Global loss of fine-scale chromatin architecture and rebalancing of gene expression during early colorectal cancer development

Yizhou Zhu[1], Hayan Lee[1], Annika K. Weimer[1], Aaron Horning[1], Stephanie A. Nevins[1], Edward D. Esplin[1], Kristina Paul[1], Gat Krieger[3], Zohar Shipony[3], Meredith Mills[2], Rozelle Laquindanum[2], Uri Ladabaum[2], Roxanne Chiu[1],Teri Longacre[4], Jeanne Shen[4], Ariel Jaimovich[3], Doron Lipson[3], Anshul Kundaje[1,5], William J. Greenleaf[1,6], Christina Curtis[1,2], James M. Ford[1,2], Michael P. Snyder[1,*]

[1]Department of Genetics, Stanford School of Medicine, Stanford, CA, United States
[2]Department of Medicine, Stanford School of Medicine, Stanford, CA, United States
[3]Ultima Genomics, Newark, CA, United States
[4]Department of Pathology, Stanford School of Medicine, Stanford, CA, United States
[5]Department of Computer Science, Stanford University, Stanford, CA, United States
[6]Chan Zuckerberg Biohub, San Francisco, CA, United States

*Corresponding author: mpsnyder@stanford.edu (M.P.S.)

Abstract

Although 3D genome architecture is essential for long-range gene regulation, the significance of distal regulatory chromatin contacts is challenged by recent findings of low correlation between contact propensity and gene expression. To better understand the role of long-range interactions between distal regulatory elements during the early transformation from healthy colon to colorectal cancer, here we performed high resolution chromatin conformation capture for 33 samples including non-neoplastic mucosa, adenomatous polyps and adenocarcinomas, mostly from Familial Adenomatous Polyposis (FAP) patients. We identified hundreds of thousands of chromatin micro-structures, such as architectural stripes and loops, which originated from active cis-regulatory elements. Surprisingly, these structures progressively decayed throughout cancer progression, particularly at promoters. Meta-analyses revealed that this decay was independent of alterations in DNA methylation and chromatin accessibility. Interestingly, the degree of interaction loss was poorly correlated with gene expression changes. Instead, genes whose expression were disproportionately lower and higher than their relative promoter interaction in mucosa shifted their expression in polyps and adenocarcinomas to yield a more direct relationship between strength of interaction and gene expression. Our work provides the first high resolution 3D conformation maps during early cancer formation and progression, and provides novel insights into transcriptional readouts associated with fine-scale chromatin conformation alterations.

Introduction

Transcriptional programs in mammalian cells are governed by coordinated function of proximal and distal regulatory elements. Under the constraint of 3D chromatin folding structures known as Topologically Associating Domains (TADs)[1-3], long-range physical interactions are formed between cis-regulatory elements (CREs). The process of direct promoter-enhancer (P-E) looping has long been proposed as essential for gene regulation[4]. Consistent with this concept, artificial interventions of P-E interactions through forced looping or disruption of TAD insulation have demonstrated direct impact on gene expression[5-7]. However, the necessity of stable P-E loops for gene regulation has been challenged by the observations from several recent studies. Fluorescent imaging of functionally paired promoters and enhancers indicated that their distance did not necessarily decrease during gene activation[8,9]. In addition, disruption of TAD structures through depletion of CTCF/cohesin or their regulators caused a surprisingly modest impact on the global transcriptome[10,11]. Furthermore, drastic expression changes either between cell types or induced by heat shock have been accompanied with only subtle changes in chromatin structure[12,13]. Together, these observations raised questions about the extent to which transcription levels are determined by contact propensities between CREs.

Our knowledge of high order chromatin folding has been largely driven by the advent of chromatin conformation capture (3C) methodologies. *In situ* Hi-C, the current state-of-the-art global 3C method, robustly identifies TAD structures but inefficiently captures fine structures such as promoter-enhancer interactions due to resolution and sequencing depth restraints[14,15]. To address this caveat, a wide variety of techniques enriching CRE regions through chromatin-binding proteins/histones[16], sequence-based capture[17], and chromatin accessibility[18] were developed. However, such approaches have systematic biases that cannot distinguish contact propensity from the availability of corresponding enrichment markers. In contrast, recent low bias mapping techniques, such as micro-C[19] and MCC[20], detect CRE interactions through improving resolution to sub kilobase levels. We previously found that a similarly high resolution could be obtained through increasing the number of restriction enzymes (REs)[21], and such multi-RE strategy required substantially lower material input and protocol complexity compared to MNase digestion. Consistent with micro-C, the multi-RE digested Hi-C (mHi-C) revealed highly detailed interaction networks among CREs including regions that lacked strong binding of CTCF/cohesin, which are considered crucial in mediating interactions in the loop extrusion model[22,23].

The development of colorectal cancer (CRC) typically involves a standard progression from normal colorectal mucosa to the formation of precancerous polyps that ultimately undergo malignant transformation[24,25]. Over 80% of colorectal carcinomas are initiated by loss-of-function mutations of APC, a key component in the cytosolic complex that targets β-catenin for destruction and suppresses Wnt signaling[26,27]. In patients with germline APC mutations, a condition also known as Familial Adenomatous Polyposis (FAP), tens to thousands of precancerous polyps are formed in adolescence or early adulthood[28,29]. The large heterogeneous polyp collection arising from the same germline genetic background provides an ideal system to study epigenetic molecular events during pre-malignant to malignant transformation.

In this study, we take the advantage of mHi-C to analyze the small sample amounts of FAP polyps and investigate fine-scale conformation changes during early stages of CRC development. Using mHi-C, we generated high resolution chromatin contact maps for 33 colon samples representing different stages of CRC progression, including non-neoplastic mucosa, adenomatous polyps, and adenocarcinoma, from 4 FAP patients and 6 sporadic CRC patients.

As part of the Human Tumor Atlas Network (HTAN)[30], we also collected the transcriptome, methylome, and chromatin accessibility profiles from the same patient group. The mHi-C robustly revealed distal interaction activities of CREs, demonstrating that instead of CTCF-binding sites found at TAD boundaries, unmethylated promoter regions most actively formed interaction stripes, i.e. high frequency interaction with entire neighboring domains, and loops. The contact propensities of regulatory elements progressively decreased in polyps and further in adenocarcinomas, with the trend most significant and consistent at active transcription start sites (TSS). The degree of interaction loss was higher in hypermethylated regions and regions that lost accessibility, although hypomethylated regions and those that gained accessibility also showed trends of loss. In mucosa, promoter stripe strength did not linearly correlate with gene expression but instead, a baseline interaction level distinguished the majority of robustly expressed genes from the low-expressed ones. In polyps and cancer, the overall loss of interaction caused the majority of promoter stripes shifting closer to the baseline, and the correlation between stripe and expression was concomitantly increased. This increased correlation was driven by the reduction of the relative gap between stripe strength and expression levels, such that genes with disproportionately strong and weak interaction activities tended to be up- and down-regulated, respectively. Based on this observation, we propose that impaired CRE interaction networks in cancers results in strong stripe- and loop-forming promoters acquiring a transcriptional advantage, and we found genes with such characteristics were highly enriched in cancer-driving gene pathways. Overall, our results show how mapping fine-level chromatin structure at high resolution provides novel insights into gene expression and early stages of oncogenesis.

Results

mHi-C reveals cis-regulatory element interaction networks

We mapped 3D chromatin interactions at high resolution (200-500 bp) using mHi-C, a protocol we adapted from the *in situ* Hi-C protocol[31] (Figure **S1A**, **Methods**). In this protocol, we included a cryo-substitution preprocessing step which was found to increase detection of interactions in frozen tissues. In addition, a mild detergent condition was used, and multiple restriction enzyme digestion was performed for preservation and detection of fine interaction signals involving regulatory elements. By using Tn5 tagmentation to construct libraries with improved sample efficiency, the protocol required as low as a few milligrams tissue input, which was important for the very low sample amounts of many of the polyps. We generated mHi-C data for 33 frozen colon tissue samples (Figure **1A**). This included 7 non-neoplastic mucosa, 19 adenomatous polyps, and one adenocarcinoma from 4 FAP patients, as well as six additional adenocarcinoma samples from non-FAP individuals who developed sporadic CRC. A total of 1.59 billion unique intrachromosomal long-range (≧ 1kb) interaction contacts were mapped (Figure **S1B,C**).

The pooled data from all samples revealed distal interaction at resolutions up to 200 bp (Figure **1B**, **S1D**). At sub-kilobase resolution (200 bp – 1 kb), we visually observed interaction stripes, which indicate high frequency interactions with the entire domain, that extended from CREs such as promoters and CTCF binding sites. Along the stripes, strong interaction dots indicating long range looping were evident, where the other anchor often overlapped with a distal regulatory element. These interactions were not present in *in situ* Hi-C despite visualization at same binning resolutions (Figure **S1D**), suggesting that finer digestion of mHi-C improved the capture of CRE interactions that were previously missing.

To statistically identify interaction stripes and loops, we applied a custom peak-calling algorithm and HiCCUPS[32],respectively, and identified a total of 254,642 stripes and 279,480 loops from all

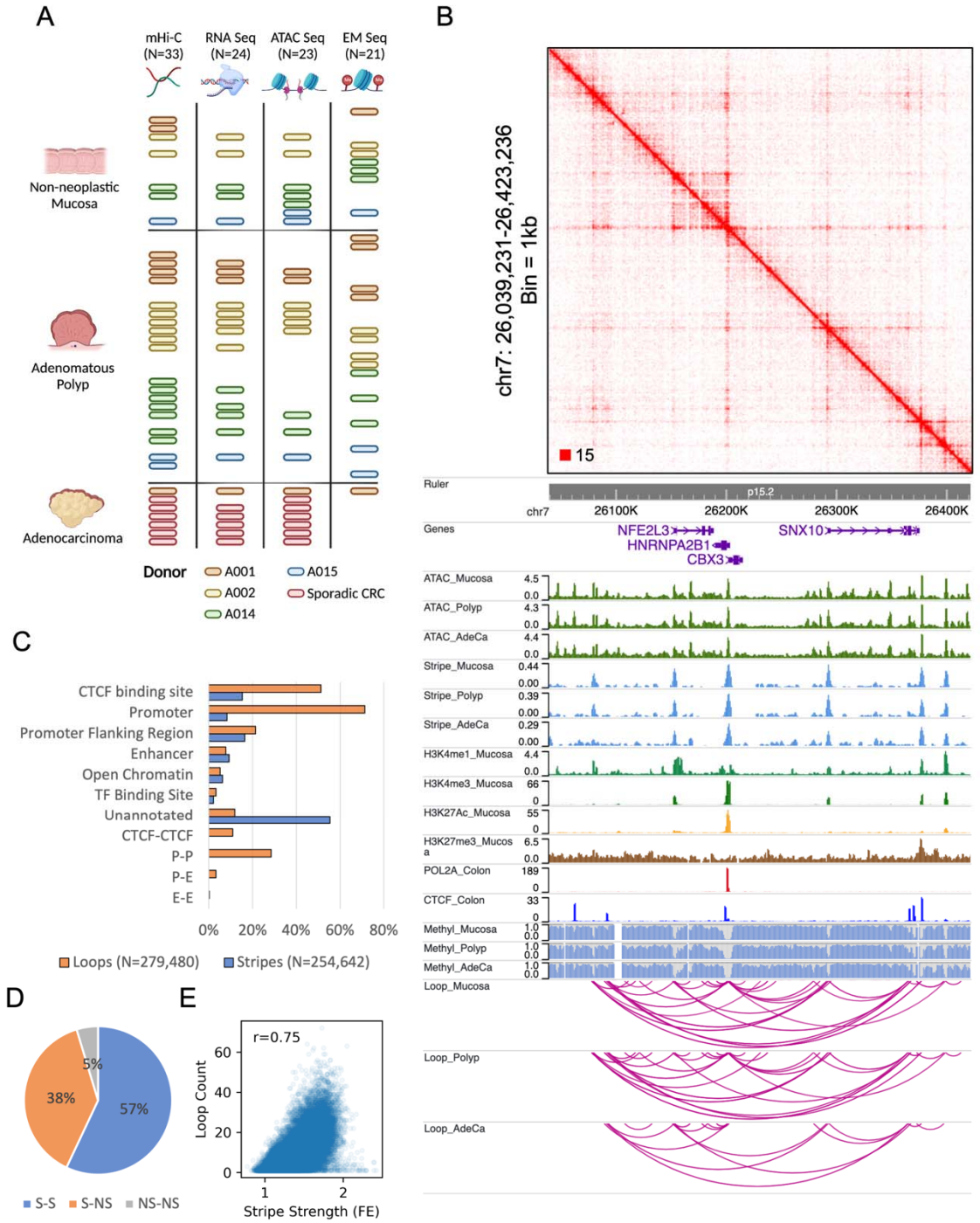

Figure 1 mHi-C reveals fine-scale chromatin structures associated with CREs. (A) Summary of colon tissue samples examined by multi-omics assays. Colors of bars indicate donors. For sporadic colorectal cancers (CRCs), each row represents a distinct donor patient. (B) An example locus showing alignment of micro-structures revealed by mHi-C (top) at 1 kb resolution with other 1D epigenomic features. Bottom panel indicates significant loop interactions at stages. (C) Annotation of stripes and loops by the regulatory element landscape in sigmoid colon. (D) Fraction of loops formed between two stripe anchors (S-S), a stripe and a non-stripe (S-NS), or two non-stripe anchors (NS-NS). (E) Scatter plot between the stripe strength and the total number of loops formed the anchor. Pearson correlation r is indicated.

samples (Figure **1C**, **Methods**). Annotation of the anchors using the Ensembl Regulatory Build database[33] revealed that 45% of stripes and 88% of loops were associated with active chromatin regions. Notably, only 39,658 (15%) stripes overlapped with CTCF sites, compared to 75,795 (30%) that intersected functional elements without the factor. Similarly, interactions between pairs of CTCF sites only accounted for a small portion (11%) of all loops. In particular, almost three fold more Promoter-Promoter (P-P) loops were identified compared to CTCF-CTCF loops (29% vs 11%), suggesting that promoters are among the highest interaction activity. Together, these results confirmed that resolution improvement of mHi-C enabled robust identification of a large number of long range interactions of active CREs.

