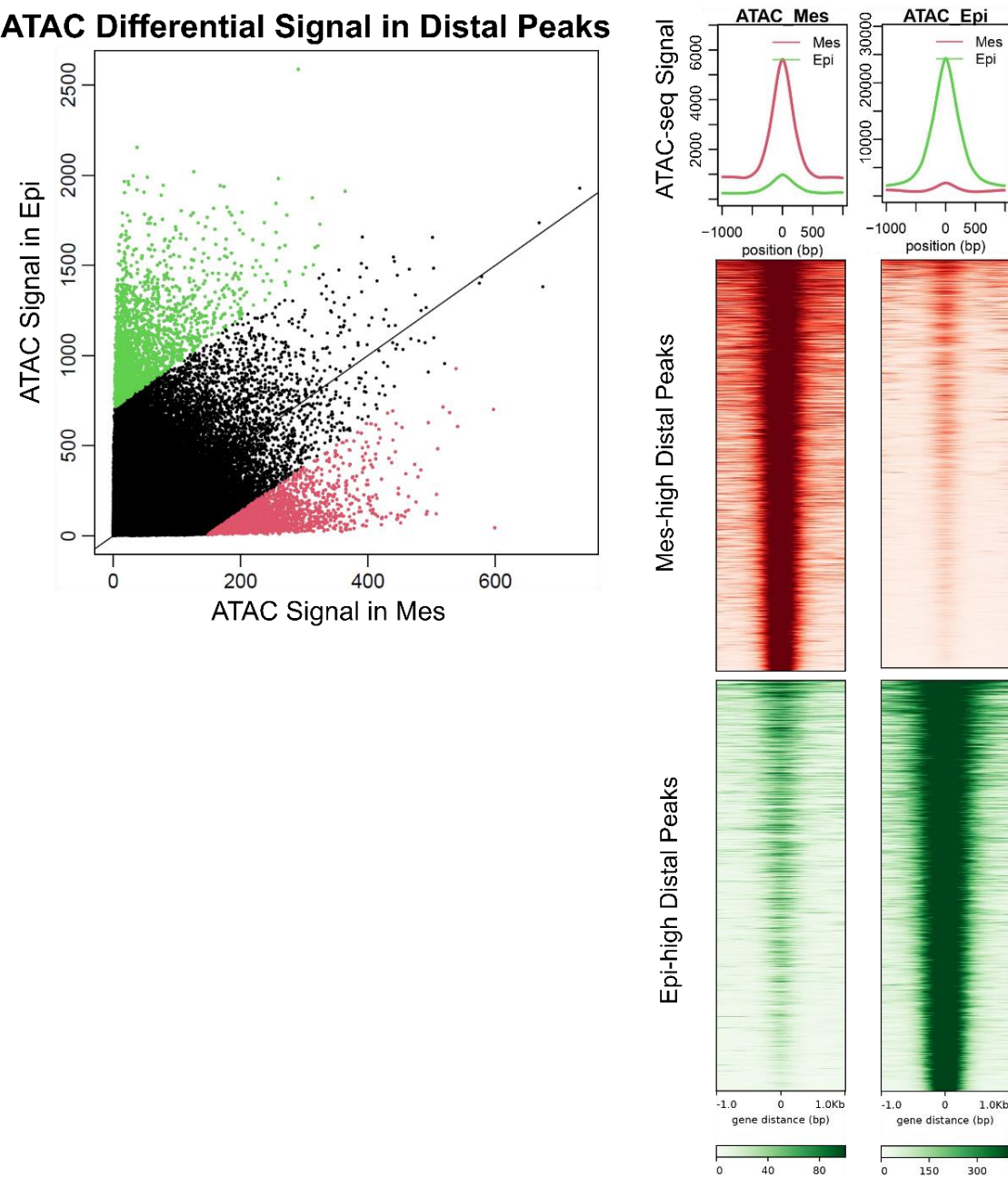
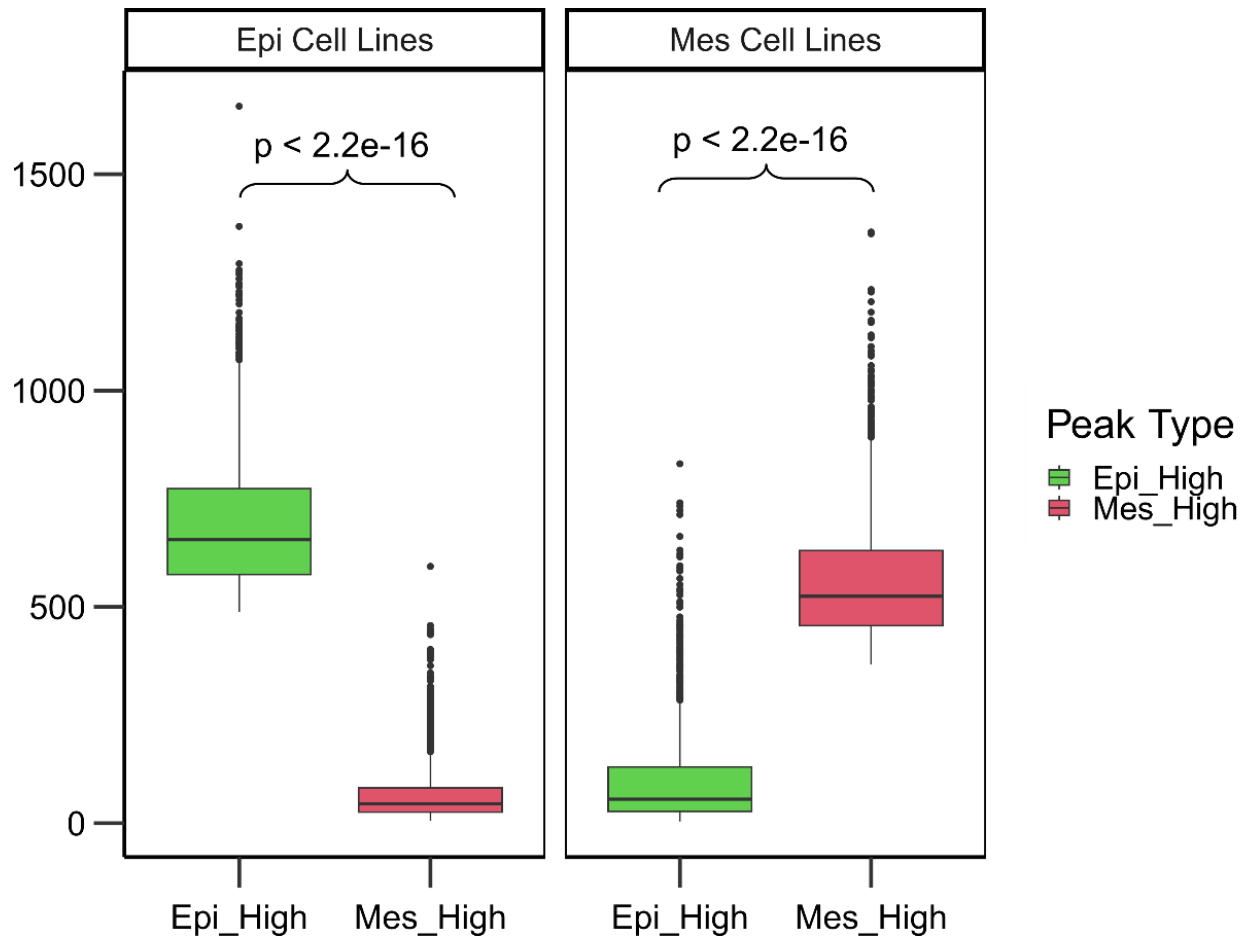


Supplemental Fig. S1) Average ATAC-seq signal in Mes-like and Epi-like differentially accessible peaks (n=2000 peaks in each set)



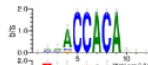

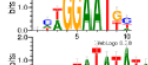
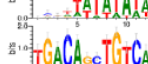

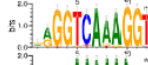
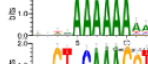
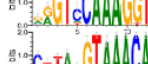
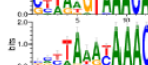
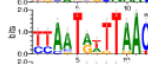
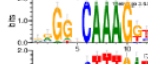
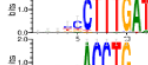
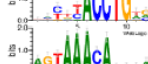


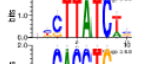
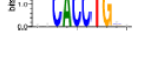
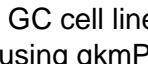
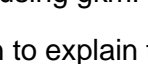
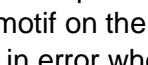
Distal ATAC peaks with the highest accessibility differences between Mes and Epi GC cell lines (2000 Mes-high & 2000 Epi-high peaks) are shown.

Supplemental Fig. S2) Statistical Validation of ATAC Signal Differences between Mes-high and Epi-high Peaks



We used linear regression to find top 2000 differentially accessible Mes-high and top 2000 Epi-high peaks. As expected, the t -test shows a significant difference between the ATAC signal of Mes-high and Epi-high peaks.

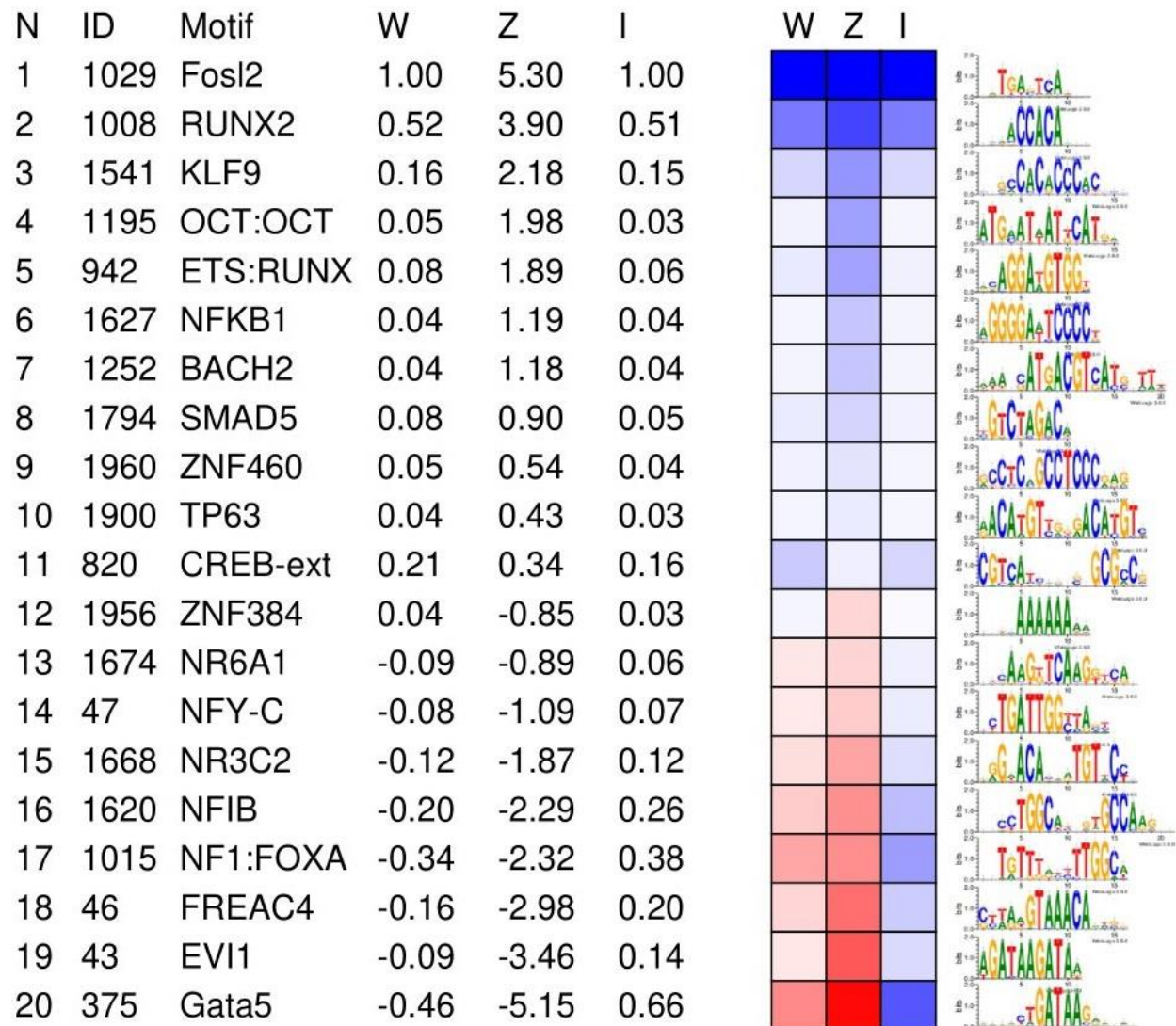
Supplemental Fig. S3) Active Motifs in Differentially Accessible Peaks in Mes vs. Epi Cell Line ATAC-seq (2000 Mes-high vs. 2000 Epi-high peaks)

N	ID	Motif	W	Z	I	W	Z	I	
1	1008	RUNX2	1.00	2.77	0.51				
2	895	AP-1	0.90	4.04	1.00				
3	1150	TEAD	0.60	2.37	0.25				
4	629	Tbp	0.30	0.61	0.13				
5	1891	TGIF2LX	0.25	0.14	0.12				
6	1428	GLI2	0.21	1.07	0.11				
7	1661	NR2F1	0.20	-0.43	0.08				
8	1956	ZNF384	0.15	0.35	0.08				
9	1464	HNF4A	-0.16	-3.10	0.08				
10	46	FREAC4	-0.20	-3.23	0.07				
11	1393	FOXE1	-0.27	-2.06	0.08				
12	89	HNF1B	-0.51	-2.92	0.40				
13	80	HNF4-01-	-0.53	-2.78	0.19				
14	222	Lef1	-0.71	-4.20	0.67				
15	339	AREB6	-0.73	-4.52	0.31				
16	1059	FOXA1:AR	-0.77	-2.29	0.36				
17	1436	GRHL2	-0.86	-3.24	0.56				
18	1146	KLF5	-1.04	-3.22	0.64				
19	1094	Gata2	-1.18	-4.24	0.99				
20	1072	E2A	-1.38	-4.81	0.91				

Top 2000 differentially accessible distal peaks between Mes-like and Epi-like GC cell lines (Supplemental Fig. S1) were compared to identify known TFBS DNA motifs, using gkmPWM.

W is the normalized weight for a particular motif found using lasso regression to explain the gapped *k*-mer (gkmSVM) weight space. *Z* is the approximate location of the motif on the gkmSVM weight distribution in terms of z-score, and *I* is the relative increase in error when removing the TF/motif from the list.

