Genome-wide profiling of active enhancers in colorectal cancer

Min Wu (wumin@whu.edu.cn)
Wuhan University https://orcid.org/0000-0003-1372-4764

Qinglan Li
Wuhan University

Xiang Lin
Wuhan University

Ya-Li Yu
Zhongnan Hospital, Wuhan University

Lin Chen
Wuhan University

Qi-Xin Hu
Wuhan University

Meng Chen
Zhongnan Hospital, Wuhan University

Nan Cao
Zhongnan Hospital, Wuhan University

Chen Zhao
Wuhan University

Chen-Yu Wang
Wuhan University

Cheng-Wei Huang
Wuhan University

Lian-Yun Li
Wuhan University

Mei Ye
Zhongnan Hospital, Wuhan University https://orcid.org/0000-0002-9393-3680

Article

Keywords: Colorectal cancer, H3K27ac, Epigenetics, Enhancer, Transcription factors

Posted Date: December 10th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-119156/v1

Version of Record: A version of this preprint was published at Nature Communications on November 4th, 2021. See the published version at https://doi.org/10.1038/s41467-021-26600-5.

Genome-wide profiling of active enhancers in colorectal cancer

Qing-Lan Li[1, #], Xiang Lin[1, #], Ya-Li Yu[2, #], Lin Chen[1, #], Qi-Xin Hu[1], Meng Chen[2], Nan Cao[2], Chen Zhao[1], Chen-Yu Wang[1], Cheng-Wei Huang[1], Lian-Yun Li[1], Mei Ye[2,*], Min Wu[1,*]

[1] Frontier Science Center for Immunology and Metabolism, Hubei Key Laboratory of Cell Homeostasis, Hubei Key Laboratory of Developmentally Originated Disease, Hubei Key Laboratory of Intestinal and Colorectal Diseases, College of Life Sciences, Wuhan University, Wuhan, Hubei 430072, China

[2]Division of Gastroenterology, Department of Geriatrics, Hubei Clinical Centre & Key Laboratory of Intestinal and Colorectal Diseases, Zhongnan Hospital, Wuhan University, Wuhan, Hubei 430072, China

[#]Equal contribution to the study.

Contact information

*Correspondence should be addressed to Dr. Min Wu, Email: wumin@whu.edu.cn, Tel: 86-27-68756620, or Dr. Mei Ye, Email: wumeiye08@163.com

Running Title

Enhancer profiling in colorectal cancer

Abstract

Colorectal cancer (CRC) is one of the most common cancers in the world. Although genomic mutations and SNPs have been extensively studied, the epigenomic status in CRC patient tissues remains elusive. Here, we profiled active enhancers genome-widely in paired CRC patient tissues through H3K27ac ChIP-Seq, together with genomic and transcriptomic analysis. Totally we sequenced 73 pairs of CRC tissues and generated 147 H3K27ac ChIP-Seq, 144 RNA-Seq, 147 whole genome sequencing and 86 H3K4me3 ChIP-Seq files. Our analysis discovered 5590 gain variant enhancer loci (VEL) and 1100 lost VELs in CRC, and 334 gain variant super enhancer loci (VSEL) and 121 lost VSELs. Multiple key transcription factors in CRC were predicted with motif analysis and core regulatory circuitry analysis. Further experiments verified the functions of 6 super enhancers governing *PHF19*, *LIF*, *SLC7A5*, *CYP2S1*, *RNF43* and *TBC1D16* in regulating cancer cell migration, and we identified KLF3 as a novel oncogenic transcription factor in CRC. Taken together, our work provides important epigenomic resource and novel functional factors for epigenetic studies in CRC.

Key words

Colorectal cancer, H3K27ac, Epigenetics, Enhancer, Transcription factors

Background

Binding of transcription factors (TFs) to enhancers is one of the critical steps in transcription activation. Recently, the development of epigenomics revealed the novel features of active and silent enhancers and shed light on the study of transcriptional regulation in multiple research fields [1-5]. Epigenetic marks on chromatin are important signatures for cell identification, which co-operate with transcription factors to regulate transcription [2,6,7]. Histone modifications mark enhancers on chromatin and are critical for their activity. H3K4me1 is the mark for primed enhancers [7,8]; H3K27ac for active enhancers and H3K27me3 for poised enhancers [1]. Though the initial discovery was concluded from ChIP-Seq of mediator subunits, now H3K27ac in the intergenic chromatin is widely used for identification of active enhancers [6,9,10]. Moreover, it was discovered that many genes are often regulated by multiple enhancers and the state of these enhancers varies in different cell types [1,4]. Therefore, it has emerged as critical questions for many fields how enhancer activity is regulated for signaling pathways and selective gene transcription.

Pioneer studies hypothesized that gain of enhancer activity is one of the common features for cancers [6,11-13], which is supported by some recent studies in patients and animal models [10,14,15]. However, it is still not clear whether it is a common feature for all the cancers or just a portion of them. Interestingly, many genes related with epigenetic regulation of enhancer activity are frequently mutated in cancer, such as MLL3/4, p300/CBP, UTX and KDM5C [12,16-20]. Moreover, inhibitors for BRD4, one reader of H3K27ac on enhancer, were shown to be effective in cancer treatment [21]. It is then urgent to clarify the roles of enhancers in cancer and the underlying mechanisms.

It has been shown that the enhancers controlling the transcription of key oncogenic genes, such as *MYC*, distinguish in different types of cancers [22]. It is probably because the active transcription factors vary in different cancer types, which causes that

different transcription factors activate and bind to different enhancers in response to variant genome mutations and upstream signals. Thus, enhancer profiling may reflect the feature of distinguished cancers and be used for classification [10,13,23-25].

Colorectal cancer (CRC) is one of the most common cancers in the world. Recent studies about aberrant DNA methylation have gain sight of the field, and epigenetic regulation becomes one of the critical regulatory factors for CRC [26-31]. Some groups have studied the genome wide distribution of active enhancers in CRC [23,32]. The early studies used H3K4me1 as a mark which was not suitable to identify functional active enhancers [23]. The recent study was mostly based on cell lines, only with very few patient tissues [32]. These data do not reflect the real clinical features of CRC patients and are not very helpful to enhancer studies in CRC.

To establish a comprehensive map for active enhancers in CRC, we performed H3K27ac ChIP-seq analysis for 73 pairs of CRC tissues (tumor tissues with paired adjacent native tissues), as well as the corresponding genomic and transcriptomic sequencing. We identified thousands of novel enhancers and multiple TFs involved in CRC, and a portion of them were experimentally verified. Our study provides important epigenomic resource and novel research candidates for future studies in CRC.

Results

Genome-wide study of enhancer distribution in CRC patient tissues

To establish a comprehensive genome-wide view of active enhancers of CRC patient tissues, we totally collected 80 pairs of tissues (cancer and their adjacent tissues) from CRC patients. We optimized the ChIP-Seq protocol and performed H3K27ac ChIP-Seq for these samples, as well as the corresponding mRNA and input DNA sequencing. Some samples failed in the study, and eventually we got high quality of sequencing data from 74 CRC tissues and 73 native tissues, among which 73 were

paired (Fig. 1A, Extended Data Fig. S1&Table 1). We also performed H3K4me3 ChIP-Seq for 43 pairs of tissues. Totally we generated 524 high quality sequencing files, including 147 H3K27ac, 86 H3K4me3, 144 RNA-Seq and 147 genomic sequencing files (Extended Data Table 2), which will provide important epigenomic information for related studies.

Our analysis revealed totally 27,156 significant enhancers in native tissues and 39,207 in tumor tissues, most of which were distributed on introns and intergenic regions as expected (Fig. 1B & Extended Data Fig. S2A), also including 9896 and 10663 promoters in native and tumor tissues, respectively (Extended Data Fig. S2A). The saturation analysis showed that the gained enhancers in tumor reached 80% when using less than 40 pairs of samples for analysis, and 90% with around 50 pairs, indicating the sample size used in our study was good enough for statistical analysis (Fig. 1C). We downloaded H3K4me1 (ENCODE, ENCFF557VIT) and BRD4 (GEO, GSM3593876) ChIP-seq data in CRC cell line HCT116 from public databases , and calculated the RPM values of H3K4me1 and BRD4 signal in significant CRC tumor enhancer loci. Our analysis showed that H3K27ac peaks of our study is nicely correlated with the published BRD4 and H3K4me1 signal (Extended Data Fig. S2B&C). Our RNA-Seq analysis identified 2226 up-regulated different expressed genes (DEGs) and 1979 down-regulated DEGs in CRC tumors (Fig. 1D). Compared with the TCGA data, many DEGs are overlapped (Extended Data Fig. S2D). The proportion of genes assigned with multiple enhancers was shown (Extended Data Fig. S2E). *MYC* is a well-known oncogene and H3K27ac track on its enhancer was shown as an example (Fig. 1E). In the adjacent native tissues, *MYC* expression was very low, and H3K27ac signal on its enhancer and eRNA were close to the background. In tumor tissues, *MYC* was highly expressed, together with the elevation of H3K27ac on its enhancer and eRNA level (Fig. 1E). One early study has reported active enhancers in multiple CRC cell lines and a few CRC samples [32]. Comparison of the two studies revealed that we identified 11796 new active enhancers in CRC, which was 32.4% of

the total enhancers (Fig. 1F&G).

Identification of variant enhancer loci in tumor

To identify significant active enhancers specific in tumors, we first compared the enhancers of all samples and found that some tissues had relatively low number of H3K27ac peaks (less than 2,500) or variant enhancer loci (VEL, less than 500) compared with the corresponding adjacent tissues (Extended Data Fig. S3A&B). We considered it may due to the sampling problem and ruled them out in the following statistical analysis. We totally identified 6690 significant VELs, including 5590 gain VELs and 1100 lost VELs (Fig. 2A, Extended Data Table 3-5) and the pipeline was shown in Extended Data Fig. S3C. At the recurrence threshold of 14 and 19 patients, 95% of the VEL candidate achieved statistical significance (q-value < 0.1, paired t-test, with Benjamini-Hochberg correction; Fig. 2B). Supporting the reliability of these analysis, significant gain VELs exhibited higher H3K27ac level in tumors than native tissues, and opposite in significant lost VELs (Extended Data Fig. S3D&E). Meanwhile, genes associated with gain VELs showed elevated expression in tumors than native tissues, while lost-VEL-associated genes were broadly repressed, and the magnitude of the change in expression positively correlated with the number of VELs per gene (Extended Data Fig. S3F). The gain VELs close to *IL20RA* and *FOXQ1* and the lost VELs close to *PPARGC1B* were shown as representative (Fig. 2C, Extended Data Fig. S3G&H). The identified VELs could nicely distinguish the native and tumor tissues (Fig. 2A). Human disease ontology and GO analysis showed that the associated genes of gain VELs were highly related with CRC (Fig. 2D & Extended Data Fig. 3I), while those of lost VELs were related with normal colon functions (Extended Data Fig. 3J). To further evaluate the potential of H3K27ac or enhancer information in distinguishing tumor and normal tissues, we clustered the adjacent native and tumor tissues using the information of gene expression, H3K27ac on enhancers, and H3K4me3 on promoters. The adjacent and tumor samples were nicely distinguished using both the gene expression or significant enhancers in PCA analysis,

while the different H3K4me3 peaks did not (Fig. 2E-G), suggesting enhancer information is useful for tumor identification.

The enhancer features of CRC subgroups

To further test the value of enhancers in tumor classification, we utilized one of the common approaches, the consensus molecular subtypes (CMS) classification of CRC tumors [33], and classify patients into four subgroups (Fig. 3A & Extended Data Fig. S4A-C). The correlation analysis based on the identified VELs and subgroup specific VELs showed that the tissues of CMS2 group had the highest correlation, while CMS4 was the lowest (Fig. 3B&C, Extended Data Fig. S5), suggesting CMS4 might be more heterogenous than others. Interestingly, when comparing the enhancers among the four subgroups, we found that CMS2 had the largest number of active enhancers, significant gain VELs and specific gain VELs (Fig. 3D-F, Extended Data Fig. S6A). The average H3K27ac signal of gain VELs in CMS2 was also higher than the other three (Fig. 3E). While, the four groups had no big difference at the amount of gain VELs among individual samples, as well as the correlation of H3K27ac across genome (Extended Data Fig. S6B-D). The above study indicates that CMS2 group is more homogenous than others; and it has more specific active enhancers, which might be a novel feature for it.