Comparative analysis showed that the activities of forming site-specific loops and non-specific interaction stripes were strongly correlated. The anchors of loops and stripes highly overlapped, with 95% of loops harboring at least one stripe-forming anchor (Figure **1D**). Consistently, the quantitative strength of stripes significantly correlated with the number of loops formed by the anchor (Figure **1E**). Thus, we consider stripe formation and strength a robust surrogate for investigating the looping activity of a locus. This surrogation is preferred for quantitative analyses since compared to the sparse 2D loop profiles, the stripe signals are typically supported by tens of folds more read counts and thereby more precise and statistically powerful.

Trajectorial decay of CRE interaction networks during early colorectal cancer development

To study the stage-dependent transition of CRE interactions, we first used DESeq2[34] to identify significant differential stripes. Using a threshold of ≧1.3 fold change and 10% FDR cutoff, we obtained 1,580 and 58,076 differential stripes between mucosa-polyp and mucosa-adenocarcinoma, respectively, (Figure **S2A,B**). The majority (77%) of stripes altered in polyps were also changed in adenocarcinoma (Figure **2A**). Comparisons of the mean fold enrichment of these stripes in individual samples revealed their higher gains and losses in adenocarcinomas than polyps (Figure **2B**). Together these results indicate extensive chromatin conformation changes in polyp formation with alterations further extended during malignant transition.

We further examined if stripe alterations were indicative of neoplastic stage progression. Principal component analysis (PCA) of differential stripes between non-neoplastic mucosa and adenocarcinoma revealed the first principal component (PC1) explaining the majority (71%) of cross-sample variation (Figure **S2C**). Under this projection, mucosa samples showed perfect separation from adenocarcinomas, whereas polyps formed a continuum between the two stages, presumably reflecting their degree of transformation (Figure **2C**). Ranking of the samples revealed a total of five samples that did not follow the normal mucosa-polyp-adenocarcinoma trajectory, and hence this simple analysis resulted in a stage classifier with 85% test accuracy. We further performed PCA separately on the gained and lost stripes, and found the sample positions on the two PC1 axes were highly consistent ($r = 0.94$, Figure **S2D**), suggesting that the overall degree of stage-dependent stripe gains and losses were comparable across samples.

We next examined if the PC1 was indicative to the heterogeneity of the polyps. Interestingly, the scores of the polyps fell into two clusters, with 4 closer to the adenocarcinomas and the rest nearest to the mucosa samples. Tracing the origin and phenotypes of these polyps, we found that the adenocarcinoma-like polyps were not larger in size, which is considered a signature of higher dysplasia, but instead sourced from the two FAP patients who developed adenocarcinoma (Figure **S2E**). Indeed, these patients harbored both cancer-like and non-

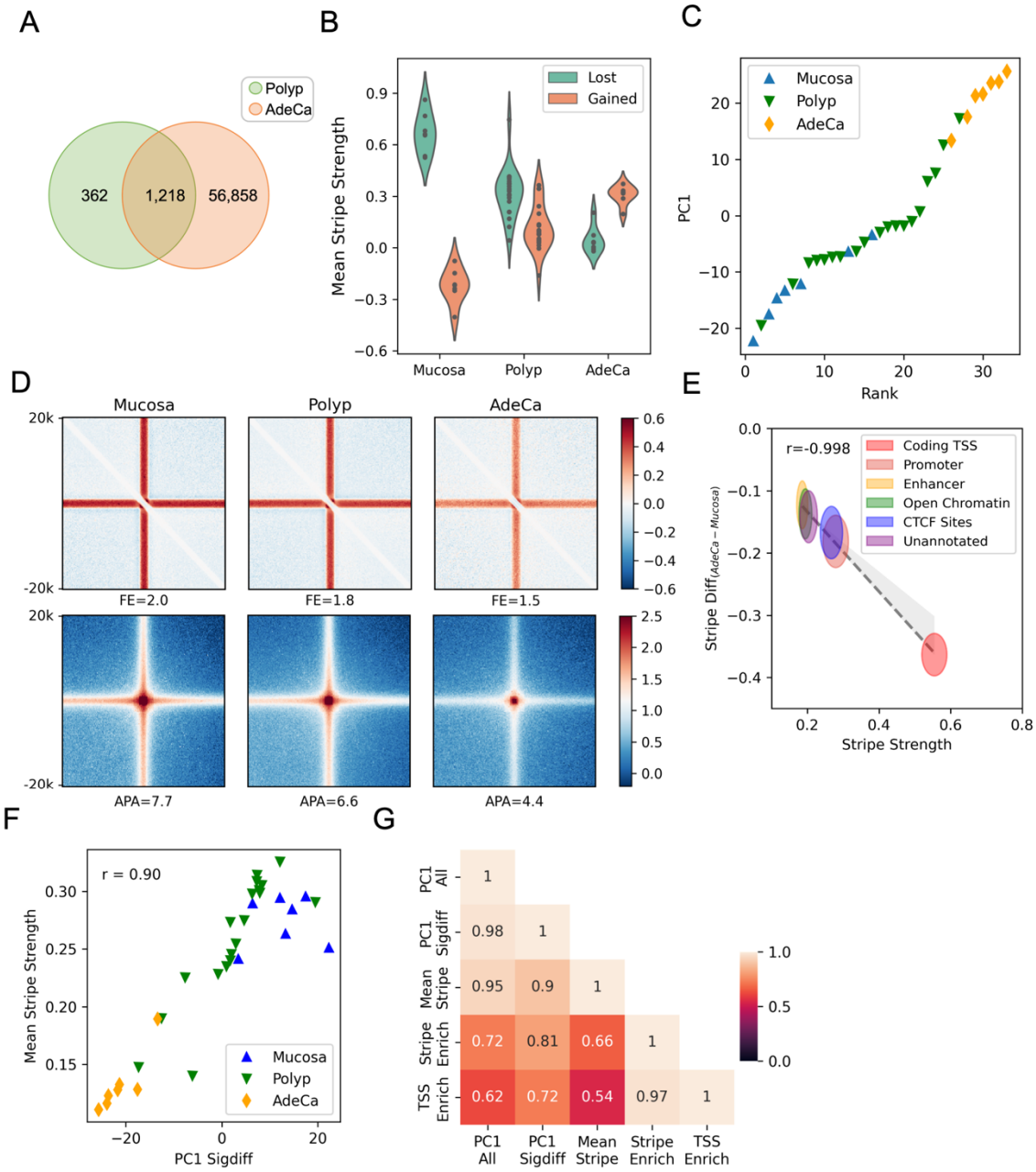

Figure 2 Progressive loss of interactions along cancer development trajectory. (A) Overlap of significant differential interaction stripes between mucosa-polyp and mucosa-adenocarcinoma. Only the overlaps that changed in the same direction are included. (B) Distribution of mean fold enrichments for stripes significantly lost (N = 982) and gained (N = 598) in polyps. (C) Rankings of samples based on PC1 from PCA analysis of significant differential stripes in adenocarcinoma. (D) Aggregated peak analyses (APA) for stripes and loops at different stages. Fold enrichment (FE) and APA scores indicate center to background signal ratios. (E) Comparison between baseline stripe strength in mucosa and degree of stripe loss in adenocarcinoma. Center and radius for each ellipse represents mean and standard deviation for stripes annotated with corresponding regulatory elements. Dotted line indicates linear regression of ellipse centers. (F) Sample correlation between PC1 values indicated in (C) and mean stripe strengths. Pearson coefficient r is indicated. (G) Pearson correlation matrix for

sample trajectories derived from (i) first PCA component of all stripes (PC1 All) (ii) first PCA component of significant differential stripes between mucosa and adenocarcinoma (PC1 Sigdiff) (iii) mean stripe strength for all anchors (Mean Stripe) (iv) fraction of intrachromosomal distal interaction reads mapped to all stripe anchors (Stripe Enrich) and (v) fraction of intrachromosomal distal interaction reads mapped to all gene promoters (TSS Enrich).

neoplastic mucosa-like polyps, suggesting their higher polyp heterogeneity than patients without adenocarcinoma at the time of colectomy, which may contribute to higher likelihood of malignant transformation.

We also found that the significantly differential stripes was due to a preferential loss of features, with the loss-gain ratio being 1.6 fold in polyps and 3.3 fold in adenocarcinomas (Figure **S2A,B**). Interestingly, such imbalance was observed in similar ratios from the entire stripe profile (Figure **S3A**). Aggregation Peak Analyses (APA)[32] revealed progressively lower average signals for both stripes and loops (Figure **2D**) at polyp and cancer stages, suggesting an overall loss of interactions at genome-wide level. This stage-dependent loss is consistent with the all feature PCA patterns (Figure **S2F,G**). To further examine the relationship between interaction decline and stage progression, we first categorized stripe anchors by their functional annotation. We found highest stripe loss at transcription start sites (TSS), followed by epigenetically annotated promoter regions, CTCF, and other regulatory elements (Figure **2E**). This order was nearly perfectly inverse-correlated with their average fold enrichments in mucosa (r = -0.998), indicating that the degree of stripe loss was proportional to their starting values (Figure **S3B-D**). We then calculated the mean stripe loss for each sample, which formed a new trajectory by ranking in order of degree of loss. Compared with the trajectory derived from differential stripes, we observed strong sample-to-sample correlation (r = 0.9, Figure **2F**). Notably, the stripe signal was comparable between non-neoplastic mucosa and mucosa-like polyps, but more reduced in adenocarcinoma and cancer-like polyps, suggesting well correspondence of the severity of stripe loss with malignant transformation.

Despite using a careful normalization against both self-ligation frequency and local background to quantify stripe strengths (**Methods**), the detection of unidirectional genome-wide peak loss is susceptible to potential experimental caveats, particularly with the unstable genomic background of cancer. Hence, we performed multiple quality control analyses to validate stripe loss by ruling out confounding factors. We first calculated the fraction of distal interaction reads in stripe anchors and gene transcriptional start sites (TSS), and observed an expected progressive fraction loss (Figure **S2H,I**). Such losses among samples were highly correlated with trajectories defined above (Figure **2G**), indicating that the degree of signal loss was consistent between measurements. To investigate the impact of cancer-associated genome instability on stripe quantification, we identified loci with copy number changes and rearrangement events in the adenocarcinoma samples (Figure **S4A,B, Methods**). Stripes in these regions were not found to be significantly different from chromosomal averages, suggesting that the impact of large chromosome abnormality on interaction quantification was controlled by normalization (Figure **S4C,D**). Lastly, to test if the reduced stripe signals were caused by cell composition changes associated with cancer, we compared promoters for genes specifically expressed in epithelial, stromal, and immune with housekeeping genes (**Table S1**). The stage-dependent stripe losses were observed for all four gene groups (Figure **S4E**), suggesting that this trend was unlikely caused by a shift in cell type compositions. These results proved the validity of the loss of interactions during cancer stage progression.

DNA Methylation and chromatin accessibility do not fully explain decays of CRE interaction

To investigate the possible underlying mechanism(s) of CRE interaction loss, we performed 23 bulk ATAC seq and incorporated 21 EM seq from a separate study[35] for the colon tissues from the same patients examined by mHi-C with high sample overlap (23/33 samples; Figure **1A**). Aligning the methylation profile with stripe strengths in mucosa, we found they were significantly inversely correlated (r = -0.52). However, such correlation was not linear; instead, a critical