Supplemental Fig. S4) Active Motifs in Differentially Accessible Peaks in TCGA-STAD Primary Tumor vs. Normal Adult Stomach ATAC-seq



Top 2000 differentially accessible distal peaks between TCGA-STAD (TCGA-BR-A4J6) and healthy adult stomach (ENCODE ENCBS441WEO) were compared to identify TFBS DNA motifs, using gkmPWM.

Supplemental Fig. S5) Active Motifs in Differentially Accessible Peaks in All TCGA-STAD Primary Tumors vs. Normal Adult Stomach ATAC-seq

TCGA-STAD ATAC (blue) vs. Normal Stomach (red, ENCODE:ENCBS441WEO)

N	ID	Motif	W	Z	I	W	Z	I	N	ID	Motif	W	Z	I	W	Z	I
1	1377	FOSL2	1.00	7.62	1.00				1	1164	KLF3	1.00	3.87	1.00			
2	1527	Klf1	0.88	3.99	0.42				2	137	NFE2	0.57	3.25	0.56			
3	1463	HNF4A	0.40	3.02	0.09				3	1628	NFKB2	0.21	2.20	0.18			
4	1464	HNF4A	0.15	2.42	0.04				4	39	ISRE	0.31	1.98	0.26			
5	1270	CDX1	0.28	1.87	0.05				5	1464	HNF4A	0.07	1.28	0.03			
6	1627	NFKB1	0.11	1.26	0.06				6	1221	ZSCAN22	0.20	1.27	0.09			
7	600	Six6	0.30	1.13	0.07				7	942	ETS:RUNX	0.18	1.19	0.07			
8	1777	Rxra	0.07	1.08	0.01				8	577	MEIS1	0.77	1.14	0.35			
9	261	NRF2	0.32	1.01	0.07				9	1947	ZNF24	0.11	0.90	0.07			
10	826	EGR1	0.19	0.92	0.06				10	86	ELK1	0.32	0.80	0.12			
11	176	Zfp410	0.19	0.81	0.03				11	54	RREB1	0.19	0.72	0.09			
12	1036	YY1	0.12	0.62	0.03				12	1923	ZBTB32	0.21	0.69	0.09			
13	59	MEIS1BHO	0.18	0.53	0.03				13	1955	ZNF382	0.21	0.29	0.12			
14	1955	ZNF382	0.17	0.41	0.04				14	1792	SMAD3	0.09	0.11	0.04			
15	35	STAF	0.21	0.33	0.04				15	46	FREAC4	-0.23	-1.69	0.12			
16	46	FREAC4	-0.19	-1.33	0.04				16	43	EVI1	-0.14	-2.43	0.09			
17	239	Gm397	-0.26	-1.44	0.06				17	1181	PGR	-0.49	-2.70	0.36			
18	1181	PGR	-0.35	-2.01	0.10				18	1015	NF1:FOXA	-0.73	-3.02	0.53			
19	1620	NFIB	-0.45	-2.36	0.21				19	1620	NFIB	-0.69	-3.06	0.88			
20	1015	NF1:FOXA	-0.77	-2.76	0.32				20	375	Gata5	-0.58	-3.47	0.32			

TCGA-CD-A48C (STAD1)

N	ID	Motif	W	Z	I	W	Z	I
1	1106	Fra1	0.85	5.55	0.67			
2	1146	KLF5	1.00	5.31	1.00			
3	1463	HNF4A	0.39	4.00	0.21			
4	80	HNF4-01-	0.11	3.18	0.02			
5	1122	HOXA2	0.26	3.08	0.12			
6	1464	HNF4A	0.06	2.52	0.02			
7	1628	NFKB2	0.13	1.74	0.11			
8	1813	Sox6	0.38	1.47	0.16			
9	1086	Cdx2	0.27	1.44	0.10			
10	1221	ZSCAN22	0.12	0.99	0.05			
11	40	PAX5	0.32	0.78	0.11			
12	59	MEIS1BHO	0.12	0.66	0.03			
13	1792	SMAD3	0.07	0.62	0.03			
14	1890	TGIF2	0.06	0.57	0.03			
15	362	Nkx3-1	-0.15	-0.90	0.05			
16	1668	NR3C2	-0.14	-1.85	0.07			
17	375	Gata5	-0.32	-2.27	0.16			
18	1015	NF1:FOXA	-0.43	-2.68	0.25			
19	1620	NFIB	-0.30	-3.28	0.21			
20	1619	NFIA	-0.50	-4.79	0.18			

TCGA-VQ-A94O (STAD2)

N	ID	Motif	W	Z	I	W	Z	I
1	1865	TCF7L2	1.00	7.78	1.00			
2	1008	RUNX2	0.43	2.89	0.12			
3	1270	CDX1	0.13	2.53	0.02			
4	316	Zfp281	0.09	2.20	0.05			
5	1628	NFKB2	0.13	2.12	0.07			
6	535	Plagl1	0.10	1.93	0.03			
7	1933	ZIC1	0.08	1.59	0.02			
8	1122	HOXA2	0.17	1.49	0.04			
9	38	POLY-C	0.09	1.03	0.01			
10	59	MEIS1BHO	0.12	0.85	0.03			
11	1835	Stat2	0.13	0.51	0.03			
12	1944	ZNF143	0.06	0.38	0.02			
13	22	GCNF	-0.05	-0.25	0.01			
14	47	NFY-C	-0.08	-1.07	0.02			
15	46	FREAC4	-0.09	-1.15	0.02			
16	1280	CLOCK	-0.17	-1.51	0.05			
17	1015	NF1:FOXA	-0.29	-1.55	0.10			
18	1748	Rarb	-0.13	-1.66	0.06			
19	1620	NFIB	-0.16	-1.66	0.06			
20	1227	Ar	-0.20	-2.21	0.09			

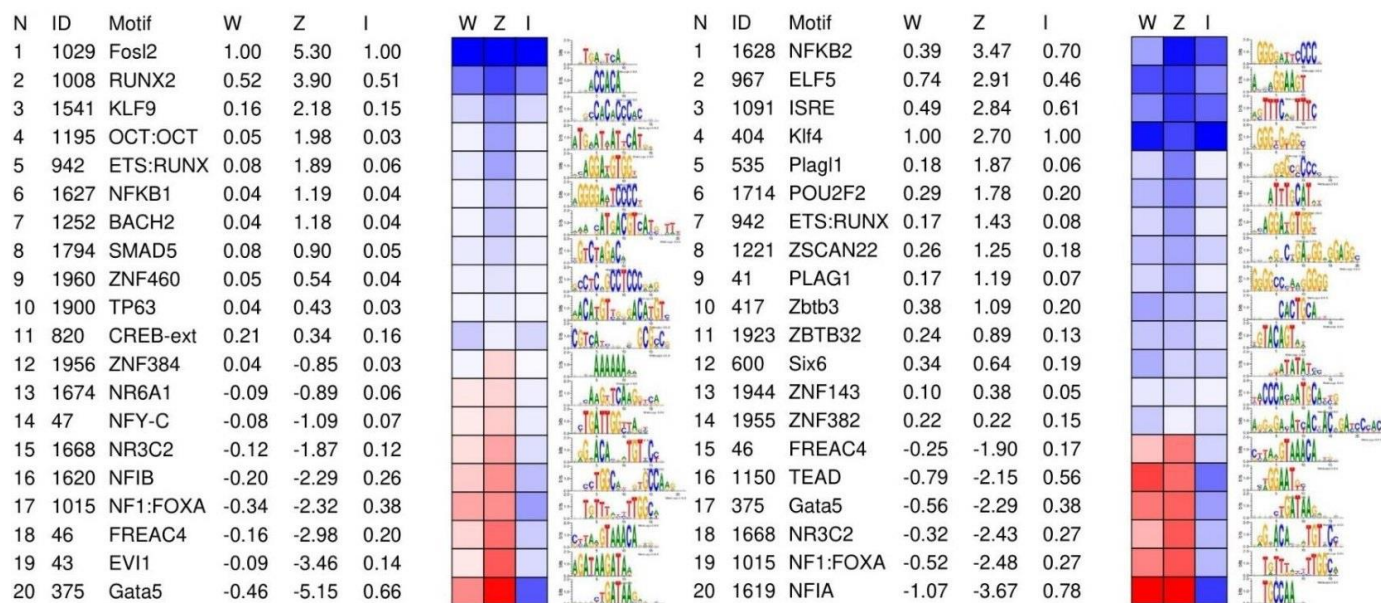
TCGA-VQ-A8PJ (STAD5)

TCGA-VQ-A91W (STAD6)

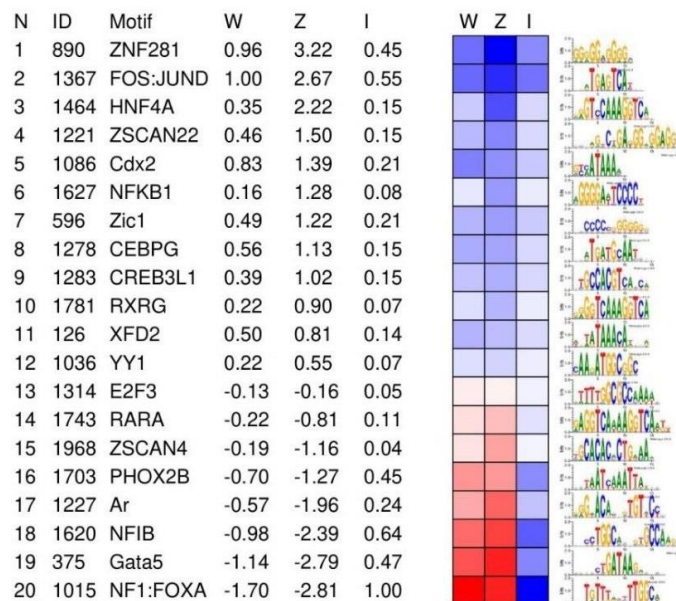
Top 2000 differentially accessible distal peaks between TCGA-STAD ATAC-seq samples and healthy adult stomach (ENCODE ENCBS441WEO) were compared to identify TFBS DNA motifs, using gkmPWM.

Supplemental Fig. S5 Continued

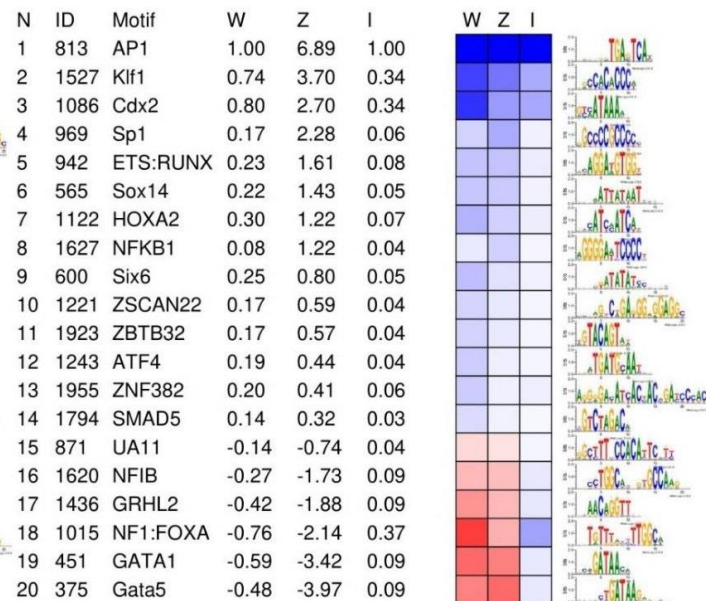
TCGA-STAD ATAC (blue) vs. Normal Stomach (red, ENCODE:ENCBS441WEO)



TCGA-BR-A4J6 (STAD12)



TCGA-BR-A4IY (STAD13)



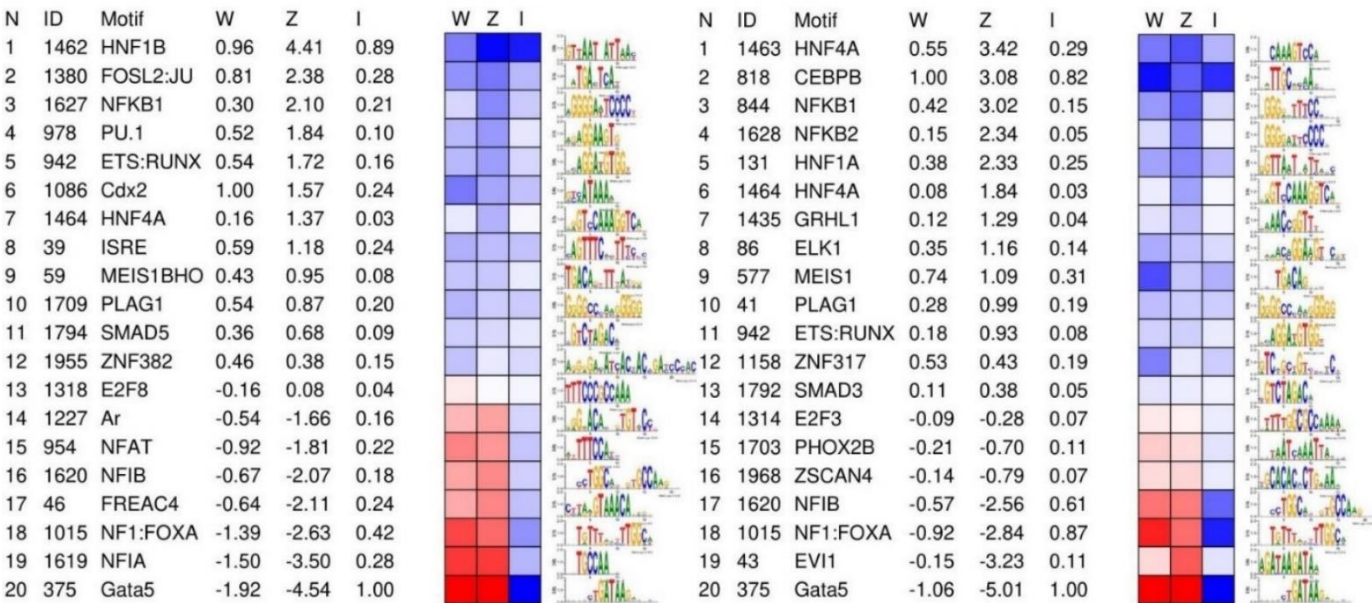
TCGA-BR-A4CS (STAD14)

TCGA-HF-A5NB (STAD16)

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. All of the 10 STAD ATAC samples with >10k distal peaks detect some activation of RUNX or AP-1 when trained against normal stomach DHS (DHS_882): 7 detect AP-1 (STAD1,2,5,12,14,16,19), and 7 detect RUNX (STAD2,6,12,13,16,19,21).