Then, we studied the function of the VEL-associated genes, and identified the enhancers and genes specifically activated in each subgroup. Some representative enhancers and genes for each group were shown (Extended Data Fig. S6E-L). For CMS2, we found its specific gain-VEL-associated genes are mainly involved in WNT signaling, cell migration and lipid metabolic process (Fig. 3G). Activation of WNT signaling and enhanced cell migration are expected, since *APC* is one of the most frequent mutated genes in CRC and cell migration is a hallmark for cancer cells [34,35]. Lipid metabolism was linked with CRC but ambiguous results from different groups exist [36-38]. Our analysis suggested dysregulation of lipid metabolic homeostasis is

possibly associated with certain CRC subgroup. Some VELs of CMS2 subgroup and their associated genes were shown, including those involved in lipid metabolism, such as *CEL* and *DPEP1* (Fig. 3H, Extended data Fig. S7).

Analysis and verification of variant super enhancer loci

Activation of super enhancers associated with oncogenes is considered as one of the important features for cancer [6]. Using the similar approaches as VEL identification, we identified the variant super enhancers loci (VSEL) in tumor tissues, including 334 gain VSELs and 121 lost VSELs, among which several well-known oncogenic targets were identified, such as *MYC*, *VEGFA*, and *LIF* (Fig. 4A, Extended Data Fig. S8A&B, Table 6&7). H3K27ac level on the gain VSELs were significantly increased and decreased on the lost VSELs as expected (Fig. 4B&C). We utilized the H3K27ac level on the identified VSELs to cluster the CRC patients, together with some normal intestinal tissues (Extended Data Fig. S8C&D). The analysis distinguished the native and tumor tissues, and classified CRC patients into three subgroups (Extended Data Fig. S8C&D). The results suggest VSELs might be useful for CRC classification.

To experimentally verify the functions of the identified VSELs, we compared the H3K27ac profiles on top gain VSELs of CRC tissues with those in HCT116 cells. The gain VSELs appearing in HCT116 were chosen and the dCas9-KRAB system was utilized to repress enhancer activity [39]. The repression of VSELs successfully reduced the expression of their neighbor genes, including *IER3*, *LIF*, *SLC7A5, CYP2S1, PHF19, RNF43, CEBPB, TBC1D16, TNFRSF6B* and *VEGFA* (Fig. 4A, Fig. 4D&E, Extended Data Fig. S9). For some enhancers, the expression of multiple close genes was repressed (Extended Data Fig. S9). We also measured H3K27ac level on some loci and found that dCas9-KRAB/sgRNA effectively repressed H3K27ac on the enhancers of *CEBPE*, *CYP2S1*, *IER3*, *PHF19*, *RNF43* and *TBC1D16* (Extended Data Fig. S10A). The chromatin interaction between promoters and super enhancers faraway from genes, such as *CEBPB*, *VEGFA* and *CYP2S1,* were confirmed with 3C

assay (Extended Data Fig. S10B). Moreover, we established stable cell lines of repressed enhancers in HCT116 and studied their proliferation and migration ability. The difference of proliferation was not very significant (Data not shown), however, quite a few cell lines exhibited attenuated migration ability, including *PHF19, LIF, SLC7A5, CYP2S1, RNF43, VEGFA* and *TBC1D16* (Fig. 4E).

Predication and verification of functional transcription factors

To investigate the potential transcription factors (TFs) playing key roles in CRC, the DNA sequences of VELs was used for prediction by HOMER software. The top hits of gain and lost VELs were listed (Fig. 5A and Extended Data Fig. S11A&B). The hypothesis of core regulatory circuitry was raised to identify core TFs in cells [40,41]. To improve TF prediction, we utilized the method to identify key TFs in CRC tissues (Fig. 5B&C, Extended Data Fig. S11C). ASCL2 was predicted as a CRC specific TF with the highest score (Fig. 5C). The H3K27ac level of *ASCL2* enhancer greatly increased in tumors (Extended Data Fig. S11D), and the gene expression analysis based on TCGA datasets suggested *ASCL2* was highly expressed specific in colorectal cancer (Extended Data Fig. S11E). These suggest ASCL2 is a key TF in CRC and further experiments are required to verify our prediction.

Combining the above results and the published literatures, we selected 4 TFs for experimental verification, including KLF3 and MAFK, two novel TFs, MAZ and RUNX1, two recently reported TFs functioning in CRC but not well characterized [42-45],. The significance of the selected TFs in each patient was calculated (Extended Data Fig. S11F). Knockdown of these genes did not affect cell proliferation in cell proliferation assay; while *KLF3, MAZ* and *RUNX1* knockdown repressed, and *MAFK* knockdown did not affect cell migration (Fig. 5D-F, Extended Data Fig. S12). Our results identified KLF3 as a novel TF involved in CRC, and confirmed the reported roles of RUNX1 and MAZ. MAFK was found at the side of native tissues in core regulatory circuitry analysis (Fig. 5B), and our experimental results did not find its

role in proliferation or migration, suggesting MAFK not involved in CRC although it was predicted in DNA motif assay.

Discussion

CRC is one of the most common cancers in the world. Though early screening greatly improves the curative ratio, novel classification approaches and drugs are still urged to be developed. The current study provides a comprehensive map of H3K27ac and active enhancers in CRC patients. The early studies used cell lines to determine the enhancers [32]. Our study uses paired patient tissues, which provides much more reliable data and identifies many novel CRC specific enhancers. Moreover, we experimentally confirmed the roles of more than 10 SEs in CRC. These provide important information for future CRC research.

Our analysis predicted many TFs functioning in CRC, some of which have never been reported before, such as KLF3. We confirmed the function of KLF3 in HCT116 cells, which supported its oncogenic role in CRC. H3K27ac on ASCL2 is highly increased in tumor tissues and it is specific high expressed in CRC. Further studies about these TFs will provide important information for CRC research.

Interestingly, verification of the identified VSELs and TFs indicated that most of them were more related with cell migration, but less with proliferation. The CRC samples we collected are mostly at relative late stages (Extended Data Fig. S1A & Table 1), and the tumor cells were probably at the metastasis stage or ready for it. The difference of enhancers and genes governing migration was probably much more significant than other genes between paired tissues, and the chosen VSELs were all among the most significant ones.

It was believed that gain of H3K27ac on the oncogene enhancers is a common feature for cancers [21]. Our analysis indicated that only the CMS2 group has the obvious

feature of enhancer activity elevation, and the other subgroups have much less gained VELs. Our data indicate that in CRC, global increase of active enhancers is an important feature for just one subgroup, not for all.

The homeostasis of lipid metabolism has been linked with CRC for many years, however, the detailed mechanisms is not clear and the use of statin analogues failed in CRC treatment [36]. We found that H3K27ac significantly increased on the enhancers of genes related with lipid metabolism. The bioinformatics analysis also pointed out that the gain VELs of CMS2 subgroup were enriched with genes involved in lipid metabolic processes. So, it is possible that the dysregulation of lipid homeostasis is only associated with CMS2 group, which should be explored by the future studies.

Methods

Reagents and cell lines

Antibodies recognizing H3K4me3 (Millipore, EMD-04-745) and H3K27ac (Abcam, ab4729, RRID: AB_2118291), were purchased from indicated commercial sources. Protein G-Sepharose beads were from GE Healthcare. PCR primers were custom synthesized by BGI and siRNAs by GenePharma. HCT116 Cell line was purchased from Cell Bank of Chinese Academy and cultured under recommended conditions according to the manufacturer's instruction with 10% FBS.

ChIP assay and ChIP-sequencing

ChIP assay was performed as previously described [46]. Briefly, around sixty milligrams of each tissue were cut into 1 mm^3 pieces in PBS with protease inhibitor. Tissue pieces were cross-linked for 10min at room temperature with 1% formaldehyde and then quenched with 0.125 M of glycine for 5 min. Cross-linked tissues were triturated for 30s and then centrifuged at 12000 rpm, 4 °C for 5 min. Supernatant with massive oil was discarded and the precipitates were lysed with 1 mL lysis buffer (50 mM Tris-HCl pH 8.0, 0.1% SDS, 5 mM EDTA) for 4 min with gentle rotation. After

centrifugation at 12000 rpm, 4 °C for 2 min, the pellet was washed once with digestion buffer (50 mM Tris-HCl pH8.0, 1 mM CaCl2, 0.2% Triton X-100), incubated in 630-μL digestion buffer with 1 μL MNase (NEB, M0247S) at 37°C for 20min and quenched with 8 μL 0.5M EDTA. The resulted mixture was sonicated and the pellet was discarded after centrifugation. 30 μL supernatant was taken for checking the efficiency of digestion. Immunoprecipitation was performed with 150 μL sheared chromatin, 2 μg antibody, 50 μL Protein G beads and 800 μL dilution buffer (20 mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.1% SDS) overnight at 4 °C. Next day, the beads were washed once with Wash buffer I (20 mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.1% SDS), once with Wash buffer II (20mM Tris-HCl pH 8.0, 500 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.1% SDS), once with Wash buffer III (10mM Tris-HCl pH 8.0, 250 mM LiCl, 1 mM EDTA, 1% Na-deoxycholate, 1% NP-40) and twice with TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA). The beads were eluted twice with 100 μL elution buffer (1% SDS, 0.1M NaHCO3, 20mg/mL Proteinase K) at room temperature. The elution was incubated at 65 °C for 6 h and then purified with DNA purification kit (TIANGEN DP214-03). Primers for ChIP-qPCR were listed in Extended Data Table 8.

ChIP-seq libraries were constructed with ChIP and input DNA using VATHS Universal DNA Library Prep Kit for Illumina (Vazyme ND606). Briefly, 50 μL of DNA (8-10 ng) was end-repaired for dA tailing, followed by adaptor ligation. Each adaptor was marked with a barcode of 8 bp DNA. Adaptor-ligated DNA was purified by AMPure XP beads (1:1) and then amplified by PCR of 9 cycles with the primer matching with adaptor universal part. Amplified DNA was purified again using AMPure XP beads (1:1) in 35 μL EB elution buffer. For multiplexing, libraries with different barcode were mixed with equal molar quantities (30-50 million reads per library). Libraries were sequenced by Illumina Nova-seq platform with pair-end reads of 150 bp.

RNA-sequencing

Around 40 mg tissue was used for RNA extraction using Ultrapure RNA Kit (CWBIO, CW0581M). Briefly, tissues were triturated for 30s in 1 mL TRIzon provided in the kit, incubated at room temperature for 5 min, added with 200 μL chloroform and shaken drastically. After centrifugation at 12000 rpm, 4 °C for 10 min, the upper water phase was moved into an adsorption column provided by the kit. The column was then eluted with 50 μL RNase-free water. RNA-seq libraries were constructed by NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB E7490) and NEBNext Ultra II Non-Directional RNA Second Strand Synthesis Module (NEB E6111). Briefly, mRNA was purified with poly-T magnetic beads and first and second strand cDNA was synthesized. The resulted cDNA was purified by AMPure XP beads (1:1) and eluted in 50 μL nucleotide-free water. The subsequent procedures were the same as described in ChIP-seq library construction, except that the sequencing depth was 20 million reads per library. RNA-seq libraries were sequenced by Illumina Nova-seq platform with pair-end reads of 150 bp.

ChIP-seq data processing

The adaptor sequence was removed using Cutadapt (version 1.16) to clean ChIP-seq raw data. Cleaned reads were mapped into human reference genome (hg19) using BWA (version 0.7.15) with default settings. Peak calling for tissues was performed by MACS2 with a p-value threshold of 1E-10. The patients with the peak number less than 2,500 were excluded from further analysis, no matter in native or tumor tissue (Patient 20, 21, 22 and 24 were excluded).