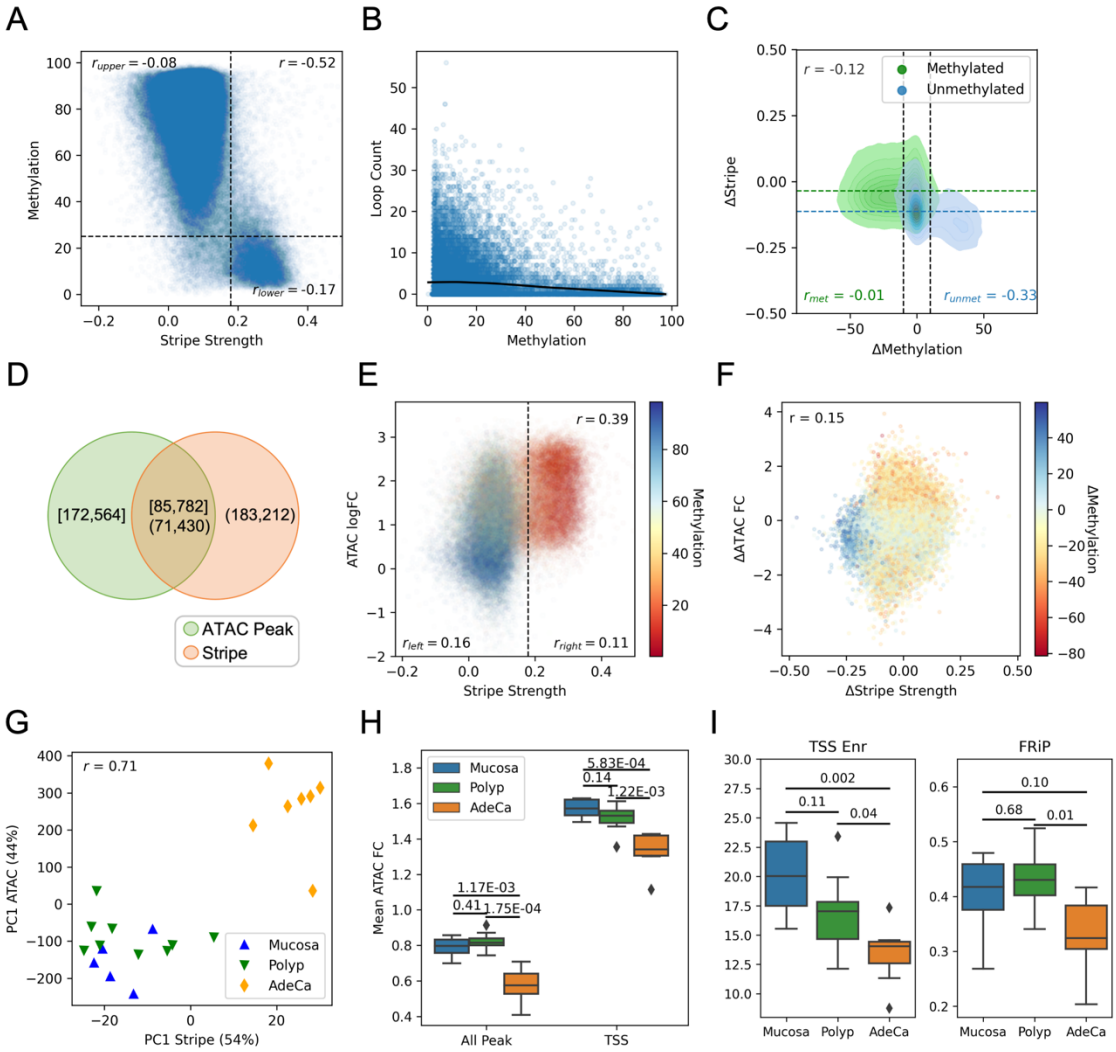

Figure 3 Association between contact propensity and DNA methylation/chromatin accessibility (A) Scatter plot between interaction stripe strength (log10 fold enrichment) and DNA methylation degree (percent) for all stripe anchors with measurable methylation sites (N=230,658). Pearson correlation was calculated for all dots as well as those falling in the upper-left and bottom-right threshold zones indicated by the dash lines. (B) Scatter plot between DNA methylation (percent) and loop count for all promoters (N=14,833) in mucosa. Black line indicates Lowess regression. (C) KDE plot showing differential methylation versus stripe strength between mucosa and adenocarcinoma for stripe anchors that are unmethylated (<25%, N=14,543) and methylated (>50%, N=200,115). Pearson correlation r is indicated for each comparison. (D) Overlap between stripe anchors (N = 254,642) and ATAC peaks (N = 258,346). (E) Scatter plot between stripe and ATAC peak strengths in mucosa and (F) their differential changes in adenocarcinoma (N = 70,301). Pearson correlation r was indicated. Colors indicated (differential) methylation. (G) Correlation between PC1 values from PCA of whole stripe and ATAC seq profiles for samples examined by both assays (N = 21). (H) Boxplots showing mean fold change and (I) fraction of reads in all and TSS ATAC peaks by sample stages. Significant p values were calculated using Mann-Whitney U test.

threshold of methylation fraction (25%-30%) distinguished strong stripes from the remaining (Figure **3A**). For the low methylation regions (<25%), further loss of methylation was found to weakly correlate with higher stripe strength (r = -0.17). Consistent with the stripe signals, lower methylation was also associated with higher loop counts (Figure **3B**). These results indicated that high interaction activity is associated with DNA unmethylation.

During polyp/adenocarcinoma transition, alteration of DNA methylation is a mix of global hypomethylation of normally methylated (>40%) loci and specific hypermethylation of a subset of unmethylated (<25%) CpG sites (Figure **S5A**), a typical pattern for cancers[36-38]**.** This trend was more significant at the adenocarcinoma stage than in polyps, and also at TSSes relative to all stripe anchors. We found that hypermethylation of normally unmethylated regions in adenocarcinoma was associated with increased stripe loss (Figure **3C**), consistent with the known repressive role of DNA methylation. However, among normally methylated regions, 72% also showed stripe reduction and such change was independent to the degree of their hypomethylation (r = -0.01). Furthermore, among the loci showing low methylation changes (< ±10%), unmethylated regions experienced higher stripe loss than methylated (Figure **3C**), suggesting that the degree of loss was associated with the absolute methylation level instead of its change during cancer development.

We next investigated whether stripe losses were accompanied with chromatin accessibility changes. Using ATAC-seq, we identified a combined 258,346 significant open chromatin sites in 23 samples (Figure **1A**). Thirty-three percent (85,782) of these sites overlapped with 28% stripe anchors (Figure **3D**). Compared to methylation, the fold enrichment of ATAC peaks was much less correlated with stripe signals (Figure **3E**), suggesting distinct underlying factors for the two signals. Correspondingly, alterations of ATAC peaks and stripes in adenocarcinoma were only weakly associated (r = 0.15, Figure **3F**), whereas both signals correlated stronger with DNA methylation (Figure **3C, S5B**). Interestingly, despite low linear correlation, the PCA analyses of ATAC peaks surprisingly similar to the stripes, in which polyps formed a continuum between clearly separated non-neoplastic mucosa and adenocarcinomas (Figure **3G**). Furthermore, consistent with stripes, the ATAC peak signals showed global loss in adenocarcinomas, although not in polyps (Figure **3H,I**). We further found that the mild gain of ATAC signals in polyps was a mix of loss on promoters and gain at other CREs such as enhancers and TF binding sites (Figure **S5C,D**). Altogether, these results showed that the strong loss of contact propensity on promoters was accompanied with loss of accessibility, while alterations of the two signals were less consistent on distal CREs.

Deviation of gene expressions from promoter distal interaction predicts their alterations during early cancer progression

To understand the functional inferences of the interaction decays, we performed 24 sample-matching bulk RNA seq experiments (Figure **1A**), identifying 4,497 and 9,011 significant differentially expressed genes in polyps and adenocarcinomas. PCA analyses of significant genes or the whole transcriptome resulted in a consistent pattern similar to that found for mHi-C, where samples showed trajectorial cancer stage progressions on PC1 (Figure **S6A,B**). The differentially expressed genes in polyps largely overlapped with adenocarcinomas (Figure **S6C**), and the majority of overlapping genes showed higher fold change in the same direction (Figure S6D). These results indicate that a significant portion of oncogenic expression alterations occurred at the early polyp stage.

To investigate the relationship between expression and chromatin interaction, we first compared promoter stripes and genes that were significantly different in both polyps and adenocarcinomas

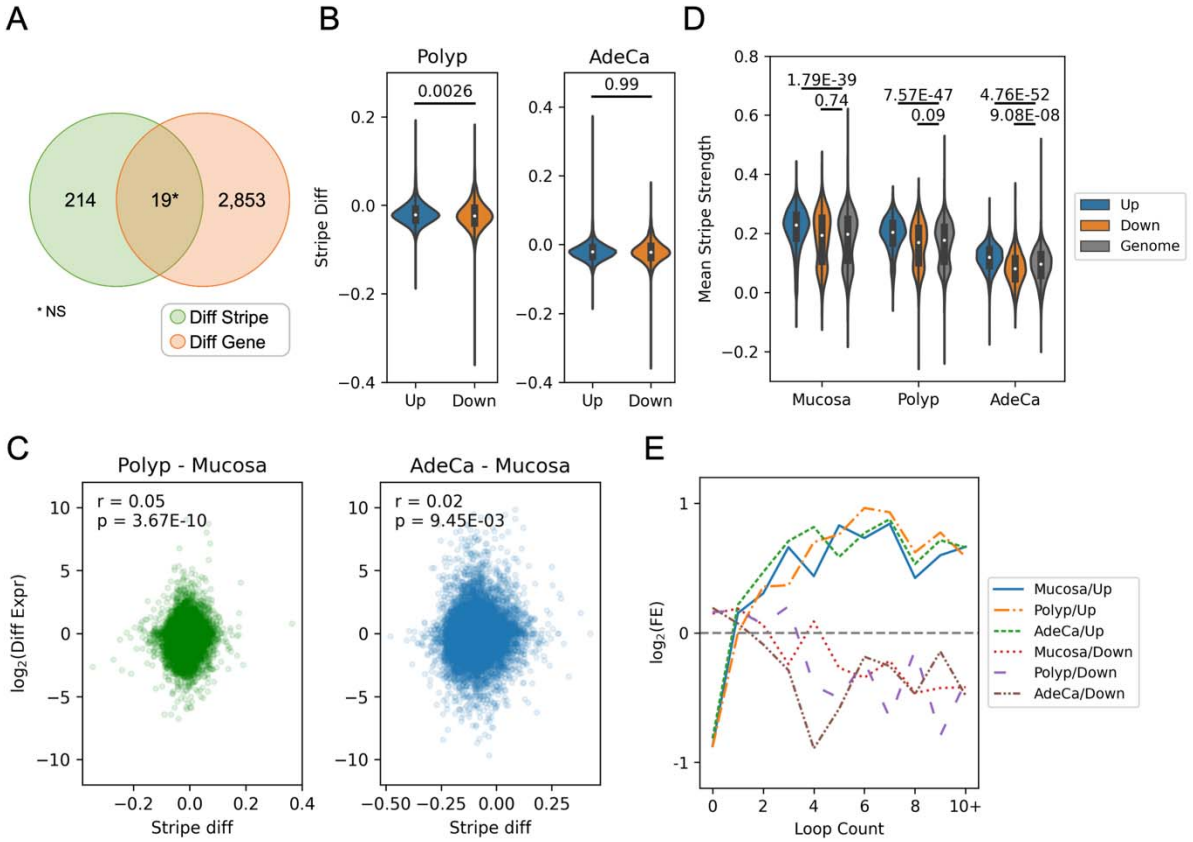

Figure 4 Alterations of gene expression in polyps and adenocarcinoma are associated with relative promoter interaction activities but not their changes. (A) Venn diagram showing overlap between promoter stripes and gene expression that are significantly different in both polyps and adenocarcinomas against mucosa. Significance of overlap was assessed by chi-squared test (N.S.: not significant). (B) Violin plots showing differential stripe strengths for genes up- (N = 1,210) or down-regulated (N = 856) in both polyps and adenocarcinoma. Significant differential strengths compared to genome average were determined by t test, with p values indicated on top of the figures. (C) Scatter plots between differential stripes and transcription fold changes determined by DESeq2 (N = 18,925). Pearson correlation r and its significance are indicated. (D) Violin plot showing mean stripe strengths of consistently up- and down-regulated genes compared to whole genome. Student t test was used to test significant difference. (E) Fraction of genes with designated loop counts for commonly up- and down-regulated genes, measured as log2 fold difference compared to genome average, at different stages.

and found their overlap was low and insignificant (Figure **4A**). Consistent with this observation, the stripe loss for up- and down-regulated gene promoters were strikingly equivalent (Figure **4B**), and at genome-wide level, the quantitative stripe difference showed no linear correlation with expression change (Figure **4C**). These results together suggested that changes of stripes had little predictable impact on differential expression. Importantly, however, we found that at all stages, the stripe signals of up-regulated and down-regulated gene promoters were consistently higher and lower, respectively, than the genome average (Figure **4D**). This trend was further confirmed from other measurements of distal interaction activities, including fraction of stripe-forming genes (Figure **S7A**), loop counts (Figure **4E,S7B**), and loop strengths (Figure **S7C**). These results demonstrated that cancer-associated differential expression is associated with the relative interaction activity of promoters, which remained largely invariant during tumorigenesis.

To further understand how higher and lower promoter interaction activity could respectively be associated with up- and down-regulation in cancer, we compared the stripe fold enrichments and transcription levels (TPM) of genes at different stages. In mucosa, the majority of high-expressed and a small cluster of low-expressed (TPM < 0.25) genes were associated with stripe levels above and below a baseline threshold, respectively (Figure **S7D**). This separation was almost exclusively responsible for the overall correlation between stripes and expression ($r = 0.43$) as neither of the two gene clusters showed positive correlation beyond the baseline. Along with cancer progression, loss of stripes resulted in the majority of high-expressed genes approaching and eventually bypassing the baseline (Figure **S7E**). Correspondingly, in adenocarcinoma, up- and down-regulated genes were respectively enriched at the two sides of the baseline (Figure **S7F**). These observations suggest a possible scenario where stripe strength correlates with expression at a specific dynamic range near the baseline, and when genes shift into such range due to interaction reduction, gene expressions are shifted toward better correlation with their promoter interaction levels.

Contrary to their exceptionally high and low interaction activities, the expression levels of up- and down-regulated genes were similar in mucosa (Figure **S6E**), suggesting a discrepancy between the two parameters. We examined whether expression changes during tumorigenesis could be associated with such discrepancies. By modeling the relationship between gene expression and stripe signal (Figure **5A**) or loop count (Figure **S7G**), we observed a consistent pattern where distal interactions of up- and down-regulated genes predicted by the corresponding gene expression levels were higher and lower, respectively, than the genome average. While such a gap persisted in polyps and adenocarcinoma, the overall rank difference found in mucosa was diminished along with altered gene expression (Figure **5B**). This trend of regression between stripe and expression was even more clearly illustrated at genome-wide level, where the Spearman correlation consistently increased with stages (Figure **5C**). Correspondingly, the net shift of stripe and expression along cancer progression was consistently directed toward the reduction of their difference, and the degree of such reduction showed significant linear correlation with the degree of the mismatch (Figure **5D**). Thus, we conclude that the transcriptomic changes in polyps and cancer can, at least to a certain extent, be described as a process of rebalancing of gene expression levels toward their promoter contact propensities.

Finally, if we postulate that the reduction of transcription-distal interaction discrepancy is a driving force for differential gene expression during cancer development, we investigated if any gene categories are trended for up- or down-regulation in cancer conditions based on their degree of mismatch (**Methods**). A total of 25 significant up-trend pathways were identified, most of which were well known for oncogenesis, including cell cycle, MAPK, Ras, Wnt, PI3K/Akt, growth factor RTK, and mesenchymal transition (Figure **5E**). This enrichment was consistent

with the ontology of actual up-regulated gene pathways, as expected (Figure **S6G**). In contrast, no down-trend pathway was identified despite several well-known down-regulated important immune genes (Figure **S6H**). This discrepancy may be explained by the fact that down-regulated genes with imbalance were only a subset of the whole pathways.

