Molecular Cancer

RESEARCH

Open Access

Unraveling the key role of chromatin structure in cancer development through epigenetic landscape characterization of oral cancer

Yue Xue[1†], Lu Liu[2†], Ye Zhang[3,4,5†], Yueying He[1], Jingyao Wang[1], Zicheng Ma[2], Tie-jun Li[4,5], Jianyun Zhang[4,5*], Yanyi Huang[1,6,7*] and Yi Qin Gao[1,2,6*]

Abstract

Epigenetic alterations, such as those in chromatin structure and DNA methylation, have been extensively studied in a number of tumor types. But oral cancer, particularly oral adenocarcinoma, has received far less attention. Here, we combined laser-capture microdissection and muti-omics mini-bulk sequencing to systematically characterize the epigenetic landscape of oral cancer, including chromatin architecture, DNA methylation, H3K27me3 modification, and gene expression. In carcinogenesis, tumor cells exhibit reorganized chromatin spatial structures, including compromised compartment structures and altered gene-gene interaction networks. Notably, some structural alterations are observed in phenotypically non-malignant paracancerous but not in normal cells. We developed transformer models to identify the cancer propensity of individual genome loci, thereby determining the carcinogenic status of each sample. Insights into cancer epigenetic landscapes provide evidence that chromatin reorganization is an important hallmark of oral cancer progression, which is also linked with genomic alterations and DNA methylation reprogramming. In particular, regions of frequent copy number alternations in cancer cells are associated with strong spatial insulation in both cancer and normal samples. Aberrant methylation reprogramming in oral squamous cell carcinomas is closely related to chromatin structure and H3K27me3 signals, which are further influenced by intrinsic sequence properties. Our findings indicate that structural changes are both significant and conserved in two distinct types of oral cancer, closely linked to transcriptomic alterations and cancer development. Notably, the structural changes remain markedly evident in oral adenocarcinoma despite the considerably lower incidence of genomic copy number alterations and lesser extent of methylation alterations compared to squamous cell carcinoma. We expect that the comprehensive analysis of epigenetic reprogramming of different types and subtypes of primary oral tumors can provide additional guidance to the design of novel detection and therapy for oral cancer.

†Yue Xue, Lu Liu and Ye Zhang contributed equally to this work.

*Correspondence:
Jianyun Zhang
jianyunz0509@aliyun.com
Yanyi Huang
yanyi@pku.edu.cn
Yi Qin Gao
gaoyq@pku.edu.cn

Full list of author information is available at the end of the article

Keywords Oral squamous cell carcinomas, Oral adenocarcinoma, Chromatin structure, DNA methylation

Introduction

Oral squamous cell carcinomas (OSCC) and adenoid cystic carcinoma (AdCC), of which major primary sites are epithelium and salivary glands, respectively, make up the majority of oral malignancies. OSCC, which is also the most common category of head and neck squamous cell carcinoma (HNSCC), accounts for 1.8% of newly diagnosed cancer cases each year globally [1]. The 5-year overall survival of OSCC patients is about 50%. AdCC is a relatively rare cancer type, accounting for only 1% of head and neck tumors and 10% of salivary gland tumors [2]. Although AdCC progresses relatively slowly, it is also a relentless cancer. The 5-year patient survival rate is approximately 60% [2]. Up to now, surgery remains the gold standard for the treatment of AdCC patients, and only limited therapy response to AdCC marker inhibitors was observed [2, 3]. Microscopically, tumor cells of AdCC are mainly arranged in two morphological forms, cribriform and solid, respectively [3, 4]. Little is known about the molecular mechanism of AdCC, particularly its epigenetic landscape.

Over the years, genetic and epigenetic alternations have been extensively investigated in tumor cells. They were found to be associated with tumor development in many aspects. Genetic alternations, such as single point mutations, gene fusions, duplications, deletions, as well as large range of copy number variations and translocations, all have the potential to directly impact the expression level of genes or alter downstream protein functions. In OSCC, frequently mutated genes are dominated by tumor suppressor genes (TSGs) [5], such as TP53, CDKN2A, FAT1, NOTCH1, while AdCC possesses a low rate of exonic somatic mutations [6]. Aside from genome instability, nonmutational epigenetic reprogramming has also been regarded as an important cancer hallmark [7]. Aberrant DNA methylation was frequently observed in tumor cells, and many studies focused on the transcriptional repression of tumor suppressor genes associated with hypermethylation and the activation of oncogenes related to hypomethylation [8–11]. In addition, it is increasingly evident that chromatin three-dimensional architecture plays a vital role in regulating gene functions and cell states, as well as cancer development [12–15]. In colon cancer, compromised spatial partitioning of the open and closed genome compartments was observed, which may repress stemness and invasion programs [16]. Subtype-specific chromatin structure was also identified in T-lineage acute lymphoblastic leukemia (T-ALL) [17] and acute myeloid leukemia (AML) [18]. Furthermore, hijacked enhancers, which are induced by structural variation, were demonstrated to be associated with AML cell growth.

Although epigenetic reprogramming was partially characterized in several types of malignancies, such as lung cancer [19], colon cancer [16, 20] and leukemia [18], less was understood about oral cancer, especially for AdCC [21]. In addition, since tumor tissues are known to be highly heterogeneous, high throughput sequencing measurement on bulk tissue can raise concerns about the ambiguous composition of cell types. Compared to primary tissues, cell lines have the advantages of high cell purity and ease of accessibility and are widely used in cancer research. However, it was reported that their nature differs from that of primary cancer cells in a number of aspects [22]. Therefore, here we combined laser-capture microdissection (LCM) and muti-omics mini-bulk sequencing, taking advantage of accurate sampling to systematically characterize the epigenetic landscape of oral cancer. We performed in situ high-throughput chromosome conformation capture (Hi-C), whole-genome enzymatic methylation sequencing (EM-seq), Cleavage Under Targets and Tagmentation (CUT&Tag) for H3K27me3, and RNA-seq, in 401 primary tumor, paracancerous normal gland and epithelial samples. These samples are from 27 AdCC patients and 24 OSCC patients, allowing us to comprehensively understand the epigenetic reprogramming from normal to tumor cells, the differences between the two types of oral malignancies, and the distinct subtypes of AdCC. We found that both chromatin structure and DNA methylation showed tissue- and tumor-type specificity. Chromatins in oral cancer cells exhibit decreased long-range contact and weakened compartmentalization, resulting in altered spatial gene-gene interactions (GGIs). In a previous study, we found that the tissue-specific GGI network was closely associated with transcriptional regulation and downstream protein-protein interactions [23]. The latter have been found to directly affect cancer cell phenotypes [24–26]. Therefore, we characterized the blueprint of abnormal GGIs for AdCC and OSCC, as well as the centrality characteristics from a network perspective. We also identified conserved copy number alteration (CNA) events. Interestingly, CNAs show a close connection with chromatin structure in that they tend to occur in genomic regions that are spatially isolated in not only cancerous but also normal samples. In terms of DNA methylation, we found that OSCC was characterized by a global aberrant DNA methylation, while little change was observed in AdCC. We also sought to investigate the causes of abnormal methylation changes, including

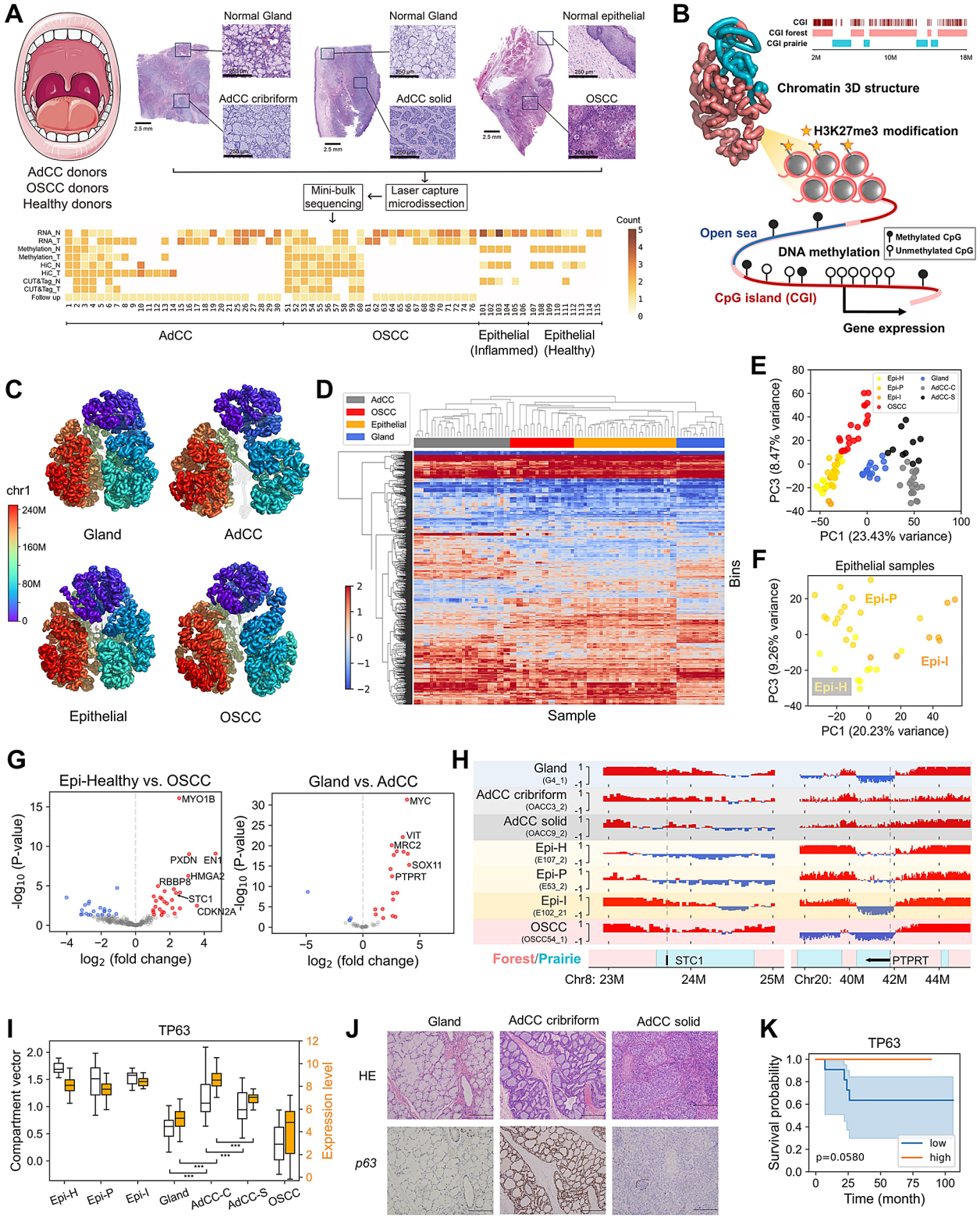

Fig. 1 (See legend on next page.)

(See figure on previous page.)

Fig. 1 Overview of experimental procedure and chromatin compartmentalization of various types of normal and tumor cells. **(A)** Hematoxylin and eosin (H&E) stained images of representative photomicrographs lesions of normal gland, AdCC cribriform and AdCC solid tissues from AdCC donors, as well as normal epithelial and OSCC tissues from an OSCC donor. Mini-bulk samples were dissected by LCM, following by multiple genomic profiling assays. Summary of the dataset is shown in the bottom heatmap plot. **(B)** The landscape of epigenetic markers and gene expression measured and analyzed in this study. Multi-scale sequence features of the genome were focused throughout the investigation. **(C)** Chromatin 3D structures of chromosome 1 for normal gland, AdCC, normal epithelial and OSCC (G6, OACC6, E53 and OSCC53 are presented as examples). **(D)** Hierarchical clustering for compartment vectors of bins with top 10% compartment variation cross all samples. Each column represents a sample and each row is a 40-kb bin. Red or blue pixel indicates vector value greater than 0 or less than 0, respectively. **(E)** The first principal component (PC1, x-axis) and third principal component (PC3, y-axis) of Principal Components Analysis (PCA) based on compartment index of all samples. Epi-H, Epi-P, Epi-I, OSCC, Gland, AdCC-C, AdCC-S represent healthy epithelial, paracancerous epithelial, inflamed epithelial, OSCC, paracancerous gland, cribriform AdCC and solid AdCC, respectively. **(F)** The first principal component (PC1, x-axis) and third principal component (PC3, y-axis) of PCA based on compartment index of healthy, paracancerous and inflamed epithelial samples. **(G)** From healthy epithelial to OSCC, expression changes of genes which switch from compartment B in epithelial to compartment A in OSCC (left panel). From normal gland to AdCC, expression changes of genes which switch from compartment B in gland to compartment A in AdCC (right panel). **(H)** Snapshots for compartment vector of regions around STC1 (left panel) and PTPRT (right panel). The bottom row is annotated with positions of gene, CGI forest and CGI prairie domains. **(I)** Boxplots for compartment vector (white box) and normalized expression count (orange box) of TP63 in seven kinds of tissues. The box extends from the first quartile (Q1) to the third quartile (Q3) of the data, with black line at the median. The whiskers extend from the box by 1.5 × the inter-quartile range (IQR). Outliers beyond the extreme ends of the whiskers are represented as individual points. *** represnets $p < 0.001$ by Mann-Whitney U test. **(J)** H&E (upper panel) stained images and p63 immunofluorescence (IF) micrographs (lower panel) of a normal gland, a cribriform AdCC sample and a solid AdCC sample. **(K)** Kaplan-Meier plots of overall survival for AdCC patients with high TP63 expression (higher than median level, $n = 10$) and low TP63 expression (lower than median level, $n = 11$). The shaded areas indicate the 95% confidence intervals, HR = 3.59, $p = 0.014$ by log rank test

methyltransferase level, cell division history, chromatin structure, and spatial H3K27me3 distribution.

Result

Primary sample collection and data overview

To obtain a multi-omics epigenetic blueprint of oral cancer, we collect samples from 66 donors, including 27 AdCC patients, 24 OSCC patients, and 15 healthy donors (Fig. 1A, S1A, S1B, Table S1). Pathologically, the laser-captured cell bulks of cribriform AdCC consist of well-formed tubules and tumor cell nests with microcystic-like spaces, and samples of solid AdCC were collected with tumor sheets composed of basaloid cells lacking tubular or cribriform formations (Fig. 1A, S1A-C). The OSCC microdissected samples contained large nests, cords, and islands of cells with pink cytoplasm, hyperchromatic nuclei, and prominent intercellular bridging. Considering the different origins of the two types of tumors, adjacent normal gland and epithelial cells were measured and analyzed as controls of AdCC and OSCC, respectively. Normal epithelial samples were collected on the gingival of oral masticatory mucosa, mainly including the basal, prickle cell, granular, and keratinized layers. To ensure the reliability of the control samples, we collected epithelial samples from both cancer patients and non-cancer donors, and paracancerous gland samples from both AdCC and OSCC donors. We used laser capture microdissection (LCM) to acquire small-scale tissue samples (200−3000 cells) from the oral epithelial, gland, and cancerous primary tissues, followed by Hi-C, EM-seq, CUT&Tag and RNA-seq (Fig. 1B). Multi-omics measurements on samples were taken along the z-axis from the same location to ensure that the data across multiple omics are matched (Fig S1A). We carefully sampled tumor cells and normal cells with distinct morphological signatures, guaranteeing high cell purity. (Fig. S1B, S1C). Ki67 immunohistochemistry staining of the slides shows a higher rate of cell division in solid AdCC compared to cribriform AdCC (Fig. S1D). Moreover, we used RNA-seq data to identify potential biases assign cell cycle phases [27] and found that the majority of samples are at G1 phase (Fig. S1E).