We calculated the normalized RPM as the ChIP-seq signal in specific region. Briefly, ChIP-seq reads aligning to the region were extended by 200 bp and the density of reads per bp was calculated using Python package HTSeq (version 0.9.1). The density of reads in each region was normalized to the total number of million mapped reads,

producing read density in units of reads per million mapped reads per bp (RPM per bp).

Plotting meta representation of ChIP-seq signal

Considering the sample number of our patient data, we utilized a way of calculating the mean to compactly display the integrated H3K27ac ChIP-seq signal in specific groups. For an individual region, we calculated the aligned read number per bp within this region using the R package HTSeq mentioned above, and then normalized to RPM. H3K27ac ChIP-seq signal is smoothed using a simple spline function and plotted as a translucent shape or a line in units of RPM per bp.

RNA-seq data processing and DEG identification

The adaptor sequence was removed using Cutadapt (version 1.16) to clean RNA-seq raw data. Cleaned reads were aligned to the human reference genome (hg19) using HISAT2 (2.1.0) with default settings. Uniquely aligned reads were counted at gene regions using the package featureCounts (version 1.4.6) based on Gencode v19 annotations. Differential gene expression analysis between native and tumor tissue was performed using the R/Bioconductor package DESeq2 (version 1.26.0) with contrast adjustment for multiple groups comparison. Genes whose log2FC < 1 and p.adj < 1E-2 were identified as differential expressed genes (DEGs).

Promoter, enhancer and super enhancer analysis

For both H3K4me3 and H3K27ac ChIP-seq data, peaks that could not be identified in at least two same kind of tissues were excluded from further analysis. H3K4me3 peaks located within the region surrounding ±2.5 kb of transcriptional start sites (TSS) were identified as promoters; and H3K27ac peaks away from the ±2.5 kb flank region of TSS were identified as enhancers. The promoters and enhancers of each samples were merged into one single set. Super enhancers were identified as following: firstly, super enhancers (ROSE) algorithm was used to classify and rank

sets of two or more H3K27ac peaks (detected by MACS2, p-value < 1E-10) within 12.5 kb distance and further than 2.5 kb from a transcriptional start site; secondly, a plot was graphed and a tangent line of the curve was drawn with the slope value of 1; finally, the enhancers above the point of tangency were defined as super enhancers. HOMER module annotatePeaks.pl was used to calculate the number of enhancers located in different chromatin elements.

Identification of VELs

To identify the significant VELs between native and tumor tissue, we first identified all VELs in paired native and tumor tissues. Individual sample VEL were defined as enhancers whose H3K27ac fold change (FC) was larger than > 2 between native and tumor tissues. The patients with VEL number (GAIN + LOST VELs) less than 500 were excluded from further analysis. We merged all VELs into one single coordinate file, and calculated the recurrence and significance (Benjamini–Hochberg corrected p-value) for all VELs. We used recurrence of 14 and 19 as significance threshold for gain and lost VELs, respectively, because gain and lost VELs achieved the significant percentage cut-off (0.95) when recurrence larger than these numbers.

Identification of VSELs

For variant super enhancer loci (VSEL), the identifying procedure was similar as described above in “Identification of VELs”. If the VSELs number in individual patient was less than 10, the patient would be excluded from further analysis (Patient 52 and 67 were excluded). And the significant percentage cut-off was changed to 0.9.

Identification of genes associated with VSELs

SE-associated Genes were identified by rose2 (https://github.com/linlabbcm/rose2) software and all these genes were merged into a single list. We considered the variation of a SE-associated gene by calculating the variant recurrence generated by its recurrence in tumor minus which in native tissue.

PCA analysis

We performed PCA analysis for gene expression, enhancer H3K27ac and promoter H3K4me3 in native and tumor tissues. For gene expression, we quantified sequencing fragments as reads per kilobase per million (FPKM) in each sample. And for two ChIP-seq signals, we used RPM. R package FactoMineR (version 2.3) was used to perform PCA analysis.

Human disease ontology and GO analysis

The coordinate file of GAIN and LOST VELs were submitted to GREAT website (version 3.0.0) and the results of human disease ontology and GO analysis (biological process) were obtained for plotting.

CMS classification for tumor tissues

Consensus Molecular Subtype (CMS) classification of tumor tissues was performed by a R package CMScaller (version 0.99.2). With an integrated CRC tumors RNA-seq result file, this package could classify all samples into 5 subgroups (CMS1/2/3/4 and no group). The samples excluded in previous steps were not analyzed here.

Identification of CMS subgroup specific gain VELs

For an individual gain VEL, if the H3K27ac signal on corresponding region in a CMS subgroup was 1.5 times higher than other 3 subgroups, we called it a specific gain VEL for this subgroup. For all CMS subgroups, significant GAIN VELs were identified as the procedure described above in “Identification of VELs” and the significant percentage cut-off was changed to 0.9.

Pathway analysis for CMS2 specific GAIN VELs

Functional characterization of CMS2 specific gain-VEL-associated genes was conducted using the ClueGO plugin for Cytoscape (version 3.8.0). These tested genes

were queried against a compendium of gene sets from GO (Biological Process), KEGG and REACTOME to identify significantly enriched processes and pathways. Analyses were performed using the GO Term Fusion option in ClueGO and only processes/pathways with a p-value < 0.01 (right-sided hypergeometric test) following p-value correction (Bonferroni step down) were visualized.

Prediction of enriched TFs on VELs

HOMER software plugin findMotifsGenome.pl was used to calculate the significance of TFs enrichment. For VELs of all patients, the coordinate files of gain and lost VELs were used for calculation and the size parameter is 200. For VELs of individual patient, TFs enrichment significance were calculated using nucleosome free regions (NRFs) within VEL. NFRs were generated by PARE (version 0.08) with default settings.

Core regulatory circuitry for super enhancer associated TFs

To quantify the interaction network of transcription factor regulation, we calculated the inward and outward binding degree of all SE-associated TFs. All SE-associated genes annotated to encode a transcription factor were considered as the node-list for network construction. For any given TFi, the IN degree was defined as the number of TFs with an enriched binding motif at the proximal SE or promoter of TFi. The OUT degree was defined as the number of TF-associated SEs containing an enriched binding site for TFi. The IN and OUT degree were generated by crc software (https://github.com/linlabcode/CRC) and the total degree was defined as IN degree plus OUT degree.

CRISPR-Cas9-KRAB mediated repression of VSELs

Site-specific single guide RNAs (sgRNAs) targeting VSELs were designed with publicly available filtering tools (https://zlab.bio/guide-design-resources) to minimize off-target cleavage. For CRISPR interference, sgRNAs were cloned into the pLH-

spsgRNA2 (Addgene, #64114) through the BbsI site according to the protocol recommended by Addgene. Lentivirus was generated by transfecting HEK293T cells with sgRNA expression cocktails or pHAGE dCas9-KRAB-MeCP2, together with helper plasmids, psPAX and pMD2.G. After 12 hours, cells were washed twice with PBS and fresh medium was added. Medium containing virus was collected 48 or 72 hours after transfection, and filtered with 0.45 μm filters (Millipore). Stable cell lines were generated by infecting HCT116 with lentivirus expressing dCas9-KRAB-MeCP2 and sgRNAs. Cells were then screened with puromycin (1 μg/ml, Amresco) and hygromycin (200 μg/ml, Roche) for 48 hours, and examined by western blot and RT-qPCR.

Chromosome conformation capture (3C) assay

Approximately 1×10^6 cells were crosslinked with 1 % formaldehyde for 10 min and quenched by glycine. The cells were washed with PBS and lysed in cell lysis buffer (10 mM Tris-HCl, pH 7.5, 10 mM NaCl, 5 mM MgCl2, 0.1 mM EDTA, 1 × complete protease inhibitor) at 4 °C for 30 min. Nuclei were collected after centrifugation at 400g at 4 °C for 5 min and removing the supernatant. The collected nuclei were digested with 400 U DpnII restriction enzyme (NEB) at 37 °C overnight. The digested nuclei were then added with 100 U of T4 DNA ligase (NEB) and incubated for 4 h at 16 °C followed by 30 min at room temperature. The samples were then added with 300 μg of proteinase K, incubated at 65 °C overnight for de-crosslinking and purificated with DNA purification kit (TIANGEN DP214-03). The relative crosslinking frequencies between the enhancer and promoter were determined by qRT-PCR. The relative cross-linking frequencies are calculated by normalizing to a primer pair (3C-N F, 3C-N R) without crossing the DpnII cut sites. Assays were repeated at least three times. Data were shown as average values ± SD and p-value was calculated using Student's t test. The sequences of primers are in Extended Data. Table 8.

Reverse transcription and quantitative PCR

Cells were scraped down and collected with centrifugation. Total RNA was extracted with RNA extraction kit (Aidlab) according to the manufacturer's manual. Approximately 1μg of total RNA was used for reverse transcription with a first strand cDNA synthesis kit (Toyobo). The resulted cDNA was then assayed with quantitative PCR. β-actin was used for normalization. The sequences of primers are in Extended Data Table 6. Assays were repeated at least three times. Data were shown as average values ± SD of at least three representative experiments. P value was calculated using student's t test.

Cell proliferation assay

The proliferation of colorectal cancer cells in vitro was measured using the MTT assay. Briefly, 1,000 cells were seeded into 96-well plate per well. Six well of each group were detected every day. MTT (0.25 μ g) was put into each well and incubated at 37℃ for 4 hours. The medium with the formazan sediment was dissolved in 50% DMF and 30% SDS (pH4.7). The absorbance was measured at 570 nm. Assays were repeated at least three times. Data were shown as average values ± SD of at least three representative experiments and p value was calculated using student's t test.

Transwell assay

1×10^5 HCT116 cells were plated in medium without serum or growth factors in the upper chamber with a Matrigel-coated membrane (24-well insert; pore size, 8 μm; BD Biosciences), and medium supplemented with 10% fetal bovine serum was used as a chemoattractant in the lower chamber. After 36 h of incubation, cells that did not invade through the membrane were removed by a cotton swab. Cells on the lower surface of the membrane were stained with crystal violet and counted. Assays were repeated at least three times. Data were shown as average values ± SD of at least three representative experiments and p value was calculated using student's t test.

Ethics approval and consent to participate

A total of 80 pairs primary tumor tissues and corresponding adjacent tissues were collected from patients who received surgical treatment at Zhongnan Hospital of Wuhan University (Wuhan, China) between August 2017 and February 2018. Samples of the collected tissues were preserved in liquid nitrogen. Clinical case data of patients was also collected. The collection procedures of clinical specimens were approved by the Clinical Research Institution Review Committee and Ethics Review Committee of Zhongnan Hospital, and consent of each patient was obtained before collection.

Availability of data and materials

All the deep sequencing data have been submitted to GEO database, with the Acc. NO. GSE156614.

Competing interests

All the authors of the manuscript do not have any financial interest related to this work.

Funding

This work was supported by grants from Ministry of Science and Technology of China (2016YFA0502100), National Natural Science Foundation of China to L.L. (31670874) and M.W. (81972647 and 31771503), Science and Technology Department of Hubei Province of China (CXZD2017000188).

Author contributions

LQL performed ChIP-Seq; LX, CL, HQX and HCW verified enhancers and TFs; LQL, ZC and WCY did bioinformatics, YYL, CM, CN and YM collected tissues and patient data; WM and LLY directed the project; LQL and WM wrote the manuscript.