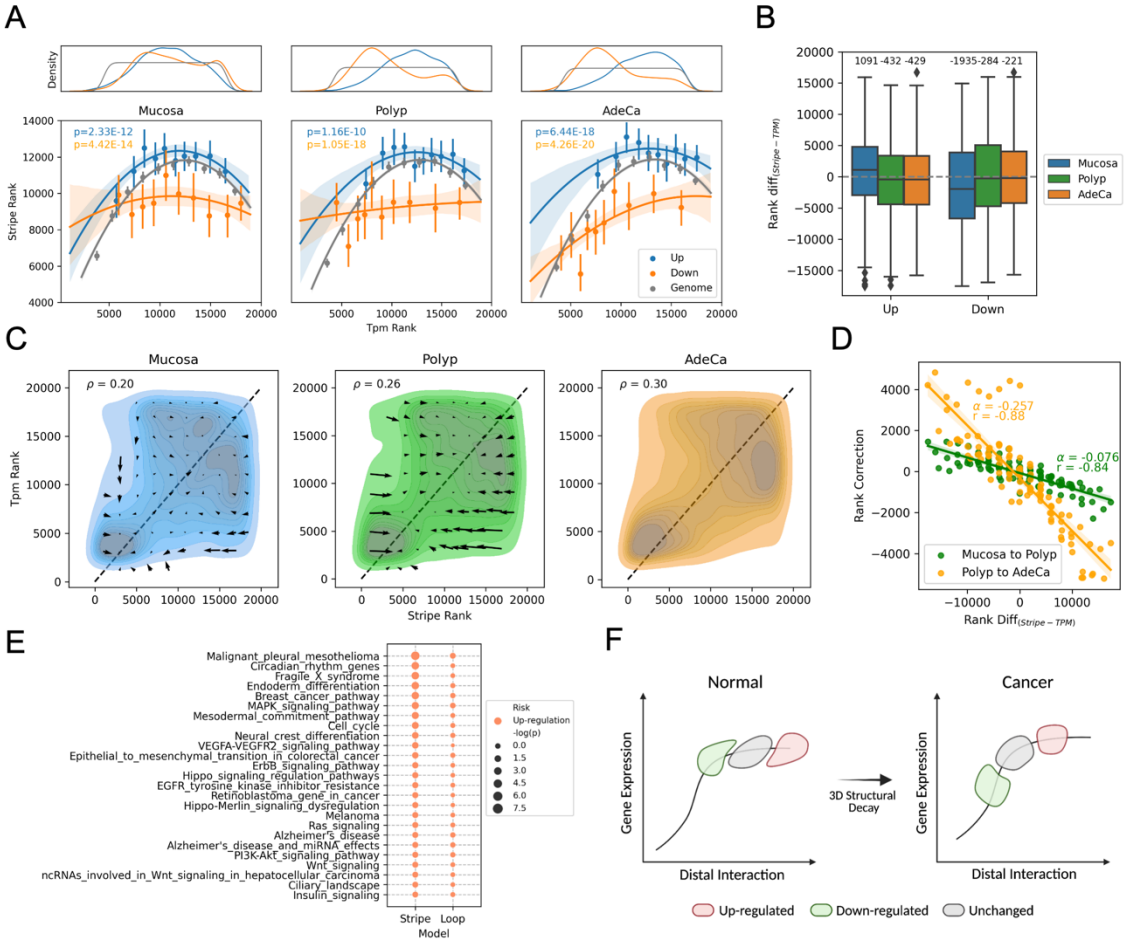

Figure 5 Discrepancies between interaction activity and gene expression predicts their rebalancing. (A) Regression plot between stripe rank and transcription (TPM) rank in different stages. A second-order polynomial model was used to fit the up- and down-regulated genes compared to the whole genome. Dots represent average stripe ranking with standard errors for genes at each one-tenth TPM quartile. Significance of alternative models are indicated. (B) Boxplot indicating distributions of rank difference between stripe and expression for differentially expressed genes at different stages. Median values are indicated on top. (C) KDE plot comparing expression and stripe ranks for all genes (N = 18,925). Spearman correlation ρ for each stage is indicated. Quiver matrices indicate average directionalities of rank shifts in the next stage for genes at each arrow position. Dashed lines indicate equal ranking. (D) Comparison between rank difference (y - x) for the quivers indicated in (C) versus their net rank correction (Δy - Δx) in the next stage. Correlation was estimated using first order linear regression, with coefficient alpha and Pearson coefficient r indicated. (E) Gene pathways predicted with overall up- or down-regulation trends based on similarity in the stripe/loop-expression imbalance compared to the up- or down-regulated genes.(F) A schematic model summarizing the relationship between interaction and gene expression based on the results from this study.

Discussion

In this study, we examined long range chromatin interaction in colon polyps and cancers using mHi-C, a high resolution chromatin conformation capture that revealed extensive interaction activities of cis-regulatory elements without target enrichment. This improved methodology enabled a deep analysis of the limited amount of FAP polyp material, and we were able to identify hundreds of thousands of interaction stripes and loops associated with promoters and distal regulatory elements in non-neoplastic and adenomatous colon tissues from FAP patients. Our rich data provide a valuable resource for 3D interactions during early stages of colorectal cancer.

We observed two major trends regarding the alterations of 3D interactions during early CRC development. First, the CRE contact propensities significantly decreased with neoplastic stages. Second, alterations of gene expression were associated with the relative interaction strengths of promoters, instead of their degree of losses. To reconcile these observations, we propose a simple sigmoid-like relationship between transcription and distal interaction (Figure **5F**). In this model, under normal conditions, most gene promoters maintain high levels of contact propensity, which is abundant and not a rate-limiting factor for transcription (Figure **S7D**). In cancer, however, the overall loss of distal interaction results in higher dependence of gene expression on the interaction strength, thereby yielding a stronger interaction-expression correlation (Figure **S7E**).

Consistent with our model, genes with significantly higher or lower distal expression compared to their promoter interaction strengths tended to shift their expression toward their interaction rank in polyps and cancer. This rebalancing was more significant for genes with larger rank differences (Figure **5C,D**), suggesting that the correction toward linearity between interaction and expression was more consistent for the outliers. Interestingly, our sigmoid model implies an asymmetric pattern of expression alteration between up- and down-regulated genes. Since the down-regulated genes are initially closer to the linear phase of the sigmoid curve, they are expected to be more sensitive to the loss of interactions than the up-regulated genes and thus experience a more significant transition from asymptotic to linear correlation (Figure **5F**). Strikingly, our differential expression result showed that the fold changes of down-regulated genes were overall higher than up-regulated (Figure **S6F**), matching with the predicted asymmetry. Together, our model fits well with experimental results.

We quantitatively compared the contact propensities of CREs with other 1D epigenetic landscapes. We found that while the correlation between stripe signal and chromatin accessibility was only moderate at each individual locus, their shifts during cancer development were surprisingly similar at genome-wide level. Consistent with the interaction losses, the overall accessibility at promoters showed similar progressive reduction. Furthermore, we also found the rebalancing effect between gene expression and ATAC peak strengths was similar to that for the interaction (Figure **S8**), although the trend was less significant. Further investigation is required to uncover the common and distinct regulators affecting chromatin accessibility and interaction activity.

With high coverage EM-seq, for the first time we found that anchors of highly interactive stripes largely overlapped with unmethylated DNA. Previous studies proposed that formation of architectural stripes and loops involved cohesin-mediated loop extrusion by the CTCF-binding anchors[22,39]. However, this model is insufficient to explain the strong stripe signals and high loop counts we identified on unmethylated CpG-rich promoters, which often lack CTCF and cohesin binding. In fact, the robust interaction activity of promoters was typically not clearly shown by

previous Hi-C methods until the recently available high resolution micro-C[19]. In the micro-C study, the CTCF/cohesin binding could not fully explain the newly observed interactions, and proposed Pol II as an alternative looping factor. However, based on our observations, the stripes often extended far beyond the gene body, and for stripes that showed directionality, the side with more robust interaction did not always agree with the direction of transcription (Figure **1B,S1D**). These patterns are difficult to explain by the Pol II mechanism and instead suggest the presence of other interaction-mediating factors. Intuitively, the overlap of stripes with unmethylated CpG island promoters suggests that certain high GC motif binding factors could contribute to distal interaction, although the exact candidates remain to be explored.

The development of cancer is a chronic process involving environmental, genetic, and epigenetic alterations[40]. Common oncogenic mutations often occur in genes involved in the regulation of cell proliferation, growth, and differentiation[41-43]. Interestingly, genes in these pathways are also frequently dysregulated to cause proliferative advantages of cancer cells[44]. While such coincidence could be due to a consequence of natural selection, i.e. only the most proliferative and surviving cells would continue on the course of malignant transformation, our study suggests that the exceptionally high distal interaction activities of cancer driving genes (Figure **5E**) could be a potential underlying factor for their upregulation. During cancer progression, the transcriptional machinery of these genes is likely to be more resilient to the decay of chromatin architecture, which results in expression advantages of the genes in such conditions. Altogether, our study demonstrated that the progressive CRE interaction loss and its rebalance with gene expression suggested a transition of fine-level chromatin structures from colorectal mucosa to adenoma to the malignant colon cancers. Further research will examine whether this pattern is generally applicable to other cancer types, and whether interaction decay is a broad hallmark of early cancer development.

Author Contributions

Conceptualization, Y.Z., M.P.S.; mHi-C and RNA seq, Y.Z.; ATAC seq: K.P., A.W.; EM seq, H.L.,G.K., A.J., D.L., Z.S.; Data Analysis, Y.Z.; Sample Collection, R.L., M.M., A.H.,U.L., E.D.E., R.C., J.S.; Original Draft, Y.Z., M.P.S., Review and Editing, All Authors; Funding Acquisition, A.K., C.C., E.D.E, J.M.F., W.J.G., M.P.S; Supervision, E.D.E., A.K., J.M.F., C.C., W.J.G., M.P.S.

Acknowledgements

Illumina sequencing of mHi-C, ATAC seq, and RNA seq were performed by the Stanford Genomics center. The data was generated with instrumentation purchased with NIH funds S10OD025212 and 1S10OD021763. Some illustration figures were created with BioRender.com.

Ethics Declarations

G.K, A.J., D.L. and Z.S. are employees and shareholders of Ultima Genomics. M.P.S. is a cofounder and scientific advisor for Personalis, Qbio, January.ai, Filtricine, Mirvie, Protos and an advisor for Genapsys.

Data Availability

Raw and processed data are available at NCBI Gene Expression Omnibus (GEO), accession number GSE207954.

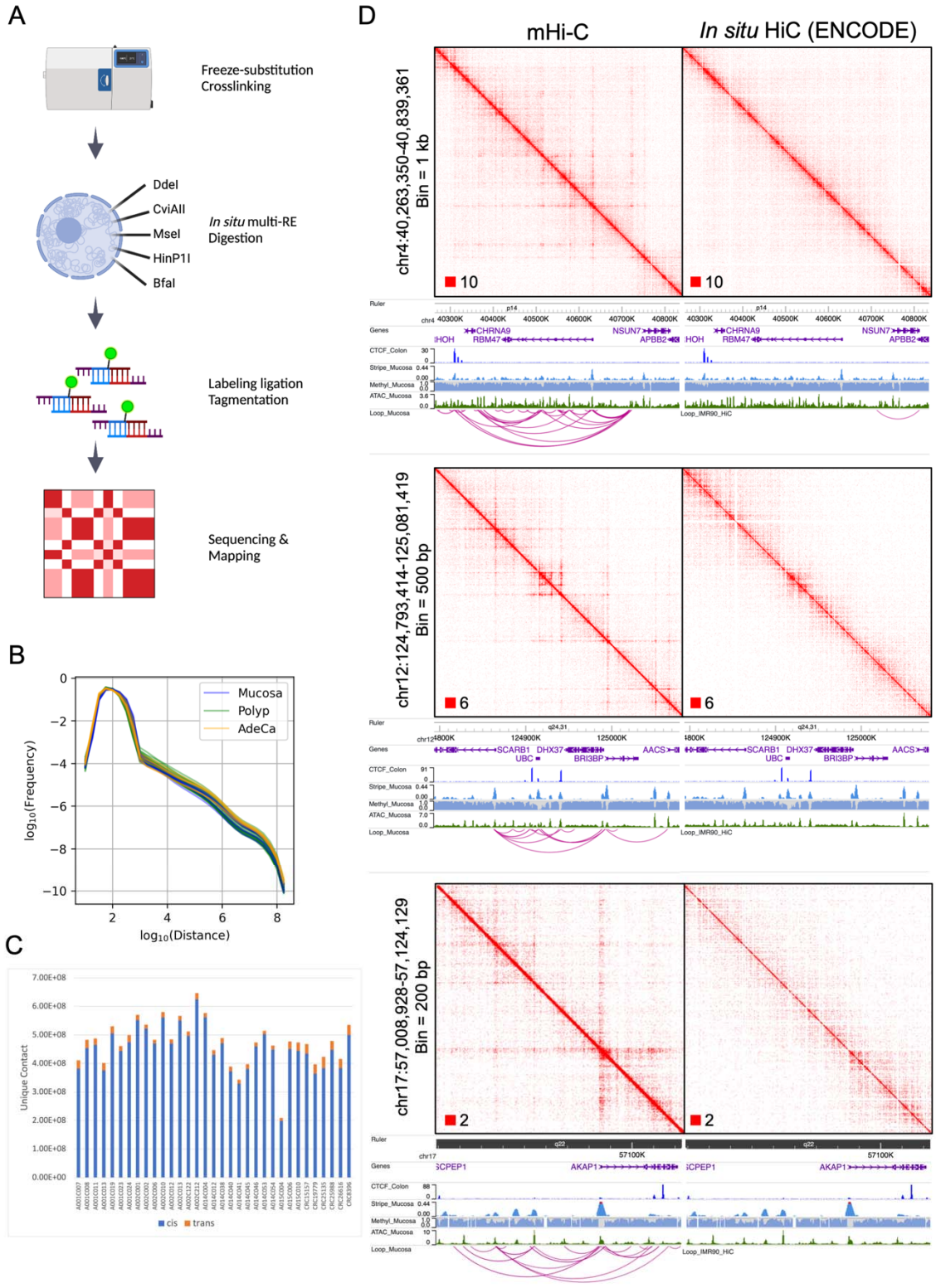

Figure S1 mHi-C design and results. (A) Schematic for the mHi-C workflow. (B) Intrachromosomal interaction frequencies by distance for samples. (C) Number of unique interaction contacts by samples. (D) Comparison of mHi-C with *in situ* HiC (ENCODE) at various resolutions.