Our analysis and understanding of the diverse epigenetic characteristics are grounded in the intrinsic multi-scale sequence features of the genome (Fig. 1B). At the kilobase scale, the genome comprises CpG islands (CGIs), which are clusters of CpGs, and the open sea regions that lie distal to these CGIs. At the megabase scale, the genome comprises forest and prairie domains, which are enriched in CGIs and depleted from CGIs, respectively [28]. In the following, we first focused on the alterations in chromatin spatial structure within two distinct types of oral tumors.

Chromatin reorganization in cancer cells follows a regular pattern

To visually present chromatin structures and their changes in carcinogenesis, a three-dimensional reconstruction was performed based on the Hi-C probability matrices (see Methods, Fig. 1C). Our analysis revealed discernible structural discrepancies between normal and cancer cells. Both AdCC and OSCC exhibit significantly increased short-range interactions compared to normal cells, accompanied by a reduction in long-range interactions (Supplementary Text, Fig. 1C, S2A-S2C).

We analyzed the chromatin compartmentalization, which is a Mb-scale structural characteristic representing spatial interaction of euchromatin and heterochromatin (see Methods), for all samples. Unsupervised hierarchical clustering of compartment vectors show that the compartmentalization of cells is both tissue- and cancer-specific (Fig. 1D). The similarity between adjacent glands

from different cancer types, and the similarity between epithelial cells from diseased and healthy donors demonstrate the reliability of using adjacent non-tumorous tissues as controls (Fig. 1E). Furthermore, chromatin structure is also cancer-subtype and cell-state specific. Distinct structural discrepancies were observed between solid AdCC and cribriform AdCC (Fig. 1E, S2D), as well as among healthy, paracancerous and inflamed epithelial samples (Fig. 1F).

To investigate the linkage between chromatin structure and cell function, we first analyzed non-cancerous cells, namely normal glands and epithelial. We found that regions undergoing compartment switch are significantly enriched with differentially expressed genes (Supplementary Text, Fig. S2E, Table S2), revealing a close relationship between chromatin structure and the realization of tissue specificity. As for the change from healthy epithelial to OSCC, 28 genes switch from compartment B to compartment A and meanwhile are transcriptionally activated (Fig. 1G, Table S2). These genes are functionally related to proliferation, cell cycle, and cell adhesion, including MYO1B, STC1, CDKN2A, HMGA2, RBBP8, and PRNP. From normal gland to AdCC samples, 19 genes significantly move to compartment A and become activated (Fig. 1G, Table S2), which are functionally enriched in cell proliferation and development, such as MYC, SOX11, BMP7, and MYT1L, consistent with the dis-regulated cell cycle and dedifferentiation of tumor cells. Furthermore, we found that tissue-specific chromatin structures become dysregulated in cancer cells. For instance, STC1, a CpG-poor and prairie gene, is located in well-defined compartment A in the gland while located in compartment B in healthy epithelial samples (Fig. 1H), consistent with its higher expression in the former than the latter samples (Fig. S2F). Interestingly, STC1 moves towards compartment B in AdCC and, in contrast, from B to A in OSCC development. This B-to-A switch also occurs in inflamed epithelial samples, suggesting some possible structural similarity between inflammation and carcinogenesis may exist. Another example is PTPRT, which is specifically repressed in the gland but activated in AdCC cells from both structural and transcriptional perspectives (Fig. 1H, S2F).

We further investigated the subtype specificity of AdCC samples and observed that both the expression level and compartment vector of TP63 in cribriform are significantly higher than solid-type samples (Fig. 1I). We performed immunohistochemistry and validated that *p63* is indeed a specific biomarker for distinguishing between two subtypes of AdCC (Fig. 1J). Meanwhile, AdCC patients exhibiting higher levels of TP63 expression have a more favorable prognosis (Fig. 1K). In addition, we found that in all samples, including both normal and malignant cells, the expression level of TP63 is closely correlated with its compartment value, indicating that chromatin structure could be an important epigenetic regulator for its transcription (Fig. 1I). Inspired by this observation, we further identified 254 genes whose transcriptions are most correlated to, and thus likely regulated by, their structural compartmentalization (Table S2, see Methods). We then explored the sequential properties of these genes. Among these compartment-regulated genes, a notably higher percentage exhibit CpG-depleted promoters and are located within prairie domains, surpassing the background proportions (35% vs. 21%, 34% vs. 19%, respectively). These observations indicate that the expression of genes with low CpG promoters and prairie genes are more sensitive to chromatin structure than that of other genes, consistent with our previous findings [15, 29].

Of note, the establishment of chromatin structure is closely correlated with intrinsic CpG density in both normal and cancer cells. Compartments A and B are mainly composed of forest and prairie domains, respectively (Fig. S2G). Although some regions undergo an A-B switch in carcinogenesis, those conserved in A or B possess significantly high or low CpG density, respectively (Fig. S2H). Regions with a propensity for compartment switching possess intermediate levels of CpG density (Fig. S2H). The compartment strength (see Methods) shows that the extent of compartmentalization decreases successively for gland, cribriform AdCC, and solid AdCC (Fig. S2I). The same trend is observed for healthy epithelial, paracancerous epithelial, and OSCC (Fig. S2I).

Cancer genes are involved in changes in the gene-gene interaction network

In the following, to investigate the detailed and specific gene-gene interactions (GGIs) and their biological relationships, we performed $C_T G$ [23] (Hi-C To Geometry, see Methods) analyses on Hi-C data. In addition to effectively eliminating systematic bias, the $C_T G$ matrix was highly consistent with imaging data generated by the FISH technique, providing reliable genome-wide analysis of proximal genes in chromatin architecture [23]. We first evaluated the similarity between each pair of $C_T G$ distance matrices by L1 distance, followed by hierarchical clustering. Clear differences were found among non-cancerous gland, epithelial, and AdCC, OSCC samples. At the same time, high consistency was also found within each sample type. These observations demonstrate the tissue- and cancer-specific nature of the $C_T G$ matrix (Fig. 2A). Taking chromosome 17 as an example, a heatmap of the $C_T G$ matrix shows that both OSCC and AdCC cells exhibit weakened long-range contacts compared to corresponding normal cells (Fig. 2B and C, annotated in pink frames), which is in line with the aforementioned analyses on the decay of contact probability. Additionally,

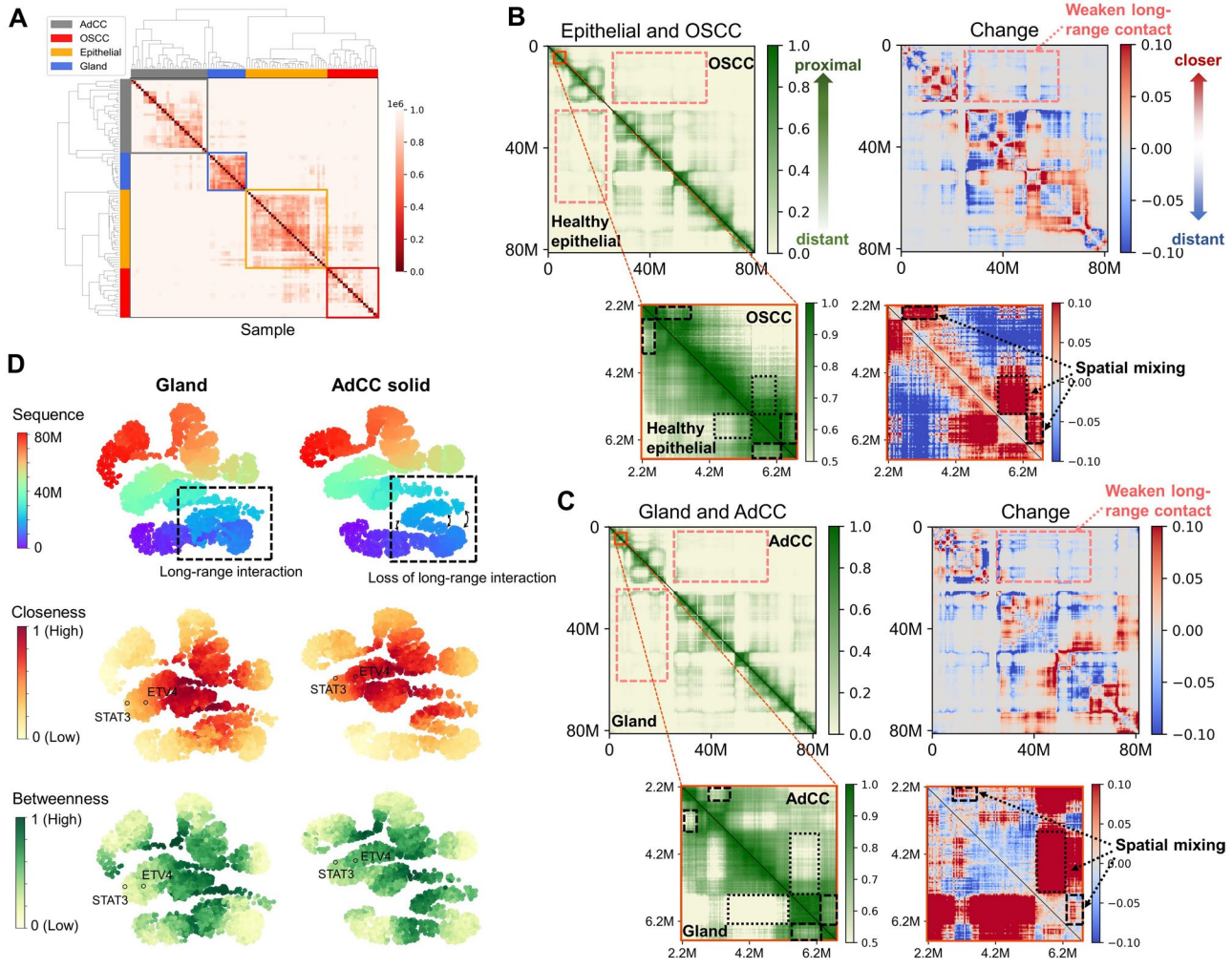

Fig. 2 Gene-gene interaction of normal and tumor cells. **(A)** Hierarchical clustering of L1 distances between the C_TG matrices of all samples. **(B)** Average C_TG matrices of healthy epithelial and OSCC (left panel) and their difference (right panel). **(C)** Average C_TG matrices of gland and AdCC (left panel) and their difference (right). **(D)** Two-dimensional layouts of average C_TG for gland (upper panel) and AdCC solid (bottom panel). Each dot represents a 40-kb bin in chr17 and the color of each in the left, middle and right panels represent its genomic location, closeness centrality and betweenness centrality, respectively

normal gland and epithelial cells possess clear and fine plaid contact patterns, reflecting the high order in the regulation of gene-gene distances (Fig. 2B and C, annotated in black frames). In contrast, AdCC and OSCC cells exhibit a significantly blurred contact diagram, resulting in the spatial mixing of domains that were isolated in normal cells (Fig. 2B and C, annotated in black frames).

To better visualize the changes in GGI, we used the Fruchterman-Reingold algorithm [30] to generate a two-dimensional force-directed layout of the genomic structure. We then used Monte-Carlo sampling to ensure the reproducibility of the two-dimensional layout (see Methods). In layouts, loss of long-range interactions in cancer cells can also be visually observed (Fig. 2D, S2A, S3B). Considering gene-gene interaction as a network, there is a set of genes showing a "hub" feature, namely to be spatially connected to a large number of genomic loci. Meanwhile, a group of genes shows high betweenness, indicating their high capability of bridging different structural modules. In order to characterize the biological importance of these genes, we treated the C_TG matrix as an adjacency matrix of the graph (G) and calculated the betweenness centrality and closeness centrality for each bin of the genome (see Methods, Fig. 2D, S3A). To evaluate the conservation of the gene interaction network, we first divided the genes into 10 groups based on the average level of centrality in normal gland or epithelial. Then we calculated the conservation within each interval among one type of cancer samples. For both betweenness and closeness centrality, the highest and lowest groups are more conserved among different samples than other groups. The corresponding distribution diagram shows a "U" shaped trend (Fig. S3C, S3D). Of note, the high similarity (more than 50% conservation) seen for the group

with the highest centrality indicates a set of genes stably function as essential "hubs" in chromatin architecture. As a comparison, the similarity between normal and tumor samples is lower but also exhibits a U-shape trend (Fig. S3C, S3D).

To investigate the connection between chromatin structure changes and cancer development, we analyzed the function of genes that undergo both significant structural centrality and gene expression changes in carcinogenesis (Table S3). For AdCC, genes with increased closeness centrality and expression level are enriched in functions pertaining to the cell cycle (105 out of 645). Besides cell cycle genes, genes with increased betweenness centrality and expression are also significantly involved in chromatin remodeling, development, and the Notch signaling pathway. Of note, the genes that possess high closeness and betweenness centrality in AdCC are also enriched in cancer genes by a 1.48-fold and 1.68-fold enrichment compared to the background, respectively. For instance, STAT3 and ETV4 show increased closeness, betweenness, as well as expression (Fig. 2D). For OSCC, genes of increased closeness centrality and expression level tend to be functionally related to cell junction disassembly and cell proliferation. Genes showing enhanced betweenness centrality and expression are functionally related to the collagen catabolic process and extracellular matrix organization. The enrichment of cancer genes for these two sets of genes are 1.46 and 2.06, respectively.

Genomic alternation in cancer highly correlates with chromatin structure

Copy number alteration (CNA), a crucial hallmark of cancer, encompasses large-scale genomic rearrangements such as DNA duplications, deletions, and insertions. It was reported that oral leukoplakia with higher CNA frequency was more likely to transfer to malignant OSCC, although the original leukoplakia had been removed [31]. We calculated copy number according to sequencing data generated from the Hi-C experiment for all samples (see Methods, Fig. 3A and B and S4A, Table S5) and demonstrated the reliability of the results from two aspects: First, whole genome sequencing (WGS) experiments were performed on 6 sets of matched cancer and normal samples. WGS-derived copy numbers showed high concordance with those obtained from Hi-C data (Fig. S4A, Table S4). Second, we compared the copy number of OSCC to that of oral leukoplakia with dysplasia (DOL) [31], considered a precancerous lesion of OSCC. A similar CNA pattern was observed for DOL and OSCC, such as a high frequency of copy number gains in chromosomes 3q, 5p13, and 8 and copy number losses in chromosomes 3p and 4. Whereas frequent chromosomal losses or gains characterize OSCC, AdCC exhibits far fewer CNA events than OSCC (Fig. 3A and B). The median length of gain and loss regions for OSCC samples are 259 Mb and 294 Mb, respectively. In contrast, those for AdCC are 53 Mb and 26 Mb, respectively (Fig. S4B), indicating that AdCC is more genetically stable than OSCC.