Reference

1. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**, 272-86 (2014).
2. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol Cell* **49**, 825-37 (2013).
3. Nizovtseva, E.V., Todolli, S., Olson, W.K. & Studitsky, V.M. Towards quantitative analysis of gene regulation by enhancers. *Epigenomics* **9**, 1219-1231 (2017).
4. Rickels, R. & Shilatifard, A. Enhancer Logic and Mechanics in Development and Disease. *Trends Cell Biol* **28**, 608-630 (2018).
5. Medina-Rivera, A., Santiago-Algarra, D., Puthier, D. & Spicuglia, S. Widespread Enhancer Activity from Core Promoters. *Trends Biochem Sci* **43**, 452-468 (2018).
6. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-47 (2013).
7. Herz, H.M. *et al.* Enhancer-associated H3K4 monomethylation by Trithorax-related, the Drosophila homolog of mammalian Mll3/Mll4. *Genes Dev* **26**, 2604-20 (2012).
8. Wang, C. *et al.* Enhancer priming by H3K4 methyltransferase MLL4 controls cell fate transition. *Proc Natl Acad Sci U S A* **113**, 11871-11876 (2016).
9. Dowen, J.M. *et al.* Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374-387 (2014).
10. Roe, J.S. *et al.* Enhancer Reprogramming Promotes Pancreatic Cancer Metastasis. *Cell* **170**, 875-888 e20 (2017).
11. Murakawa, Y. *et al.* Enhanced Identification of Transcriptional Enhancers Provides Mechanistic Insights into Diseases. *Trends Genet* **32**, 76-88 (2016).
12. Yao, J., Chen, J., Li, L.Y. & Wu, M. Epigenetic plasticity of enhancers in cancer. *Transcription* **11**, 26-36 (2020).
13. Flavahan, W.A., Gaskell, E. & Bernstein, B.E. Epigenetic plasticity and the hallmarks of cancer. *Science* **357**(2017).
14. Yuan, J. *et al.* Super-Enhancers Promote Transcriptional Dysregulation in Nasopharyngeal Carcinoma. *Cancer Res* **77**, 6614-6626 (2017).
15. Li, Q.L. *et al.* The hyper-activation of transcriptional enhancers in breast cancer. *Clin Epigenetics* **11**, 48 (2019).
16. Shen, H. *et al.* Suppression of Enhancer Overactivation by a RACK7-Histone Demethylase Complex. *Cell* **165**, 331-42 (2016).
17. Wang, L. & Shilatifard, A. UTX Mutations in Human Cancer. *Cancer Cell* **35**, 168-176 (2019).
18. Fagan, R.J. & Dingwall, A.K. COMPASS Ascending: Emerging clues regarding the roles of MLL3/KMT2C and MLL2/KMT2D proteins in cancer. *Cancer Lett* **458**, 56-65 (2019).
19. Sze, C.C. & Shilatifard, A. MLL3/MLL4/COMPASS Family on Epigenetic Regulation of Enhancer Function and Cancer. *Cold Spring Harb Perspect Med* **6**(2016).
20. van Haaften, G. *et al.* Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. *Nat Genet* **41**, 521-3 (2009).
21. Loven, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320-34 (2013).
22. Lancho, O. & Herranz, D. The MYC Enhancer-ome: Long-Range Transcriptional Regulation of MYC in Cancer. *Trends Cancer* **4**, 810-822 (2018).
23. Akhtar-Zaidi, B. *et al.* Epigenomic enhancer profiling defines a signature of colon cancer.

Science **336**, 736-9 (2012).
24. Chen, H. *et al.* A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples. *Cell* **173**, 386-399 e12 (2018).
25. Ooi, W.F. *et al.* Epigenomic profiling of primary gastric adenocarcinoma reveals super-enhancer heterogeneity. *Nat Commun* **7**, 12983 (2016).
26. Tse, J.W.T., Jenkins, L.J., Chionh, F. & Mariadason, J.M. Aberrant DNA Methylation in Colorectal Cancer: What Should We Target? *Trends Cancer* **3**, 698-712 (2017).
27. Hinoue, T. *et al.* Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res* **22**, 271-82 (2012).
28. Sahnane, N. *et al.* Aberrant DNA methylation profiles of inherited and sporadic colorectal cancer. *Clin Epigenetics* **7**, 131 (2015).
29. Wang, H.Y. *et al.* Histone demethylase KDM3A is required for enhancer activation of hippo target genes in colorectal cancer. *Nucleic Acids Res* **47**, 2349-2364 (2019).
30. Yao, J. *et al.* GLIS2 promotes colorectal cancer through repressing enhancer activation. *Oncogenesis* **9**, 57 (2020).
31. Jagle, S. *et al.* SNAIL1-mediated downregulation of FOXA proteins facilitates the inactivation of transcriptional enhancer elements at key epithelial genes in colorectal cancer cells. *PLoS Genet* **13**, e1007109 (2017).
32. Cohen, A.J. *et al.* Hotspots of aberrant enhancer activity punctuate the colorectal cancer epigenome. *Nat Commun* **8**, 14400 (2017).
33. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat Med* **21**, 1350-6 (2015).
34. Fodde, R. The APC gene in colorectal cancer. *Eur J Cancer* **38**, 867-71 (2002).
35. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-74 (2011).
36. Lochhead, P. & Chan, A.T. Statins and colorectal cancer. *Clin Gastroenterol Hepatol* **11**, 109-18; quiz e13-4 (2013).
37. Jacobs, R.J., Voorneveld, P.W., Kodach, L.L. & Hardwick, J.C. Cholesterol metabolism and colorectal cancers. *Curr Opin Pharmacol* **12**, 690-5 (2012).
38. Bardou, M., Barkun, A. & Martel, M. Effect of statin therapy on colorectal cancer. *Gut* **59**, 1572-85 (2010).
39. Thakore, P.I. *et al.* Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nat Methods* **12**, 1143-9 (2015).
40. Saint-Andre, V. *et al.* Models of human core transcriptional regulatory circuitries. *Genome Res* **26**, 385-96 (2016).
41. Boyer, L.A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947-56 (2005).
42. Li, Q. *et al.* RUNX1 promotes tumour metastasis by activating the Wnt/beta-catenin signalling pathway and EMT in colorectal cancer. *J Exp Clin Cancer Res* **38**, 334 (2019).
43. Triner, D. *et al.* Myc-Associated Zinc Finger Protein Regulates the Proinflammatory Response in Colitis and Colon Cancer via STAT3 Signaling. *Mol Cell Biol* **38**(2018).
44. Zhang, X. *et al.* Somatic Superenhancer Duplications and Hotspot Mutations Lead to Oncogenic Activation of the KLF5 Transcription Factor. *Cancer Discov* **8**, 108-125 (2018).
45. Wei, X. *et al.* Ascl2 activation by YAP1/KLF5 ensures the self-renewability of colon cancer

progenitor cells. *Oncotarget* **8**, 109301-109318 (2017).

46. Zhu, K. *et al.* SPOP-containing complex regulates SETD2 stability and H3K36me3-coupled alternative splicing. *Nucleic Acids Res* **45**, 92-105 (2017).

Figure legends

Figure 1 The annotation of active enhancers in CRC patient tissues. (A) Experimental workflow for studying the enhancer landscapes of tumor and native tissues from CRC patients. **(B)** Genomic distribution of enhancer elements in tumor and native tissues from CRC patients. **(C)** Saturation analysis showing the percentage of newly gained enhancers comparing with total significant enhancers along with increasing number of tumor sample. **(D)** Fold change (FC) and p.adj of human gene expression comparing tumor and native tissues. Red dots represent tumor up-regulated genes, blue dots for native tissue up-regulated genes and grey dots for genes not changed. **(E)** Normalized ChIP-seq and RNA-seq Meta tracks showing H3K27ac and mRNA signal on *MYC* promoter and enhancer loci. **(F)** Overlap of enhancer loci between our patient data and 20 COAD cell lines (GSE77737, Andrea J. Cohen et al.). **(G)** Percentage of novel enhancers in CRC identified in our study.

Figure 2 Identification of variant enhancer loci in CRC. (A) Relative H3K27ac signals of lost and gain VELs in all tumor and native tissues. **(B)** The required recurrence for gain and lost VELs meeting statistical significance (q-value < 0.05). The two vertical dashed lines at left highlights the recurrence of gain and lost VELs when achieve the cut-off (0.95, black dashed line) of significant percentage, and the two lines at right highlights the highest recurrence in tumor or native tissue of gain and lost VELs, respectively. **(C)** Representative H3K27ac tracks of gain VEL on *IL20RA* loci. **(D)** The human disease ontology in which gain VELs participated detected by GREAT software (version 3.0.0). The red bars represent CRC related diseases and the black bars represent other diseases. **(E-G)** PCA analyses to classify tumor and native tissues using gene expression (E), all significant enhancers (F) and promoters (G) information identified using our patient RNA-seq and ChIP-seq data.

Figure 3 The feature of enhancers in CMS subgroups. (A) The consensus molecular subtypes (CMS) classification of CRC samples using R package

CMScaller. **(B)** Correlation of H3K27ac signal on the regions of gain VELs in all tumor samples of CMS1-4 subgroups. Correlations were calculated by Spearman correlation coefficient. **(C)** The required recurrence for gain VELs in each CMS subgroup to meet statistical significance (q-value < 0.05) at different cut-offs. The dashed lines highlight the recurrence of gain VELs when achieve the cut-off (0.9, black dashed line) of significant percentage. **(D)** The number of subgroup significant gain VELs in four CMS subgroups. **(E)** The average H3K27ac signal (RPM) on the regions of gain VELs in four CMS subgroups. **(F)** The number of subgroup specific gain VELs in each CMS. The subgroup specific gain VELs were identified when the mean RPM of one VEL in one CMS subgroup were 1.5 times higher than other three. **(G)** Functional annotation of target genes associated with CMS2 specific gain VELs based on their significant overlap with gene sets annotated in Gene Ontology (Biological Process) and pathway database (Reactome). **(H)** Meta tracks of normalized H3K27ac on *CEL* and *DPEP1* gene loci in four CMS subgroups.

Figure 4 Functions of tumor-specific super enhancers in CRC. (A) The genes associated with top super enhancers (SEs) ranked by recurrence. Red dots represent tumor specific SE genes and blue dots represent native tissue specific SE genes. Top 10 tumor and native tissue specific genes were listed. **(B-C)** The average H3K27ac signal (RPM) at the regions of gain VSELs (B) and lost VSELs (C) in tumor and native tissues. **(D)** Meta normalized H3K27ac tracks at IER3 gene loci. The green track on the top represents H3K27ac signal in HCT116, and the black and grey lines at the bottom represent the average signal of tumor and native tissues, respectively. The pink lines indicate the target positions of dCas9-KRAB sgRNAs. **(E)** Bar plot showing the relative mRNA level of *LIF*, *SLC7A5*, *CYP2S1*, *PHF19*, *RNF43*, *CEBPB*, *TBC1D16*, *TNFRSF6B*, *VEGFA* and *IER3* in control and sgRNA groups. * $p < 0.05$. **(F)** Transwell assays for HCT116 cell lines stably transfected with dCas9-KRAB sgRNAs of the enhancers mentioned in Fig. 5E. * $p < 0.05$.

Figure 5 Prediction of functional transcription factors in CRC. (A) DNA motifs enriched within nucleosome-free regions (NFRs) of tumor gain VELs determined by HOMER motif analysis. **(B)** Heatmap of transcription factors ranked by predicted core regulatory circuitry (CRC) total degrees (Tumor - Native tissue). Top 30 tumor and native specific TFs were listed. **(C)** Scatter plot showing the total degree (Tumor - Native tissue) and expression FC (Tumor / Native tissue) of the specific TFs listed in Fig. 5B. Blue dots represent top 30 tumor specific TFs, and red dots represent top 30 native specific TFs. Circle size indicates the mean expression (FPKM) of TFs in its specific tissue. **(D&E)** Transwell assay for HCT116 cell line with *KLF3* knockdown. ** $p < 0.01$. **(F)** Relative mRNA level for HCT116 cell line after KLF3 knockdown. * $p < 0.05$, ** $p < 0.01$.