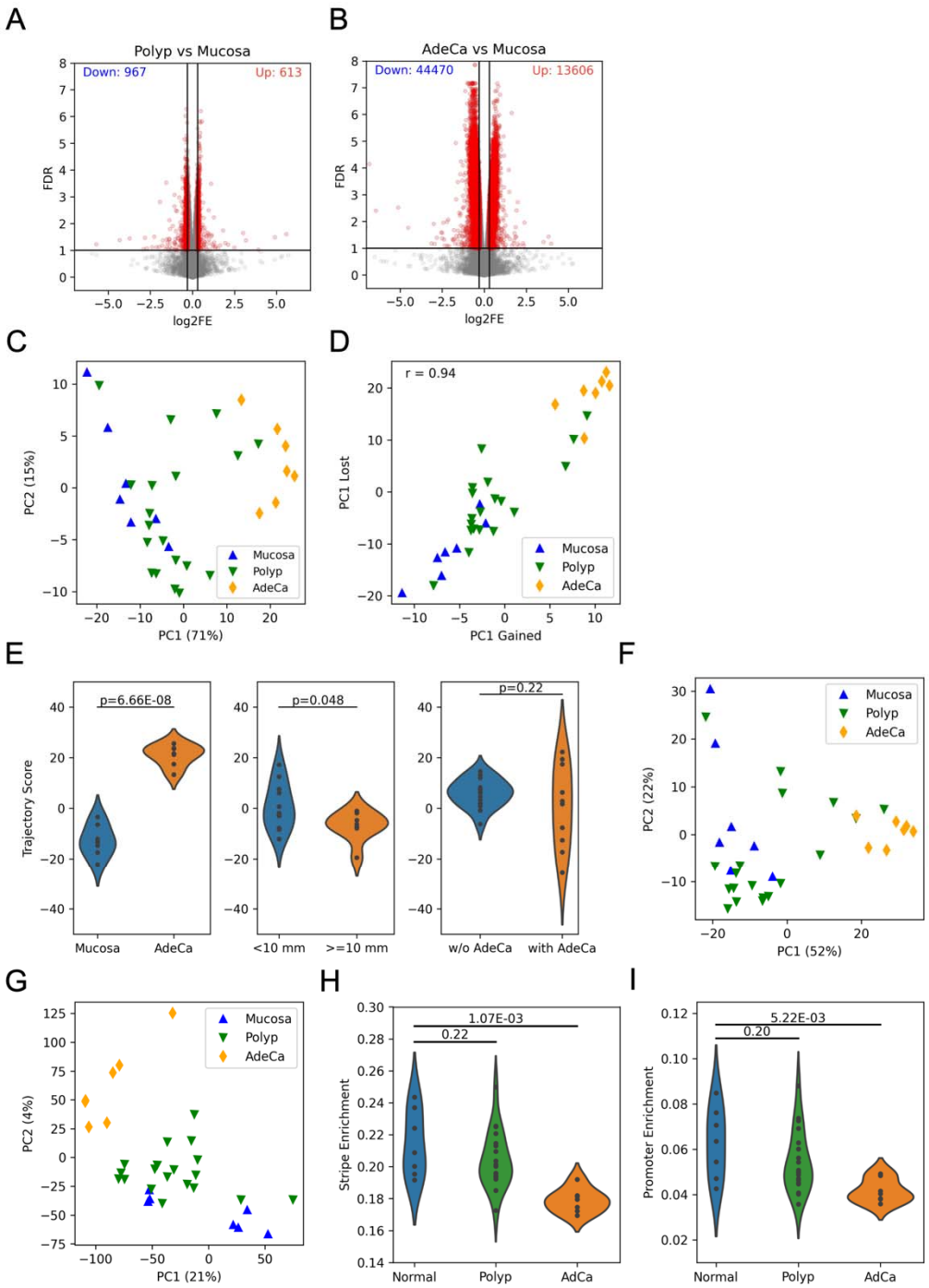

Figure S2 Sample stratification by mHi-C. (A) Volcano plots for differential analyses of stripes between mucosa-polyp and mucosa-adenocarcinoma. (B) PCA analysis of stripes significantly differential between mucosa and adenocarcinoma. (C) Sample correlation using PC1 values from PCAs of gained and lost stripes in adenocarcinoma. (D) Comparisons of PC1 values shown in (B) by sample phenotypes between i) mucosa and adenocarcinoma, ii) small (<10 mm) and large (>= 10mm) polyps, and iii) polyps from cancer-developed FAP patients (A001, A015) versus without (A002, A014). (E) PCA analysis using the whole stripe profile. (F) PCA analysis using the whole loop strength profile. (G) Fraction of distal (> 1.5 kb) intrachromosomal reads on all stripes and (H) gene TSS for samples.

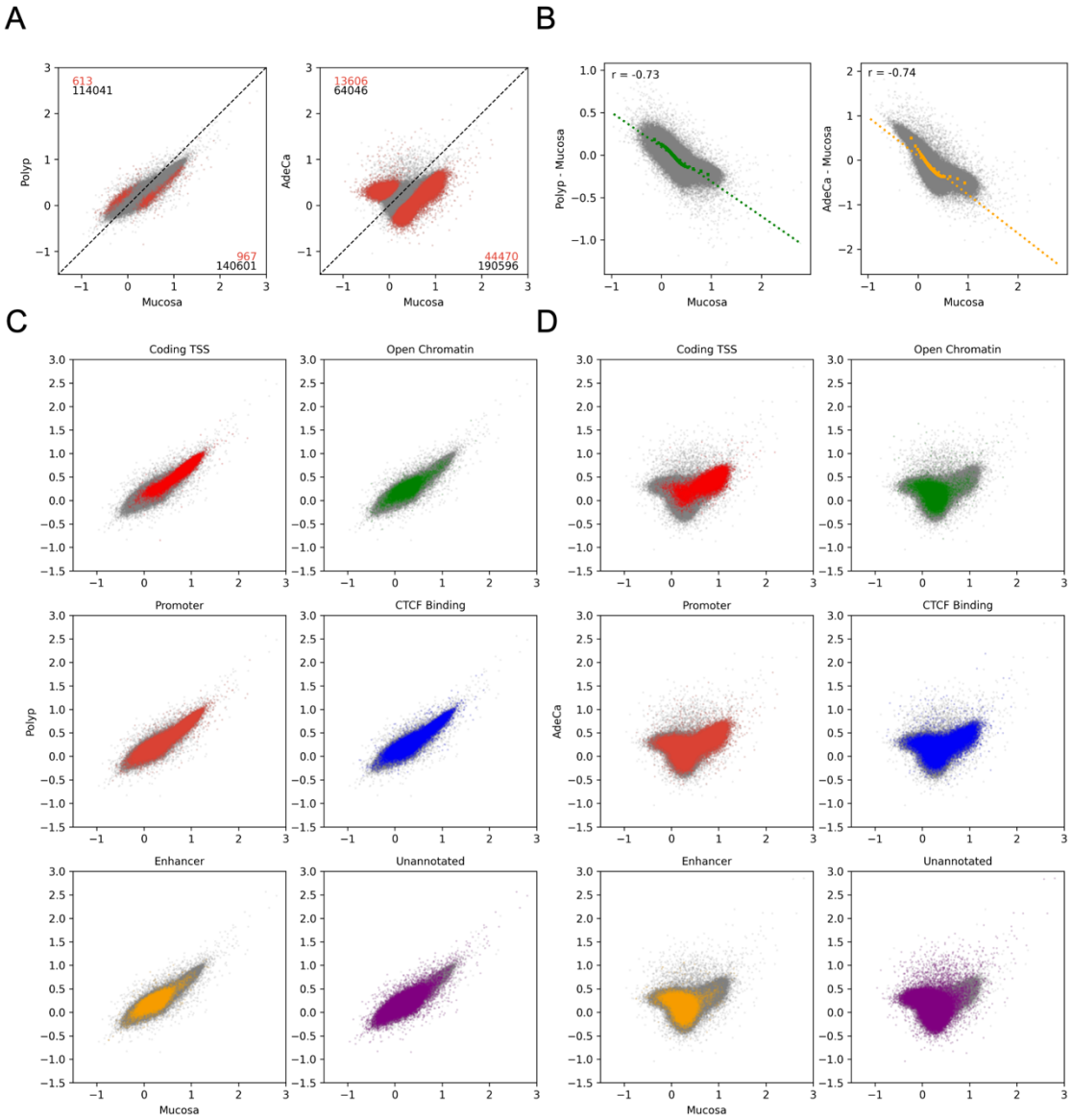

Figure S3 Alterations of stripe strengths in different cancer stages. (A) Scatter plots comparing stripe strengths between different stages. Red numbers indicate significantly differential stripes above and below the line of y = x, while black numbers indicate that for all stripes. (B) Scatter plots comparing basic stripe strengths in mucosa versus changes in advanced stages. Linear regression was performed and Pearson correlation r was indicated. (C) Scatter plots of stripe strengths by functional annotation in polyps and (D) adenocarcinoma versus mucosa.

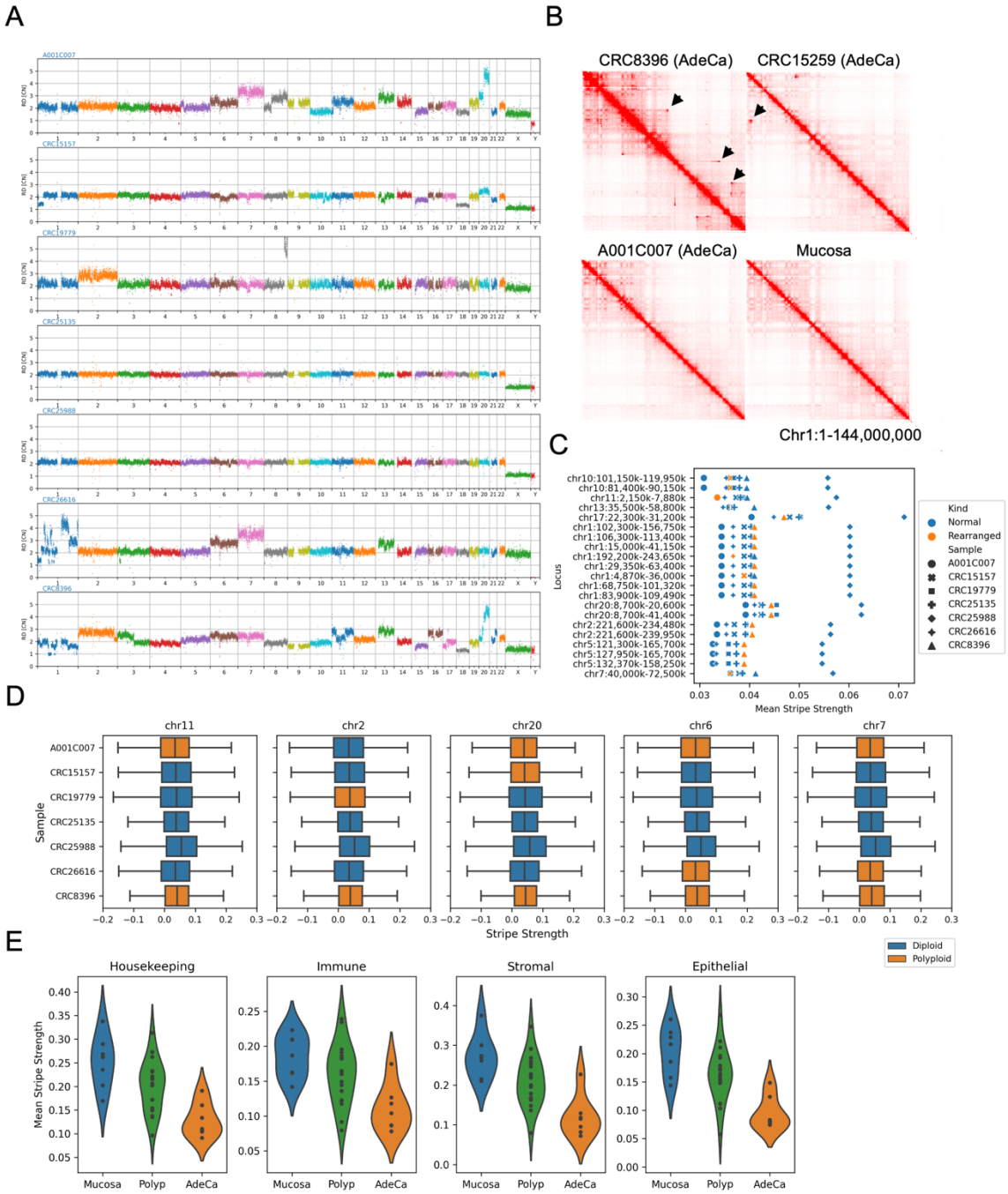

Figure S4 Impact of confounding factors on stripe quantification. (A) Manhattan plot showing chromosomal copy number variations in adenocarcinoma samples. (B) Examples of visually inspected chromosomal rearrangements. (C) Comparison of mean fold enrichments for stripes located within 100kb to the rearrangement junction sites in adenocarcinoma samples with and without the indicated rearrangement. (D) Comparisons of stripe strength distributions in chromosomes with polyploidy in subsets of adenocarcinoma samples. Paired t test was performed between mean stripe fold enrichment in samples with diploid and polyploid. No significant difference ($p<0.05$) was identified for any of five chromosomes. (E) Violin plots showing sample mean stripe fold enrichment for genes that are housekeeping or specific to immune, stromal, or epithelial cells.