Although cancer cells are known to be heterogeneous, we found here that CNAs are partially conservative. For instance, more than 50% of OSCC samples lose chromosome 3p and gain 3q (Fig. 3B). Quantitatively, the average Pearson correlation coefficient is 0.38 among samples whose total CNA lengths belong to the top 50% of all OSCC samples. CNA can lead to aberrant gene expression (Supplementary Text, Fig. S4C-F, and Table S5). It was reported that whole genome doubling (WGD) could confer malignant phenotypes of cells by reducing chromosomal segregation, such as TADs and compartments [32]. We compared two samples from the same OSCC patient which exhibit different copy numbers and found that, compared to the diploid chromosome, the strength of compartmentalization in the amplified chromosomes was slightly reduced. On the other hand, the insulation strength of TAD boundaries showed no significant difference (Fig. S4G-I).

Although CNA has been found to be associated with the development of cancer phenotypes, the underlying causes of their propensity to occur at specific genomic locations remain unclear. It is thus interesting to interrogate the relation between CNAs and structural features of the chromatin. We defined common CNA breakpoints for AdCC and OSCC, respectively, as one breakpoint bin occurred in more than 2 tumor samples (see Methods). In the Hi-C contact matrix, we observed that these common breakpoints are often located between two spatially separated domains. For instance, the breakpoint downstream of the chr9p gain observed in two AdCC samples exhibits strong insulating strengths (Fig. 3C). Notably, this strong insulation can also be observed in cancer samples without CNA in this locus and even in normal glands (Fig. 3C). Figure 3D presents the three-dimensional structural modeling results for an AdCC sample without CNAs at chr9p, revealing that the boundaries of CNA, namely breakpoints observed in other AdCCs, indeed coincide with the borders between domains. Comparable instances are observed in OSCC: multiple samples, irrespective of the presence or absence of CNA, exhibit pronounced insulation at the frequent CNA loci on chr16p (Fig. 3E). These results imply a close connection between chromatin spatial structure and CNA, and the former might influence the propensity of the latter. To quantitively describe the relationship between chromatin structure and CNA, we calculated insulation score along the genome with a window size of 800 kb (see Methods). We found that the insulation scores of common breakpoints are significantly lower than the average, reflecting

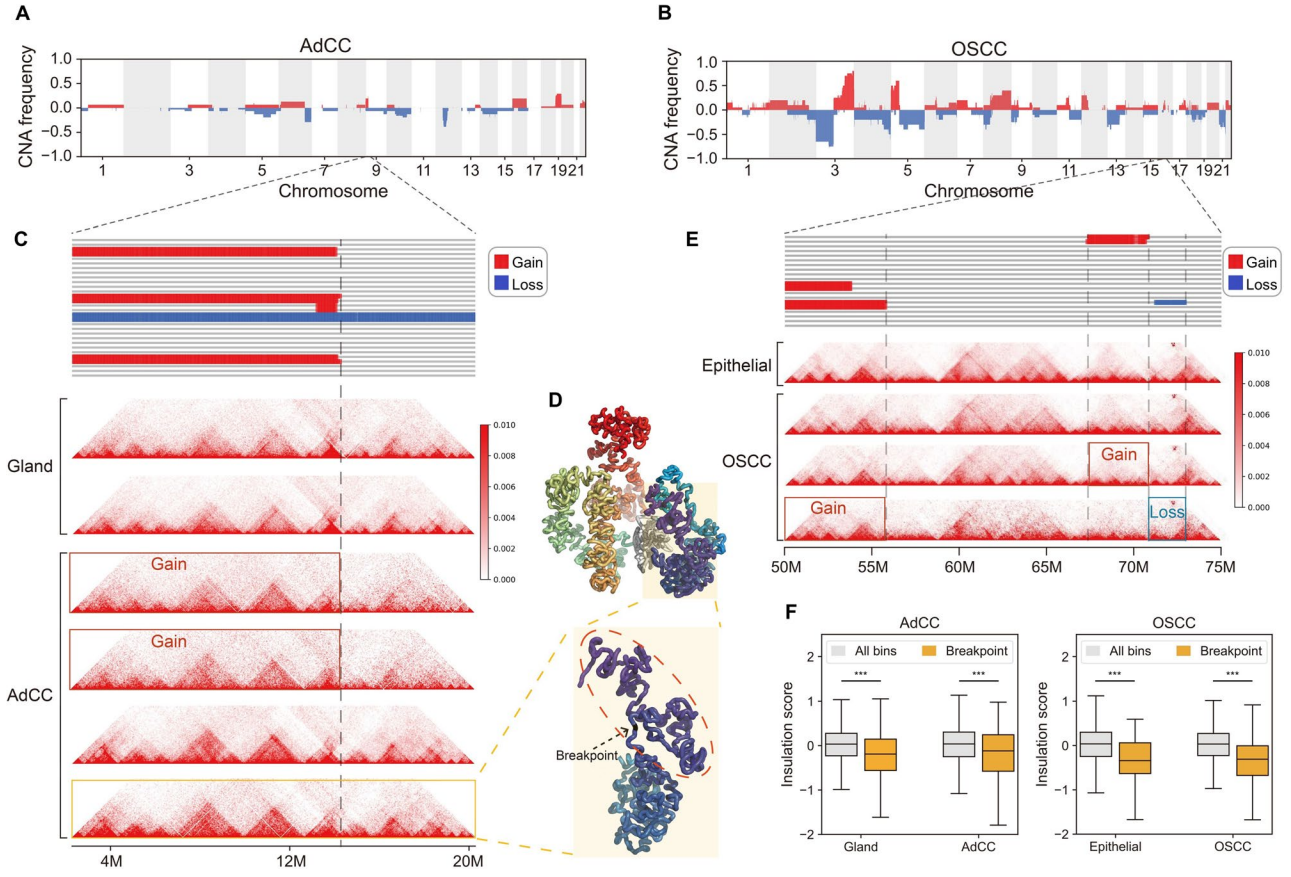

Fig. 3 Copy number alteration of tumor samples and its association with chromatin structure. **(A)** The frequencies of copy number gain and loss of AdCC samples and **(B)** OSCC samples. **(C)** Copy number of all AdCC samples on chromosome 9 (upper), as well as representative Hi-C contact heatmaps within the same region (lower panel, representative samples are arranged from top to bottom as G2_1, G4_1, OACC2_1, OACC7_2, OACC4_1 and OACC1_1, respectively). The breakpoints are marked with gray dashed lines. **(D)** Chromatin 3D structures of chromosome 9 for OACC1 (upper). The area identical to Figure **(C)** is magnified and displayed below. **(E)** Copy number of all OSCC samples on chromosome 16 (upper), as well as the Hi-C contact heatmaps within the same region (lower panel, representative samples are arranged from top to bottom as E58_1, OSCC53_1, OSCC51_1, OSCC58_1, respectively). The breakpoints are marked with gray dashed lines. **(F)** Boxplots for insulation scores in gland and AdCC samples (left panel), epithelial and OSCC samples (right panel). Grey boxes represent all bins and yellow boxes represent copy number breakpoints occurred in AdCC (left panel) and OSCC (right panel). *** represents $p < 0.001$ by Mann-Whitney U test

their stronger insulation strength (Fig. 3F). Again, this insulation property is observed for the same loci in corresponding normal samples (Fig. 3F). In conclusion, we found that a subset of cancer samples are characterized by conserved CNA events, which tend to occur in regions that are inherently spatially separated.

DNA methylation changes in AdCC is distinct from OSCC

Changes of DNA methylation (5mC) in cancer development are frequently observed in tumor cells [8, 33], including hypermethylation of CGIs and hypomethylation of open seas [34]. Owing to the LCM and mini-bulk sequencing, we are able to investigate the methylation pattern in tumor cells of high purity in AdCC and OSCC patients, respectively.

We first calculated the average CpG methylation level at 1-kb resolution along the genome. Subsequent PCA shows significant differences among normal gland, epithelial, AdCC, and OSCC (Fig. 4A). Next, we calculated the correlation of DNA methylation level at various genomic distances, averaging methylation level with a 200-bp window size (Fig. S5A). Methylation autocorrelations exhibit nearly power-law decays in normal epithelial and gland samples. While the decay pattern in OSCC cells differs from that in normal cells, with significantly higher autocorrelation at various genomic distances (Fig. S5A). Meanwhile, the inflamed epithelial positions between normal and cancerous states, probably showing some connection between inflamed states and the latter. In contrast, the differences of methylation autocorrelation features between normal glands and AdCC are much smaller than those between epithelial and OSCC (Fig. S5A).

Considering the distinct sequential and methylation characteristics of CGI and open sea, we calculated their methylation levels separately (Fig. 4B). In normal

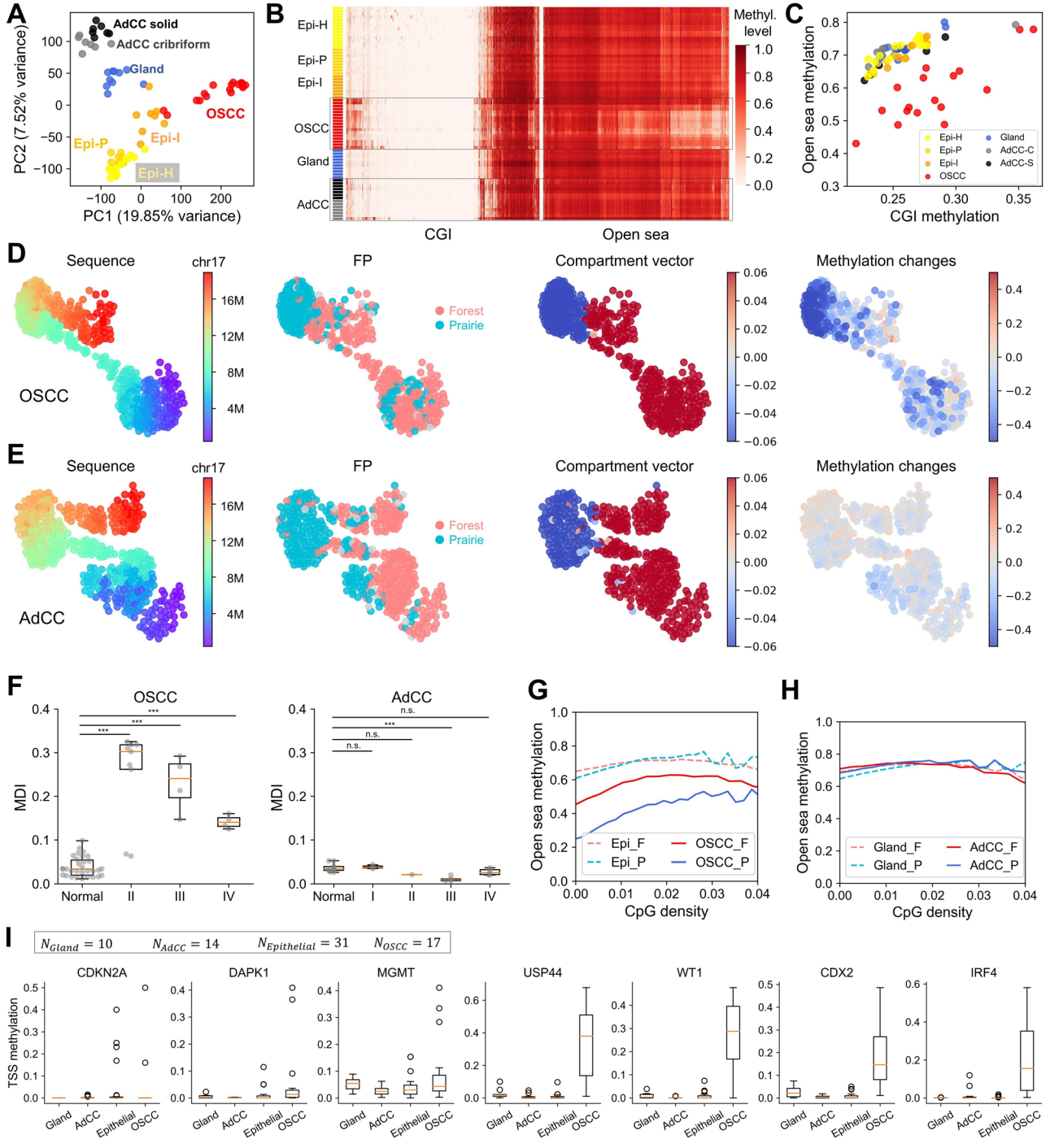

Fig. 4 DNA methylation changes in carcinogenesis and their associations with multiscale CpG densities and chromatin structures. **(A)** PCA based on the methylation levels of 1-kb bins in chr1 for various types of tissues. **(B)** Heatmap of methylation level of CGIs (left panel) and open seas (right panel). Each row represents a sample and each column represents a CGI or open sea. **(C)** Each dot represents the average levels of CGIs (x-axis) and open sea (y-axis) for a sample. **(D)** Two-dimensional layouts of three-dimensional C_TG matrix of OSCC54. Each dot represents a 40-kb bin in chr17p and from left to right panel, the color of each subfigure indicates its genomic location, whether it belong to CGI forest or prairie domains, compartment vector in OSCC54 and open sea methylation change from E54 to OSCC54, respectively. **(E)** Two-dimensional layouts of three-dimensional C_TG matrix of OACC4. Each dot represents a 40-kb bin in chr17p and from left to right panel, the color of each subfigure indicates its genomic location, CGI forest/CGI prairie, compartment vector in OACC4 and open sea methylation change from G4 to OACC4, respectively. **(F)** Boxplot for Forest-Prairie open sea methylation differences (MDIs) of normal epithelial and OSCC in different pathological stages (left) and normal gland, AdCC in different pathological stages (right). *P*-values are calculated by Mann-Whitney U test. ***, $p < 0.001$. *n.s.*, not significant. **(G)** The average open sea methylation level of 1-kb bins with different CpG densities (x-axis). The methylation levels are calculated separately according to whether bins belong to forest (F) or prairie (P) domains. E54, OSCC54 and **(H)** G4, OACC4 serve as representative samples for epithelial, OSCC, gland and AdCC, respectively. **(I)** Boxplots for TSS methylation levels (y-axis) in four types of tissues (x-axis)

epithelial cells, CGI and open seas are lowly and highly methylated, respectively. In contrast, every OSCC sample undergoes various degrees of CGI hypermethylation and/or open sea hypomethylation (Fig. 4C), in line with earlier discoveries in HNSCC [5]. Unexpectedly, such canonical global methylation alterations are rare in AdCC (Fig. 4B and C). Specifically, although some CGIs are hypermethylated in OACC, more are hypomethylated, leading to low overall CGI methylation levels (Fig. S5B, 4C). At the same time, their open seas are highly methylated, akin to normal gland and epithelial cells (Fig. 4B). These features are seen in all solid subtype samples and most of the cribriform samples examined. The non-negative matrix factorization results of CGI and open sea methylation demonstrate that OSCC is indeed similar to many other types of malignancies, while AdCC is distinctly different and more similar to normal cells (Fig. S5C).