Fig. 1

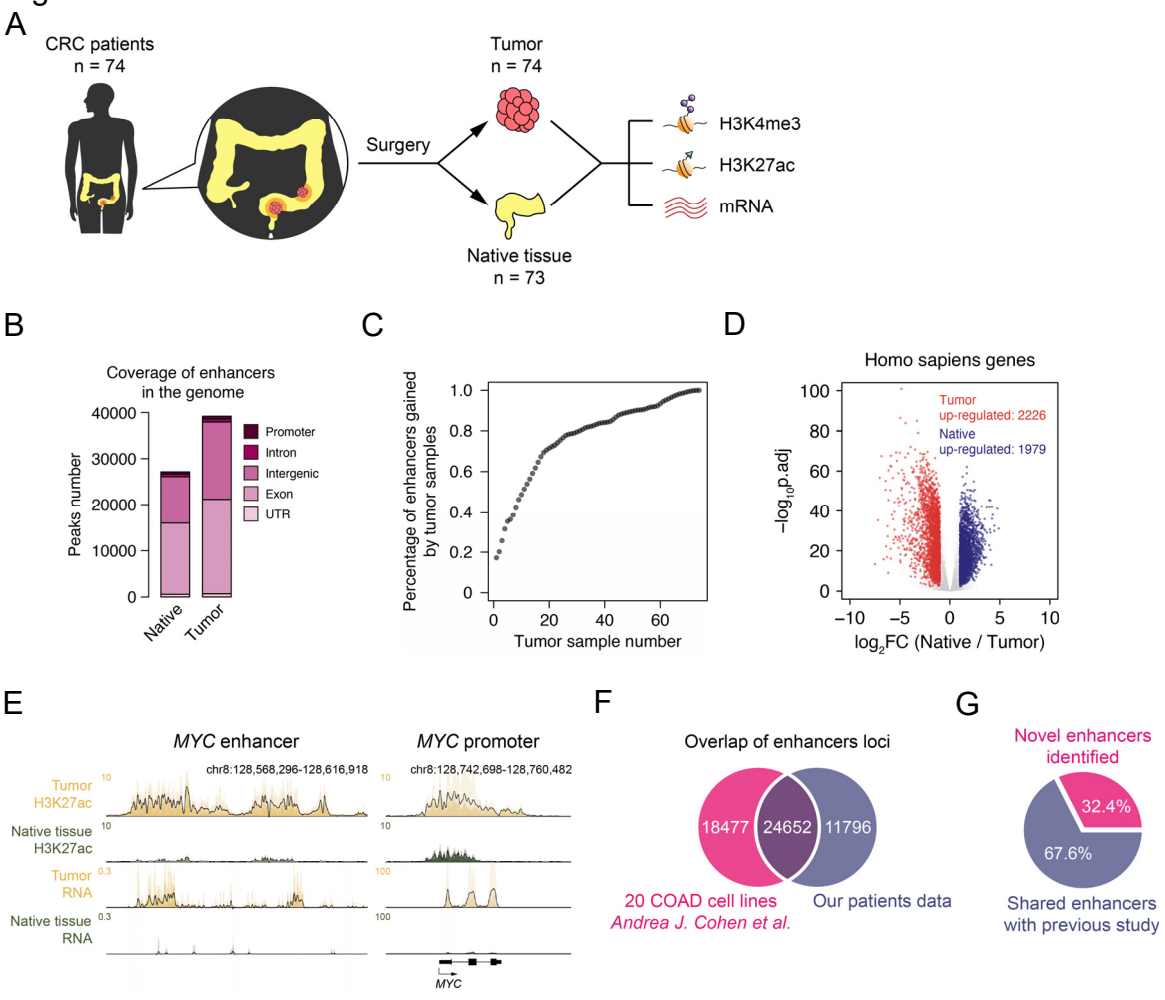

A
CRC patients
n = 74
Surgery
Tumor
n = 74
Native tissue
n = 73
H3K4me3
H3K27ac
mRNA
B
Coverage of enhancers
in the genome
Peaks number
40000
30000
20000
10000
0
Promoter
Intron
Intergenic
Exon
UTR
Native
Tumor
C
Percentage of enhancers gained
by tumor samples
1.0
0.8
0.6
0.4
0.2
0
0
20
40
60
Tumor sample number
D
Homo sapiens genes
Tumor
up-regulated: 2226
Native
up-regulated: 1979
−log10p.adj
100
80
60
40
20
0
−10
−5
0
5
10
log2FC (Native / Tumor)
E
MYC enhancer
MYC promoter
chr8:128,568,296-128,616,918
chr8:128,742,698-128,760,482
Tumor
H3K27ac
Native tissue
H3K27ac
Tumor
RNA
Native tissue
RNA
MYC
F
Overlap of enhancers loci
18477
24652
11796
20 COAD cell lines
Andrea J. Cohen et al.
Our patients data
G
Novel enhancers
identified
32.4%
67.6%
Shared enhancers
with previous study

Fig. 2

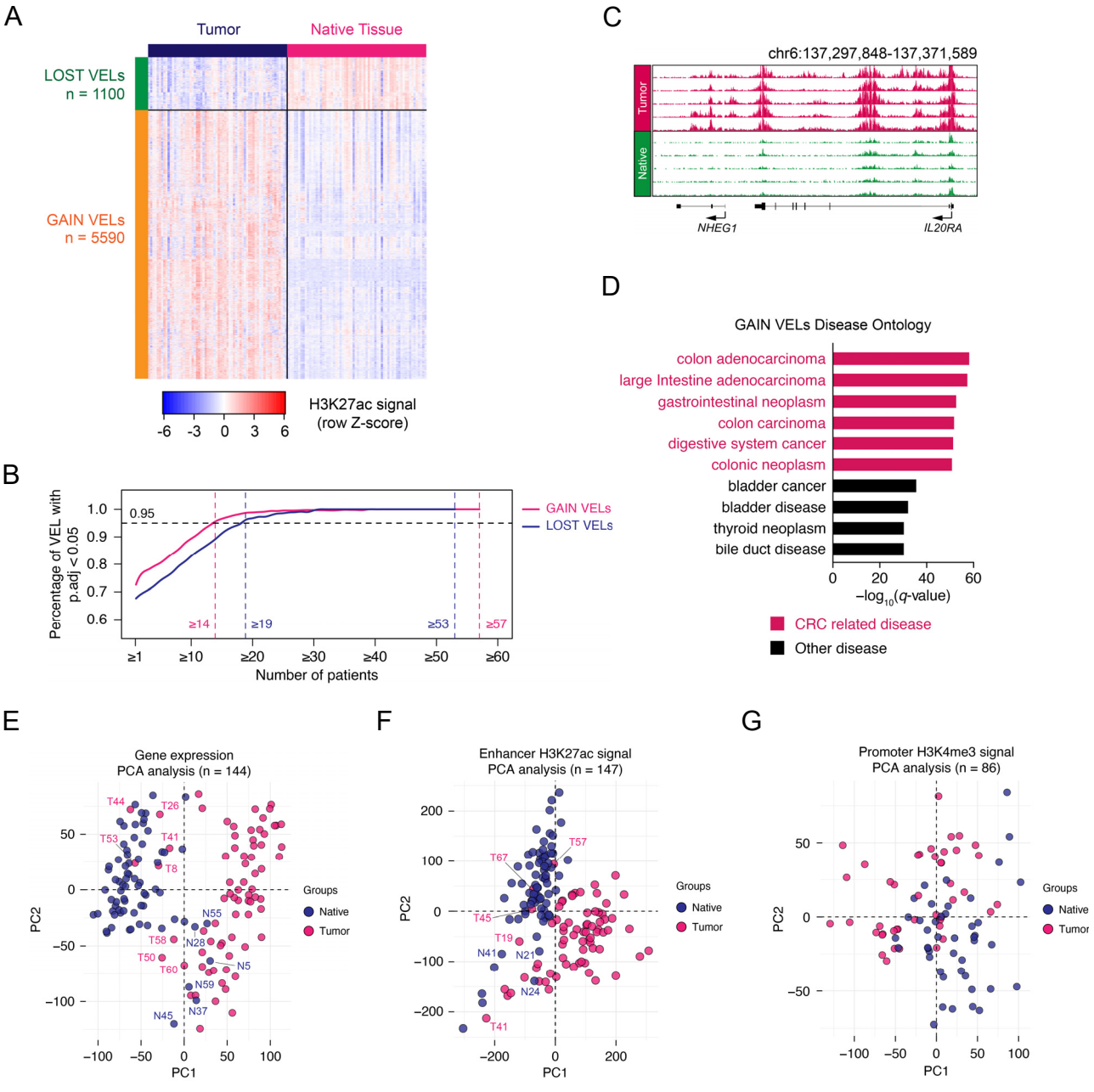

A
Tumor
Native Tissue
LOST VELs
n = 1100
GAIN VELs
n = 5590
H3K27ac signal
(row Z-score)
-6 -3 0 3 6
B
Percentage of VEL with
p.adj < 0.05
0.95
GAIN VELs
LOST VELs
≥14
≥19
≥53
≥57
≥1 ≥10 ≥20 ≥30 ≥40 ≥50 ≥60
Number of patients
C
chr6:137,297,848-137,371,589
Tumor
Native
NHEG1
IL20RA
D
GAIN VELs Disease Ontology
colon adenocarcinoma
large Intestine adenocarcinoma
gastrointestinal neoplasm
colon carcinoma
digestive system cancer
colonic neoplasm
bladder cancer
bladder disease
thyroid neoplasm
bile duct disease
−log10(q-value)
CRC related disease
Other disease
E
Gene expression
PCA analysis (n = 144)
PC1
PC2
Groups
Native
Tumor
F
Enhancer H3K27ac signal
PCA analysis (n = 147)
G
Promoter H3K4me3 signal
PCA analysis (n = 86)