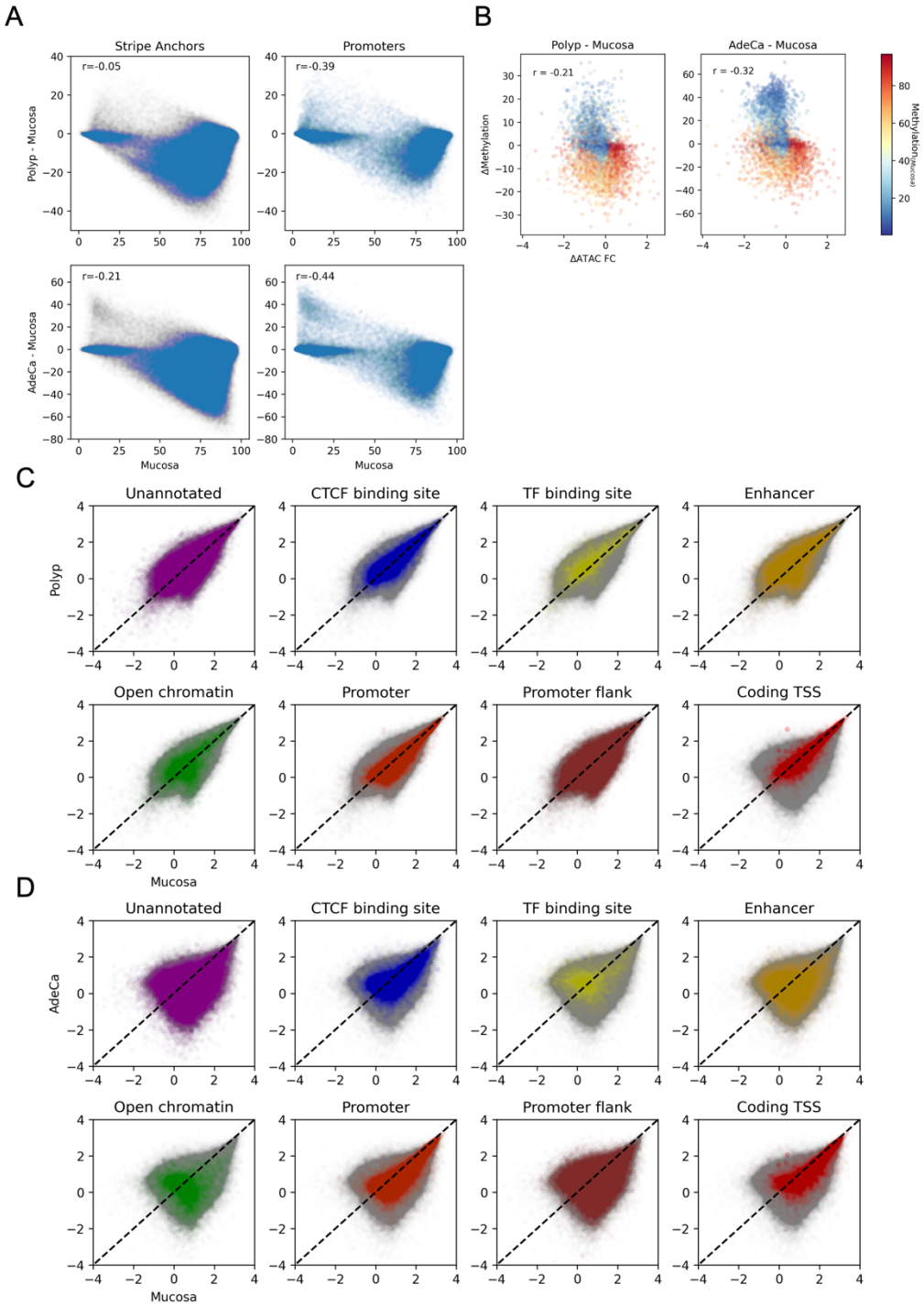

Figure S5 Association between DNA methylation and open accessibility during CRC development. (A) Scatter plots between methylation percentage in mucosa and differential methylation in polyps and adenocarcinoma for all stripe anchors (N = 230,658) TSS regions (N = 55,093) Colors indicate methylation degrees in mucosa. Pearson correlation r is indicated for each comparison. (B) Scatter plots comparing differential open accessibility and methylation for gene TSS (N = 15,387). Pearson correlation r is indicated for each comparison. (C) Scatter plots comparing strengths of ATAC peak with various annotations between mucosa and polyp or (D) adenocarcinoma (N = 258,346).

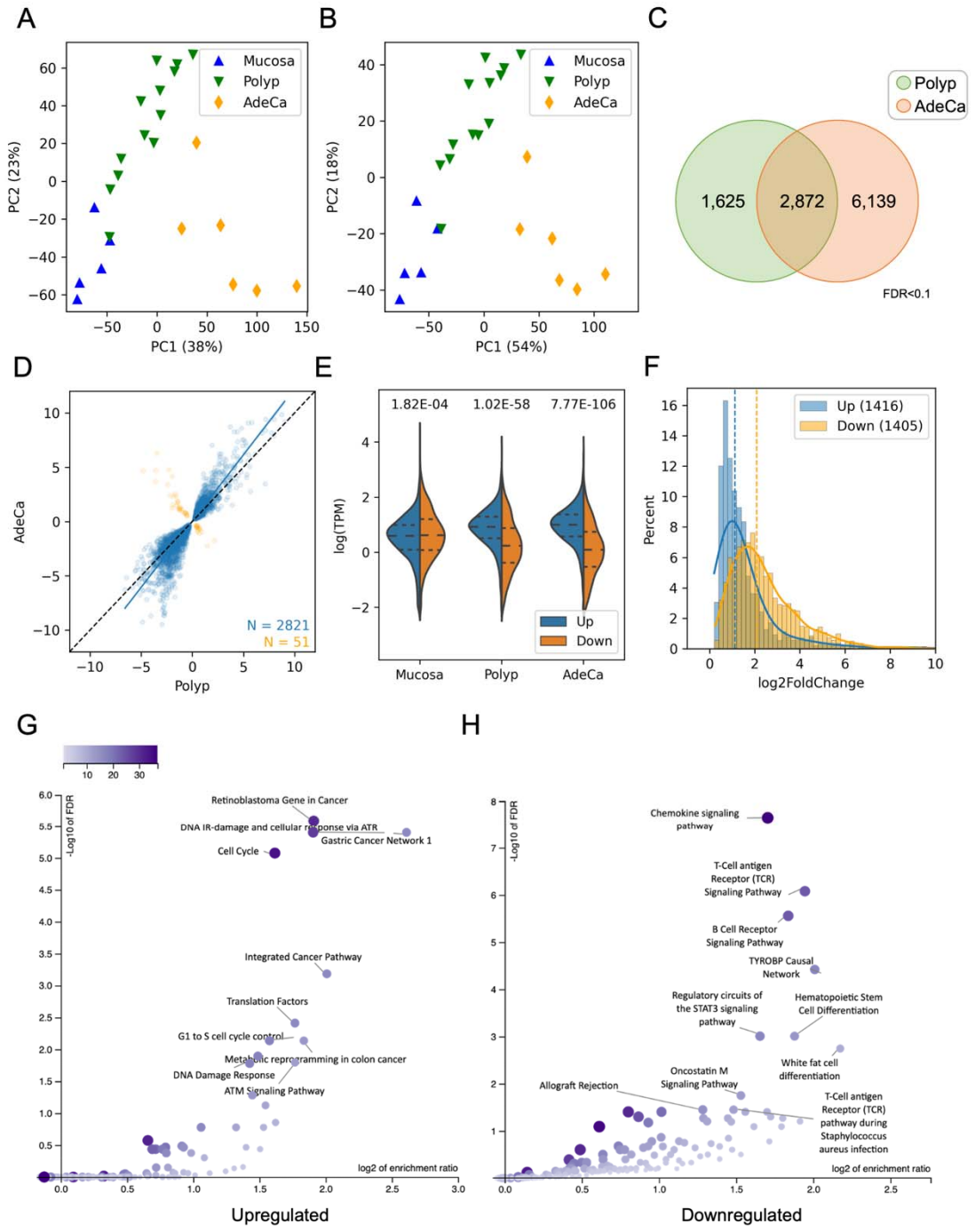

Figure S6 Alterations of transcriptomes during CRC development. (A) PCA analysis for the whole transcriptome and (B) differential genes between mucosa and adenocarcinoma. (C) Overlap of differential genes at stages compared to mucosa. (D) Scatter plot comparing differential expression (log2 fold change) in polyp and cancer. (E) Violin plot showing distributions of expression levels (log TPM) for commonly up- and down-regulated genes. Lines within the plot indicate quarter quartiles. P values from two-side t tests are indicated on top. (F) Distributions of log differential expressions for commonly up- and down-regulated genes in adenocarcinoma. Dotted lines indicate medians. (G) Enriched pathways for commonly up- and (H) down-regulated genes.

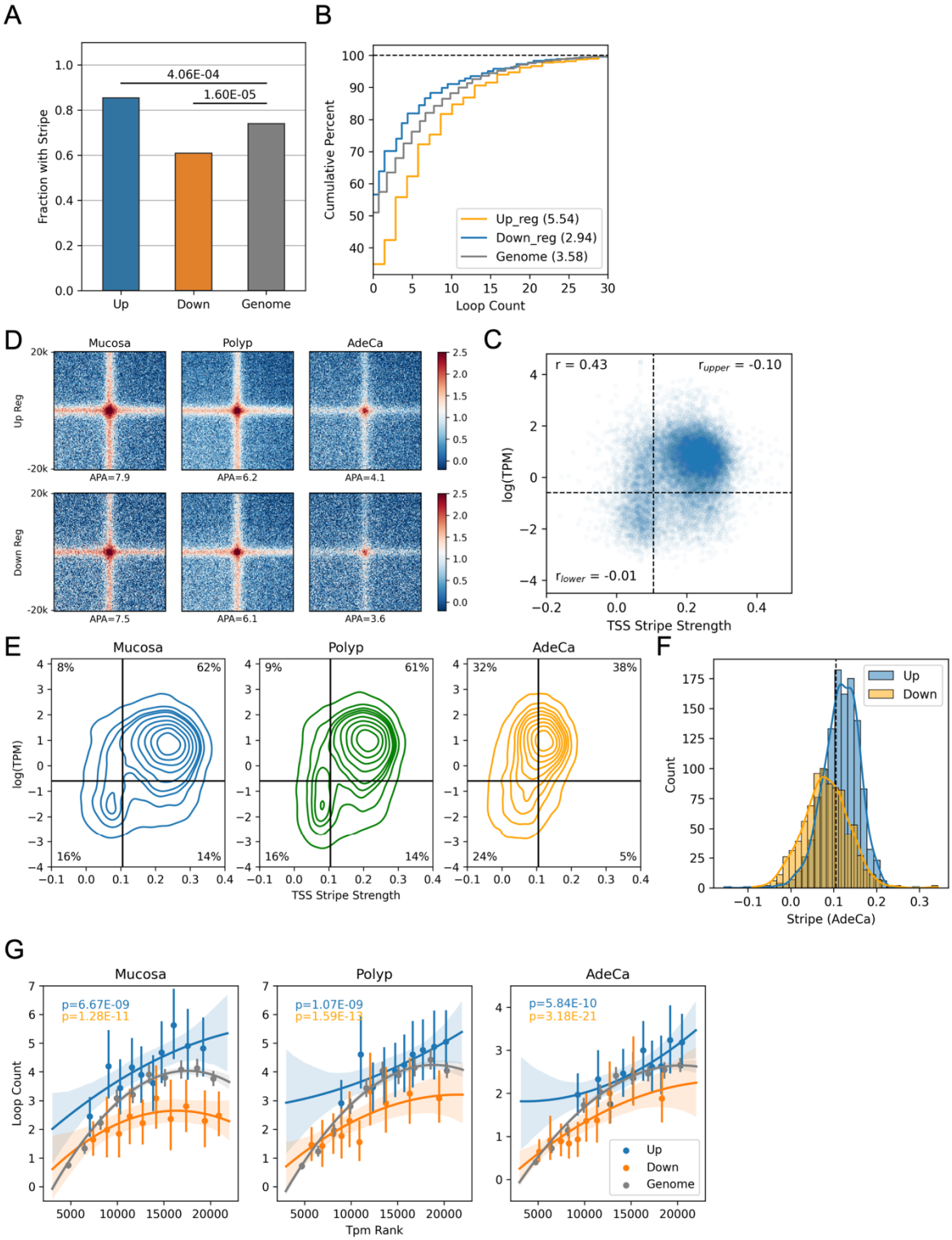

Figure S7 Relationship between promoter distal interaction and gene expression. (A) Comparison of fraction with promoter interaction stripes for up-regulated, down-regulated or all expressing genes in mucosa. Numbers on top indicate significance of differential fraction tested by the 2x2 Chi-squared test. (B) Cumulative distributions of loop counts for up-, down-regulated or all genes in mucosa. Numbers in legends indicate mean loop counts. (C) APA analyses of

the complete loop profiles for up- and down-regulated genes at different stages. (D) Scatter plot comparing the promoter stripe strength with expression in mucosa. Dotted lines indicate thresholds separating the low stripe-low expression and high stripe-high expression clusters. Pearson correlation r was calculated for the whole plot as well as the lower left and upper right sections. (E) KDE plots comparing the promoter stripe strength with expression at different stages. Lines indicate the same thresholds shown in (D). (F) Distribution of stripe strengths for commonly differential expressed genes in adenocarcinomas. Dotted lines indicate the stripe threshold defined in (D). (G) Regression plot between TSS loop count and transcription (TPM) rank in different stages. Similar to Figure 5A, a second-order polynomial model was used to fit the up- and down-regulated genes compared to the whole genome. Significance of alternative models are indicated.

