# Comparative profiling of chromatin open state of various cells and tissues for dissecting the epigenetic causes of various biological processes

**Jian Wu**
Southeast University
**Jinliang Gao**
Southeast University
**Shuyan Zhang**
Southeast University
**Tao Luo**
Southeast University
**Wei Dai**
Southeast University
**Jianhui Ling**
Southeast University
**Jinke Wang** ( wangjinke@seu.edu.cn )
https://orcid.org/0000-0002-3352-4690

**Research article**

**Keywords:** SALP-seq, chromatin state, profiling, cells, comparative

**Posted Date:** January 20th, 2020

**DOI:** https://doi.org/10.21203/rs.2.21232/v1

# Abstract

Background The comparative study of chromatin open state of various cells and tissues is helpful for dissecting the epigenetic causes of various biological and pathological processes such as development and tumorigenesis. This study comparatively characterized the chromatin open regions (CORs) of two normal human cell lines (HL7702 and MRC-5), nine human cancer cell lines (GM12878, HepG2, PANC-1, A549, HT29, SKOV3, SiHa, C-33A, and HeLa), one mouse cancer cell line (Hepa1-6), and healthy mouse liver tissue with SALP-seq.

Results This study therefore made a set of systematic and useful COR profiles of human and mouse cells and tissues. These COR profiles were used to explore the possible epigenetic bases of several interesting scientific problems, including how chromatin state changes contribute to tumorigenesis, how the cancer-related mutation hotspots are formed, why fibroblast was most widely used to prepare iPSCs, and how HPV subtypes differently affect the chromatin structure of cervical cells. The results revealed that the comparative COR profiling can shed new insights into the potential molecular mechanisms underlying these important biological and pathological processes. The comparative COR profiling also demonstrated transcription factors differentially dominate various biological and pathological processes. At last, the comparative COR profiling uncovers new potential markers or targets for cancer diagnosis and therapy.

Conclusions The comparative COR profiling with our developed SALP-seq technique can be used to uncover epigenetic bases and molecular mechanisms underlying various biological processes.


# Background

Chromatin state is an important aspect of epigenetics, which plays critical roles in various biological and pathological processes, such as development, differentiation, and disease occurrence. For example, epigenetic abnormality is a common feature of all human cancers and epigenetic aberrations play profound and ubiquitous roles in cancers [1]. As an important aspect of epigenetics, chromatin accessibility is the degree of nuclear macromolecules physically contacting with chromatinized DNA. Restrictive chromatin state may prevent appropriate induction of tumor suppressor programs or block differentiation, while permissive states may allow stochastic oncogene activation [2]. Identification of chromatin accessibility landscapes in different types of cancers can also provide clues for understanding of tumorigenesis [3], which is a process of oncogenic reprogramming actually with many dysregulated chromatin remodeling complexes in cancer cells [4, 5]. Therefore, characterization of chromatin state or epigenetic changes of various cells can provide critical insights into molecular mechanisms underlying a variety of biological and pathological processes. Characterization of chromatin state can also provide useful information to the identification of regulatory DNA elements and transcription factors, which is important for elucidating the regulators of various genes.

Due to these key roles of chromatin state, an important technique, Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) [6, 7], was developed and widely used to characterize chromatin

state. This techniques can be used to rapidly identify those accessible chromatin regions (i.e. open chromatin). For example, Corces et al., have recently used ATAC-seq to map the chromatin accessibility landscape of 23 types of primary human cancer using 410 samples derived from The Cancer Genome Atlas (TCGA) [8]. By identifying over 500,000 accessible DNA elements, the authors greatly expanded the scope of known DNA regulatory elements of cancer genomes and offered new insights into inherited risk loci for cancer. Besides its key role in characterization of chromatin accessibility, ATAC-seq is in fact a strong technique to construct libraries of next generation sequencing (NGS) with double-strand DNA (dsDNA).

To overcome some limitations of dsDNA-based NGS library construction and sequencing techniques, we have recently developed a new single strand-based NGS library preparation and sequencing technique, Single Strand Adaptor Library Preparation-sequencing (SALP-seq) [9]. When combined with the Tn5 tagmentation of ATAC-seq technique, SALP-seq can be used to detect the chromatin state of different cell samples in a high-throughput format [9]. Except overcoming a key limitation of ATAC-seq technique, SALP-seq can be used to comparatively characterize the chromatin states of various cells by using differently barcoded Tn5 transposome adaptors. Moreover, the cell-free DNA can even be used to effectively construct the NGS library and sequenced with SALP-seq [10], from which the chromatin accessibility of both healthy and cancer individuals can be obtained.