To envision the connections of open sea demethylation in carcinogenesis to sequence and structure, we plotted the 2-dimensional layout of spatial chromatin structure. We colored it according to DNA sequence, forest and prairie domains, compartment vector and open sea methylation changes (Fig. 4D and E). As discussed earlier, forests and prairies are spatially separated, respectively making main contributions to compartments A and B. At the same time, open sea demethylation predominantly occurs in prairie/compartment B in OSCC, leading to enlarged Forest-Prairie open sea methylation difference index (MDI, see [28, 35]) in tumor cells (Fig. 4F). Compared to the methylation levels of CGI and open sea, MDI reflects better the differences between normal and cancer (Fig. 4F, S5D). Notably, little tumor stage dependence is found for either CGI methylation, open sea methylation, or MDI level (Fig. 4F, S5D), suggesting that the methylation change, if it occurs in a cancer type, occurs at very early stages of cancer. Overall, genomic regions with lower local CpG density are more likely to be hypomethylated (Fig. 4G), showing the close association between the extent of demethylation and CpG density distribution. In contrast, in AdCC, the low CpG regions remain highly methylated (Fig. 4H), and the differences between forest and prairie domains become even smaller in several tumor samples (Fig. 4F).

To further investigate the reason for the non-canonical methylation in AdCC, we performed CUT&Tag for H3K27me3 in corresponding samples. It was well known that H3K27me3-rich regions, which are often CpG-rich, tend to undergo hypermethylation in carcinogenesis [10, 36]. Therefore, we examined the methylation level of gene promoters heavily marked by H3K27me3 in both normal and cancer samples. We found that these genes are unmethylated in normal epithelial cells and undergo significant hypermethylation in OSCC, as expected (Fig. S5E). In contrast, most of the genes marked by H3K27me3 in normal glands remain unmethylated in AdCC (Fig. S5E). In OSCC, genes with promoters consistently hypermethylated (see Methods) are functionally related to embryonic development and synaptic signaling (Fig. S5F). Meanwhile, the majority of these promoters possess high CpG density (Fig. S5G). Several tumor suppressor genes were reported to be hypermethylated in OSCC [11], such as p16 (CDKN2A) [37], DAPK1, MGMT [38]. A subset of OSCC samples, but almost no AdCC samples, are found here to undergo hypermethylation in these promoters (Fig. 4I). Other tumor suppressor genes which are found significantly hypermethylated in our data, including USP44, WT1, CDX2 and IRF4, also exhibit low methylation level in AdCC samples (Fig. 4I). For AdCC, the majority of hypermethylated promoters possess low CpG density (Fig. S5G), and is functionally related to immune process (Fig. S5F).

Transcription start sites (TSSs) that undergo methylation changes in OSCC are prone to be located in the more repressive spatial environment (compartment B) in both epithelial and OSCC, compared to TSSs of all coding genes (Fig. S6A). Consistently, these genes are also transcriptionally silenced in epithelial samples, and become even more repressed in OSCC samples (Fig. S6B). However, the correlation between expression and methylation changes is only 0.059 for genes that undergo hyper or hypo methylation. For AdCC, in contrast, there is no significant difference in terms of compartment among all TSSs, hypermethylated TSSs, and hypomethylated TSSs in both gland and AdCC (Fig. S6C). In AdCC, hypomethylated TSSs marginally move to compartment A and become transcriptionally activated compared to the background (Fig. S6C, S6D). Meanwhile, hypermethylated TSSs become significantly repressed in AdCC (Fig. S6D). The correlation between expression and methylation changes for these genes is -0.29. In conclusion, methylation changes in OSCC mainly occur in spatially repressive regions with little effect on expression, whereas in AdCC, methylation changes are more closely associated with expression regulation.

To understand why AdCC does not undergo canonical methylation changes like OSCC or other cancers studied previously, we focused on the methylation enzymes, including DNA methyltransferases (DNMTs) and the ten-eleven translocation (TET) family (Fig. S6E). Whole Exome Sequencing (WES) performed on 6 pairs of matched cancer and normal samples revealed no mutations within these enzymes (Table S4). Next, we analyzed their expression level and found that compared to OSCC, AdCC possesses higher expression levels of DNMT1 and DNMT3A, especially the latter. DNMT1 is a maintenance enzyme responsible for methylation during replication, and DNMT3A was also reported to be critical for maintaining global highly methylated status [39]. For all

tumor samples (AdCC and OSCC), the open sea methylation level positively correlates with DNMT3A expression level (with a Spearman correlation of 0.4, Fig. S6F). On the other hand, the expression level of demethylation enzyme TET1 in AdCC was also found to be much higher than that of OSCC (Fig. S6E), and consistently, the methylation level of CGI is negatively correlated with TET1 expression (spearman correlation is -0.62, Fig. S6F). In accordance with this correlation, it was reported that knocking out TET1 could cause hypermethylation of promoters for genes that are functionally related to development and nervous system [39], indicating that the maintenance of low methylation level of these genes requires the demethylation enzyme.

Interplay among chromatin structure, DNA methylation and H3K27me3 modification

Considering that Polycomb complex could function as silencers via chromatin interactions [40, 41], and that they are associated with DNA methylation alterations in carcinogenesis, we next try to delineate the relationship between H3K27me3 signal, methylation changes, and chromatin structures. Firstly, consistent with earlier studies, H3K27me3 markers are found to be unevenly distributed along the genome [41](Fig. 5A, S7A). Furthermore, we observed spatial clustering of H3K27me3-rich regions (MRRs) in the three-dimensional structural models (Fig. 5B). We calculated the spatial H3K27me3 density (see Methods) for each bin and found that MRRs indeed tend to be surrounded by bins of higher H3K27me3 densities than background (Fig. 5C, S7B), which suggests that H3K27me3-rich domains could not only spread linearly along DNA, but also spatially cluster.

Next, we combined genomically adjacent MRRs and further defined long MRRs (see Methods, Fig. 5D, S7C). Notably, although located in compartment A, long MRR domains can act as potent silencers, as evidenced by the considerably lower expression levels of genes in long MRRs than those in compartment B (Fig. 5E and F, S7D, S7E). These potent silencers are observed in both normal gland, epithelial, OACC, and OSCC samples, consistent with previous findings in cell lines [41]. Additionally, MRRs of different lengths tend to differentiate in gene functions (Fig. 5G): Short MRRs are enriched in synapse organization. Long MRRs are functionally enriched in pattern specification, cell fate commitment, and embryonic development, which are usually profoundly repressed in differentiated cells, implying function-specific epigenetic regulation.

MRRs can form interactions over large genomic distances. It was reported that hypermethylation, often observed in MRRs, could lead to disrupted chromatin loops [42]. However, in this study, the chromatin loops are found barely influenced by hypermethylation. On the one hand, the proportions of loops mediated by methylated CGIs are similar between normal and tumor samples (Fig. S7F, 5H). On the other hand, compared to other loops, including unmethylated CGI-mediated loops, loops found in normal epithelial cells with at least one hypermethylated anchor in OSCC showed unchanged contact probability in the latter (Fig. S7G). For instance, HOXA genes and upstream regions in chromosome 7 (Fig. 5I), which are modified by high levels of H3K27me3, spatially interact with each other in normal epithelial. In OSCC, although CGIs in these regions are hypermethylated, their interactions remain unaltered. In AdCC, these loci also undergo hypermethylation, although to a lesser extent than in OSCC, their spatial interactions are also similar to those in normal glands. Similar findings in another region of chr20 show that neither hypermethylation (in OSCC) nor lack thereof (in AdCC) significantly affects the spatial contacts.

We also explored the correlation between changes in H3K27me3 level, spatial H3K27me3 density, methylation, and expression (Fig. 5J). In AdCC, taking G6-OACC6 as a pair of examples, MRRs in which H3K27me3 signals decrease in carcinogenesis are spatially isolated from the MRR environment, and concurrently, related genes are generally activated. Conversely, MRRs with elevated H3K27me3 signals are spatially adjacent to other MRRs, showing an overall decrease in gene expression in cancer. It is noteworthy that the regions with changes in H3K27me3 levels exhibit little methylation change. In the instance of the OSCC (E52-OSCC52), MRRs with decreased H3K27me3 levels in cancer are accompanied by decrescent spatial H3K27me3 densities, as well as a more pronounced rise in methylation than the MRRs of increased levels of H3K27me3. The difference in gene expression changes between these two groups of regions is more minor in OSCC compared to AdCC. In conclusion, the local H3K27me3 spatial density variation is likely accompanied by changes in its neighbor regions. Changes in H3K27me3 for AdCC correlate with changes in gene expression, whereas changes in H3K27me3 in OSCC are more related to the degree of hypermethylation.

Dysregulated gene expression in cancer cells and its association with epigenetic reprogramming

In the following, we try to gain insights into the related factors of abnormal gene expression in carcinogenesis, including those biomarker genes implicated for AdCC [3, 43] and OSCC [5] (Fig. 6A and B). For these genes contributing to cancer development, we performed linear regression analyses between expression changes and each epigenetic factor by ordinary least-squares (OLS) models. Regression coefficients show that compartment changes are most closely associated with abnormal expression,

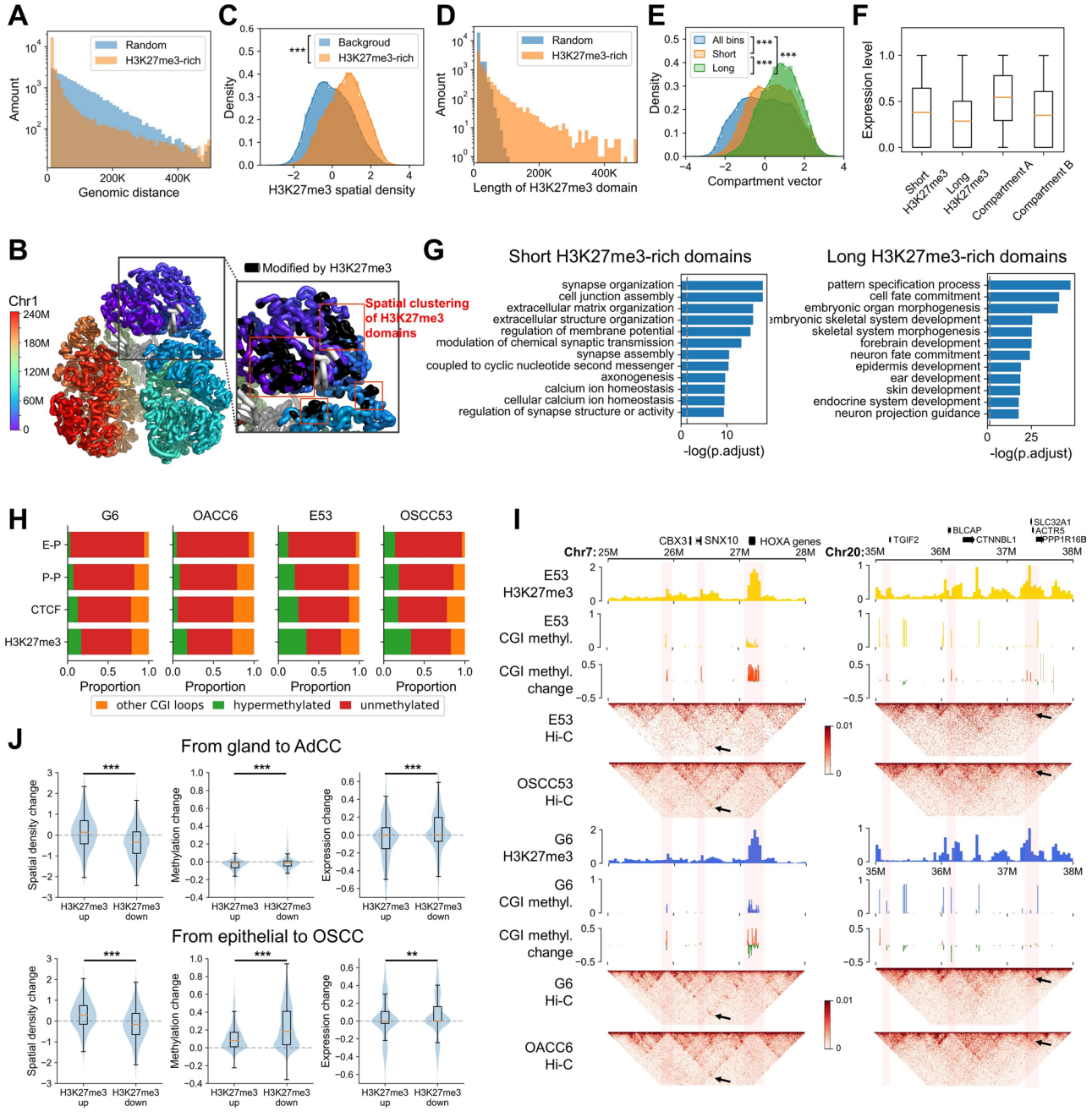

Fig. 5 Associations among chromatin structure, DNA methylation H3K27me3 modification and gene expression. **(A)** Distribution of genomic distances between all pairs of adjacent H3K27me3-rich domains (MRRs) in G6 (normal gland). **(B)** Chromatin structure of chromosome 1 in gland (G6). H3K27me3-rich domains are colored in black. **(C)** Probability density of H3K27me3 spatial density for MRRs (yellow) and all bins (blue). **(D)** The distribution of length for merged MRRs. **(E)** Probability density of compartment vector for all bins (blue), MRRs with length between 10 Kb and 40 Kb (yellow) and MRRs that longer than 40 Kb (green). **(F)** Boxplots for expression level of various types of domains in G6. **(G)** GO analysis for genes in short MRRs (left) and long MRRs (right) in G6. **(H)** Proportion of loops of which anchors keep unmethylated or become hypermethylated in carcinogenesis for all CGI loops in different samples. E-P, P-P, CTCF, H3K27me3 refer to enhancer-promoter loops, promoter-promoter loops, CTCF-mediated loops and H3K27me3-rich mediated loops, respectively. **(I)** Snapshots of two example regions showing H3K27me3, CGI methylation, CGI methylation change and Hi-C data in matched representative samples (E53, OSCC53, G6 and OACC6). **(J)** From left to right, boxplots for changes of spatial H3K27me3 density, TSS methylation and expression for MRR genes that show increased or decreased H3K27me3 level from gland to AdCC (upper panel) and from epithelial to OSCC (lower panel). ** and *** represent *P*-value < 0.01 and *P* < 0.001 by Mann-Whitney U test, respectively