Fig. 3

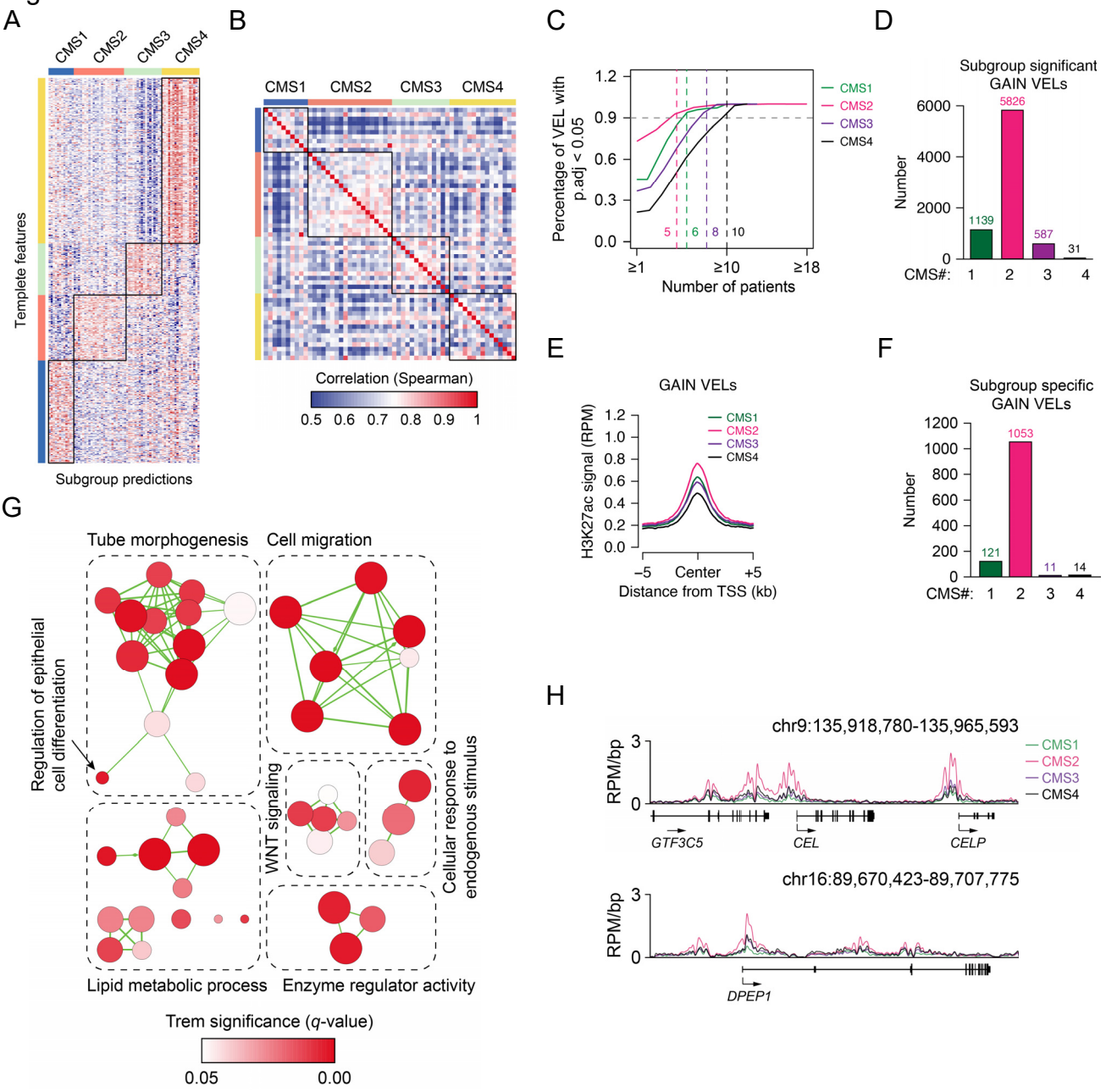
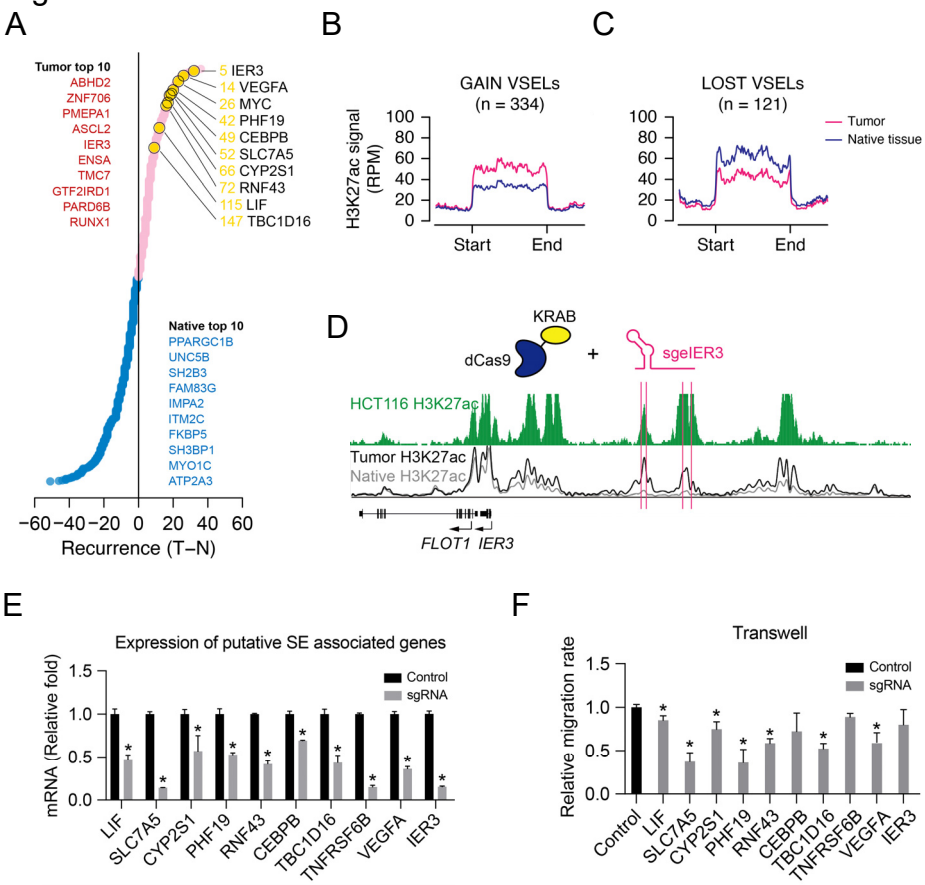

A
CMS1
CMS2
CMS3
CMS4
Templete features
Subgroup predictions
B
CMS1
CMS2
CMS3
CMS4
Correlation (Spearman)
0.5 0.6 0.7 0.8 0.9 1
C
Percentage of VEL with p.adj < 0.05
1.2
0.9
0.6
0.3
0.0
5
6
8
10
≥1
≥10
≥18
Number of patients
CMS1
CMS2
CMS3
CMS4
D
Subgroup significant GAIN VELs
Number
6000
4000
2000
0
5826
1139
587
31
CMS#: 1 2 3 4
E
GAIN VELs
H3K27ac signal (RPM)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
CMS1
CMS2
CMS3
CMS4
−5 Center +5
Distance from TSS (kb)
F
Subgroup specific GAIN VELs
Number
1200
1000
800
600
400
200
0
1053
121
11
14
CMS#: 1 2 3 4
G
Tube morphogenesis
Cell migration
Regulation of epithelial cell differentiation
WNT signaling
Cellular response to endogenous stimulus
Lipid metabolic process
Enzyme regulator activity
Trem significance (q-value)
0.05
0.00
H
chr9:135,918,780-135,965,593
RPM/bp
3
0
CMS1
CMS2
CMS3
CMS4
GTF3C5
CEL
CELP
chr16:89,670,423-89,707,775
RPM/bp
3
0
DPEP1

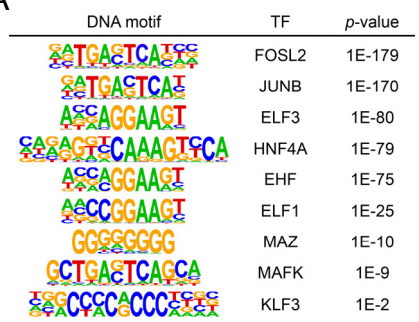
Fig. 4


A
B
C
Tumor top 10
ABHD2
ZNF706
PMEPA1
ASCL2
IER3
ENSA
TMC7
GTF2IRD1
PARD6B
RUNX1
5 IER3
14 VEGFA
26 MYC
42 PHF19
49 CEBPB
52 SLC7A5
66 CYP2S1
72 RNF43
115 LIF
147 TBC1D16
Native top 10
PPARGC1B
UNC5B
SH2B3
FAM83G
IMPA2
ITM2C
FKBP5
SH3BP1
MYO1C
ATP2A3
−60 −40 −20 0 20 40 60
Recurrence (T−N)
GAIN VSELs
(n = 334)
H3K27ac signal
(RPM)
100
80
60
40
20
0
Start
End
LOST VSELs
(n = 121)
Tumor
Native tissue
D
KRAB
dCas9
+
sgeIER3
HCT116 H3K27ac
Tumor H3K27ac
Native H3K27ac
FLOT1 IER3

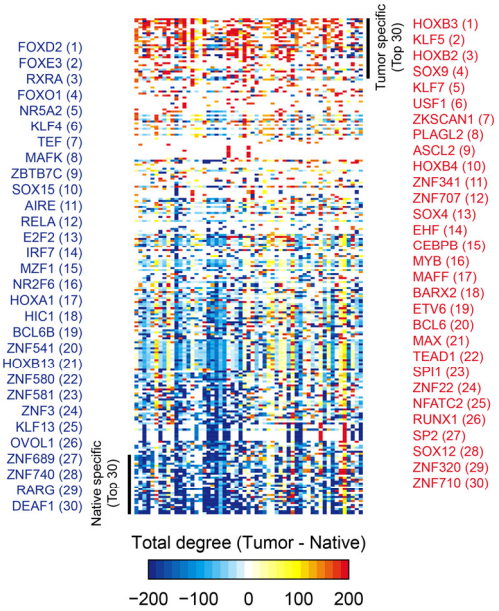
E
Expression of putative SE associated genes
mRNA (Relative fold)
1.5
1.0
0.5
0.0
Control
sgRNA
LIF
SLC7A5
CYP2S1
PHF19
RNF43
CEBPB
TBC1D16
TNFRSF6B
VEGFA
IER3
F
Transwell
Relative migration rate
1.5
1.0
0.5
0.0
Control
sgRNA
Control
LIF
SLC7A5
CYP2S1
PHF19
RNF43
CEBPB
TBC1D16
TNFRSF6B
VEGFA
IER3

Fig. 5

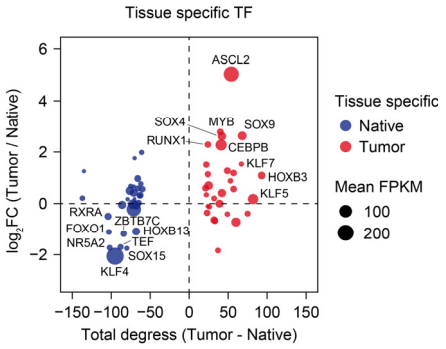
A


DNA motif
TF
p-value
FOSL2
1E-179
JUNB
1E-170
ELF3
1E-80
HNF4A
1E-79
EHF
1E-75
ELF1
1E-25
MAZ
1E-10
MAFK
1E-9
KLF3
1E-2

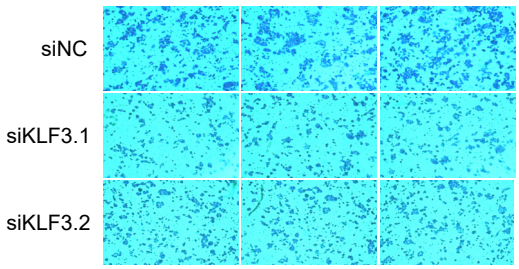
B


FOXD2 (1)
FOXE3 (2)
RXRA (3)
FOXO1 (4)
NR5A2 (5)
KLF4 (6)
TEF (7)
MAFK (8)
ZBTB7C (9)
SOX15 (10)
AIRE (11)
RELA (12)
E2F2 (13)
IRF7 (14)
MZF1 (15)
NR2F6 (16)
HOXA1 (17)
HIC1 (18)
BCL6B (19)
ZNF541 (20)
HOXB13 (21)
ZNF580 (22)
ZNF581 (23)
ZNF3 (24)
KLF13 (25)
OVOL1 (26)
ZNF689 (27)
ZNF740 (28)
RARG (29)
DEAF1 (30)
Native specific (Top 30)
Tumor specific (Top 30)
HOXB3 (1)
KLF5 (2)
HOXB2 (3)
SOX9 (4)
KLF7 (5)
USF1 (6)
ZKSCAN1 (7)
PLAGL2 (8)
ASCL2 (9)
HOXB4 (10)
ZNF341 (11)
ZNF707 (12)
SOX4 (13)
EHF (14)
CEBPB (15)
MYB (16)
MAFF (17)
BARX2 (18)
ETV6 (19)
BCL6 (20)
MAX (21)
TEAD1 (22)
SPI1 (23)
ZNF22 (24)
NFATC2 (25)
RUNX1 (26)
SP2 (27)
SOX12 (28)
ZNF320 (29)
ZNF710 (30)
Total degree (Tumor - Native)
−200 −100 0 100 200

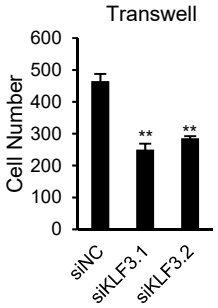
C


Tissue specific TF
log2FC (Tumor / Native)
Total degress (Tumor - Native)
ASCL2
SOX4
MYB
SOX9
RUNX1
CEBPB
KLF7
HOXB3
KLF5
RXRA
ZBTB7C
FOXO1
HOXB13
NR5A2
TEF
SOX15
KLF4
Tissue specific
Native
Tumor
Mean FPKM
100
200