Supplemental Fig. S5 Continued

TCGA-STAD ATAC (blue) vs. Normal Stomach (red, ENCODE:ENCBS441WEO)

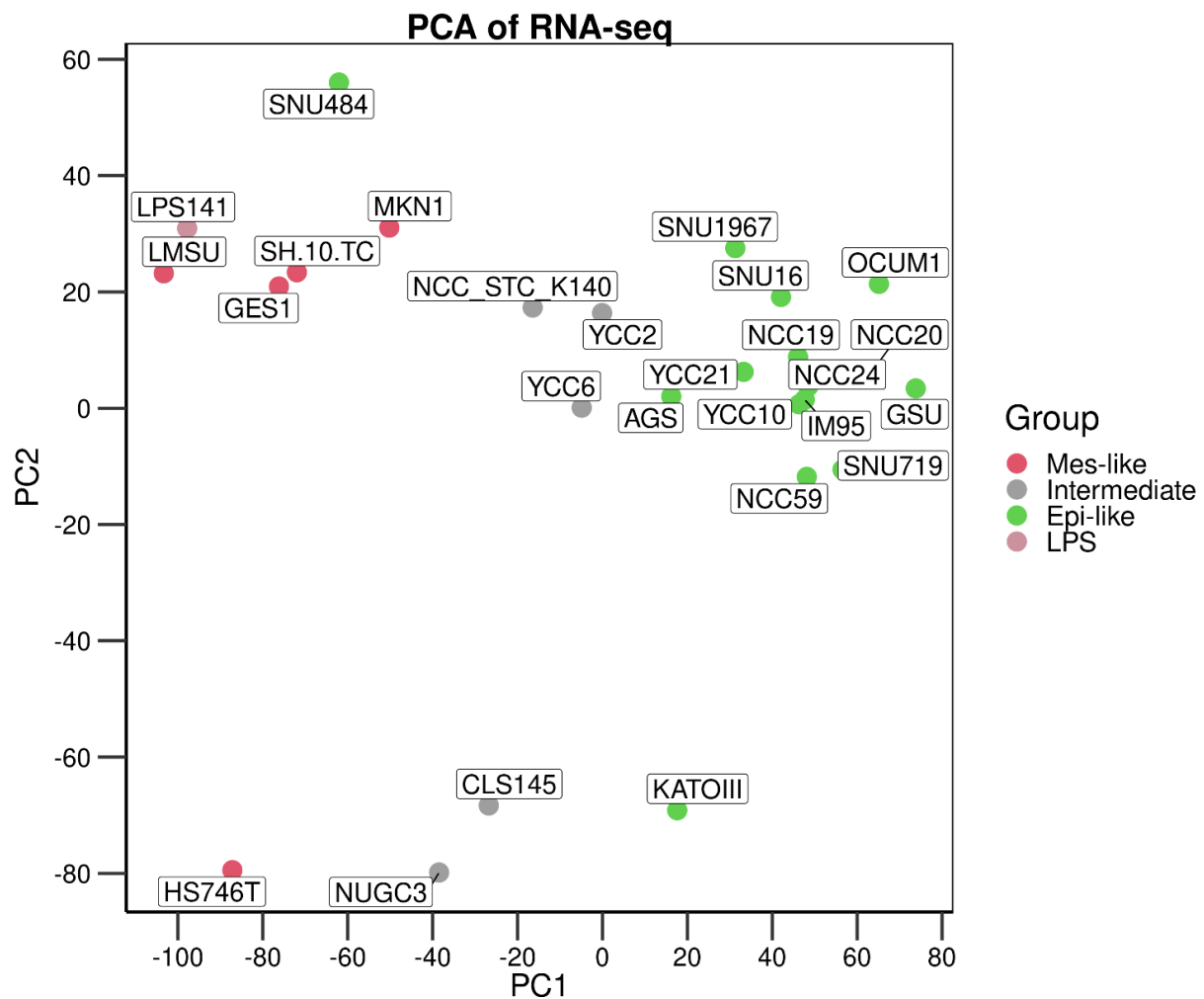


TCGA-BR-A4J4 (STAD19)

TCGA-CD-A486 (STAD21)

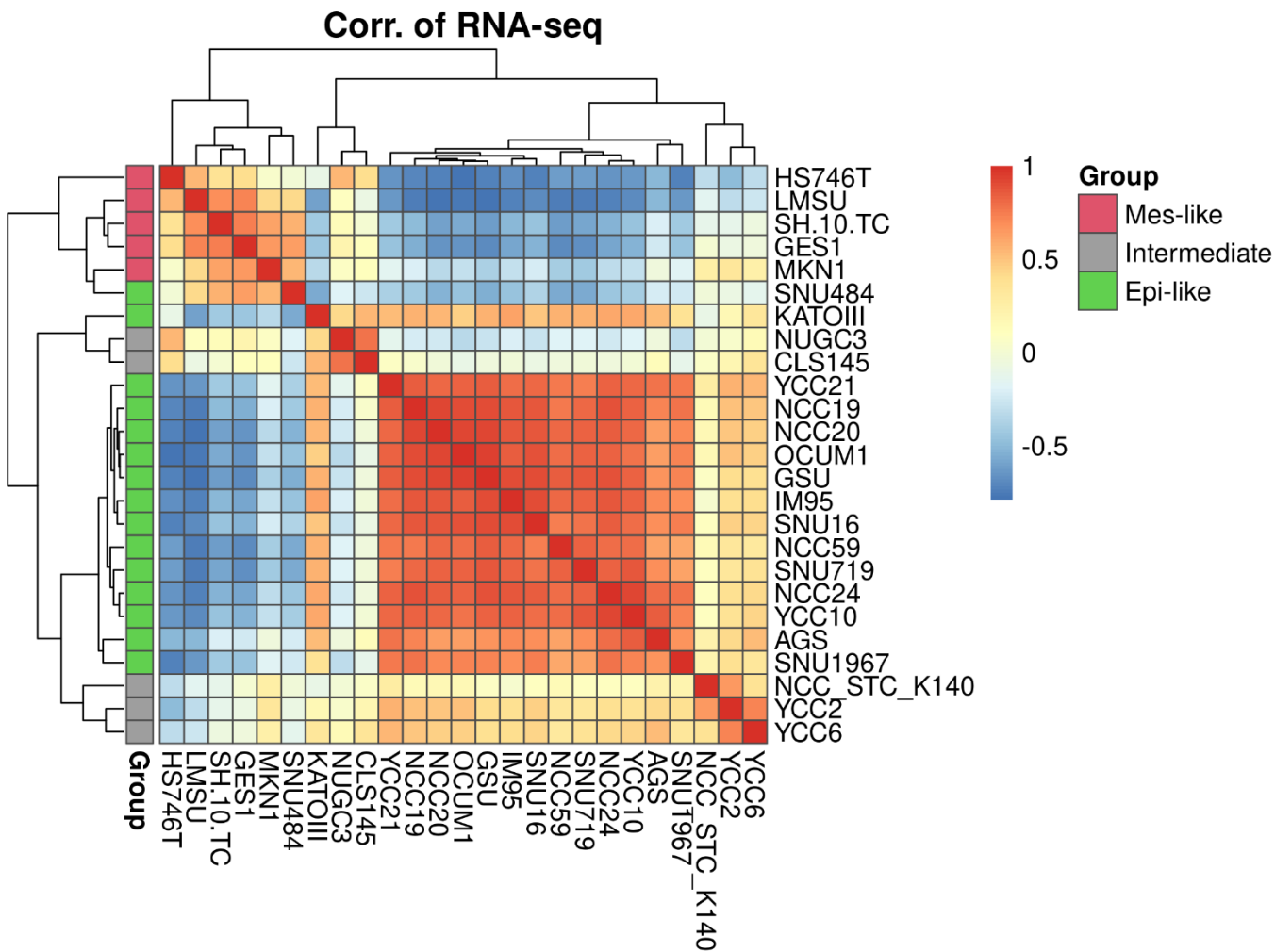
W is the normalized weight for a particular motif found using lasso regression to explain the gapped *k*-mer (gkmSVM) weight space. Z is the approximate location of the motif on the gkmSVM weight distribution in terms of z-score, and I is the relative increase in error when removing the TF/motif from the list.

Supplemental Fig. S6) LPS141 Transcriptional Profile Compared to GC Cell Lines



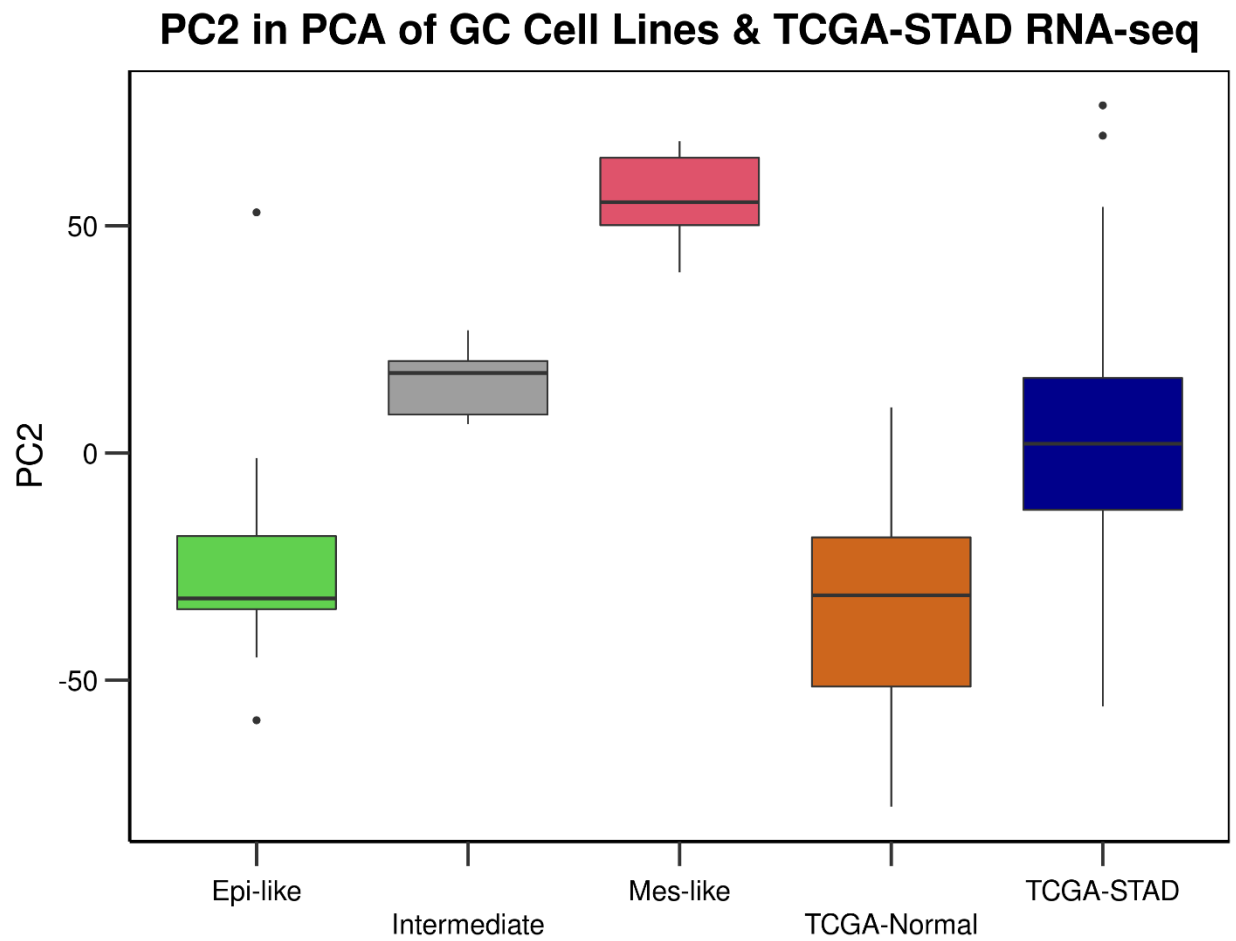
PCA of RNA-seq over ~11,300 tissue-specific (see Methods) protein-coding genes. LPS141 mesenchymal liposarcoma cell line has a very similar transcriptional profile to Mes-like GC cell lines.

Supplemental Fig. S7) Correlation Heatmap of RNA-seq in GC Cell Lines



Correlation heatmap of RNA-seq profiles is consistent with the ATAC-seq clustering.