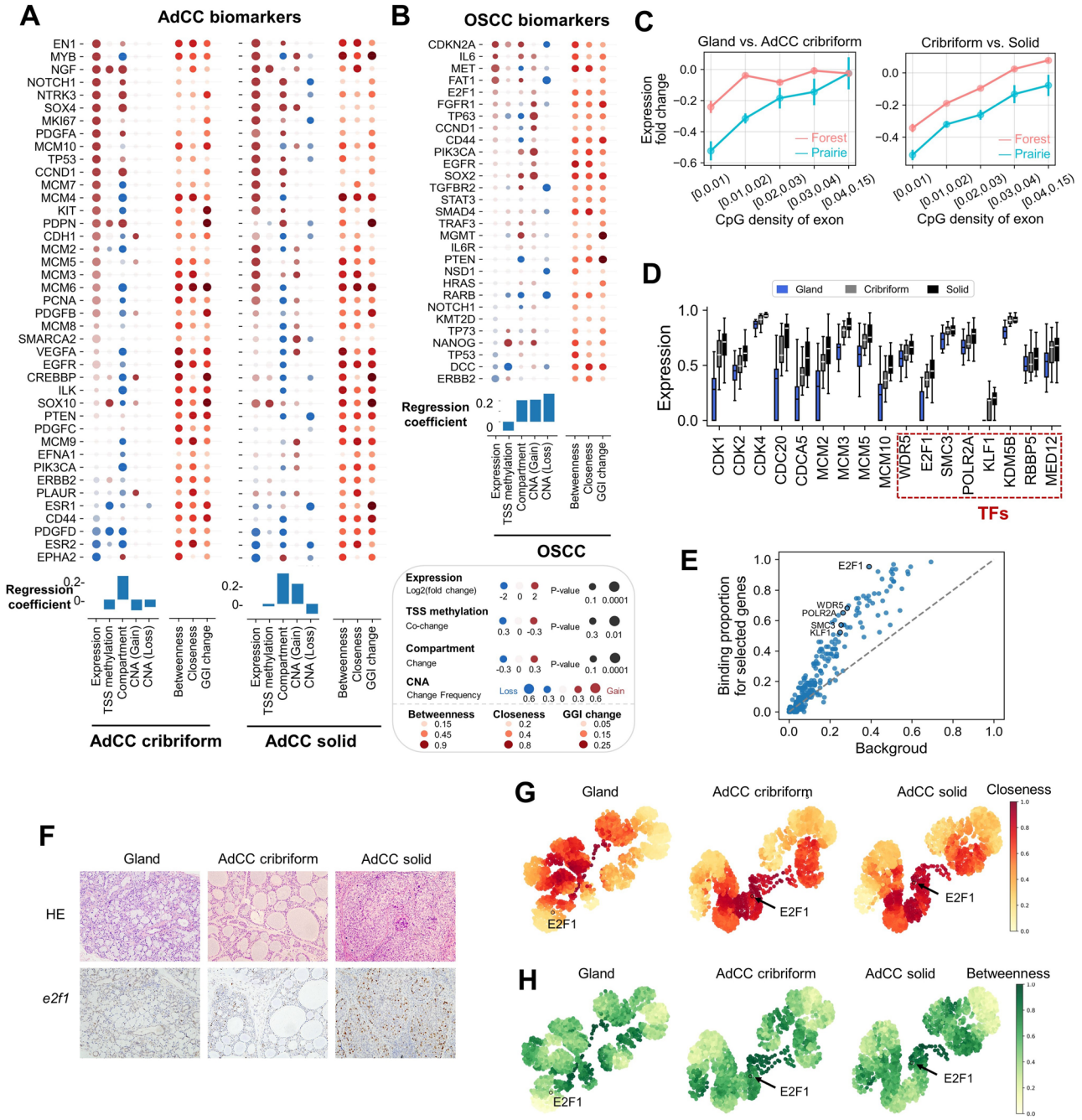

Fig. 6 Changes of gene expression, as well as genomic and epigenomic factors for AdCC and OSCC biomarkers. **(A)** Changes of gene expression, TSS methylation, compartment vector and CNA for AdCC biomarkers and **(B)** OSCC biomarkers are represented by circles, the colors of which represent the directions of changes and the sizes of which indicate the significances of the changes. The degrees of betweenness centrality and closeness centrality in cancer cells, as well as the extent of GGI changes in cancer development are characterized by both colors and sizes of circles. **(C)** Mean expression fold changes (calculated by DESeq2, y-axis) from gland to AdCC cribriform (left panel) and from AdCC cribriform to AdCC solid (right panel) for genes that possess certain level of exon CpG densities (x-axis). Error bars are shown on dots. **(D)** Boxplots for expression ranks of a set of cell cycle genes in gland, cribriform and solid samples. **(E)** Each dot represents a TF and x-axis, y-axis represent the proportion of genes that could be bound by this TF for all genes, and for cell cycle genes which are collectively up-regulated during AdCC development. **(F)** H&E (upper panel) stained images and e2f1 immunofluorescence (IF) micrographs (lower panel) of a normal gland, an AdCC cribriform sample and an AdCC solid sample. **(G)** Two-dimensional layouts colored by closeness centrality and **(H)** betweenness centrality for chromosome 20 in gland, AdCC cribriform and AdCC solid

followed by TSS methylation. However, CNA is the least conserved factor, exhibiting inconsistent effects on expression changes. Notably, these factors are insufficient to account for all gene transcription alterations, necessitating further exploration of other important regulating factors [44]. For instance, from a 3-D genomic perspective, significant changes of chromatin spatial neighbors have been observed, like MCM6, CREBBP, SOX10 in AdCC and MGMT, PTEN in OSCC. Besides, altered betweenness centrality and closeness centrality are also observed for these genes, suggesting the possible roles of the chromatin structure network in carcinogenesis.

Next, we investigated whether tumor cells possess stage-specific transcription features. We identified genes differentially expressed between normal and tumor samples at stage I, and found their expression levels at later stages are similar to stage IV (Fig. S8A, S8B). Similarly, the expression levels of genes identified to differentially express between normal and stage IV tumor samples are found to have undergone such changes at the early stages of the tumor (Fig. S8C, S8D). Meanwhile, the two sets of genes largely overlap (Fig. S8E). These results show that overall gene expression changes occur at early stages for both OSCC and the two AdCC subtypes, with little further changes as the carcinogenesis progresses. As discussed earlier, similar patterns are also observed for DNA methylation changes in OSCC.

Next, we examined whether the gene expression changes in carcinogenesis are DNA sequence-dependent. For normal gland and AdCC cribriform, we found that genes with higher CpG density are more likely to be up-expressed in the cancer cells (Fig. 6C). Meanwhile, genes located in forest domains exhibit a greater degree of upregulation than those found in prairie domains, indicating a close relationship between gene expression alterations in carcinogenesis and large-scale sequence features. This trend is also observed when AdCC cribriform is compared with AdCC solid subtypes (Fig. 6C). These findings show that solid tumors are likely at a more advanced stage than cribriform tumors, in accordance with their clinical phenotypes. For OSCC, the correlation between sequence properties and the extent of expression up-regulation is weaker than AdCC. However, the forest genes also tend to be up-regulated more significantly than prairie genes (Fig. S8F).

Functional analysis shows that up-expressed genes in OSCC cells are highly enriched in extracellular matrix organization, Wnt signaling pathway, and cell division. Meanwhile, genes related to epithelial functions, such as cornification and skin development, are significantly repressed, indicating the loss of tissue specificity in carcinogenesis (Fig. S8G). Comparing the gland to AdCC cribriform and cribriform to solid form, one sees that genes related to cell division and cell cycle regulation are significantly up-expressed (Fig. S8G). These genes include CDK1/2/4, CDC20, CDCA5, and minichromosome maintenance proteins MCM2/3/5/10 (Fig. 6D). The expression levels of these genes become elevated with the aggravation of malignant degree, but not cancer stages (Fig. S8H). Notably, a number of these genes have been used as valuable clinical markers in diagnosing oral cancer. For instance, the expression of MCM2, which is involved in initiating DNA replication, was used for the differential diagnosis of adenoid cystic carcinoma and polymorphous low-grade adenocarcinoma [45]. It also serves as the marker of poor prognosis in several types of tumors, such as breast cancer [46], ovarian cancer [47] and non-small-cell lung cancer [48].

To further explore why these cell cycle genes are collectively up-regulated, we examined whether they are regulated by common transcription factors (TFs). We downloaded the TF-gene regulation data from hTFtarget database [49] and found these genes shared more TFs than random (Fig. 6E). We then identified 20 "common TFs" that bind to more than half of the genes in the gene set, and these binding enrichment levels are more than twice the average value of all genes. Of note, 8 of 20 common TF genes, including WDR5, E2F1, SMC3, POLR2A, KLF1, KDM5B, RBBP5, and MED12, show increasing transcription following the order of normal gland, cribriform, and solid tumors (Fig. 6D). We also performed immunohistochemistry of *e2f1* to validate its higher abundance in solid AdCC (Fig. 6F). Among them, E2F1 and WDR5 also showed increased expression levels in OSCC than in epithelial cells, while other TFs did not (Fig. S8I), indicating the specificity of gene regulation between these two different types of tumors. Notably, we observed that E2F1 exhibits significantly higher closeness and betweenness centrality in the chromatin structure of AdCC compared to that of normal cells (Fig. 6G and H). High expression level of E2F1 also exhibits a disadvantage effect on survival in pan-cancer samples in TCGA (Fig. S8J) [50]. Consequently, E2F1 may not only act as a transcription factor to upregulate target genes, but also play a more pivotal regulatory role through the GGI network, thus holds potential as a cancer biomarker and therapeutic target.

Overview of epigenetic reprogramming

As indicated by our findings above, AdCC and OSCC, and the two subtypes of AdCC show discrepancies in several important genetic and epigenetic aspects, as shown in Fig. 7A (see Methods). Both AdCC subtypes exhibit much lower frequencies of large-scale genetic alteration and little changes in DNA methylation compared to OSCC. In contrast, OSCC and AdCC undergo comparable changes in chromatin structure (compartment and C_TG contact) and gene expression. Meanwhile, the solid

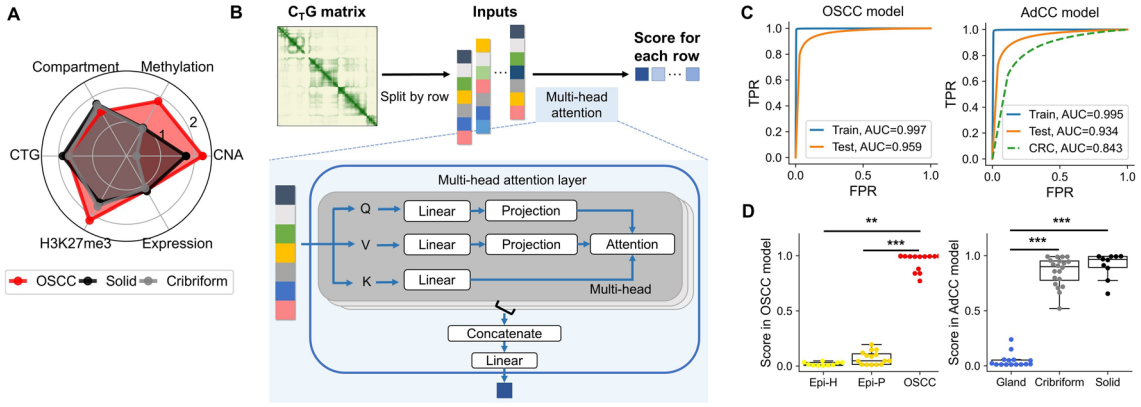

Fig. 7 Multi-omics epigenetic blueprint of oral cancer and predictive models of carcinogenic status based on chromatin structure. **(A)** Radar plot shows the average extents of epigenetic changes for three types of cancers. The value of CNA is 10 times of average proportion of regions that undergo copy number gain or loss. **(B)** Schematic diagram of the transformer model predicting carcinogenic status by C_TG. **(C)** ROC curve (receiver operating characteristic curve) of OSCC model (left) and AdCC model (right) which are used to classify the normal or cancer state for each locus in C_TG matrices. X-axis and y-axis represent false positive rate (FPR) and true positive rate (TPR), respectively. CRC refers to Hi-C samples of colorectal adenocarcinoma. **(D)** Boxplot for cancer risk scores of healthy epithelial (Epi-H), paracancerous epithelial (Epi-P), OSCC samples in OSCC model (left panel) and gland, AdCC cribriform, AdCC solid samples in AdCC model. ** and *** represent $p < 0.01$ and $p < 0.001$ respectively by Mann-Whitney U test

subtype AdCC shows more drastic changes than the cribriform subtype in features such as compartment strength and expression level of cell cycle genes, consistent with their pathological characteristics. We didn't observe significant differences among cancer samples of various stages, and all markers we analyzed show change at the very early cancer stage, if they do change in carcinogenesis (Fig. S8K). Previous published works also found that epigenetic changes, such as aberrant methylation, appear early during tumorigenesis [51, 52].

Given the crucial role of chromatin structure in gene regulation and obvious difference between the C_TG matrix of normal and cancer cells, we aimed to develop a structure-based model for cancer detection. We trained transformer models for two types of cancers to predict whether a region is in cancerous state or normal state from the perspective of structure features (Fig. 7B, see Methods). For OSCC, the AUC scores for training and test sets are 0.997 and 0.959, respectively (Fig. 7C). For AdCC, these numbers are 0.995 and 0.934, respectively (Fig. 7C). To examine the applicability of this mode to other cancer types, we also tested the pan-cancer performance of AdCC model and the AUC for colon adenocarcinoma [16] is 0.843. These results indicate that there exists a common chromatin structure change for different cancer types so that a model based on chromatin structure possesses pan-cancer predictive power. The performance of traditional model, such as SVM, was not as effective as that of the transformer model (Fig. S8L), indicating that the transformer model is more adept at handling high-dimensional and hierarchical interactions.

We can obtain a cancer risk factor for each sample by averaging the predicted score of all loci. For healthy epithelial and paracancerous samples, their risk scores are significantly lower than OSCC and AdCC samples (Fig. 2F). Notably, the score of paracancerous epithelial is slightly higher than healthy ones, indicating the trend of some paracancerous samples to go through cancer transformation, although they are still phenotypically normal.

Discussion

In this study, we performed large-scale multi-omics analysis on the epigenetic landscape of two types of oral cancer, AdCC and OSCC. We found that alterations in the three-dimensional chromatin structure are central to cancer progression. First, the chromatin structure is cell-type and cell-state dependent from multiple perspectives, including the organization of overall structure, compartmentalization, and gene-gene interaction network derived from genome distance matrix. Chromatin structures are shown to closely associate with the execution of cell-specific functions. Second, structural changes in AdCC and OSCC are significant and exhibit similar trends. In contrast, the frequency of CNAs and methylation changes differ greatly between the two types of cancer. Compared to normal cells, cancer cells exhibit weakened long-range interactions, blurred contact-confined domains, and altered compartmentalization, which are closely associated with the dysregulation of tissue specificity and the activation of cancer genes. Third, chromatin structure shows a higher correlation with gene dysregulation in various cancers than other genetic and epigenetic factors we examined. Furthermore, we developed transformer models that successfully identified the cancer state at single-locus resolution. Notably, the chromatin structures of paracancerous epithelial cells deviate from those of normal cells, with a subset of samples exhibiting changes towards tumor cells, as evidenced by

patterns in both GGI network and contact decay features, suggesting that chromatin structure changes have the potential to be used as sensitive and early markers for carcinogenesis. Studies on phenotypic effects of chromatin structure, combined with biochemical experiments, are expected to provide additional in-depth insights into the molecular mechanisms of gene regulation in disease development.