In order to further mine the worth of SALP-seq technique in comparative characterization of chromatin state and show its useful applications in this field, this study performed a systematic comparative characterization of chromatin accessibility of as many as eleven human cell lines and two mouse cells, including human pancreatic cancer cell line (PANC-1), human alveolar basal epithelial cells (A549), human colon cancer cell line (HT-29), human liver cancer cell line (HepG2), ovarian cancer cell line (SKOV-3), three cervix cancer cell lines (SiHa, C-33A and HeLa), human normal liver cell line (HL7702), human embryonic lung fibroblast (MRC-5), human lymphoblastoid cell line (GM12878), mouse liver cancer cell line (Hepa1-6), and normal liver cells isolated from BALB/C mouse. These cells represent main kinds of human cancers and normal cells as controls. Based on the obtained chromatin accessibility profiles, this study explored the potential epigenetic bases of several interesting scientific problems, including how chromatin state changes contribute to tumorigenesis, how the cancer-related mutation hotspots are formed, why fibroblast was most widely used to prepare iPSCs, and how HPV subtypes differently affect the chromatin structure of cervical cells.

# Methods

### Cultivation of cells

PANC-1, SiHa, C-33A, MRC-5 and Hepa1-6 were grown in Dulbecco's Modified Eagle's Medium (DMEM, GIBCO), A549, HT-29, SKOV-3, HL7702 cells were cultured in RPMI 1640, which were all supplemented with 10% fetal bovine serum (FBS, GIBCO) with 100 μg/mL streptomycin and 100 units/mL penicillin at 37°C in 5% $CO_2$.

### Preparation of mouse liver single cell suspension

Fresh BALB/C mouse liver was washed with PBS for three times then cut into 1 $mm^2$ in PBS. After transferred into 15 mL tube, the tissue was spun at 1,000 rpm for 10 min then the supernatant was removed. The tissue was digested with 5 mL pancreatin for 30 min and during digestion process the reactions were gently mixed per 5 min to improve the digestion efficiency. After digestion, 5 mL medium with serum were added to stop the reaction. The mixture was centrifuged at 1000 rpm for 10 min, and washed with 5 mL PBS for one time then resuspended the precipitate with 1 mL PBS. The cell density was counted with hemocytometer and diluted to suitable density with PBS.

### Preparation of various adaptors

The preparation process of adaptors was described in previous paper [9]. Briefly, oligonucleotides were all synthesized by Sangong Biotech (Shanghai) (Table S1). To prepare the barcoded Tn5 adaptors (BTAs), barcode and ME oligos were dissolved in $ddH_2O$ to a final concentration at 20 μM, and then mixed in equimolar in PCR tube. For preparing single strand adaptors (SSAs), SSA-PN and SSA-PNrev oligo were dissolved in $ddH_2O$ at the concentration of 20 μM, and then mixed in equimolar in PCR tube. Finally, all oligo mixtures were denatured in the water bath for 5 minutes at 95 °C and gradually cooled to 25 °C for annealing into various adaptors.

### Preparation of barcoded Tn5 transposome

Briefly, according the instruction of the Tn5 transposase (Robust Tn5 Transposase, Robustnique Corporation Ltd.), 4 μL of BTA (10 μM) was mixed with 2 μL 10× Tn5 transposome assemble buffer (TPS), 1 μL Tn5 transposase and 13 μL $H_2O$. The reaction was gently mixed then incubated at 25 °C for 30 min to generate Tn5 transposome. The transposome was stored at -20 °C until use.