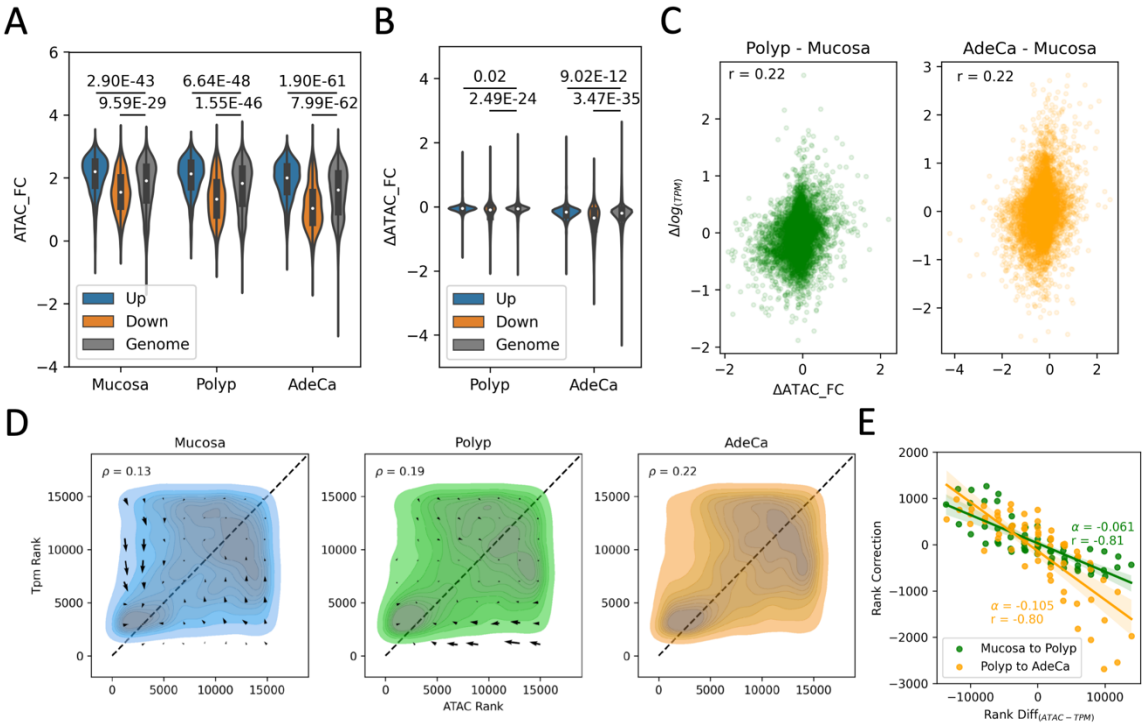

Figure S8 Association between differential expression and open accessibility. (A) Violin plots showing promoter ATAC seq peak strengths (log fold change) and (B) changes in polyps and adenocarcinoma for up- and down-regulated genes. Significance of differential distribution compared to the genome was examined with the two-sided t test. (C) Scatter plots showing genome-wide association between ATAC peak alteration and differential gene expression in polyps (N = 14,805) and adenocarcinoma (N = 14,672). Pearson correlation r is indicated. (D) KDE plot comparing expression and ATAC peak ranks for all genes (N = 15,626). Spearman correlation ρ for each stage is indicated. Quiver matrices indicate average directionalities of rank shifts in the next stage for genes at each arrow position. Dashed lines indicate equal ranking. (E) Comparison between rank difference (y - x) for the quivers indicated in (D) versus their net rank correction (Δy - Δx) in the next stage. Correlation was estimated with first order linear regression, with coefficient alpha and Pearson coefficient r indicated.

Methods

Sample collection

FAP tissues were collected at the time of partial or full colectomies for 4 patients. Immediately after colectomy, patient-matched non-neoplastic colorectal mucosa, adenomatous polyps, and adenocarcinomas were snap frozen and preserved in liquid nitrogen. One FAP adenocarcinoma (A001-C-007) was embedded in the optimal cutting temperature compound (OCT) prior to storage at -80°C. Six sporadic CRCs were obtained from the Stanford Tissue Bank. Tissues were examined for histopathology to confirm their disease states. All collection procedures were conducted under IRB protocol 47044.

Multi-digested Hi-C (mHi-C)

As a derived protocol from Tri-HiC[21], mHi-C was performed as described previously with minor modifications. Briefly, 5-10 mg of snap-frozen tissue was loaded in tissueTUBE-TT05 (Covaris 520071) and cryopulverized using the Covaris CP02 cryoPREP Automated Dry Pulverizer following manufacturer's procedure. Pulverized tissues were treated with freeze substitution[45] by submerging in 1ml -80 ^{o}C 0.01% formaldehyde (ThermoFisher 28906), 97% ethanol, and 2% water. Samples were incubated in dry ice for 3 hours on a rotor with spinning speed around 100 rpm, and then placed in a CoolCell Container (Corning), transferred to a -20 ^{o}C freezer for overnight incubation. On day 2, the Container was transferred to a 4 ^{o}C cold room, spinning on a rotor around 100 rpm for 1 hour to bring the sample temperature above freezing point.

The tissue samples were separated from ethanol solution by centrifuging at 300 g for 5 min in a 4 ^{o}C microcentrifuge. Crosslinking was performed by incubating with 1ml 1% TBS-formaldehyde for 10 min at room temperature. The solution was quenched by adding 80 µl 2.5 M glycine and incubated for 5 min. Samples were centrifuged, washed once with 1 ml TBS (pH 7.5), and resuspended with 250 µl Hi-C lysis buffer (10 mM Tris-HCl, pH8.0, 10 mM NaCl, 0.2% Igepal CA630) plus 50 µl proteinase inhibitor cocktail (Sigma P8340). Nuclei extraction was performed on ice by squeezing the samples with 1.5 ml disposable pellet pestles (Fisher Scientific 12-141-368) for 15-20 times.

The crude suspension was centrifuged at 1500 g, 4 ^{o}C for 5 min, resuspended in 800 µl Hi-C lysis buffer and passed through a 100 µm strainer (Sysmex). After centrifugation, purified nuclei were resuspended in 170 µl 10 mM Tris-HCl containing 0.5% Triton X-100 (Sigma 93443), and incubated at room temperature with rotation for 15 min. Ten microliter of 1% SDS, 20 µl Cutsmart buffer (NEB), 3 µl HinP1I (NEB R0124S), 3 µl DdeI (NEB R0175L), 3 µl CviAII (NEB R0640L), 3 µl FspBI (ThermoFisher ER1762), and 0.6 µl MseI (NEB R0525M), were added to the suspension in the indicated order. The mixture was incubated at 25 ^{o}C and then 37 ^{o}C for 2 hours each with rotation.

To stop restriction digestion, the suspension was incubated in a 62 ^{o}C heating block for 20 min followed by cool down. End repair was carried out by adding a 30 µl solution containing 0.5 mM biotin-14-dATP (Active Motif 14138), 0.5 mM biotin-14-dCTP (AAT bio 17019), 0.5 mM dTTP, 0.5 mM dGTP, and 4 µl Klenow DNA polymerase (NEB M0210L) to the mixture, and incubated for 1 hour at 37 ^{o}C with rotation. For ligation, a 750 µl solution containing 1x NEB T4 DNA ligase buffer (NEB B0202), 120 µg BSA (ThermoFisher AM2616), and 2000 U T4 DNA ligase (NEB M0202M) was added. The mixture was incubated at room temperature for 90 min, then 4 ^{o}C overnight, and then room temperature for additional 60 min with rotation.

To perform reverse crosslinking, the mixture was centrifuged at 1500 g for 5 min, and the supernatant was replaced with a mixture of 300 µl 1x T4 ligase buffer, 30 µl 20 mg/ml proteinase K (ThermoFisher 25530049), 30 µl 10% SDS, and 40 µl 5 M NaCl. The suspension was then incubated at 66 ^{o}C for 4 hours. DNA content was then purified by phenol-chloroform extraction and resuspended in 20 µl 10 mM Tris-HCl.

To generate mHi-C sequencing library, 300 ng of purified DNA was tagmented with 2.5 µl Tn5 transposase (APExBIO K1155, discontinued) loaded with equimolar Mosaic Ends containing Illumina Nextera i5 and i7 extensions following the manufacturer's protocol. Tagmentation was performed in a 100 µl buffer containing 10% DMF, 10 mM Tris-HCl, and 150 mM NaCl, at 55 ^{o}C for 10 min. The product was column purified (Zymo D4014) and PCR amplified for 2 cycles using the NEBNext master mix (NEB M0544L) with Illumina Nextera primers and conditions. Biotin enrichment was then performed by adding 20 µl Dynabeads MyOne Streptavidin C1 (ThermoFisher 65001) and incubating at room temperature for 30 min with rotation. The magnetic beads were washed three times with 1x wash buffer (10 mM Tris-HCl pH7.5, 1 mM NaCl, 0.5 mM EDTA) and one time with 10 mM Tris-HCl. Final libraries were obtained by amplifying the beads with additional 8 cycles of PCR, followed by purification with SPRI (Beckman B23318) size selection at 0.5x-1.1x range. The 33 samples were combined to 2 pools and sequenced using 2 NovaSeq (Illumina) S4 200 cycle flow cells.

RNA seq

Total RNA was extracted from ~5-10 mg of frozen tissues using Zymo Quick-RNA Miniprep (Zymo R1054), following the manufacturer's instructions. After purification, DNA digestion was performed using the DNA-free DNA Removal Kit (ThermoFisher AM1906). Sequencing libraries of mRNA were prepared from 1 µg total RNA using the NEBNext Ultra un-stranded preparation kit (E7775S,E7490S), following the manufacturer's protocol. Samples were sequenced using a NovaSeq S1 flow cell for 50 bp pair-end sequencing. This achieved an average of 86.3 million raw paired reads.

ATAC seq

ATAC seq was carried out using the latest ENCODE tissue protocol as described[46]. Sequencing was carried out on a NovaSeq S1 flow cell using 50 bp pair-end sequencing. On average 53.2 million unique fragments were mapped for each sample.

EM seq

The Enzymatic Methyl seq was performed as described[35]. Libraries were constructed by using the NEBNext Enzymatic Methyl-seq Kit (NEB), following manufacturer's guidance. Sequencing was conducted using the novel ultrahigh throughput UG-100 (Ultima Genomics) sequencer.

Data processing for mHi-C

Initial processing of mHi-C data was performed using the distiller pipeline (https://github.com/open2c/distiller-nf) with default parameters set for SLURM cluster. The deduplicated pair files were fed to Juicer pre[32] to generate KR balanced .hic matrices at resolutions of 200, 500, 1k, 2k, 5k, 10k, 20k, 50k, 100k, 250k, 500k, and 1mbp, with a quality score filter at 30. For generating piled up master matrices for stages and all samples, pair files were first merged and sorted using pairtools (https://github.com/open2c/pairtools).

Stripe calling was performed as previously described[21] with minor modifications on parameters. Briefly, long-range (>1.5 kb) and short-range (<1 kb) mapped read pairs were sorted to two .bam files using awk and samtools. Using Bedtools, these reads were mapped to two binning bed tracks, a local one with 2 kb window and a background one with 50 kb window, both with a 100 bp sliding size. Using MACS2 bdgcmp -m qpois, the local long-range read count for each bin was examined for statistical significance of enrichment against the expected count number, calculated as $(long_{bg} / short_{bg}) * short_{local}$. The log fold change signal (stripe strength) was calculated with the same formula by feeding the actual and expected values to MACS2 bdgcmp -m logFE. A pseudo-count of 1 was added to avoid NaN errors.

After stripe q values were determined, each 100 bp bin was counted for the number of samples showing significance (FDR <0.01). We considered bins with at least three sample hits significant. These bins were merged, and only windows with at least 500 bp size were included as final stripe anchors. The anchor stripe strengths for each sample was derived from the mean of logFE signal of bins within the windows. Stripe peaks overlapping ENCODE blacklist regions (https://sites.google.com/site/anshulkundaje/projects/blacklists) were removed. To avoid gender variations among patients, only autosomal chromosomes were included for downstream analyses.

For loop calling, the HiCCUPS algorithm from Juicer tools[32] was applied with the following parameters: -r 500,1000,2000,5000,10000 -f 0.1 -p 4,2,2,2,2 -i 20,10,10,6,6 -t 0.1,1.25,1.75,2 -d 2000,2000,4000,10000,20000. Because library complexity enormously affected loop calling power, the analysis was not performed for each individual sample, but instead on pooled libraries of 1) all samples, 2) all mucosa, 3) all polyps, and 4) all adenocarcinomas. Post-processed loop pixels from all profiles at different resolutions were merged in the order of high resolution > low resolution from combined> mucosa > polyp > adeca, where a loop with lower priority was filtered if both anchors overlapped with a higher priority loop. This master loop list was then applied to each sample to perform individual loop quantification. To calculate loop strengths, read counts in the identified loops were divided by the expected count from the donut background and log transformed. A pseudo count of 1 was added to avoid NaN error as needed. For stage-specific counting, loops with average loop strengths greater than 1.2 fold in samples of the specified stage were considered positive.

For annotations of stripe anchors, features from Ensembl regulatory build[33] and TSS from Gencode were mapped using Bedtools. Loop anchors were first annotated with stripe anchors from the analysis. If multiple stripes resided in a loop, the annotation matched the one closest to the loop center. The functional annotation of loop anchors were then considered equivalent to the overlapping stripe anchor. This indirect annotation is underpowered to a certain extent as loop anchors without stripes may also overlap with functional CREs. However, due to the broad size of the loop calling algorithm (1-5 kb), it is sometimes difficult to distinguish true loop signals from potential random overlaps. By contrast, stripe anchors were determined in a much higher resolution (100 bp), and hence the annotation reflects the direct source of the signal. Since the main interest of this study was to identify the genome-wide behaviors of regulatory elements, we decided to maintain stringent annotation.

To identify differential stripes between stages, raw read counts for each stripe were adjusted with the local background by dividing by a background coefficient calculated as the ratio between the total background reads of the examined sample and average background read count from all samples. The adjusted read count was thereby a pseudo fold-change signal against its background, but multiplied by the average coverage of the locus. The adjusted read count matrix was processed with DESeq2[34] using default settings. Comparisons were made

between mucosa-polyp and mucosa-adenocarcinoma. Significant differential stripes were defined by greater than 1.3 fold difference between stages and adjusted p value < 0.1.