In addition to the dysregulation of gene expression, chromatin structure was also found to have a close relationship with essential hallmarks of cancer: CNA and DNA methylation. In particular, we found CNA events tend to occur in the spatially insulated domains, thus connecting genetic alterations and the 3-D genome. Most OSCC samples undergo global hypomethylation and localized hypermethylation in H3K27me3-rich regions, similar to many other cancers. These changes exhibit dependencies on multi-scale sequence features and chromatin structure. Both hyper- and hypo-methylation tend to occur in regions that spatially reside in repressive domains. Surprisingly, AdCC was associated with nonconventional methylation changes, a rare occurrence in other types of tumor cells. It was reported that deletion of TET genes results in hypermethylation of bivalent genes related to development, nervous system, and cell communication [39]. Meanwhile, DNMTs are required for maintaining a global high methylation level [53]. The faster division rate of cancer cells has been argued to prevent the full restoration of DNA methylation, especially for regions with low CpG density [54, 55]. AdCC, which was reported to typically grow more slowly compared with other malignancies pathologically [56] (Fig. S1D), was found in this study to overexpress methylation "writers" and "erasers", thus likely maintains a largely "normal" DNA methylation pattern. Moreover, the extent of the difference between forest and prairie for OSCC samples is positively correlated with HIST1H1A expression level (spearman correlation is 0.69, Fig. S6F). It was reported that H1 histones could influence local chromatin compaction to further control the epigenetic landscape [57, 58]. In line with our findings, increased HIST1H1A concentrations may amplify the methylation gap between forest and prairie domains by enlarging the local compaction differences between these two types of domains. To investigate the significant differences in methylation changes between AdCC and OSCC, we measured H3K27me3 modification and further examined the interplay between methylation, H3K27me3 signals, and chromatin structure. It is important to note that this study has certain limitations, as it does not include various types of activating histone modifications and transcription factor binding events, which are critical for understanding gene interactions and regulatory networks. Moreover, although we identified a number of consistent epigenetic changes and correlations, their causal relationships require further biochemical experiments and data exploration.

Methods

Materials and experiments

Sample collection

Human tissue samples were obtained with consent from the patients of Peking University Hospital of Stomatology or from healthy donors. The research was approved by the Ethics Committee of the Peking University Hospital of Stomatology (2021-NSFC-43). Tumor and matched normal tissues of 27 AdCC patients and 24 OSCC patients were collected. Normal epithelial samples were collected from the residual tissue of 15 healthy donors after the tooth extraction. Detailed sample information is shown in Table S1.

Preparation of tissue sections

Freshly frozen tissues were embedded in optimal cutting temperature (OCT) medium (SAKURA, #4583) at −25 °C. A 3-μm-thick tissue slice was sectioned to perform a histopathological examination using Leica Cryostat (#CM-1900). The slice was stained by standard hematoxylin and eosin (H&E) procedure and then analyzed under light microscopy (Olympus, #BX51). This slice was used as a morphological and histological reference during laser capture microdissection. For each tissue, two adjacent 10-μm-thick serial sections were obtained. One unstained frozen section was used to capture cell samples for RNA-seq and CUT&Tag. The other section was first fixed in 4% PFA (Solarbio, #P1110) at room temperature for 10 min and quenched by 1.25 M Glycine (Sigma-Aldrich, #G7403-100G) at RT for 10 min. The fixed slice was then processed to H&E staining and ready for laser capture microdissection.

Laser capture microdissection (LCM)

A laser microdissection microscope (Leica, #LMD7000) with 100× magnification was used to perform cell sample capture with proper laser settings. To perform RNA-seq and CUT&Tag, around 3,000 target cells were laser captured from the unstained slice to the empty cap of a nuclease-free 0.2-ml PCR tube for each sample. The adjacent fixed slice was used to extract cell samples for the library construction of Hi-C and methylation sequencing. Cells with the same spatial coordinates as those sampled for RNA-seq and CUT&Tag were collected. For Hi-C library construction, each sample contained approximately 1,000 cells. For methylation sequencing, each sample contained around 200 cells.

Hi-C library preparation

Hi-C experiments were performed following the method described in Ref. [59] with some modifications. Briefly, the LCM slice of each sample was lysed in 100 μL Hi-C lysis buffer on ice for at least 30 min followed by the incubation in 0.5% SDS (Invitrogen, #15553027) at 65℃ for 20 min and the quench of 10% Triton X-100 (Sigma-Aldrich, #T8787). Chromatin digestion by MboI enzyme (NEB, #R0147L) was carried out at 37℃ with rotation for 24 h. After fill-in reaction, the ligation reaction was carried out by incubating at room temperature with rotation for 24 h. After ligation, DNA fragments were released by Proteinase K (Qiagen, #19133) and purified by Ampure XP beads (Beckman Coulter, #A63881). Tagemtation was then performed using TTE Mix V50 Tn5 enzyme (Vazyme, #TD501). Dynabeads M-280 streptavidin beads (Invitrogen, #11206D) were used for the capture of ligation junctions at room temperature overnight with rotation. Ten cycles of PCR amplification were carried out. Post-PCR purification was performed using Ampure XP beads according to the manufacturer's instructions.

RNA-seq

Geo-seq protocol was used to perform small-bulk RNA-seq [60]. After preamplification, to assess the quality of amplified cDNA, size distribution was determined using the 5200 Fragment Analyzer System (Agilent, #M5310AA). Qualified cDNA was diluted to the desired amount (50 ng) and used to construct a library using TruePrep DNA Library Prep Kit V2 for Illumina (Vazyme, TD501) following the manufacturer's instructions.

Methylation library preparation

To construct methylation libraries using LCM mini-bulk samples, we applied NEB EM-seq method [61] (NEB, #E7125) with modifications made based on T-WGBs protocol described in Ref. [62]. LCM samples were first lysed and performed tagmentation as described in T-WGBs protocol. DNA fragments were purified by AMPure XP beads. Methylation conversion was then conducted using NEBNext Enzymatic Methyl-seq Conversion Module (NEB, #E7125L) following the manufacturer's instructions. Converted DNA libraries were amplified by NEBNext Q5U Master Mix (NEB, #M0597) with 12 cycles of PCR reaction.

CUT&Tag

To perform CUT&Tag experiments on unstained LCM frozen slices containing about 3,000 cells each sample, we used the Hyperactive Universal CUT&Tag Assay Kit (Vazyme, #TD903-02) following the protocol. H3K27me3 (abcam, #ab195477) was used as the primary antibody with 1:50 dilution. Rabbit IgG (H&L) secondary antibody (Rockland, #611-201-122) was used with 1:100 dilution.

Whole genome library preparation and sequencing

The LCM samples each containing approximately 600 cells were lysed using a low-temperature protocol to eliminate artifacts in somatic mutation calling as described before [63]. Briefly, each biopsy sample was lysed in customized lysis buffer containing 15 μg/μl native cold-active Bacillus licheniformis Protease (Creative Enzymes, Cat. No. NATE-0633). The lysis reaction was conducted at 6 °C for 1 h. The released genomic DNA was further tagmented by Tn5 transposome (Vazyme, TTE Mix V50 in Cat. No. TD501) followed by 21 cycles of PCR reaction to amplified the library molecules. The qualified libraries were sequenced by 2×150 bp paired-end run on a Novaseq 6000 System (Illumina).

Whole exome library preparation and sequencing

Every four WGS libraries were pooled together for whole exome probe capture using the SureSelectXT Human All Exon V7 (Cat. No. 5191−4005) following the manufacturer's guidelines. The products were quality checked and sequenced with Novaseq 6000 System (Illumina), generating 2×150 bp paired-end reads.

Library QC and sequencing

The libraries were quantified using Qubit 1x dsDNA HS Assay kits (Invitrogen, #Q33230) and the size distribution was assessed using 5200 Fragment Analyzer System (Agilent, #M5310AA). The qualified libraries were then quantified by qPCR and sequenced by 2×150 bp paired-end run on a Novaseq 6000 System (Illumina).

Informatics analysis

Genome data

CGI forest domains and prairie domains are defined according to Ref. [28]. Cluster 1, 2, 3 genes, which were classified based on the distribution of CpG density around TSS, were obtained from Ref. [29]. In this work, we analyzed all protein coding genes which are downloaded from GENECODE release 19 (https://www.gencodegenes.org). Data used for analysis about transcription factors was obtained from hTFtarget database [49]. CGI coordinates were derived from UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables). Cancer genes are acquired from Cosmic database (https://cancer.sanger.ac.uk/cosmic).

RNA-seq analysis

RNA-seq sequence mapping

After adaptor trimming using Cutadapt (version 2.10) [64]. RNA reads were processed by Kallisto (version 0.46.0) [65] to obtain TPM matrix of gene expression.

To calculate the gene count matrix, trimmed reads were mapped to hg19 genome by STAR (version 2.7.6a) [66] and counted by HTSeq 2.0 (version 2.0.1) [67] using htseq-count command with GENCODE v38lift37 annotation. Gene expression level was further analyzed in two format, rank and normalized count, to perform intra-sample comparison and inter-sample comparison, respectively. For the calculation of gene rank, raw counts were converted to TPM (transcripts per million) format and then re-ranked to [0,1]. The rank of highest expression level is 1 and silenced genes are 0. For inter-sample analysis, normalized count was calculated by DESeq2 (Love et al. 2014).

The identification of differentially expressed genes

Differential expression analysis was performed using DESeq2 [68]. Genes with log2(expression fold change) > 1 and FDR < 0.05 were defined as up-expressed genes, and genes with log2(expression fold change) < -1 and FDR < 0.05 were regarded as down-expressed genes.

Gene function analysis

The clusterProfiler package [69] and DAVID (https://david.ncifcrf.gov/) were used in this study for gene function analysis. The background for GO analysis is all genes in orgDB by default.

Tissue specificity for gene

The normalized RNA-seq data was downloaded from GTEx project [70, 71]. The tissue specificity of gene i in tissue t was defined as

$$s_i^t = \frac{\epsilon_i^t - \mu_i^{all}}{\mu_i^{all}}$$

where ϵ_i^t and μ_i^{all} are the mean expression level of gene i in tissue t and all tissues examined, respectively. A gene with a tissue specificity value greater than 2 was defined as a tissue-specific gene.

Hi-C analysis

Hi-C sequence mapping

Paired-end reads were first under adaptor trimming using Cutadapt (version 2.10) [64] with default arguments. Reads shorter than 20 bp were filtered out after adapter trimming. Trimmed reads were mapped to Genome Reference Consortium Human Build 37 (hg19, downloaded from UCSC, https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips) and processed by HiC-Pro (version 2.11.4) [72] using default settings. The contact matrix extracted by HiC-Pro were then used in downstream analysis.

Chromatin 3D structure modeling

A comprehensive description of our methodology is available in our preceding publication [73]. In summary, we employed a coarse-graining approach to represent a chromosome as a series of beads. The equilibrium spacing between adjacent beads was determined by translating contact frequencies into spatial distances. We initiated the process with a randomly generated structure, which was subsequently refined through Molecular Dynamics (MD) simulations. This iterative optimization continued until the root-mean-square deviation (RMSD) of the resultant structure reached a state of convergence.

Compartment identification

This study regards the Hi-C matrix as an adjacency matrix and proposes a physically more interpretable compartmentalization method based on spectral clustering. The main steps are as follows:

(1) Compute the Laplacian matrix corresponding to the Hi-C matrix of each individual chromosome.
(2) Perform eigenvalue decomposition on the Laplacian matrix.
(3) Select the top *k* eigenvectors for Linear Discriminant Analysis (LDA) dimensionality reduction [74].
(4) The dimensionally reduced output serves as indicator vectors for compartments A and B, where a positive or negative sign indicates compartment A or compartment B, respectively.

LDA can effectively utilize prior knowledge for dimensionality reduction while minimizing the impact of the choice of *k* on classification. Additionally, LDA dimensionality reduction being linear helps preserve the inherent features of the matrix, reducing the risk of overfitting. The selection of *k* is robust and set to 50. Subcompartments are inferred by dcHiC [75].

ENCODE hg19 blacklist regions were filtered out for analysis. For a group of gland samples or healthy epithelial samples (amount is N), conserved compartment A was defined as those belong to compartment A in no less than N-1 samples. Similar definition applies to conserved compartment B. For inflamed epithelial, AdCC and OSCC, considering the high heterogeneity, the criterion of conserved A/B was adjusted to more than 0.5*N samples.

The compartment index (CI) of bin i was calculated as

$$\boldsymbol{CI_i} = \boldsymbol{ln}\left(\frac{C_{i-A}}{C_{i-B}}\right)$$

where C_{i-A} and C_{i-B} are the average normalized contact probabilities between bin i and compartment A bins, and between bin i and compartment B bins, respectively.

A higher value of CI indicates that one region is located in a more open environment.

Calculation of compartment strength

Compartment strength at the quantile i (ranging from 1 to 99) is

$$strength\,(quantile\,i) = \frac{\begin{array}{c}\mathrm{hom}\,otypic\mathrm{inter}actions\\ within\,region_x\,and\,region_y\end{array}}{\begin{array}{c}heterotypic\,\mathrm{inter}actions\\ between\,region_x\,and\,region_y\end{array}}$$

$region_x$ and $region_y$refer to the regions with top i% and bottom i% compartment vector, respectively.

Identification of compartment switch

When comparing the compartments of any pair of tissues, e.g., tissue 1 and tissue 2, a region defined as A-to-B meets the following conditions:

(1) Average compartment vectors for samples belongs to tissue 1 is positive.
(2) Average compartment vectors for samples belongs to tissue 2 is negative.
(3) The compartment vectors of samples belonging to tissue 1 are significantly higher than those of samples belonging to tissue 2, namely $p < 0.05$ by Mann-Whitney U test.

Identification of compartment regulated genes

Firstly, we calculated the spearman correlation between compartment vectors and normalized expression counts among all samples. Then we calculated spearman correlation between averaged compartment vectors and averaged expression count for seven kinds of tissues, including epithelial samples from healthy, inflamed and cancer donors, gland, OSCC and AdCC in solid and cribriform subtypes. A gene was defined as closely regulated by its compartmentalization when both of two kinds of correlations were larger than 0.5.

Force-directed algorithm in C_TG two-dimensional layout

The two-dimensional visualization of chromatin structure is realized based on Fruchterman-Reingold algorithm [30] and can help study the differences of the three-dimensional structure of chromatin during carcinogenesis from a global perspective.