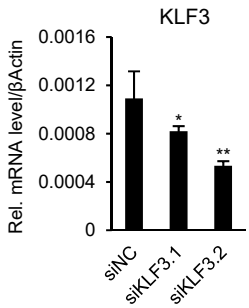
D


siNC
siKLF3.1
siKLF3.2

E


Transwell
Cell Number
siNC
siKLF3.1
siKLF3.2

F


KLF3
Rel. mRNA level/βActin
siNC
siKLF3.1
siKLF3.2

Figures

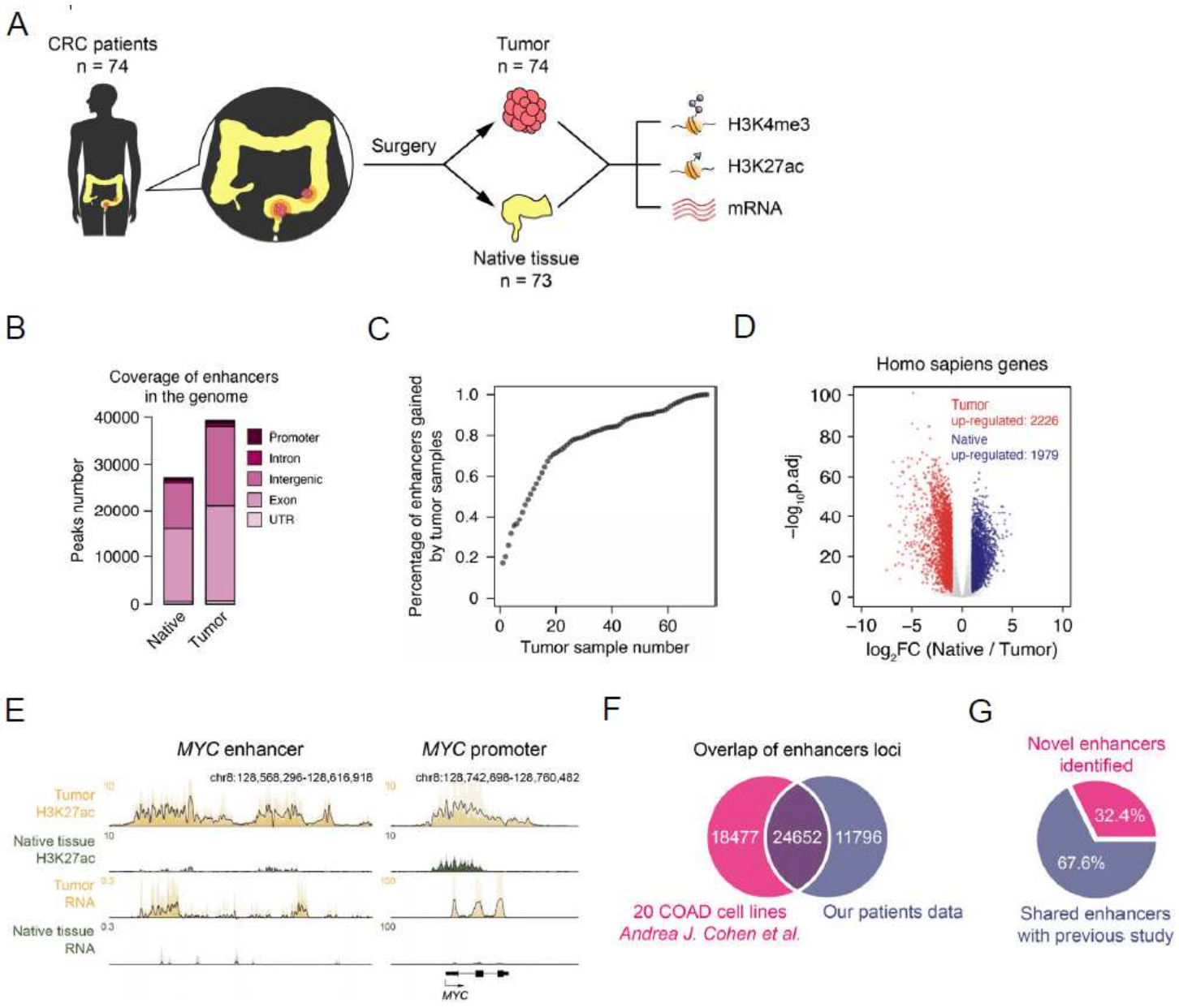


Figure 1

The annotation of active enhancers in CRC patient tissues. (A) Experimental workflow for studying the enhancer landscapes of tumor and native tissues from CRC patients. (B) Genomic distribution of enhancer elements in tumor and native tissues from CRC patients. (C) Saturation analysis showing the percentage of newly gained enhancers comparing with total significant enhancers along with increasing number of tumor sample. (D) Fold change (FC) and p.adj of human gene expression comparing tumor and native tissues. Red dots represent tumor upregulated genes, blue dots for native tissue up-regulated genes and grey dots for genes not changed. (E) Normalized ChIP-seq and RNA-seq Meta tracks showing H3K27ac and mRNA signal on MYC promoter and enhancer loci. (F) Overlap of enhancer loci between our

patient data and 20 COAD cell lines (GSE77737, Andrea J. Cohen et al.). (G) Percentage of novel enhancers in CRC identified in our study.

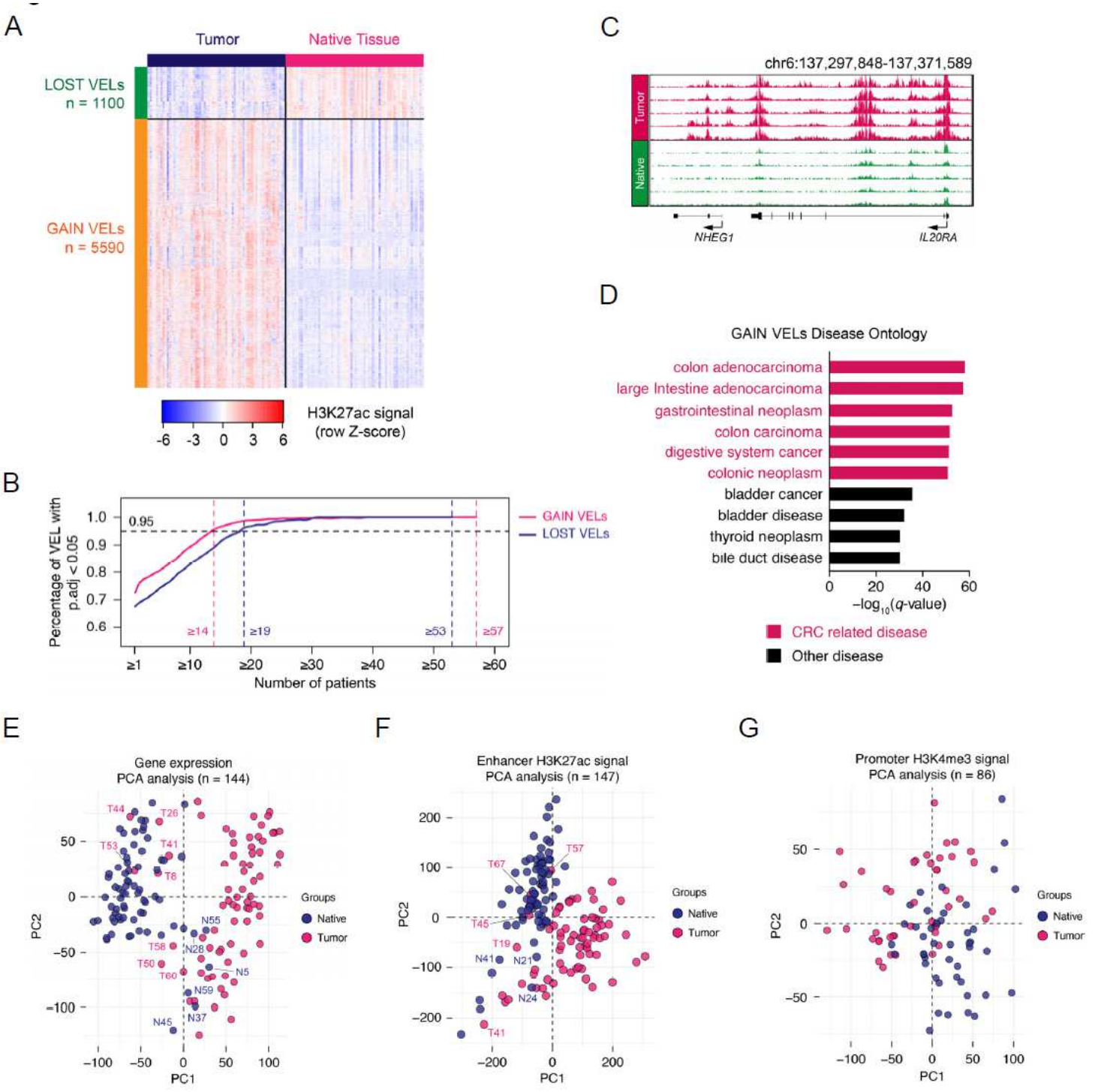


Figure 2

Identification of variant enhancer loci in CRC. (A) Relative H3K27ac signals of lost and gain VELs in all tumor and native tissues. (B) The required recurrence for gain and lost VELs meeting statistical significance (q-value < 0.05). The two vertical dashed lines at left highlights the recurrence of gain and lost VELs when achieve the cut-off (0.95, black dashed line) of significant percentage, and the two lines

at right highlights the highest recurrence in tumor or native tissue of gain and lost VELs, respectively. (C) Representative H3K27ac tracks of gain VEL on IL20RA loci. (D) The human disease ontology in which gain VELs participated detected by GREAT software (version 3.0.0). The red bars represent CRC related diseases and the black bars represent other diseases. (E-G) PCA analyses to classify tumor and native tissues using gene expression (E), all significant enhancers (F) and promoters (G) information identified using our patient RNA-seq and ChIP-seq data.