Supplemental Fig. S8) PC2 of RNA-seq PCA



t-test p-values are:

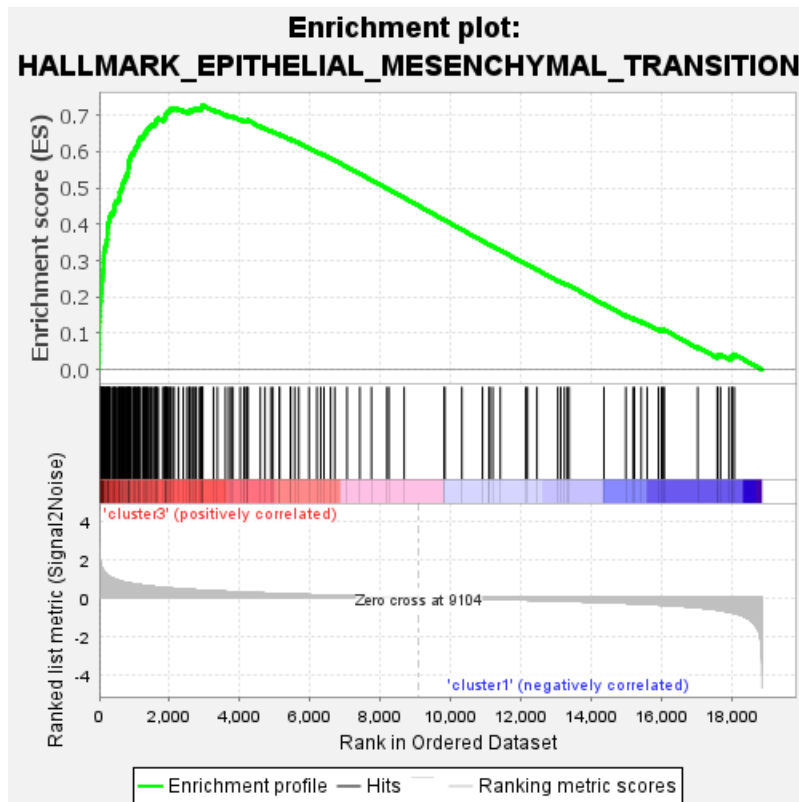
- (Mes vs. Intermediate) = 9×10^{-5}
- (Intermediate vs. Epi) = 6×10^{-5}
- (TCGA-STAD vs. TCGA-Normal) = 1×10^{-5}

Supplemental Fig. S9) Gene Set Enrichment Analysis (GSEA) of Hallmark Gene Sets for Mes vs. Epi GC Cell Lines

NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	199	0.7285	1.9075	0	0.0388071	0.032	2945
HALLMARK_UV_RESPONSE_DN	144	0.5945	1.7767	0	0.0689343	0.094	2439
HALLMARK_ANGIOGENESIS	36	0.5243	1.599	0.007874	0.1885999	0.32	2712
HALLMARK_TGF_BETA_SIGNALING	54	0.4646	1.5919	0.0281955	0.1491298	0.335	3497
HALLMARK_HYPOXIA	200	0.4703	1.5498	0.0246085	0.1554014	0.414	2932
HALLMARK_INFLAMMATORY_RESPONSE	199	0.4039	1.457	0.0108108	0.2272618	0.597	4181
HALLMARK_MITOTIC_SPINDLE	197	0.3839	1.4352	0.1172023	0.2202722	0.644	5495
HALLMARK_TNFA_SIGNALING_VIA_NFKB	197	0.4306	1.3232	0.1384298	0.3746146	0.852	4163
HALLMARK_COMPLEMENT	197	0.3386	1.3231	0.0857788	0.3331756	0.852	3708
HALLMARK_HEDGEHOG_SIGNALING	36	0.4529	1.3229	0.1175115	0.3004615	0.853	3598
HALLMARK_MYOGENESIS	199	0.3844	1.276	0.1338384	0.3441439	0.903	3337
HALLMARK_ALLOGRAFT_REJECTION	198	0.3235	1.2282	0.1296296	0.3867421	0.946	3816
HALLMARK_KRAS_SIGNALING_UP	197	0.3266	1.2196	0.1008403	0.3703505	0.951	1824
HALLMARK_APICAL_JUNCTION	198	0.3294	1.2093	0.2004831	0.3589813	0.956	2002
HALLMARK_IL2_STAT5_SIGNALING	198	0.287	1.1636	0.1428572	0.4041026	0.973	2930
HALLMARK_INTERFERON_GAMMA_RESPONSE	198	0.319	1.1524	0.2579909	0.3981969	0.975	4018
HALLMARK_APOPTOSIS	159	0.2736	1.1442	0.2038835	0.3860222	0.978	1957
HALLMARK_G2M_CHECKPOINT	198	0.396	1.1142	0.4158965	0.4062013	0.986	5503
HALLMARK_IL6_JAK_STAT3_SIGNALING	87	0.3074	1.0972	0.2819843	0.4092465	0.991	3530
HALLMARK_APICAL_SURFACE	44	0.3613	1.082	0.3340961	0.4099661	0.993	5733
HALLMARK_COAGULATION	136	0.3046	1.0599	0.3333333	0.4221038	0.995	2712
HALLMARK_INTERFERON_ALPHA_RESPONSE	96	0.306	1.0076	0.4318658	0.4756718	0.998	3747
HALLMARK_E2F_TARGETS	200	0.3688	0.9369	0.5620301	0.5585469	0.999	6861
HALLMARK_SPERMATOGENESIS	135	0.2288	0.7839	0.8423529	0.7690128	1	4479

GSEA between Mes and Epi GC cell lines was performed using “Hallmark” gene sets.

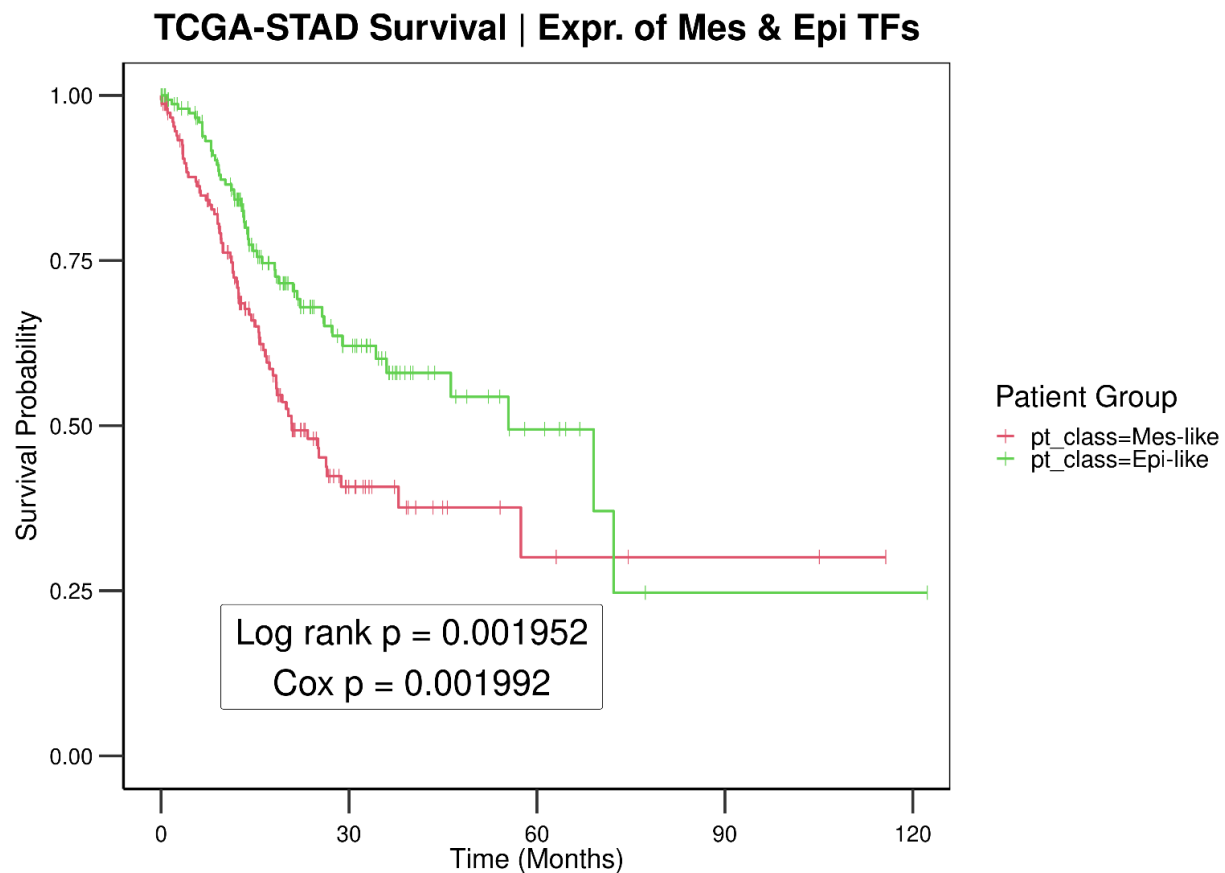
Supplemental Fig. S10) GSEA Enrichment Plot for Hallmark of EMT Gene Set



In the enrichment plot:

- Cluster 1: Epi-like GC Cell Lines
- Cluster 3: Mes-like GC Cell Lines

Supplemental Fig. S11) TCGA-STAD KM-plot Based on Mes vs Epi TF Expression

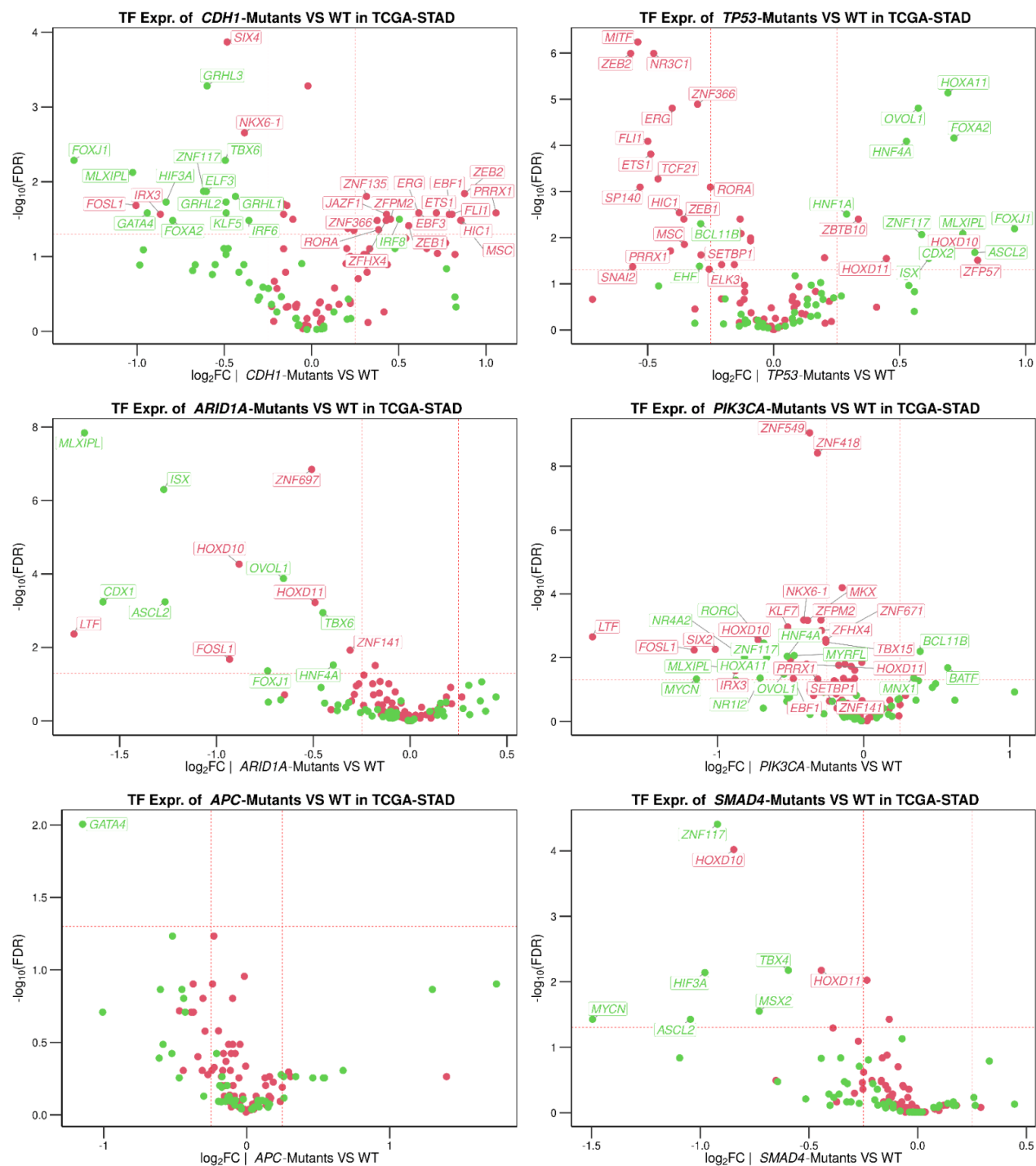


We calculated the difference between the expression of Mes and Epi differentially expressed TFs on TCGA-STAD RNA-seq. Mes and Epi TFs were defined from the differential gene expression analysis of the GC cell lines, and those differentially expressed genes that are TFs are used here for stratifying TCGA-STAD patients.

Difference = [Average(Mes TF expression) – Average(Epi TF expression)].

TCGA-STAD samples were sorted by the difference in the average expression of Mes vs. Epi TFs. Those with a differential value above the median were assigned to the STAD Mes-like patient group, and the rest of the STAD patients were assigned to the STAD Epi-like group for this survival plot (and similarly for ACRG survival plot Supplemental Fig. 2G).

Supplemental Fig. S12) Changes in TF Expression in TCGA-STAD Samples with or without Driver Gene Mutation



X-axis: $\log_2(\text{TF RNA in STAD Mutant Sample} / \text{TF RNA in STAD WT})$

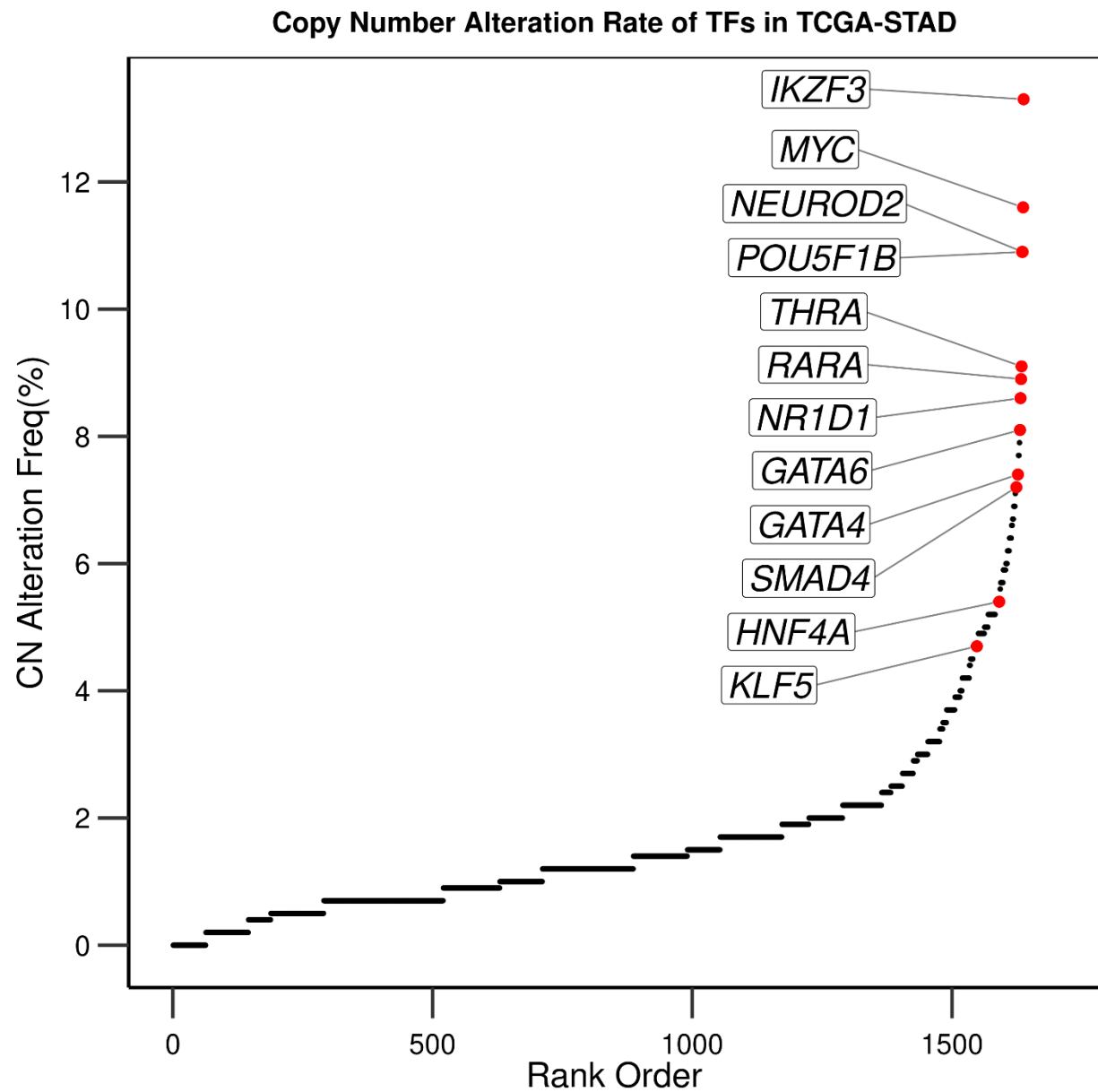
TF Expr. of *KMT2D*-Mutants VS WT in TCGA-STAD