The Fruchterman-Reingold algorithm treats nodes as "electrons" and treats edges as "springs". The force between two nodes can be defined as:

$$F_{ij} = \vec{r}_{ij}(k^2/|r_{ij}|^2 - (A_{ij}|r_{ij}|)/k) \quad (1)$$

The interaction force between nodes includes Coulomb repulsion force and the Hooke attraction force. The algorithm pushes and pulls nodes apart and searches equilibrium layout that reaches local minima of total energy by iterations. The resulted layouts are of uncertainties and repeated tests with same parameters may end up with different results. To ensure the reproducibility of the layout, we used Monte Carlo sampling to find the global minima of total energy. The total energy is defined as:

$$E_{\mathrm{total}} = \sum\nolimits_{i>j} E_{ij} \quad (2)$$

And,

$$E_{ij} = k^2 \lg |r_{ij}| - \frac{A_{ij}|r_{ij}|^3}{3k} \quad (3)$$

The detailed steps are as follows:

(1) Use simple random sampling to estimate the distribution of total energy and to design the acceptance probability function;
(2) Initialize random layout;
(3) Use force-directed algorithm to find local minima;
(4) Reject/accept;
(5) Perform random perturbations and return to step 3.

In order to get a more reasonable layout within each epoch, we generate the initial layout based on the highly connected edges, and then perform random perturbations. It helps improve the reproducibility within fewer iterations.

Betweenness, degree and closeness centrality of gene nodes

In order to quantify the importance of each node in perspective of graph theory, we calculate the betweenness centrality and closeness centrality of each node for different samples. The C_TG matrix is regarded as the adjacency matrix A of graph G with n nodes, and the interaction between the nodes i and j is denoted as A_{ij}.

The betweenness centrality of node i is defined as:

$$B_i = \sum\nolimits_{s \neq i \neq t} \frac{n^i_{s,t}}{g_{s,t}}$$

The betweenness centrality reflects the controlling of node i over the rest of nodes, and the node with high betweenness centrality in the three-dimensional chromatin structure network may be related with gene that plays key roles in the regulatory network.

The closeness centrality of node i is defined as:

$$C_i = \frac{N-1}{\sum_{j=1}^{n} d_{i,j}}$$

The closeness centrality reflects the proximity between node i and the rest of nodes, and node with high proximity centrality may be related with gene that functionally associated with more genes.

To ensure the comparability of centrality, we perform statistical analysis on sorting ranks rather than the real values.

The transformer model for cancer identification

Transformer model used the $C_T G$ contact probability between each chromatin bin and other bins belonging to the same chromosome as the input feature. Bins in 22 autosomes were merged for training and zero-filling was performed to generate 1*6232 features for each bin (equal to the length of chromosome 1). The network consists of a multi-head attention mechanism (head=8) encoding layer and a fully connected layer, and the output feature number is 2. The model used the Binary Cross Entropy Loss function and Adam optimizer, with a learning rate 5e-6. AdCC model uses autosomes from randomly selected 8 AdCC and 8 gland samples as training set (G10_1, G10_2, G10_3, G1_1, G1_2, G2_1, G2_2, G5_1, OACC4_1, OACC4_2, OACC6_1, OACC6_2, OACC8_1, OACC8_2, OACC13_1, OACC13_2) and remaining samples as test set. The OSCC model uses autosomes from 15 OSCC samples and 5 epithelial samples from healthy donors, inflamed tissues and paracancerous samples as training set (OSCC51_1, OSCC51_2, OSCC53_1, OSCC53_2, OSCC55_1, OSCC55_2, OSCC56_1, OSCC57_1, OSCC57_2, OSCC58_1, OSCC58_2, OSCC59_1, OSCC59_2, OSCC60_1, OSCC60_2, E53_1, E53_2, E58_1, E58_2, E57_2, E102, E103_1, E103_2, E104_1, E104_2, E107_1, E107_2, E112_1, E112_2, E111) and remaining samples as test set.

Loop identification

Loop calling in high-quality Hi-C sample was performed using Peakachu [76], which utilizes Random Forest classification framework to identify loops.

Analysis of genetic alterations

Copy number alteration analysis

To perform CNA calling, 10 million of mapped reads of either Hi-C or WGS sequencing data were used for each sample. The reads were tabulated into non-overlapping dynamic bins (50 kb resolution) across the genome. Lowess regression normalization was performed to reduce the GC bias of bin counts. Copy number was called by R package DNAcopy (version 1.44.0, https://bioconductor.org/packages/DNAcopy) [77] using circular binary segmentation algorithm (alpha=0.0001, min.width=5, undo.SD=2). Regions with copy number>2.2 or <1.8 in more than three non-cancerous samples were ignored for analysis. For diploid autosomes, copy number gain was defined as a copy number≥2.5, and copy number loss was defined as a copy number≤1.5. CNA breakpoints were defined as 40-kb bins where a continuous gain or loss segment starts and ends, as well as its one upstream bin and one downstream bin.

Single nucleotide variation (SNV) calling

Paired-end reads from the sequencer were aligned to the human reference genome hg38 using bwa-mem2 (version 2.2.1) [78] with default parameter settings. The aligned BAM files were then sorted using Samtools (version 1.11) [79]. To call somatic SNVs from the WES data, we followed the best practice guidelines of The Genome Analysis Toolkit (GATK) v4.3.0.0 [80]. Briefly, Picard-tools 2.27.5 was used to fix mate pairs and mark PCR duplicates [81]. Next, the base quality recalibration was performed with GATK. We used gatk Mutect2 [82] to call the somatic SNVs in each tumor sample, with the corresponding normal samples as the germline comparator. To ensure SNVs calling accuracy, we applied gatk FilterMutectCalls tool to perform filtering steps. The variants listed in the dbSNP 150 database were excluded. The filtered mutations were annotated by gatk Funcotator for the downstream analysis. De novo extraction of mutational signatures was conducted using SigProfilerExtractor [83].

CUT&Tag data analysis

CUT&Tag sequence mapping

Trimmed sequencing reads were aligned to the hg19 genome using Bowtie2 aligner (version 2.2.9) [84]. Coverage matrix was calculated using SAMtools (version 1.7) [85] by the command 'samtools depth'and converted to bedgraph format by custom script. H3K27me3 signals were processed to counts per million (CPM) format using deeptools [86] with "binSize=50". ENCODE hg19 blacklist regions were filtered out during normalization. CPM signal was then averaged by 10-kb window size or averaged by gene promoter (defined as upstream 4-kb and downstream 4-kb considering the broad distribution of H3K27me3), followed by quantile normalization.

Definition of long and short H3K27me3-rich domains

CPM signals were averaged in 10-kb window size along the genome. Regions with top 10% abundance were defined as H3K27me3-rich domains. H3K27me3-rich regions were further merged when adjacent gap was no more than 10 kb. According to Fig. 5D and S7C, the intersection point between the length distribution of merged regions and random scenarios is 40 kb. Therefore, the regions of which length longer than 40 kb and shorted

than 40 kb are defied as long and short H3K27me3 regions, respectively.

Methylation data analysis

Methylation sequence mapping

Cutadapt (version 2.10) [64] was used to trim adaptors of methylation data. Trimmed data was mapped by Bismark (version v0.23.0) [87] to hg19 genome. To deduplicate reads, Bismark function deduplicate_bismark was used. The methylation matrix was extracted by Bismark command bismark_methylation_extractor using the default setting. DNA methylation level of each CpG site was given in percentage by

$$\beta = \frac{M}{M+U} \times 100\%$$

where M and U are the signal strength of methylated and unmethylated CpG, respectively. For the calculation of TSS methylation level, the location of gene promoter was downloaded from FANTOM5 project ("https://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks"). The midpoints of CAGE peaks labeled with "p1@" are regarded as transcriptional start sites (TSSs). For each gene, average methylation level was calculated for sequences between −1 kb (upstream) and +1 kb (downstream) of TSS using the 200-bp non-overlapping window. The average methylation level of the five windows with the lowest methylation among these 10 windows was regarded as the methylation level of TSS. The up/down stream methylation level was defined as average methylation level for [-8 kb, -2 kb] and [2 kb, 8 kb].

Analysis of hypermethylation and hypomethylation

For each TSS, differences between any two non-cancerous samples were calculated and changing cutoff was set as (mean+standard variation), which is 0.l. Methylation increase more than cutoff was defined as hypermethylation and decrease more than cutoff was hypomethylation. The consistency of methylation changes c_k for given promoter k in carcinogenesis was defined as

$$c_k = \frac{\sum_{i,j} d_{ij}}{N1 \times N2}$$

$$d_{ij} = \begin{cases} 1, & m_j - m_i > cutoff \\ -1, & m_j - m_i < -cutoff \\ 0, & else \end{cases}$$

Where m_i and m_j are the methylation level of normal sample i and tumor sample j. N1 and N2 are total amounts of all normal and tumor samples, respectively.

The degree of epigenetic changes

To evaluate the extend of epigenetic changes in AdCC and OSCC samples relative to normal. We calculated Euclidean distances between epigenetic marks of any pairs of normal gland samples or any pairs of epithelial samples. The average distance is $\bar{l}_{gland}$ and $\bar{l}_{epithelial}$. Average distances between any pairs of normal and tumor samples are calculated as $\bar{l}_{gland\_cribriform}$, $\bar{l}_{gland\_solid}$ and $\bar{l}_{epithelial\_OSCC}$. For epigenetic mark k, the extent of changes c_k was defined as

$$c_{k,\ cribriform} = \frac{\bar{l}_{gland\_cribriform}}{(\bar{l}_{gland} + \bar{l}_{epithelial})/2}$$

$$c_{k,\ solid} = \frac{\bar{l}_{gland\_solid}}{(\bar{l}_{gland} + \bar{l}_{epithelial})/2}$$

$$c_{k,\ OSCC} = \frac{\bar{l}_{epithelial\_OSCC}}{(\bar{l}_{gland} + \bar{l}_{epithelial})/2}$$

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12943-024-02100-0.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Supplementary Material 6

Acknowledgements

Parts of the Fig. 1A were drawn by using pictures from Servier Medical Art. Servier Medical Art by Servier is licensed under a Creative Commons Attribution 3.0 Unported License (https://creativecommons.org/licenses/by/3.0/). The data that support the findings of this study were derived from TCGA resources (http://cancergenome.nih.gov) available in the public domain.

Author contributions

Conceptualization, Y. Q. G., Y. H., J. Z. and T. L.; Data curation and analysis, Y. Z., L. L., Y. X.; Funding acquisition, Y. H., Y. Q. G., T. L., J. Z.; Investigation, Y. X., L. L., Y. Z., Y. H., J. W., Z. M.; Supervision, Y. Q. G., Y. H., J. Z., T. L.; Visualization, Y. X., L. L., Y. Z.; Writing – original draft, Y. X., L. L., Y. Z.; Writing – review & editing, Y. Q. G., Y. H., J. Z.

Funding

This work was funded by National Natural Science Foundation of China (No. 92053202, 92353304, 22050003, 22050002 and T2188102), Beijing Municipal Science and Technology Commission Grant (Z211100003321006), CAMS Innovation Fund for Medical Sciences (2019-I2M-5-038), New Cornerstone Science Foundation (NCI202305).

Data availability

All data analyzed during this study are publicly available. The processed data reported in this paper have been deposited in the OMIX [88, 89],

China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences (https://ngdc.cncb.ac.cn/omix: accession no. OMIX007102 (RNA), no. OMIX007103 (CUT&Tag), no. OMIX007106 (Methylation), no. OMIX007107 (Hi-C)). The raw sequence data reported in this paper are deposited in the Genome Sequence Archive [89] in National Genomics Data Center [90], China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences (GSA-Human: HRA008185) that are publicly accessible at https://ngdc.cncb.ac.cn/gsa-human.

Declarations

Ethical approval

All procedures in this study were approved by the University Institutional Ethics Committee (2021-NSFC-43).

Competing interests

The authors declare no competing interests.

Author details

[1]Beijing National Laboratory for Molecular Sciences, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China
[2]Changping Laboratory, Beijing 102206, China
[3]Department of Stomatology, Beijing Chaoyang Hospital, Capital Medical University, Beijing 100020, China
[4]Department of Oral Pathology, National Center of Stomatology, National Clinical Research Center for Oral Diseases, Peking University School and Hospital of Stomatology, National Engineering Research Center of Oral Biomaterials and Digital Medical Devices, Beijing, China
[5]Research Unit of Precision Pathologic Diagnosis in Tumors of the Oral and Maxillofacial Regions, Chinese Academy of Medical Sciences (2019RU034), Beijing, China
[6]Biomedical Pioneering Innovation Center (BIOPIC), Peking University, Beijing, China
[7]Institute for Cell Analysis, Shenzhen Bay Laboratory, Shenzhen 528107, China

Received: 28 June 2024 / Accepted: 23 August 2024
Published online: 06 September 2024