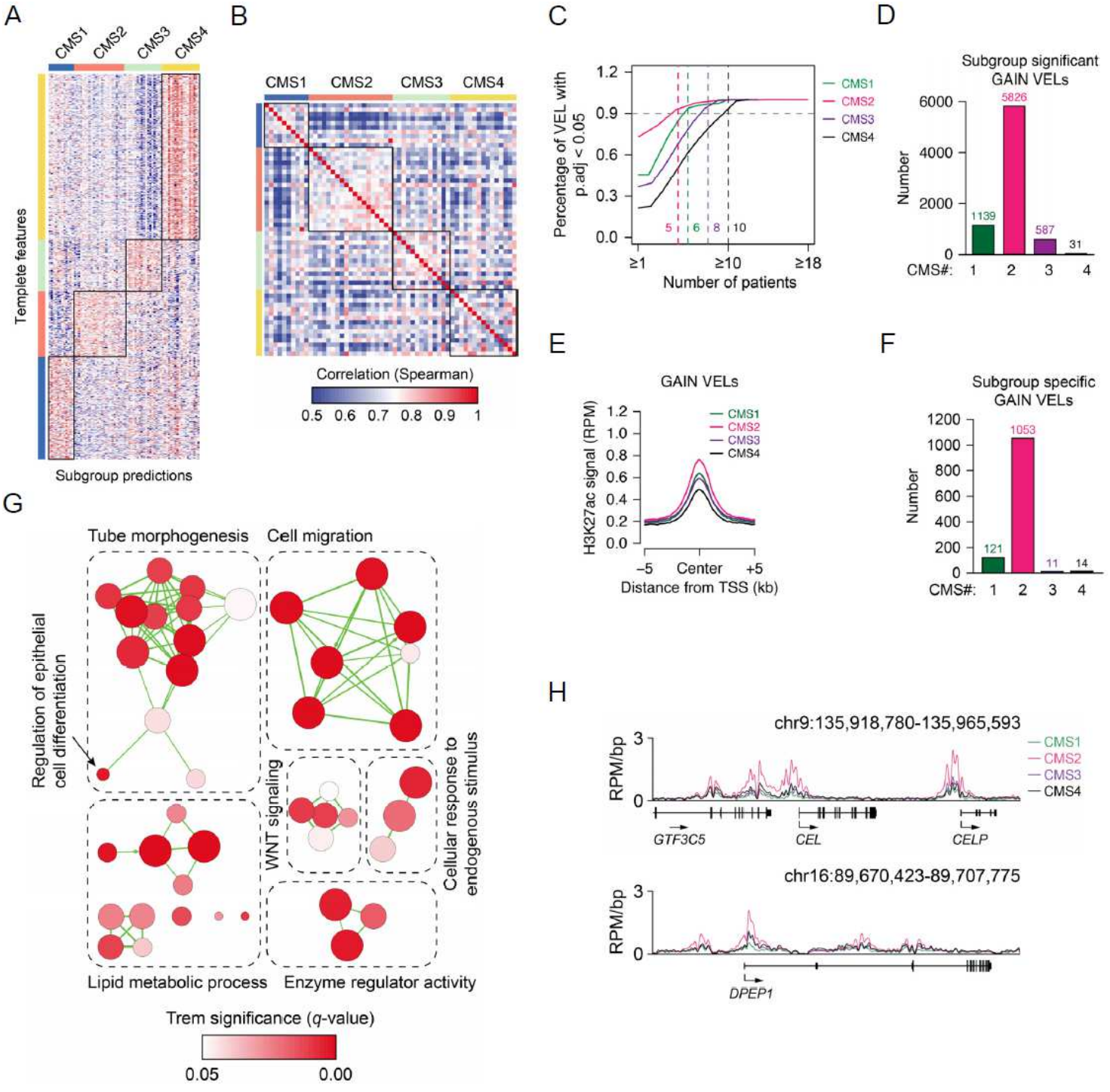


Figure 3

The feature of enhancers in CMS subgroups. (A) The consensus molecular subtypes (CMS) classification of CRC samples using R package CMScaller. (B) Correlation of H3K27ac signal on the regions of gain VELs in all tumor samples of CMS1-4 subgroups. Correlations were calculated by Spearman correlation coefficient. (C) The required recurrence for gain VELs in each CMS subgroup to meet statistical significance (q-value < 0.05) at different cut-offs. The dashed lines highlight the recurrence of gain VELs when achieve the cut-off (0.9, black dashed line) of significant percentage. (D) The number of subgroup significant gain VELs in four CMS subgroups. (E) The average H3K27ac signal (RPM) on the regions of gain VELs in four CMS subgroups. (F) The number of subgroup specific gain VELs in each CMS. The subgroup specific gain VELs were identified when the mean RPM of one VEL in one CMS subgroup were 1.5 times higher than other three. (G) Functional annotation of target genes associated with CMS2 specific gain VELs based on their significant overlap with gene sets annotated in Gene Ontology (Biological Process) and pathway database (Reactome). (H) Meta tracks of normalized H3K27ac on CEL and DPEP1 gene loci in four CMS subgroups.

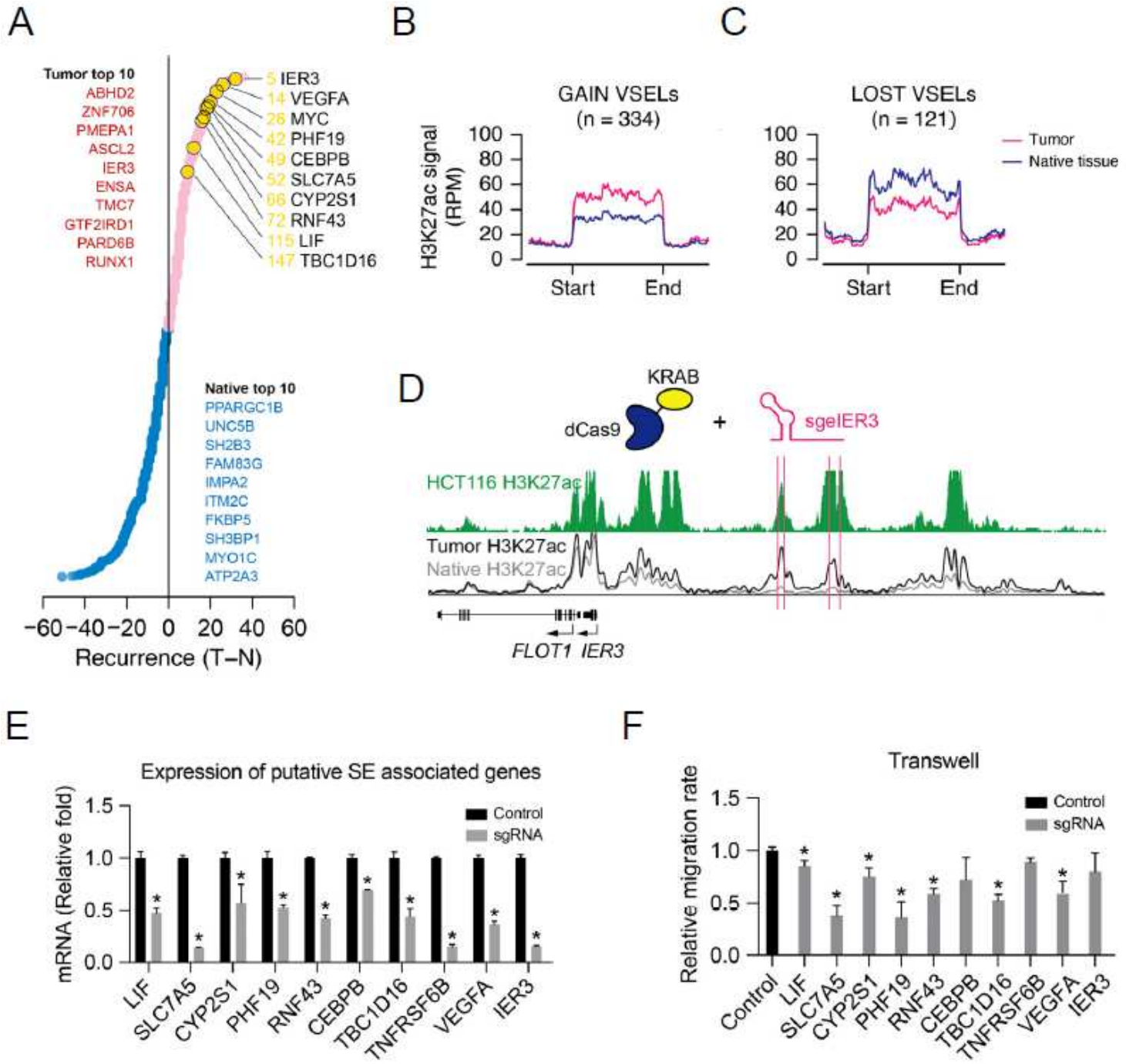


Figure 4

Functions of tumor-specific super enhancers in CRC. (A) The genes associated with top super enhancers (SEs) ranked by recurrence. Red dots represent tumor specific SE genes and blue dots represent native tissue specific SE genes. Top 10 tumor and native tissue specific genes were listed. (B-C) The average H3K27ac signal (RPM) at the regions of gain VSELs (B) and lost VSELs (C) in tumor and native tissues. (D) Meta normalized H3K27ac tracks at IER3 gene loci. The green track on the top represents H3K27ac signal in HCT116, and the black and grey lines at the bottom represent the average signal of tumor and native tissues, respectively. The pink lines indicate the target positions of dCas9-KRAB sgRNAs. (E) Bar plot showing the relative mRNA level of LIF, SLC7A5, CYP2S1, PHF19, RNF43, CEBPB, TBC1D16,

TNFRSF6B, VEGFA and IER3 in control and sgRNA groups. * p < 0.05. (F) Transwell assays for HCT116 cell lines stably transfected with dCas9-KRAB sgRNAs of the enhancers mentioned in Fig. 5E. * p < 0.05.

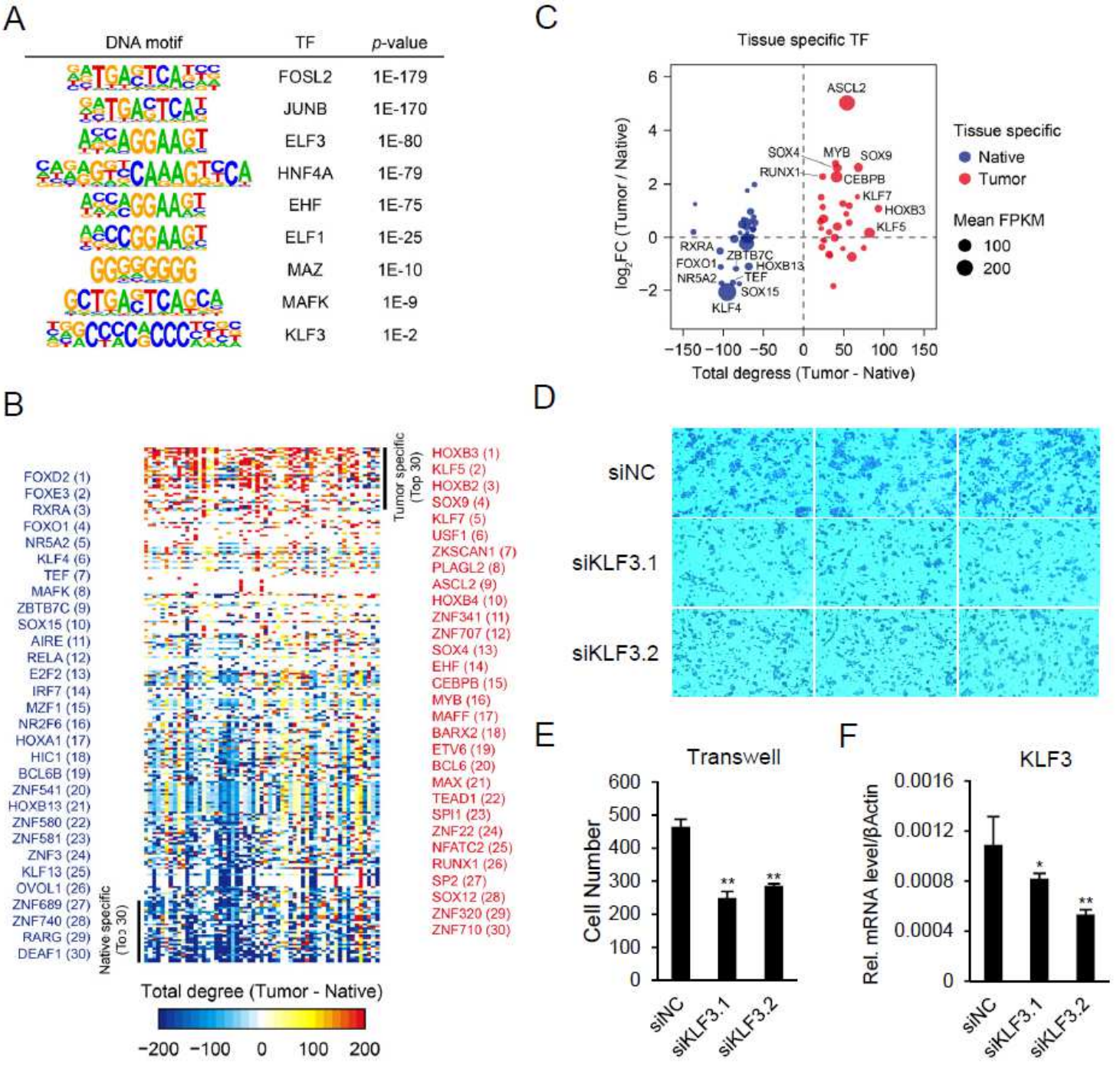


Figure 5

Prediction of functional transcription factors in CRC. (A) DNA motifs enriched within nucleosome-free regions (NFRs) of tumor gain VELs determined by HOMER motif analysis. (B) Heatmap of transcription factors ranked by predicted core regulatory circuitry (CRC) total degrees (Tumor - Native tissue). Top 30 tumor and native specific TFs were listed. (C) Scatter plot showing the total degree (Tumor - Native tissue) and expression FC (Tumor / Native tissue) of the specific TFs listed in Fig. 5B. Blue dots represent

top 30 tumor specific TFs, and red dots represent top 30 native specific TFs. Circle size indicates the mean expression (FPKM) of TFs in its specific tissue. (D&E) Transwell assay for HCT116 cell line with KLF3 knockdown. ** $p < 0.01$. (F) Relative mRNA level for HCT116 cell line after KLF3 knockdown. * $p < 0.05$, ** $p < 0.01$.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupFigures.pdf
- SupTable.xlsx