TF Expr. of *KRAS*-Mutants VS WT in TCGA-STAD

TF Expr. of *CDH11*-Mutants VS WT in TCGA-STAD

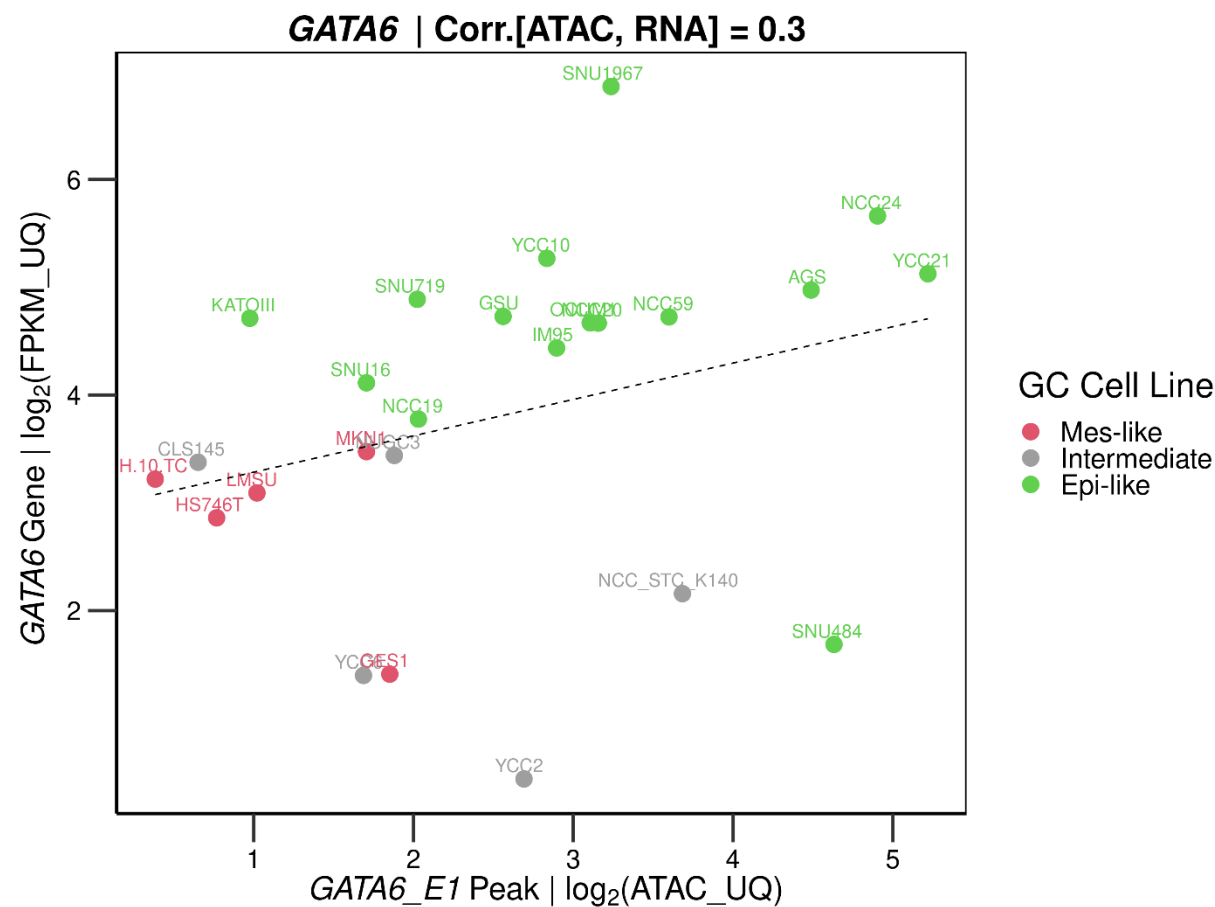
TF Expr. of *ERBB3*-Mutants VS WT in TCGA-STAD

Supplemental Fig. S13) Copy Number Alteration Rates in TCGA-STAD



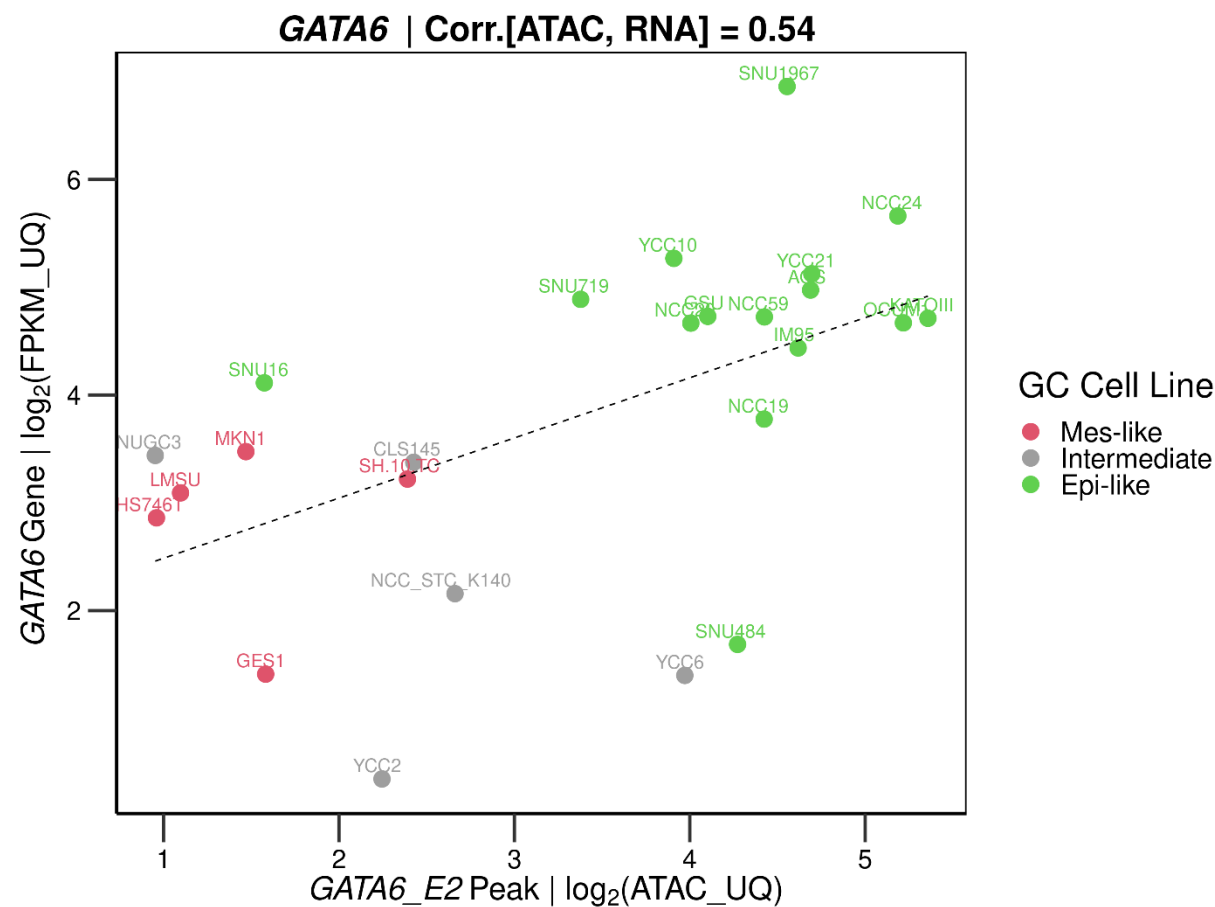
Copy number alteration = DNA copy number amplification rate + DNA copy number deletion rate

Supplemental Fig. S14) Correlation of GATA6 Enhancer (*E1*) Accessibility and *GATA6* Gene Expression



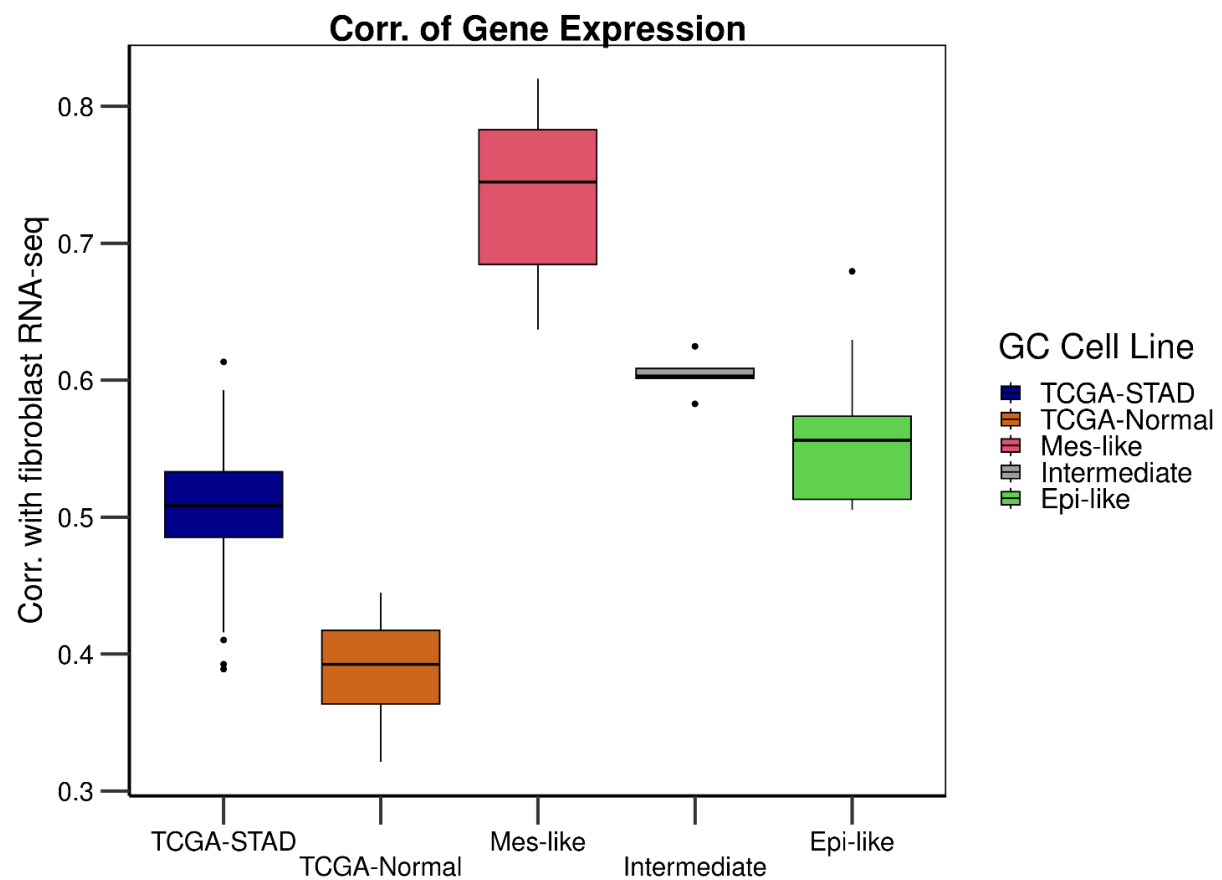
Pearson's correlation of [*GATA6* RNA, *GATA6_E1* ATAC] signals. ATAC-seq and RNA-seq values are upper-quartile normalized.

Supplemental Fig. S15) Correlation of GATA6 Enhancer (*E2*) Accessibility and *GATA6* Gene Expression



Pearson's correlation of [*GATA6* RNA, *GATA6_E2* ATAC] signals. ATAC-seq and RNA-seq values are upper-quartile normalized.