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer statistics 2020: GLOBOCAN estimates of incidence and Mortality Worldwide for 36 cancers in 185 countries. Cancer J Clin. 2021;71:209–49.
2. Sahara S, Herzog AE, Nör JE. Systemic therapies for salivary gland adenoid cystic carcinoma. In *American journal of cancer research*, vol. 11. pp. 4092–4110; 2021:4092–4110.
3. Atallah S, Marc M, Schernberg A, Huguet F, Wagner I, Mäkitie A, Baujat B. Beyond Surgical Treatment in Adenoid cystic carcinoma of the Head and Neck: A literature review. Cancer Manage Res. 2022;14:1879–90.
4. Hellquist H, Skalova A. *Histopathology of the Salivary Glands*. 2014.
5. Johnson DE, Burtness B, Leemans CR, Lui VWY, Bauman JE, Grandis JR. Head and neck squamous cell carcinoma. Nat Reviews Disease Primers. 2020;6:92.
6. Ho AS, Kannan K, Roy DM, Morris LGT, Ganly I, Katabi N, Ramaswami D, Walsh LA, Eng S, Huse JT, et al. The mutational landscape of adenoid cystic carcinoma. Nat Genet. 2013;45:791–8.
7. Hanahan D. Hallmarks of cancer: new dimensions. Cancer Discov. 2022;12:31–46.
8. Koch A, Joosten SC, Feng Z, de Ruijter TC, Draht MX, Melotte V, Smits KM, Veeck J, Herman JG, Van Neste L, et al. Analysis of DNA methylation in cancer: location revisited. Nat Reviews Clin Oncol. 2018;15:459–66.
9. Vidal E, Sayols S, Moran S, Guillaumet-Adkins A, Schroeder MP, Royo R, Orozco M, Gut M, Gut I, Lopez-Bigas N, et al. A DNA methylation map of human cancer at single base-pair resolution. Oncogene. 2017;36:5648.
10. Ehrlich M. DNA hypermethylation in disease: mechanisms and clinical relevance. Epigenetics. 2019;14:1141–63.
11. Gaździcka J, Gołąbek K, Strzelczyk JK, Ostrowska Z. Epigenetic modifications in Head and Neck Cancer. Biochem Genet. 2020;58:213–44.
12. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289.
13. Schmitt Anthony D, Hu M, Jung I, Xu Z, Qiu Y, Tan Catherine L, Li Y, Lin S, Lin Y, Barr Cathy L, Ren B. A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep. 2016;17:2042–59.
14. Quan H, Tian H, Liu S, Xue Y, Zhang Y, Xie W, Gao YQ. Progressive domain segregation in early embryonic development and underlying correlation to genetic and epigenetic changes. Cells 2021, 10.
15. Tian H, Yang Y, Liu S, Quan H, Gao YQ. Toward an understanding of the relation between gene regulation and 3D genome organization. Quant Biology. 2020;8:295–311.
16. Johnstone SE, Reyes A, Qi Y, Adriaens C, Hegazi E, Pelka K, Chen JH, Zou LS, Drier Y, Hecht V, et al. Large-scale topological changes restrain malignant progression in colorectal cancer. Cell. 2020;182:1474–89.
17. Yang L, Chen F, Zhu H, Chen Y, Dong B, Shi M, Wang W, Jiang Q, Zhang L, Huang X, et al. 3D genome alterations associated with dysregulated HOXA13 expression in high-risk T-lineage acute lymphoblastic leukemia. Nat Commun. 2021;12:3708.
18. Xu J, Song F, Lyu H, Kobayashi M, Zhang B, Zhao Z, Hou Y, Wang X, Luan Y, Jia B et al. Subtype-specific 3D genome alteration in acute myeloid leukaemia. Nature 2022.
19. Li X, Liu Y, Salz T, Hansen KD, Feinberg A. Whole-genome analysis of the methylome and hydroxymethylome in normal and malignant lung and liver. Genome Res. 2016;26:1730–41.
20. Heyn H, Vidal E, Ferreira HJ, Vizoso M, Sayols S, Gomez A, Moran S, Boque-Sastre R, Guil S, Martinez-Cardus A, et al. Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. Genome Biol. 2016;17:11.
21. Bell A, Bell D, Weber RS, El-Naggar AK. CpG island methylation profiling in human salivary gland adenoid cystic carcinoma. Cancer. 2011;117:2898–909.
22. Pastor DM, Poritz LS, Olson TL, Kline CL, Harris LR, Koltun WA, Chinchilli VM, Irby RB. Primary cell lines: false representation or model system? a comparison of four human colorectal tumors and their coordinately established cell lines. In *International journal of clinical and experimental medicine*, vol. 3. pp. 69–83; 2010:69–83.
23. He Y, Xue Y, Wang J, Huang Y, Liu L, Huang Y, Gao YQ. Computational enhanced Hi-C data reveals the function of structural geometry in genomic regulation. bioRxiv. 2022;2022(2007):2012–499232.
24. Swaney Danielle L, Ramms Dana J, Wang Z, Park J, Goto Y, Soucheray M, Bhola N, Kim K, Zheng F, Zeng Y, et al. A protein network map of head and neck cancer reveals PIK3CA mutant drug sensitivity. Science. 2021;374:eabf2911.
25. Kim M, Park J, Bouhaddou M, Kim K, Rojc A, Modak M, Soucheray M, McGregor Michael J, O'Leary P, Wolf D, et al. A protein interaction landscape of breast cancer. Science. 2021;374:eabf3066.
26. Cheng F, Zhao J, Wang Y, Lu W, Liu Z, Zhou Y, Martin WR, Wang R, Huang J, Hao T, et al. Comprehensive characterization of protein–protein interactions perturbed by disease mutations. Nat Genet. 2021;53:342–53.
27. Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, Srivastava A, Molla G, Madad S, Fernandez-Granda C, Satija R. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotechnol. 2024;42:293–304.
28. Liu S, Zhang L, Quan H, Tian H, Meng L, Yang L, Feng H, Gao YQ. From 1D sequence to 3D chromatin dynamics and cellular functions: a phase separation perspective. Nucleic Acids Res. 2018;46:9367–83.
29. Tian H, He Y, Xue Y, Gao YQ. Expression regulation of genes is linked to their CpG density distributions around transcription start sites. Life Sci Alliance. 2022;5:e202101302.
30. Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement. Software: Pract Experience. 1991;21:1129–64.
31. Li X, Liu L, Zhang J, Ma M, Sun L, Li X, Zhang H, Wang J, Huang Y, Li T. Improvement in the risk assessment of oral leukoplakia through morphology-related copy number analysis. *SCIENCE CHINA Life Sciences*.
32. Lambuta RA, Nanni L, Liu Y, Diaz-Miyar J, Iyer A, Tavernari D, Katanayeva N, Ciriello G, Oricchio E. Whole-genome doubling drives oncogenic loss of chromatin segregation. Nature. 2023;615:925–33.
33. Klutstein M, Nejman D, Greenfield R, Cedar H. DNA methylation in Cancer and Aging. Cancer Res. 2016;76:3446.
34. Sandoval J, Heyn H, Moran S, Serra-Musach J, A Pujana M, Bibikova M, Esteller M. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics: Official J DNA Methylation Soc. 2011;6:692–702.

35. Xue Y, Yang Y, Tian H, Quan H, Liu S, Zhang L, Yang L, Zhu H, Wu H, Gao YQ. Computational characterization of domain-segregated 3D chromatin structure and segmented DNA methylation status in carcinogenesis. Mol Oncol. 2022;16:699–716.
36. Schlesinger Y, Straussman R, Keshet I, Farkash S, Hecht M, Zimmerman J, Eden E, Yakhini Z, Ben-Shushan E, Reubinoff BE, et al. Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. Nat Genet. 2006;39:232.
37. Allameh A, Moazeni-Roodi A, Harirchi I, Ravanshad M, Motiee-Langroudi M, Garajei A, Hamidavi A, Mesbah-Namin SA. Promoter DNA methylation and mRNA expression level of p16 gene in oral squamous cell carcinoma: correlation with clinicopathological characteristics. Pathol Oncol Research: POR. 2019;25:1535–43.
38. Don KR, Ramani P, Ramshankar V, Sherlin HJ, Premkumar P, Natesan A. Promoter hypermethylation patterns of P16, DAPK and MGMT in oral squamous cell carcinoma: a systematic review and meta-analysis. Indian J Dent Research: Official Publication Indian Soc Dent Res. 2014;25:797–805.
39. Wang Q, Yu G, Ming X, Xia W, Xu X, Zhang Y, Zhang W, Li Y, Huang C, Xie H, et al. Imprecise DNMT1 activity coupled with neighbor-guided correction enables robust yet flexible epigenetic inheritance. Nat Genet. 2020;52:828–39.
40. Schuettengruber B, Bourbon H-M, Di Croce L, Cavalli G. Genome regulation by Polycomb and trithorax: 70 years and counting. Cell. 2017;171:34–57.
41. Cai Y, Zhang Y, Loh YP, Tng JQ, Lim MC, Cao Z, Raju A, Lieberman Aiden E, Li S, Manikandan L, et al. H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. Nat Commun. 2021;12:719.
42. Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, Lee K, Canfield T, Weaver M, Sandstrom R, et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res. 2012;22:1680–8.
43. Coca-Pelaz A, Rodrigo JP, Bradley PJ, Vander Poorten V, Triantafyllou A, Hunt JL, Strojan P, Rinaldo A, Haigentz M, Takes RP, et al. Adenoid cystic carcinoma of the head and neck – an update. Oral Oncol. 2015;51:652–61.
44. Liang W-W, Lu RJ-H, Jayasinghe RG, Foltz SM, Porta-Pardo E, Geffen Y, Wendl MC, Lazcano R, Kolodziejczak I, Song Y et al. Integrative multi-omic cancer profiling reveals DNA methylation patterns associated with therapeutic vulnerability and cell-of-origin. Cancer Cell.
45. Ghazy S, Baghdadi MH. Maspin and MCM2 immunoprofiling in salivary gland carcinomas. Diagn Pathol. 2011;6:89.
46. Wojnar A, Pula B, Piotrowska A, Jethon A, Kujawa K, Kobierzycki C, Rys J, Podhorska-Okolow M, Dziegiel P. Correlation of intensity of MT-I/II expression with Ki-67 and MCM-2 proteins in invasive ductal breast carcinoma. Anticancer Res. 2011;31:3027–33.
47. Levidou G, Ventouri K, Nonni A, Gakiopoulou H, Bamias A, Sotiropoulou M, Papaspirou I, Dimopoulos MA, Patsouris E, Korkolopoulou P. Replication protein A in nonearly ovarian adenocarcinomas: correlation with MCM-2, MCM-5, Ki-67 index and prognostic significance. Int J Gynecol Pathology: Official J Int Soc Gynecol Pathologists. 2012;31:319–27.
48. Ramnath N, Hernandez F, Tan D-F, Huberman J, Natarajan N, Beck A, Hyland A, Todorov I, Brooks J, Bepler G. MCM2 is an independent predictor of survival in patients with non–small-cell Lung Cancer. J Clin Oncology: Official J Am Soc Clin Oncol. 2001;19:4259–66.
49. Zhang Q, Liu W, Zhang H-M, Xie G-Y, Miao Y-R, Xia M, Guo A-Y. hTFtarget: a comprehensive database for regulations of human transcription factors and their targets. Genom Proteom Bioinform. 2020;18:120–8.
50. Tang Z, Kang B, Li C, Chen T, Zhang Z. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. Nucleic Acids Res. 2019;47:W556–60.
51. Lu Y, Chan Y-T, Tan H-Y, Li S, Wang N, Feng Y. Epigenetic regulation in human cancer: the potential role of epi-drug in cancer therapy. Mol Cancer. 2020;19:79.
52. Mancarella D, Plass C. Epigenetic signatures in cancer: proper controls, current challenges and the potential for clinical translation. Genome Med. 2021;13:23.
53. Lyko F. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. Nat Rev Genet. 2018;19:81–92.
54. Zhou W, Dinh HQ, Ramjan Z, Weisenberger DJ, Nicolet CM, Shen H, Laird PW, Berman BP. DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat Genet. 2018;50:591–602.
55. Ming X, Zhang Z, Zou Z, Lv C, Dong Q, He Q, Yi Y, Li Y, Wang H, Zhu B. Kinetics and mechanisms of mitotic inheritance of DNA methylation and their roles in aging-associated methylome deterioration. Cell Res 2020:1–17.
56. Dillon PM, Chakraborty S, Moskaluk CA, Joshi PJ, Thomas CY. Adenoid cystic carcinoma: a review of recent advances, molecular targets, and clinical trials. Head Neck. 2016;38:620–7.
57. Willcockson MA, Healton SE, Weiss CN, Bartholdy BA, Botbol Y, Mishra LN, Sidhwani DS, Wilson TJ, Pinto HB, Maron MI, et al. H1 histones control the epigenetic landscape by local chromatin compaction. Nature. 2021;589:293–8.
58. Yusufova N, Kloetgen A, Teater M, Osunsade A, Camarillo JM, Chin CR, Doane AS, Venters BJ, Portillo-Ledesma S, Conway J, et al. Histone H1 loss drives lymphoma by disrupting 3D chromatin architecture. Nature. 2021;589:299–305.
59. Rao Suhas SP, Huntley Miriam H, Durand Neva C, Stamenova Elena K, Bochkov Ivan D, Robinson James T, Sanborn Adrian L, Machol I, Omer Arina D, Lander Eric S. Aiden Erez L: a 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80.
60. Chen J, Suo S, Tam PPL, Han J-DJ, Peng G, Jing N. Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-Seq. Nat Protoc. 2017;12:566–80.
61. Vaisvila R, Ponnaluri C, Sun Z, Langhorst B, Saleh L, Guan S, Dai N, Campbell M, Sexton B, Marks K et al. Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. Genome Res 2021, 31.
62. Wang Q, Gu L, Adey A, Radlwimmer B, Wang W, Hovestadt V, Bähr M, Wolf S, Shendure J, Eils R, et al. Tagmentation-based whole-genome bisulfite sequencing. Nat Protoc. 2013;8:2022–32.
63. Li R, Di L, Li J, Fan W, Liu Y, Guo W, Liu W, Liu L, Li Q, Chen L, et al. A body map of somatic mutagenesis in morphologically normal human tissues. Nature. 2021;597:398–403.
64. Martin M. CUTADAPT removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 2011, 17.
65. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.
66. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
67. Putri GH, Anders S, Pyl PT, Pimanda JE, Zanini F. Analysing high-throughput sequencing data in Python with HTSeq 2.0. Bioinformatics. 2022;38:2943–5.
68. Love MI, Huber W, Anders S. Moderated estimation of Fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
69. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: J Integr Biol. 2012;16:284–7.
70. Consortium GT. The genotype-tissue expression (GTEx) project. Nat Genet. 2013;45:580–5.
71. Sonawane AR, Platig J, Fagny M, Chen C-Y, Paulson JN, Lopes-Ramos CM, DeMeo DL, Quackenbush J, Glass K, Kuijjer ML. Understanding tissue-specific gene regulation. Cell Rep. 2017;21:1077–88.
72. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, Heard E, Dekker J, Barillot E. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 2015;16:259.
73. Xie WJ, Meng L, Liu S, Zhang L, Cai X, Gao YQ. Structural Modeling of Chromatin Integrates Genome Features and reveals chromosome folding Principle. Sci Rep. 2017;7:2818.
74. Fisher RA. The use of multiple measurements in taxonomic problems. Annals Eugenics. 1936;7:179–88.
75. Chakraborty A, Wang JG, Ay F. dcHiC detects differential compartments across multiple Hi-C datasets. Nat Commun. 2022;13:6827.
76. Salameh TJ, Wang X, Song F, Zhang B, Wright SM, Khunsriraksakul C, Ruan Y, Yue F. A supervised learning framework for chromatin loop detection in genome-wide contact maps. Nat Commun. 2020;11:3428.
77. Venkatraman E, Olshen A. DNAcopy: A Package for analyzing DNA copy data. 2010.
78. Vasimuddin M, Misra S, Li H, Aluru S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. *IEEE Parallel and Distributed Processing Symposium (IPDPS)* 2019:314–324.
79. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. *GigaScience* 2021, 10:giab008.
80. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics* 2013, 43:11.10.11–11.10.33.
81. Institute B. Picard Toolkit. GitHub Repository 2019, https://broadinstitute.github.io/picard/

82. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling Somatic SNVs and Indels with Mutect2. bioRxiv; 2019.
83. Islam SMA, Díaz-Gay M, Wu Y, Barnes M, Vangara R, Bergstrom EN, He Y, Vella M, Wang J, Teague JW et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. Cell Genomics 2022, 2.
84. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
85. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. Genome Project Data Processing S: the sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
86. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016;44:W160–5.
87. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinf (Oxford England). 2011;27:1571–2.
88. Members C-N, Partners. Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. Nucleic Acids Res. 2024;52:D18–32.
89. Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, Dong L, Zhang Z, Yu C, Sun Y, et al. The genome sequence Archive Family: toward Explosive Data Growth and Diverse Data types. Genom Proteom Bioinform. 2021;19:578–83.
90. Members C-N, Partners. Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2022. Nucleic Acids Res. 2022;50:D27–38.

Publisher's note