For aggregation analysis of loops, APA from Juicer was performed with the parameters -r 200 -u -n -0 -w 500 -k KR -q 20. The enrichment score was calculated as the average intensity of 10*10 center pixels (2 kb) against the mean of 100*100 pixels from bottom left. For aggregation of stripes, the same function was performed with the parameters -r 200 -u -n -0 -w 250 -k KR -q 20. Since distance-dependent interaction decay is phenomenal at stripe vicinity, interaction intensities at specific distances (a.k.a. matrix diagonals) were normalized against the average intensity of the distance. Fold enrichment of the aggregated stripes was calculated by averaging the normalized values in the center 10 pixels. For visualizations of loop and stripe APA, matrices were log transformed before plotting to the heatmaps.

Copy number information

Similar to previously described[47], we treated mHi-C as whole-genome sequencing data to determine chromosomal copy number variations in cancer samples. The mHi-C profiles were sequenced with sufficient coverage (30x on average) to support CNV calling. Reads were mapped to the hg38 genome using bwa-mem. After deduplication, short-range read pairs, a.k.a. the self-ligation products, defined by <1000 bp distance with a normal forward-reverse mapping orientation, were selected to be processed by the CNVpytor package[48] (v1.2). Identification of CNVs and visualization were conducted with default parameters in 100 kb windows.

Chromatin rearrangement identification

A visual inspection approach of the mHi-C heatmap matrices was applied to identify rearrangements, as previously described[47]. For this analysis, we focused on intrachromosomal rearrangements only, particularly inversions, since interchromosomal fusions are likely accompanied with polyploidy, which has been included in our analysis above. Generally, rearranged loci were separated by at least 1 mbp, characterized with sharp increase of rectangular-shaped contact intensities. These aberrant interactions were examined on mucosa matrices to confirm their normal absence. While this examination does not exclusively identify all rearrangements, positive hits can be obtained with high confidence.

ATAC seq processing

ATAC seq results were processed with the ENCODE-DCC atac-seq-pipeline (https://github.com/ENCODE-DCC/atac-seq-pipeline) using default settings. Parameters such as fractions of reads in peaks and promoters were directly derived from the pipeline. To obtain the integrated peak list from all samples, a 100 bp binning track was mapped with the pseudo-replicated peak regions from each sample using Bedtools. Bins with at least three hits were considered valid. These bins were merged, and only intervals with at least 300 bp size were included as final peak sites. On peak fold enrichments of samples were then obtained from the pipeline-derived fold change bigwig tracks.

RNA seq processing

RNA seq results were processed using Tomas Bencomo's pipeline (https://github.com/tjbencomo/bulk-rnaseq), a simple workflow using salmon to quantify transcript level and DESeq2 to identify differential genes. Transcription levels (TPM) of genes were obtained by summing transcript-based TPM from salmon output (.rf).

Gene Ontology

Enrichment analysis of significantly up- or down-regulated genes in pathways was performed and visualized using the WEB-based GEne SeT AnaLysis Toolkit[49]. Method of over-representation was selected to test enrichments in the KEGG pathway against the protein-coding genome. Analysis was performed with default parameters.

DNA methylation processing

We obtained 21,175,510 CpG sites with measurable methylation ratios in all examined samples. Methylation degree of features, including mHiC hotspots, ATAC peaks, and gene promoters, were calculated by averaging the methylation percentage for all valid CpG sites within the feature.

Defining multiomic overlaps

For all analyses performed in this study, overlap of features refer to at least 1 bp of common position of two tracks without flank extension. Promoters of genes were defined as a -1500 bp to +500 bp window from the start position of the gene. Comparisons between transcription and stripe strength, loop count, or ATAC peak strength were reduced to genes, meaning that the same value was allocated to all TSS it overlapped. Comparisons between stripes and ATAC peaks were mapped to ATAC peaks, meaning that the corresponding stripe strength of each ATAC peak was equal to the mean of all overlaps. For all analyses comparing multiple omics profiles in the study, the number of features was shrunken to those with detected values from all profiles unless indicated otherwise. The only exception was loop counts, where TSS without any loop overlaps were assigned with 0 loop count.

Principal component analysis

The PCA analyses of mHi-C and RNA seq were performed using the Python sklearn.decomposition.PCA package. Input were untransformed stripe strengths (in log10 basis) for stripes, Z-score transformed log fold enrichment for loops, and untransformed DESeq2-exported normalized read count for the transcriptome.

Housekeeping and cell type-specific gene lists

The list of housekeeping and cell-specific genes (Table S1) was obtained from public literature and validated by previously reported single cell RNA-seq performed in the same cohort[50].

Modeling transcription-interaction relationship

Promoter stripe strengths (or promoter loop counts) and expression levels for up- and down-regulated genes and the whole transcriptome were rank-transformed. After filtering out genes at bottom 2000 rank transcription levels (due to their frequent missing or 0 values in samples), the transcription-stripe rank relations were fit to a 2nd order polynomial regression model. Log-likelihood ratios for differentially expressed genes between their own parameters and the whole transcriptome were calculated. Significance of the alternative model was tested using the chi-squared test following Wilks' theorem.

Ontology for interaction-transcription imbalance

WikiPathways gene sets were downloaded from the official website[51]. For each pathway, genes with valid transcription and stripe/loop count values were fed to the polynomial models described above, and calculated the log likelihood ratios between whole genome model and up-/down-regulated model. Pathways with <0.2 FDR corrected q value from the likelihood ratio test were considered having significant trend for up- or down-regulation.

Public data usage

Unmasked hg38 genome was used as reference for all analyses. Regulatory build for sigmoid colon (ver 20210107) was downloaded from Ensembl (http://www.ensembl.org) for regulatory annotations. Gencode v38 was used for RNA-seq alignment and defining TSS positions. ENCODE *in situ* Hi-C for IMR-90 cell line (ENCSR852KQC) was used for comparison with mHi-C. Roadmap histone ChIP-seq tracks for colonic mucosa (GSM1112779, GSM916043, GSM916046, GSM916045) and ENCODE CTCF (ENCSR833FWC), Pol II (ENCSR322JEO) ChIP-seqs were used for CRE visualization. Locations of CpG islands were downloaded from UCSC genome browser.

Supplementary Tables

Table S1 List of housekeeping and cell type-specific genes examined for promoter interaction changes.

Reference

1 Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381-385, doi:10.1038/nature11049 (2012).

2 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).

3 Valton, A. L. & Dekker, J. TAD disruption as oncogenic driver. *Curr Opin Genet Dev* **36**, 34-40, doi:10.1016/j.gde.2016.03.008 (2016).

4 Levine, M., Cattoglio, C. & Tjian, R. Looping back to leap forward: transcription enters a new era. *Cell* **157**, 13-25, doi:10.1016/j.cell.2014.02.009 (2014).

5 Deng, W. *et al.* Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**, 1233-1244, doi:10.1016/j.cell.2012.03.051 (2012).

6 Deng, W. *et al.* Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell* **158**, 849-860, doi:10.1016/j.cell.2014.05.050 (2014).

7 Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454-1458, doi:10.1126/science.aad9024 (2016).

8 Benabdallah, N. S. *et al.* Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation. *Mol Cell* **76**, 473-484 e477, doi:10.1016/j.molcel.2019.07.038 (2019).

9 Alexander, J. M. *et al.* Live-cell imaging reveals enhancer-dependent Sox2 transcription in the absence of enhancer proximity. *Elife* **8**, doi:10.7554/eLife.41769 (2019).

10 Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320 e324, doi:10.1016/j.cell.2017.09.026 (2017).

11 Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944 e922, doi:10.1016/j.cell.2017.05.004 (2017).
12 Ray, J. *et al.* Chromatin conformation remains stable upon extensive transcriptional changes driven by heat shock. *Proc Natl Acad Sci U S A* **116**, 19431-19439, doi:10.1073/pnas.1901244116 (2019).
13 Greenwald, W. W. *et al.* Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. *Nat Commun* **10**, 1054, doi:10.1038/s41467-019-08940-5 (2019).
14 Lajoie, B. R., Dekker, J. & Kaplan, N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods* **72**, 65-75, doi:10.1016/j.ymeth.2014.10.031 (2015).
15 Pal, K., Forcato, M. & Ferrari, F. Hi-C analysis: from data generation to integration. *Biophys Rev* **11**, 67-78, doi:10.1007/s12551-018-0489-1 (2019).
16 Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* **13**, 919-922, doi:10.1038/nmeth.3999 (2016).
17 Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* **47**, 598-606, doi:10.1038/ng.3286 (2015).
18 Luo, Z. *et al.* NicE-C efficiently reveals open chromatin-associated chromosome interactions at high resolution. *Genome Res* **32**, 534-544, doi:10.1101/gr.275986.121 (2022).
19 Hsieh, T. S. *et al.* Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. *Mol Cell* **78**, 539-553 e538, doi:10.1016/j.molcel.2020.03.002 (2020).
20 Hua, P. *et al.* Defining genome architecture at base-pair resolution. *Nature* **595**, 125-129, doi:10.1038/s41586-021-03639-4 (2021).
21 Zhu, Y. & Suh, Y. Tri-4C: efficient identification of cis-regulatory loops at hundred base pair resolution. *bioRxiv*, 743005, doi:10.1101/743005 (2019).
22 Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep* **15**, 2038-2049, doi:10.1016/j.celrep.2016.04.085 (2016).
23 Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A* **112**, E6456-6465, doi:10.1073/pnas.1518552112 (2015).
24 Fodde, R., Smits, R. & Clevers, H. APC, signal transduction and genetic instability in colorectal cancer. *Nat Rev Cancer* **1**, 55-67, doi:10.1038/35094067 (2001).
25 Aoki, K. & Taketo, M. M. Adenomatous polyposis coli (APC): a multi-functional tumor suppressor gene. *J Cell Sci* **120**, 3327-3335, doi:10.1242/jcs.03485 (2007).
26 Parker, T. W. & Neufeld, K. L. APC controls Wnt-induced beta-catenin destruction complex recruitment in human colonocytes. *Sci Rep* **10**, 2957, doi:10.1038/s41598-020-59899-z (2020).
27 Miyoshi, Y. *et al.* Somatic mutations of the APC gene in colorectal tumors: mutation cluster region in the APC gene. *Hum Mol Genet* **1**, 229-233, doi:10.1093/hmg/1.4.229 (1992).
28 Galiatsatos, P. & Foulkes, W. D. Familial adenomatous polyposis. *Am J Gastroenterol* **101**, 385-398, doi:10.1111/j.1572-0241.2006.00375.x (2006).

29 Groden, J. *et al.* Identification and characterization of the familial adenomatous polyposis coli gene. *Cell* **66**, 589-600, doi:10.1016/0092-8674(81)90021-0 (1991).
30 Rozenblatt-Rosen, O. *et al.* The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. *Cell* **181**, 236-249, doi:10.1016/j.cell.2020.03.053 (2020).
31 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).
32 Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95-98, doi:10.1016/j.cels.2016.07.002 (2016).
33 Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The ensembl regulatory build. *Genome Biol* **16**, 56, doi:10.1186/s13059-015-0621-5 (2015).
34 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).
35 Lee, H. *et al.* Ultra high-throughput whole-genome methylation sequencing reveals trajectories in precancerous polyps to early colorectal adenocarcinoma. *bioRxiv*, 2022.2005.2030.494076, doi:10.1101/2022.05.30.494076 (2022).
36 Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* **43**, 768-775, doi:10.1038/ng.865 (2011).
37 Feinberg, A. P., Koldobskiy, M. A. & Gondor, A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nat Rev Genet* **17**, 284-299, doi:10.1038/nrg.2016.13 (2016).
38 Jones, P. A. & Baylin, S. B. The fundamental role of epigenetic events in cancer. *Nat Rev Genet* **3**, 415-428, doi:10.1038/nrg816 (2002).
39 Vian, L. *et al.* The Energetics and Physiological Impact of Cohesin Extrusion. *Cell* **175**, 292-294, doi:10.1016/j.cell.2018.09.002 (2018).
40 Baylin, S. B. & Jones, P. A. Epigenetic Determinants of Cancer. *Cold Spring Harb Perspect Biol* **8**, doi:10.1101/cshperspect.a019505 (2016).
41 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674, doi:10.1016/j.cell.2011.02.013 (2011).
42 Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385 e318, doi:10.1016/j.cell.2018.02.060 (2018).
43 Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-339, doi:10.1038/nature12634 (2013).
44 Li, M., Sun, Q. & Wang, X. Transcriptional landscape of human cancers. *Oncotarget* **8**, 34534-34551, doi:10.18632/oncotarget.15837 (2017).
45 Zheng, W. *et al.* Freeze substitution Hi-C, a convenient and cost-effective method for capturing the natural 3D chromatin conformation from frozen samples. *J Genet Genomics* **48**, 237-247, doi:10.1016/j.jgg.2020.11.002 (2021).
46 Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res* **48**, D882-D889, doi:10.1093/nar/gkz1062 (2020).
47 Harewood, L. *et al.* Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biol* **18**, 125, doi:10.1186/s13059-017-1253-8 (2017).

48 Suvakov, M., Panda, A., Diesh, C., Holmes, I. & Abyzov, A. CNVpytor: a tool for copy number variation detection and analysis from read depth and allele imbalance in whole-genome sequencing. *Gigascience* **10**, doi:10.1093/gigascience/giab074 (2021).
49 Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res* **41**, W77-83, doi:10.1093/nar/gkt439 (2013).
50 Becker, W. R. *et al.* Single-cell analyses reveal a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer. *bioRxiv*, 2021.2003.2024.436532, doi:10.1101/2021.03.24.436532 (2021).
51 Martens, M. *et al.* WikiPathways: connecting communities. *Nucleic Acids Res* **49**, D613-D621, doi:10.1093/nar/gkaa1024 (2021).