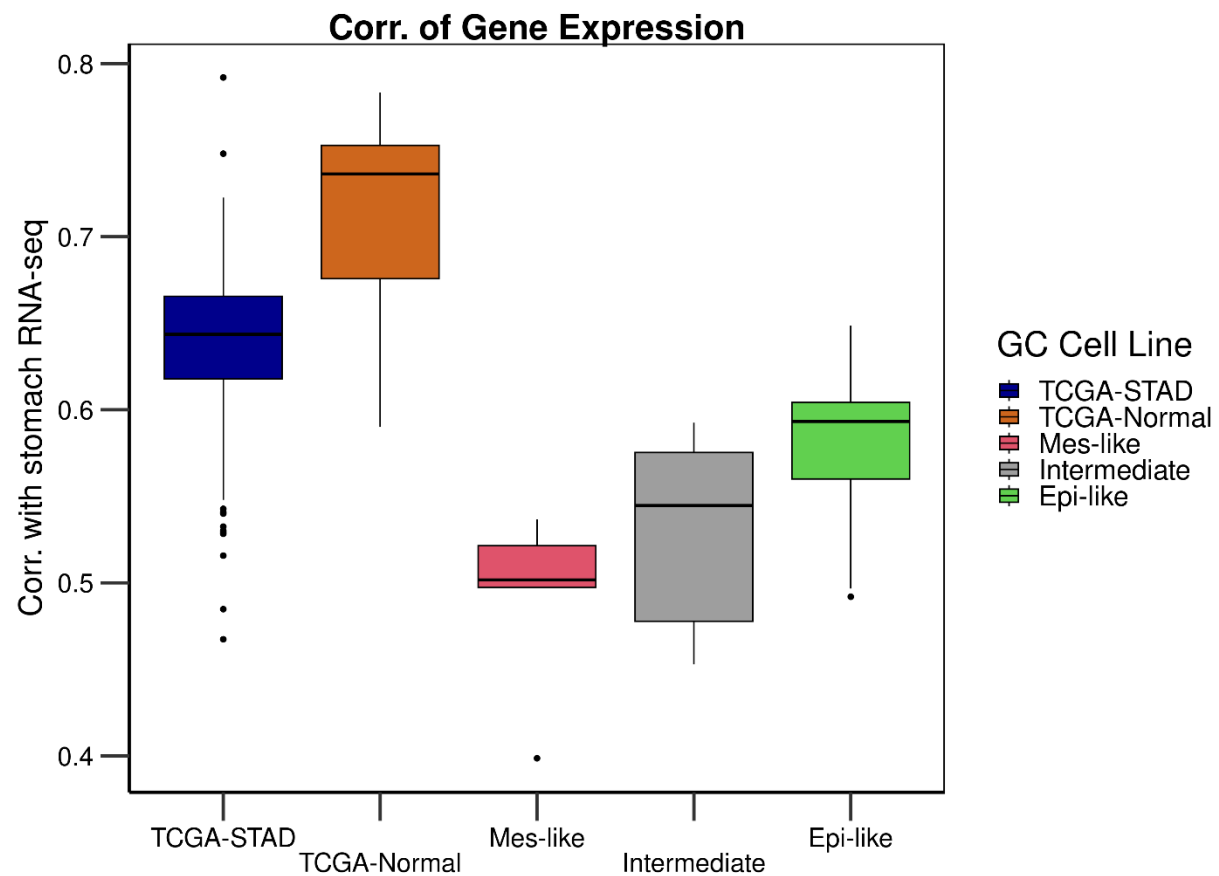
Supplemental Fig. S16) Correlation of RNA-seq for GC Cell Lines and TCGA-STAD with ENCODE Primary Lung Fibroblast



t-test p-values:

- (Mes vs. Epi) = 0.0037
- (TCGA-STAD vs. TCGA-Normal) = 1.2×10^{-10}

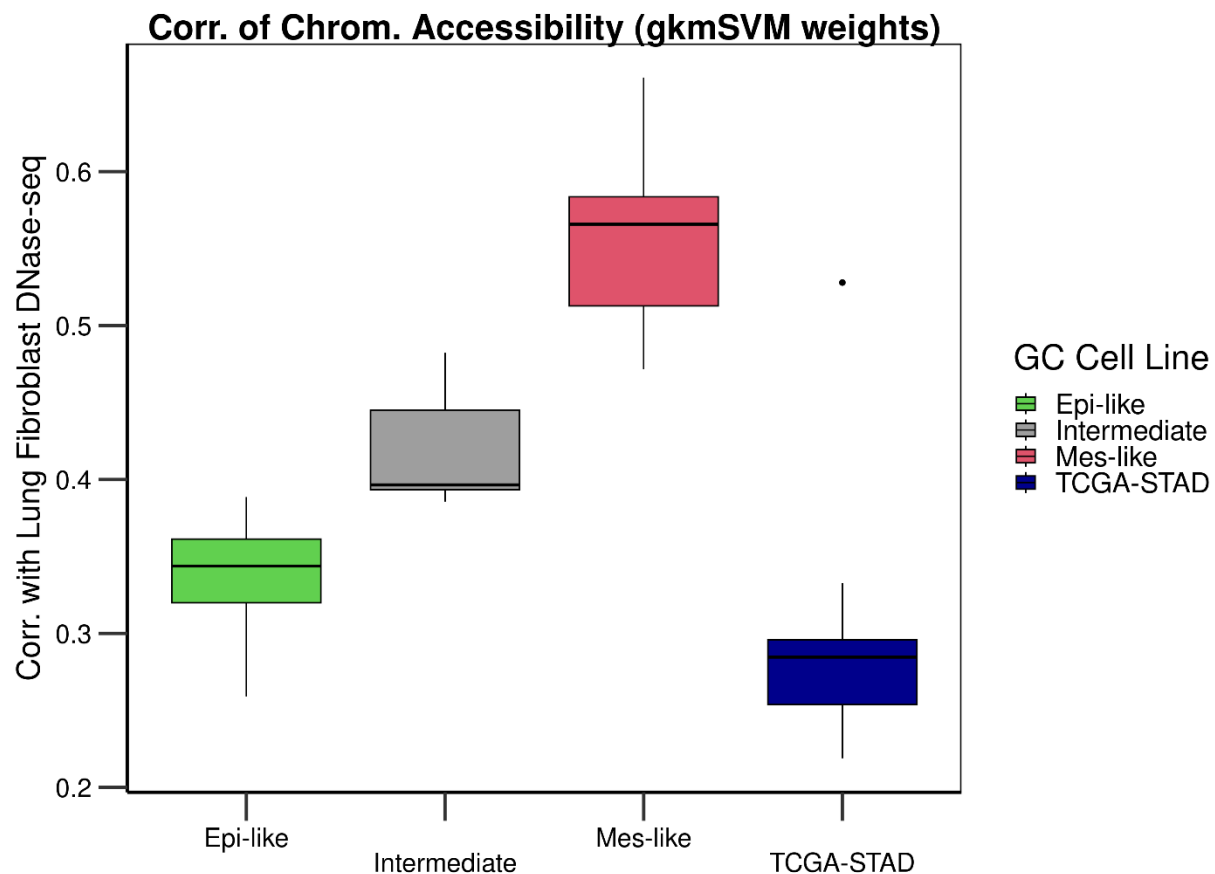
Supplemental Fig. S17) Correlation of RNA-seq for GC Cell Lines and TCGA-STAD with ENCODE Primary Stomach



t-test p-values:

- (Mes vs. Epi) = 0.0168
- (TCGA-STAD vs. TCGA-Normal) = 9.3×10^{-5}

Supplemental Fig. S18) Correlation of Chromatin Accessibility for GC Cell Lines and TCGA-STAD with ENCODE Primary Lung Fibroblast

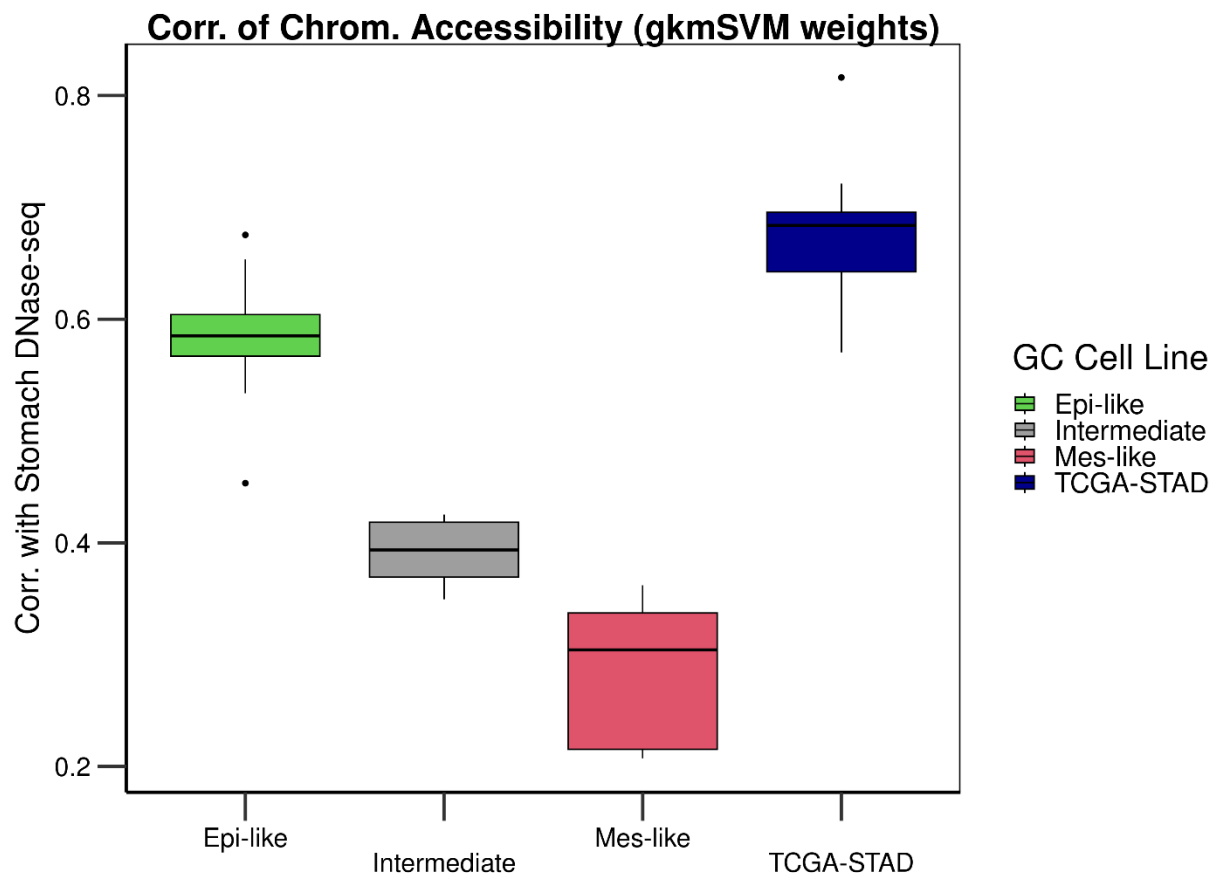


Chromatin accessibility is measured by ATAC-seq in the GC cell lines and with DNase-seq in the ENCODE healthy primary tissue. For each sample, gkm-SVM models were trained on distal enhancer DNA regions (peak length = 300). The gkm-SVM output is a weight vector for $(4^{11} / 2)$ k -mers ($k = 11$) which shows the overrepresentation and underrepresentation of that particular k -mer (transcription factor binding site) in the ATAC-seq or DNase-seq. Correlation of these weight vectors were used to measure similarity in the regulatory landscape and chromatin accessibility of different samples.

t -test p-values:

- (Mes vs. Epi) = 0.0014
- (TCGA-STAD vs. Epi) = 0.2396

Supplemental Fig. S19) Correlation of Chromatin Accessibility for GC Cell Lines and TCGA-STAD with ENCODE Primary Stomach

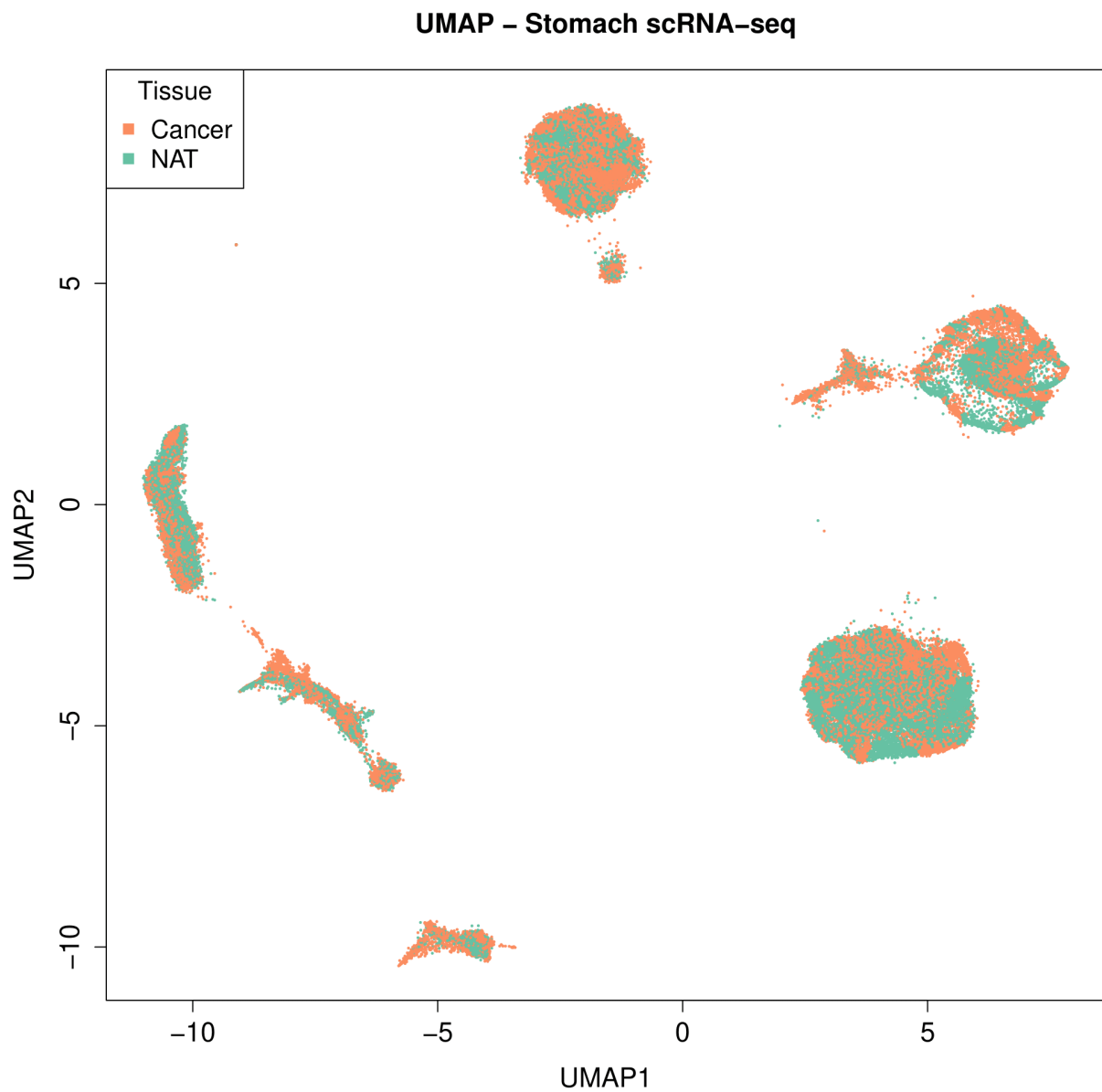


Correlation of gkm-SVM *k*-mer weight vectors was used to measure similarity in the regulatory landscape and chromatin accessibility of different samples.

t-test p-values:

- (Mes vs. Epi) = 0.00018
- (TCGA-STAD vs. Epi) = 0.0039

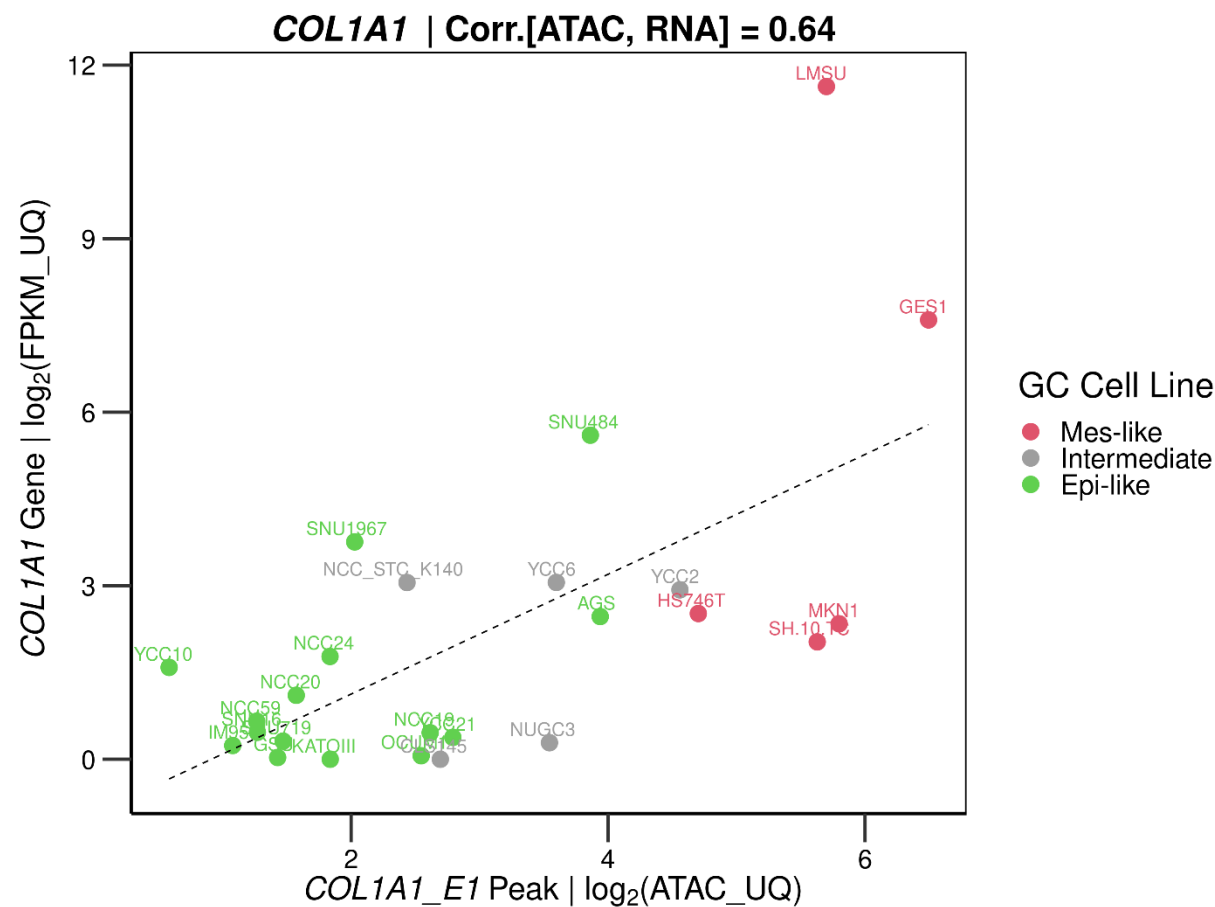
Supplemental Fig. S20) UMAP of scRNA-seq of GC Tumors and Normal Adjacent Tissue



NAT: Normal Adjacent Tissue

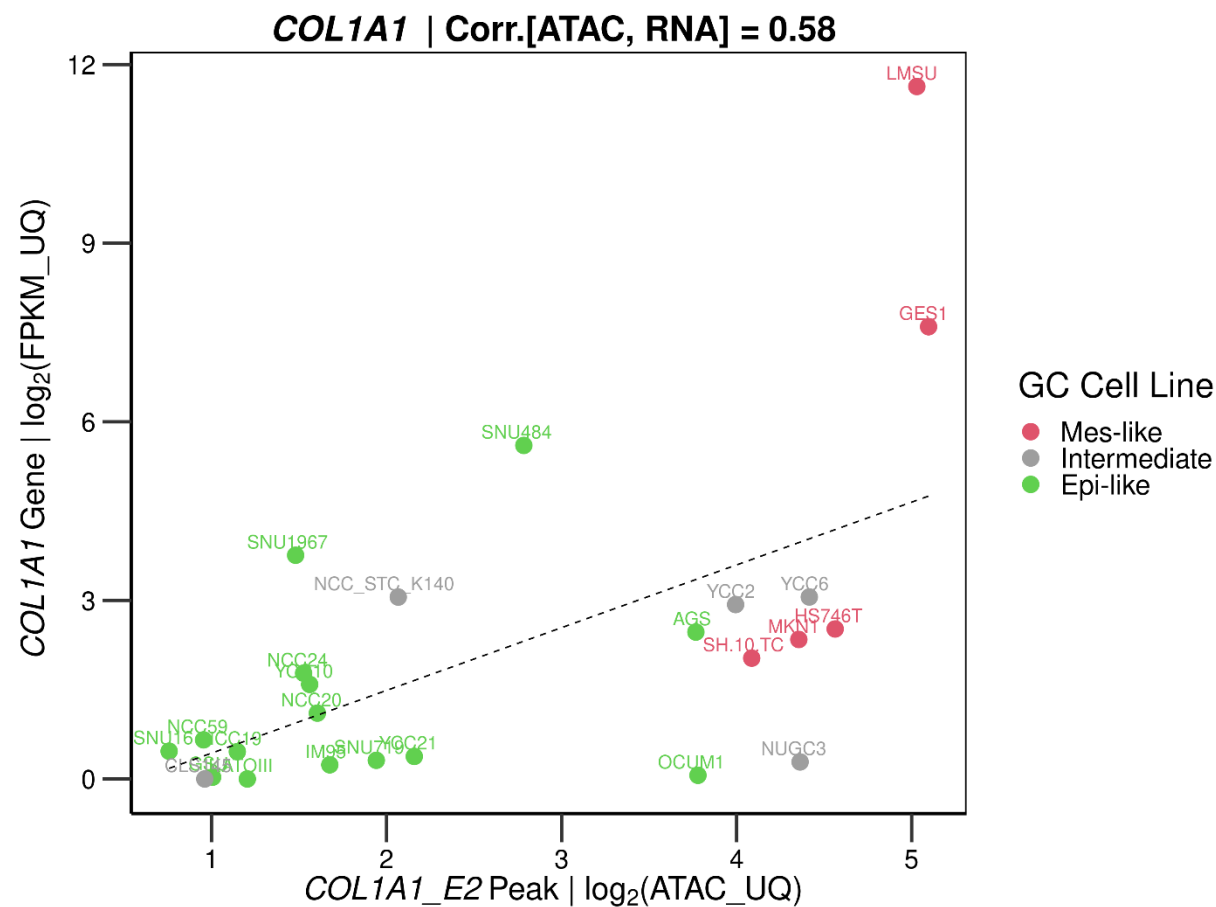
Cancer and NAT cells spread uniformly across the 6 single-cell groups detected by clustering over UMAP.

Supplemental Fig. S21) Correlation of COL1A1 Enhancer (*E1*) Accessibility and *COL1A1* Gene Expression



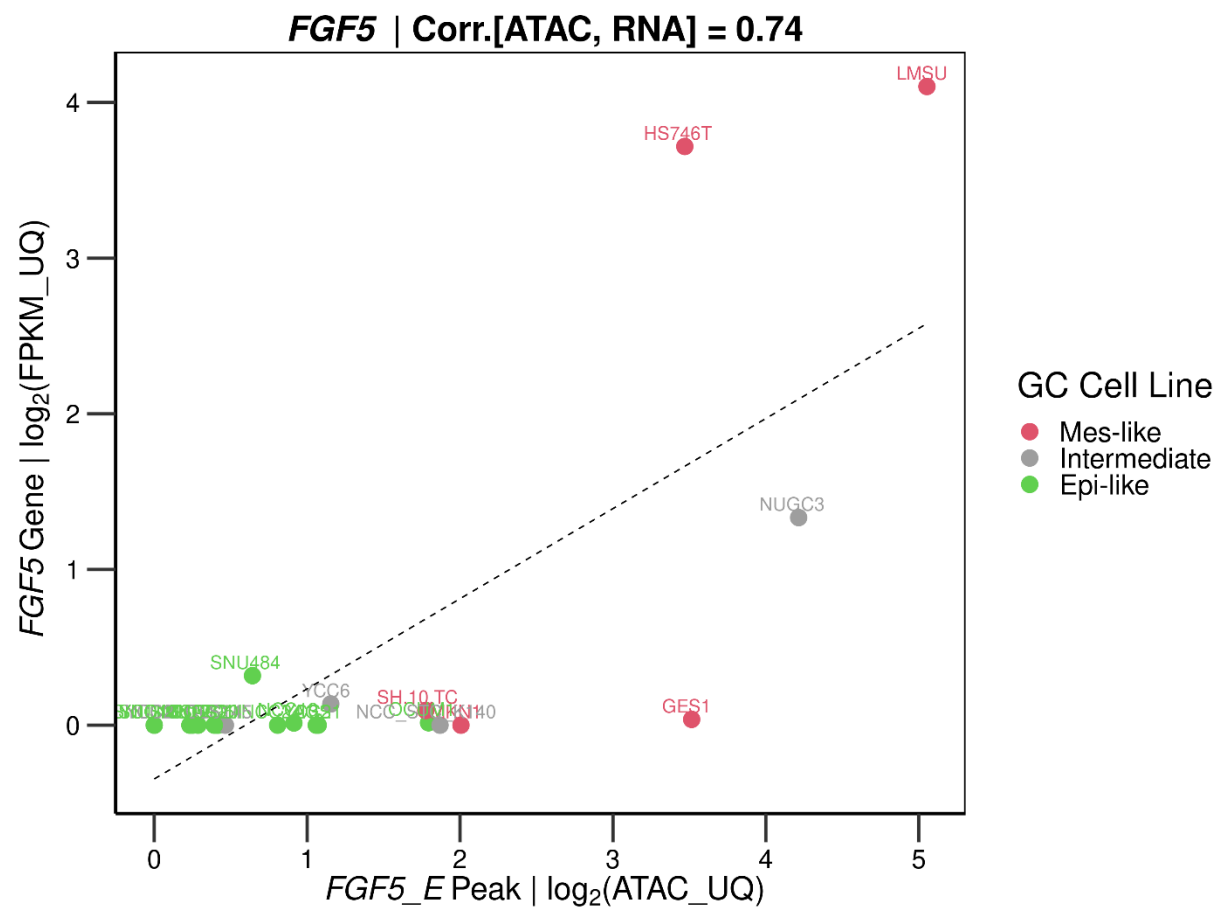
Pearson's correlation of [*COL1A1* RNA, *COL1A1_E1* ATAC] signals. ATAC-seq and RNA-seq values are upper-quartile normalized.

Supplemental Fig. S22) Correlation of COL1A1 Enhancer (E2) Accessibility and COL1A1 Gene Expression



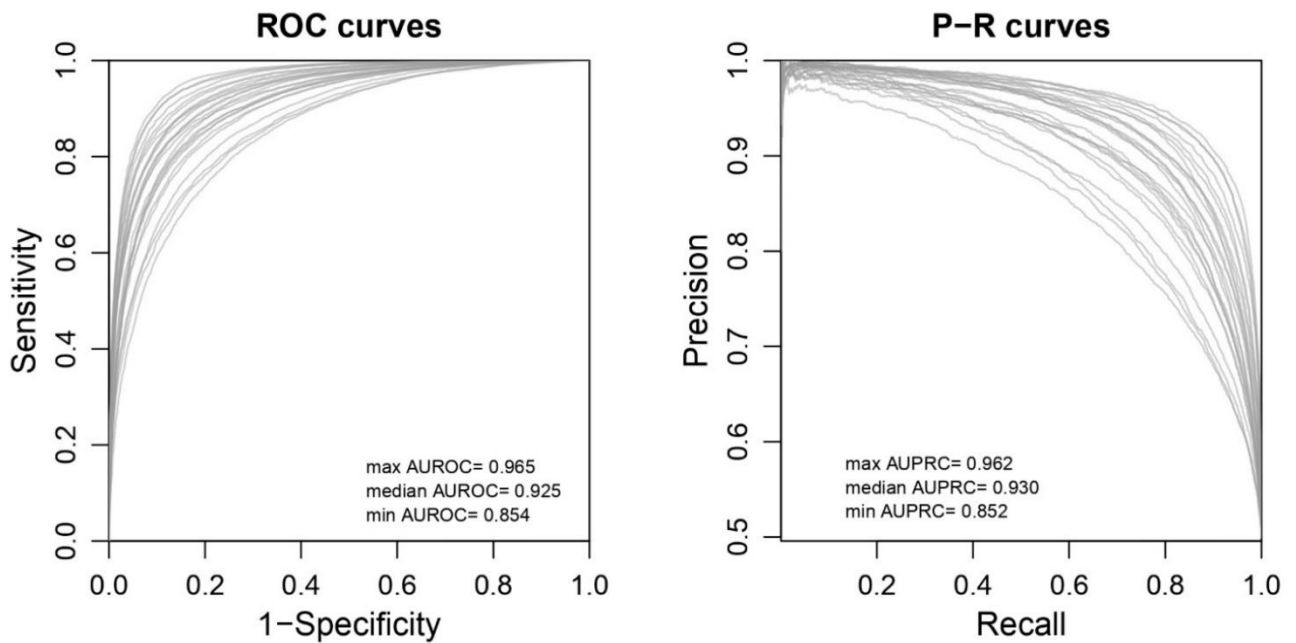
Pearson's correlation of [COL1A1 RNA, COL1A1_E2 ATAC] signals. ATAC-seq and RNA-seq values are upper-quartile normalized.

Supplemental Fig. S23) Correlation of FGF5 Enhancer (*E*) Accessibility and *FGF5* Gene Expression



Pearson's correlation of [*FGF5* RNA, *FGF5_E* ATAC] signals. ATAC-seq and RNA-seq values are upper-quartile normalized.