### Tagmentation of chromatins from various samples

We performed SALP-seq with 100,000 cells collected from 8 human cell lines, SiHa, C-33A, MRC-5, A549, HT-29, SKOV-3, HL7702, 1 mouse cell line Hepa1-6, and mouse liver tissue. Cells were collected by spinning at 500 g for 5 min at 4 °C and washed once with 50 μL of cold PBS. Cells were lysed by resuspended in cold lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$ and 0.1% IGEPAL CA-630). Cells were then spun at 500 g for 10 min at 4 °C to collect the nuclei precipitate. For tagmentation, nuclei were tagmented in a 30 μL reaction containing 20 μL of Tn5 transposome, 3 μL of DMF and 1× LM buffer. The tagmentation reactions were mixed gently and incubated at 37 °C for 30 min, during which the reactions were gently mixed per 10 min to improve the tagmentation efficiency. In tagmentation, different BTAs (Table S2) were used to tagment the different cell samples. After tagmentation, the tagmentated chromatins of PANC-1, SiHa, C-33A, Hepa1-6, A549, HT-29, SKOV-3, HL7702 and mouse liver cells were pooled together and the chromatin of MRC-5 were treated separately. The chromatin from multiple cells and MRC-5 were incubated in 65 °C for 1 hour with 0.1% SDS and 400 μg/mL Proteinase K (Sigma). After incubation, the DNA were purified with standard phenol:chloroform extraction [11].

### Preparation of SALP-seq libraries with tagmented chromatins of various samples

For preparing NGS library, the purified gDNAs were denatured at 95 °C for 5 min and chilled on ice immediately for 5 min. The denatured gDNAs were ligated with SSA in a 10 μL reaction volume with 1 μL of T4 DNA ligase (NEB, M0202L), 1× T4 DNA ligase buffer and 1 μL of SSA (5 μM) at 16 °C overnight. Then the reaction was mixed with equal volume of 2× rTaq mix (Takara) and incubated at 72 °C for 15 min, the SSA-linked gDNAs was purified with 1.2× Ampure XP beads (Beckman Coulter). The two kinds of purified gDNAs were amplified with different index primers in a 50 μL PCR reaction containing 25 μL of NEBNext® Q5® Hot Start HiFi PCR Master Mix (NEB, M0543S), 1 μL of NEBNext Universal PCR Primer (10 μM), and 1 μL of NEBNext Index Primers (10 μM). The PCR program was as follows: (i) 98 °C for 5 min; (ii) 98 °C for 10s; 65 °C for 30s; 72 °C for 1 min; 18 cycles; (iii) 72 °C for 5 min. The PCR products were run with agarose gel and the DNA fragments of 300—1000 bp were extracted with the QIAquick Gel Extraction Kit.

### Next-generation sequencing

After amplified with Illumina-compatible primers (Table S1), two NGS libraries, including NGS-L1 (tagmented chromatins of 9 cell lines), NGS-L2 (tagmented chromatins of MRC-5 cell line) were constructed by using SALP. The libraries were detected and quantified with Agilent Bioanalyzer 2100. Two libraries were sequenced using Illumina Hiseq X Ten platform (Nanjing Geneseeq).

### Analysis of SALP-seq data

The raw reads data were separated according to the index and barcode by using a homemade Perl scripts. Then the ME (19 bp) and barcode (6 bp) sequences were removed from the 5′ end of the pair-end sequencing reads 2. All reads were cut to 30 bp and aligned to the human or mouse genome (hg19 for human; mm9 for mouse) by using Bowtie program (version 1.1.2), with the default settings, except that the parameter -X 2000 was used to ensure the long fragments could be aligned to the genome. Peak calling was performed with macs2 [12], with the following parameters: -f BEDPE -keep-dup=2. Peak annotation was performed with Homer software [13].

For comparison of COS difference between cancer and normal cell lines, we combined HepG2, HeLa and GM12878 SALP-seq data sequenced in previously research [9] with 8 human cell lines to perform comprehensive analysis. All bam files were merged and recalling peaks using macs2. To get high confidence chromatin regions, 500 bp around the summit of the top 50,000 peaks were defined and reads count of these regions were calculated for each cell type using bedtools [14]. The difference between these regions was analyzed by DESeq2 [15], and regions with $p < 0.05$ were defined as significant difference chromatin regions. The significant difference chromatin regions between two liver cell lines (HepG2 and HL7702), three cervix cancer cell lines (HeLa, SiHa and C-33A), two human normal cell lines (MRC-5 and HL7702) and two mouse cell lines (Hepa1-6 and mouse liver cell) were compared with same procedure.

De novo motif analysis of significant difference chromatin regions was performed with HOMER. Gene ontology analysis was performed with DAVID by uploading the genes to the website [16]. Peak annotation was performed using HOMER to defined the closest gene to the peaks. The genomic feature of chromatin open regions (CORs) was defined with ChIPseeker [17]. All statistical analysis was performed with R program and homemade Perl scripts.

## Results

# SALP-seq libraries constructed with high throughput procedure

Using SALP method, the chromatin state of totally 10 cell types was analyzed. Based on the feature of SALP, multiple samples could be pooled together after tagmentation, and following library construction process could be performed simply. With the specific designed BTAs, sequencing data from multiple samples were divided easily according with the barcodes (Table S2). As a result, a total of 235,362,453 mapped reads were obtained, and the average mappable ratio was 82.21%.

Characterization of COS of cancer and normal cell lines

To compare the COS difference between human normal and cancer cell lines, the SALP-seq results of totally 11 human cell lines were compared comprehensively. To validate the reliability of the chromatin state identified by SALP-seq, the reads density around TSS ± 5000 bp was calculated with 100 bp window. Compared with distal regions, proximal regions around TSS show high reads density in all cell lines (Fig. 1), which means the high COL, consistent with previous reports [18, 19]. This result reveals the reliability of the chromatin state captured by SALP-seq in different cell lines. The reads distribution in whole genome scale was shown as Circos (Fig. 2A). Different cancer cell lines show great diverse reads density. MRC-5 and HL7702, as two normal cell lines derived from different tissues, also show diverse reads density. Some chromatin regions show high reads density in all cell types. In order to further compare the chromatin state between cancer and normal cell lines, the differential CORs were identified (Fig. 2B). The results show that the COL of cancer cell lines is all higher than normal cell lines, indicating that lots of chromatin regions are opened during the cancer process. The number of closed chromatin regions caused by cancer is limited (Fig. 2B). The comparison of distance between differential CORs and TSS shows that the percentage of CORs located around TSS (TSS ± 1 kb) in cancer cell lines is higher than normal cell lines, while the normal cell lines contain more CORs located in TSS downstream 10–100 kb (Fig. 2C). The genomic distribution of CORs in different chromatin regions is also distinct. The significantly different chromatin regions of cancer cell lines are mainly located in promoters, while the regions of normal cell lines are located in distal intergenic regions (Fig. 2D). In order to investigate the functional TFs in cancer and normal cell lines, de novo motif analysis was performed for chromatin regions with differential chromatin state. In cancer cell lines, the binding sites of 21 TFs newly existed due to chromatin open and the binding sites of 10 TFs were lost due to chromatin close (Fig. 2E) (Table S3). The gene ontology analysis of genes associated with the different chromatin regions also show

great diversity between cancer and normal cell lines, indicating that these cell-specific CORs served different roles in transcription regulation (Fig. 2F).

## Characterization of COS of multiple cancer cell lines

In order to define the COS difference between multiple cancer cell lines, further analysis was performed with 11 human cell lines. Similar to the reads distribution in genome scale (Fig. 2A), the enriched CORs also show great diversity in different cell lines (Fig. 3A). The numbers of CORs enriched by the same numbers of reads from multiple cell lines also show the great diversity between cell lines (Fig. 3B). It was reported that the relationship between COR and TSS plays a key role in transcription regulation [18]. The distance from the enriched CORs to TSSs was thus calculated for each cell line. The results indicate that the distance in cancer cell lines is larger than that in MRC-5 (Fig. 3C), indicating that the regulation mechanism in cancer cells is more complex than that in MRC-5. The distance distribution in HL7702 cell also shows a wide region as cancer cell lines (Fig. 3C), which may be caused by the feature of liver cell lines, in which large numbers of related genes need to be regulated and the regulation events are much more than in other normal cell lines. Further analysis of the relationship between CORs and TSSs shows that compared with MRC-5, more CORs are located in proximal promoters (TSS ± 1 kb) (Fig. 3D), indicating more expressed genes in cancer cell lines. The genomic feature of CORs in different cell lines shows that large percentage of CORs is located in intergenic regions (Fig. 3E), demonstrating that the distal regulation serves important roles in both cancer and normal cell lines. CORs provide binding sites for TFs, and the de novo motif analysis of cancer common CORs found several motifs similar to known TF motifs (Table S4). Two top enriched motifs are associated with NFY and Oct2 (Fig. 3F), which serves important roles during the cancer process [20, 21]. The GO term analysis of cancer common CORs reveals that these regions are all closely related with metabolic process, cell growth and nucleotide binding (Fig. 3G). These GO terms all play key roles in cancer. Several known motifs are also enriched by the cell-specific CORs (Figure S1). This result demonstrates that the transcription regulation process has high cell specificity.