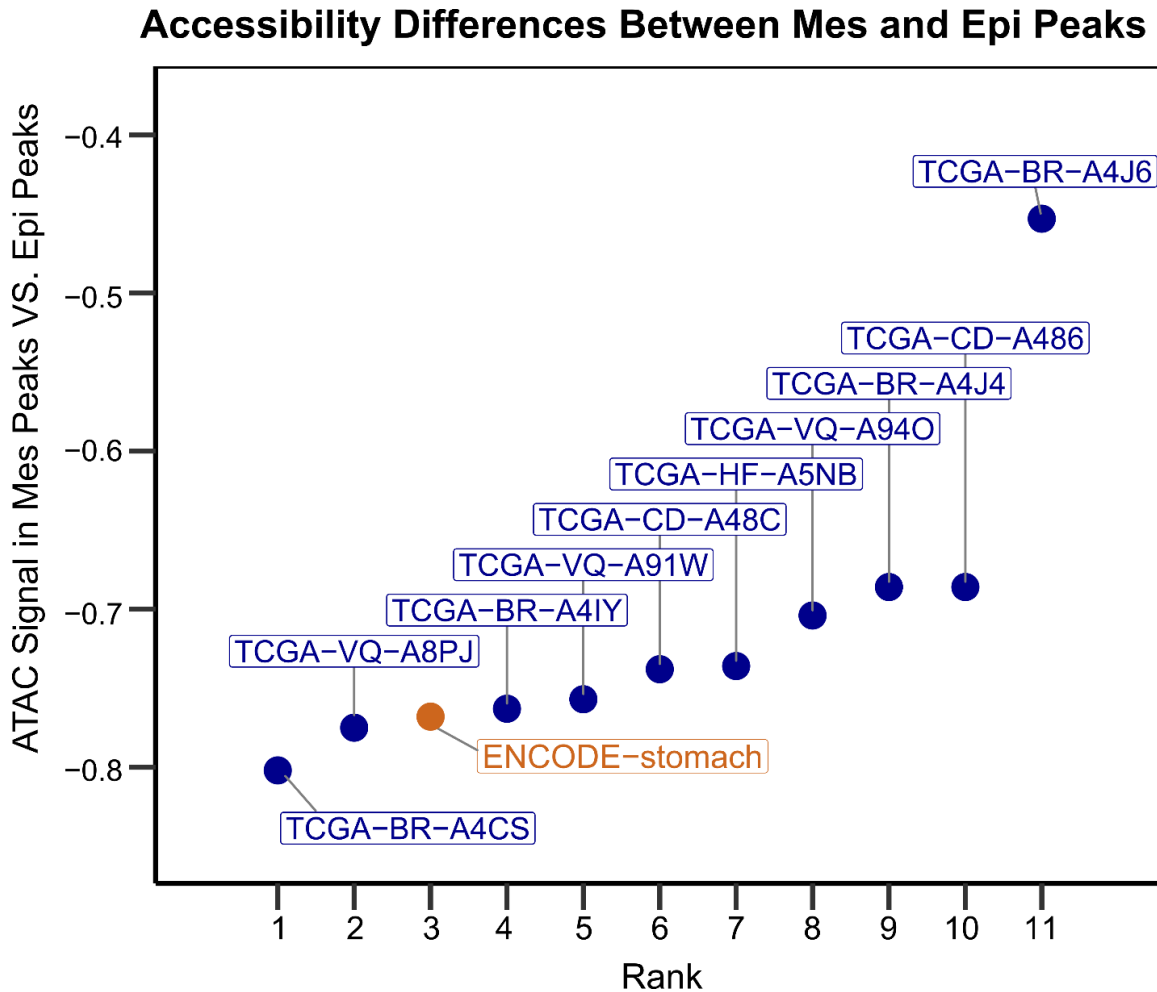
Supplemental Fig. S24) ROC Curves – GC cell lines VS. random GC-matched Genomic Regions



gkm-SVM models were trained on the top10k distal ATAC peaks of each cell line (n=25) vs. 10k random GC-matched genomic regions

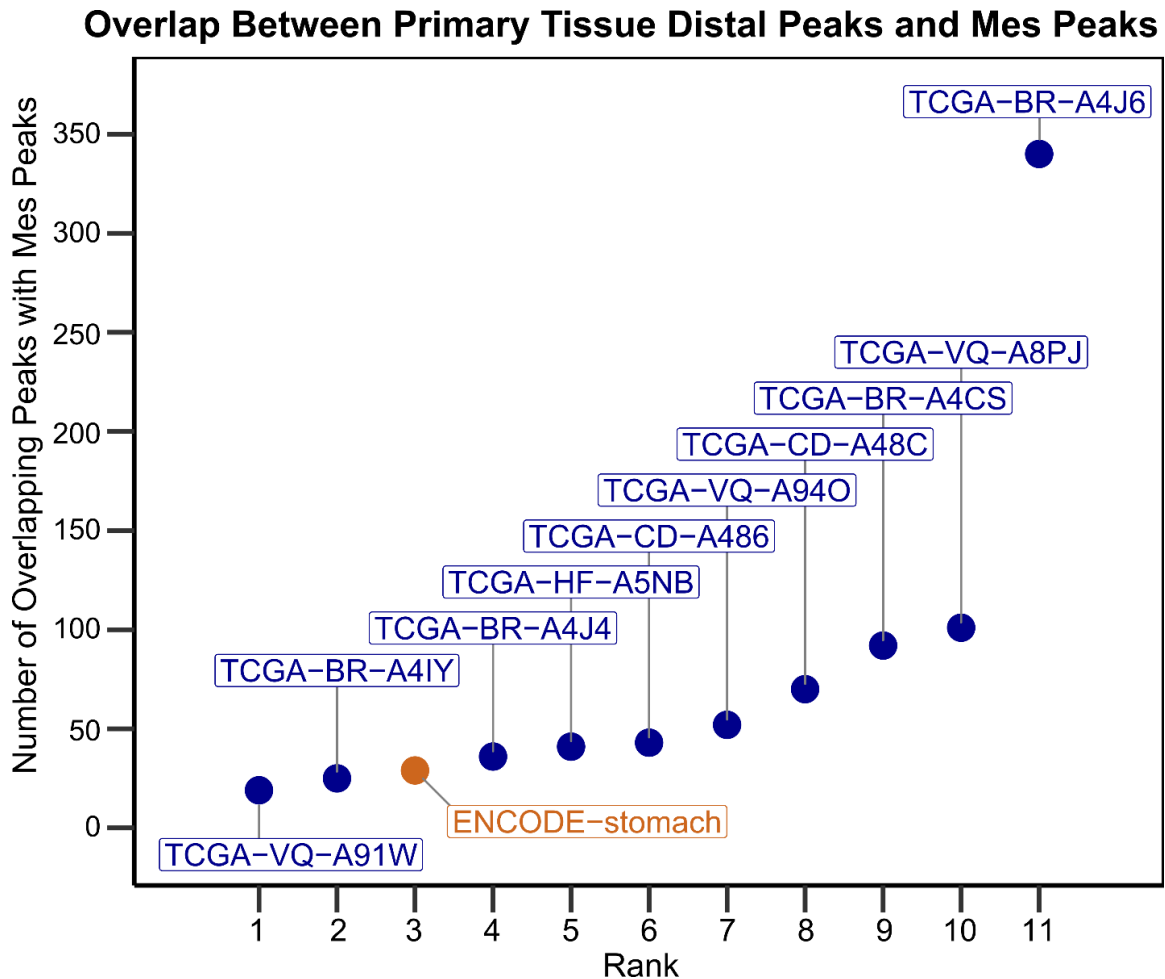
Peak length = 300bp

Supplemental Fig. S25) ATAC-seq Signal of Mes vs. Epi Peaks in TCGA-STAD



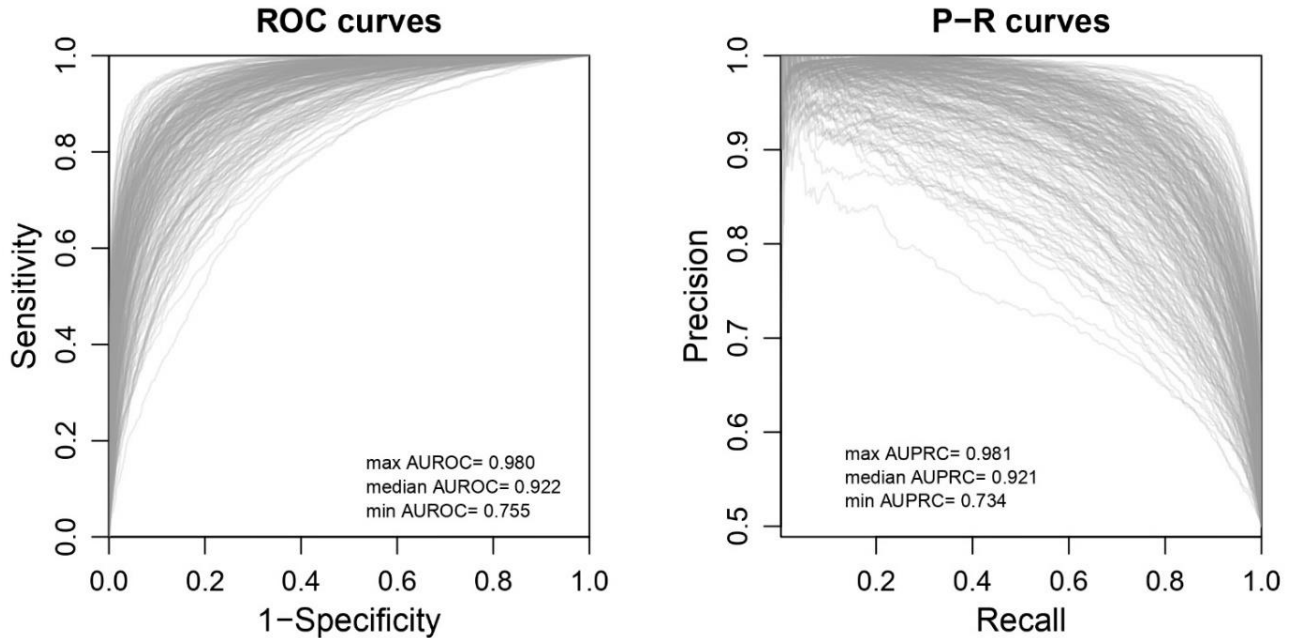
The Y-axis shows the ATAC signal values calculated over the 2000 Mes-high and 2000 Epi-high peaks as follows: $(\text{Mes} - \text{Epi}) / (\text{Mes} + \text{Epi})$, which is between $[-1, +1]$.

Supplemental Fig. S26) Number of Overlaps between Top TCGA-STAD ATAC Peaks and Mes-high Peaks



Y-axis shows the number of overlapping peaks between the top ATAC peaks in TCGA-STAD samples and the 2000 Mes-high peaks.

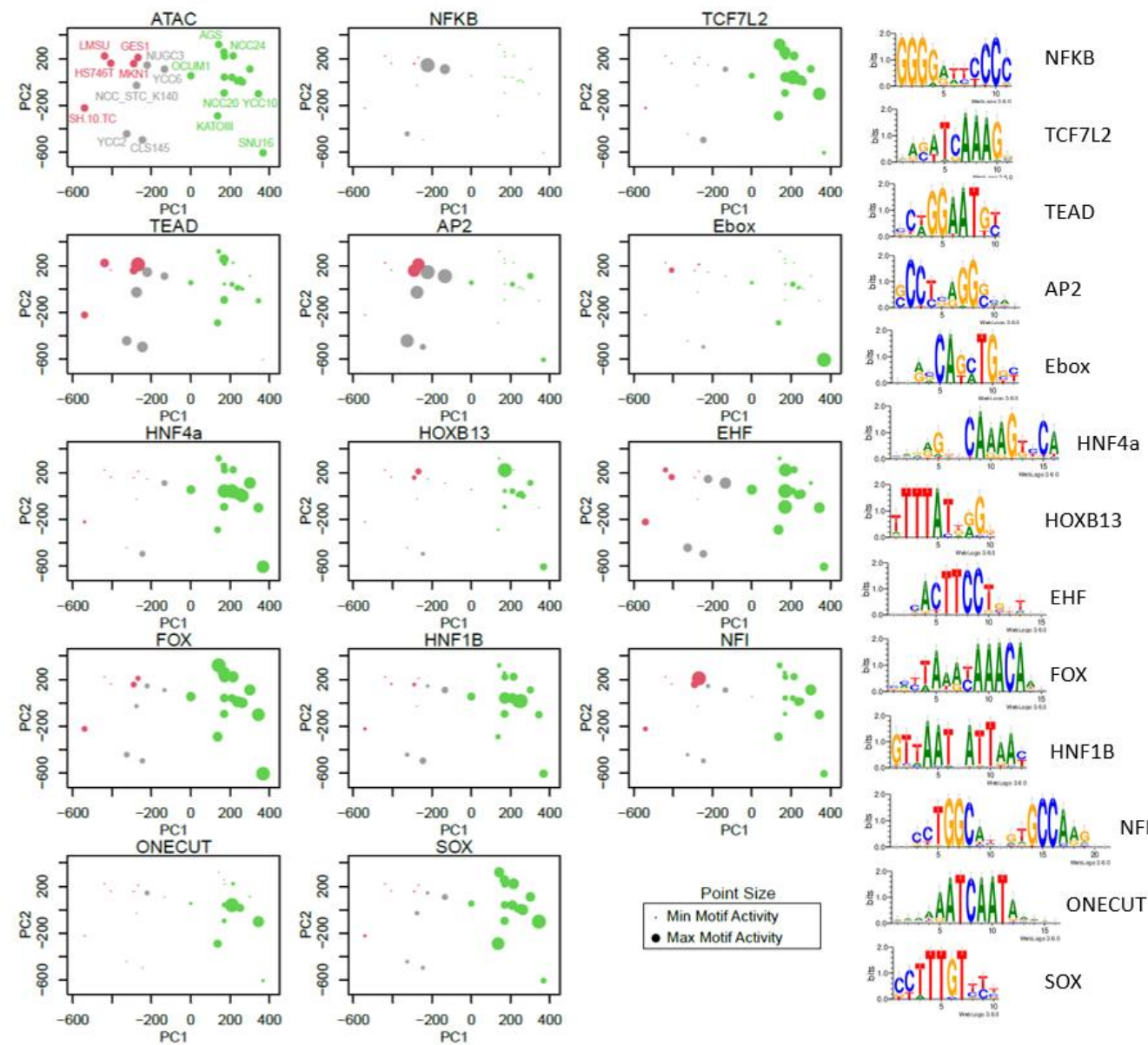
Supplemental Fig. S27) ROC Curves – Pairs of GC Cell Lines Trained Against Each Other



gkm-SVM models were trained on the most differentially accessible distal ATAC peaks (n=2000 positive peaks and n=2000 negative peaks) of each pair of GC cell lines (n=25) vs. all other GC cell lines (300 pairs = $(25 * 24) / 2$)

Peak length = 300bp

Supplemental Fig. S28) PCA of ATAC-seq | TF Motif Activity



gkm-SVM inferred activity (dot size) of TFs not shown in Fig 1E, across all samples.