## Characterization of COS of human normal and cancer liver cell lines

In order to investigate the chromatin state difference between normal and cancer cell lines derived from the same tissue, the COS of HepG2 and HL7702 were compared. The comparison of differential CORs shows that the chromatin regions with significant higher COL in HepG2 cell lines are much more than that in HL7702 cell line (Fig. 4A). This result demonstrates that the gene expression level in liver cancer cells were much higher than in normal liver cells due to the chromatin region opened. In order to validate the reliability of difference between two cell lines, Tn5 transposition cutting events were analyzed. The results indicate that there are high Tn5 transposition cut events in the HepG2-specific CORs, but no Tn5 transposition cut events in the HL7702-specific CORs (Fig. 4B), indicating the accuracy of the defined COS. The comparison of distance between COR and TSS reveals that compared with the CORs of HL7702, the CORs of HepG2 are all far from TSS (Fig. 4C), indicating that the distal regulation serves important roles in HepG2 cell. The motif analysis of differential CORs indicates that the AP-1 and HNF4A

motifs are enriched in HL7702 cell (Fig. 4D). AP-1 was reported playing roles in liver cell proliferation [22], and HNF4A is a known key TF in normal liver function [23, 24]. However, the JUNB and ATF4 motifs were enriched in HepG2 cell (Fig. 4D), suggesting that these TFs play regulatory roles in HepG2 cell by binding the HepG2-specific CORs (Table S5), consistent with previous reports [25]. The GO term analysis of the closest genes assigned to CORs reveals that the GO terms closely related to gene activity regulation are enriched in HL7702 cell, but those related to cancer process are enriched in HepG2 cell (Fig. 4E).

## Characterization of COS of human fibroblast and normal liver cell lines

Fibroblasts are the most commonly used primary somatic cell type for the generation of iPSCs [26]. In order to investigate the roles of the chromatin state during the induced process, the chromatin state of MRC-5 and HL7702 was compared. Identification of differential CORs in two cell lines reveals that more chromatin regions are significantly opened in MRC-5 cell compared with HL7702 (Fig. 5A). This result indicates that the fibroblast maintains high level of COS compared with liver cells. Moreover, high percentage of MRC-5-specific CORs is located in distal intergenic regions but most of HL7702-specific CORs are located in proximal promoter (Fig. 5B). This feature is also revealed by the distance of CORs to TSS (Fig. 5C). Because the OSKM TFs (Oct4, Sox2, Klf4 and c-Myc) and Nanog were commonly used to produce iPSCs, the motifs of these TFs were searched in CORs of two cell lines. The results reveals that the MRC-5 CORs contain much more Oct4, Sox2 and Nanog binding sites than the HL7702 CORs (Fig. 5E). However, the HL7702 CORs contain more Klf4 and c-Myc binding sites than the MRC-5 CORs (Figure S2). The motif analysis of CORs also indicates the great difference of gene regulation between the two cells (Table S6). The gene annotation of CORs-associated genes reveals that the GO terms related with normal cell metabolic process are enriched in HL7702 but several development related terms were enriched in MRC-5 (Fig. 5F). In order to further validate the COS difference between the two cells, several Oct4 target genes were selected and the chromatin state of these genes were observed with genome browser (Figure S3). The results demonstrate these genes have more CORs (i.e. SALP-seq peaks) in MRC-5 than in HL7702.

## Characterization of COS of different cervix cancer cell lines

Although HPV infection alone is not sufficient for cervical tumorigenesis even infected by high-risk type HPV [27], most high-risk type HPV infections are subclinical. In order to investigate whether the chromatin state of cervix cancer cells is influenced by the infection of different types of HPVs, the chromatin state of three cervix cancer cell lines (C-33A: HPV−; SiHa: HPV16+; HeLa: HPV18+) was characterized. The results reveal that the chromatin state of these cervix cancer cell lines are significantly different, especially between HPV positive and negative cell lines (Fig. 6A). This result demonstrates that the HPV infection can greatly change the COS of cervix cancer cell lines. The comparison shows that large percentage of CORs is located in the distal intergenic regions in both HPV positive and negative cell lines (Fig. 6B). Moreover, there are many CORs distributed in proximal promoters in HPV negative cell line C-33A (Fig. 6B). The calculation of distance of CORs to TSSs reveals that C-33A possesses more CORs around TSSs (Fig. 6C). The TF motif analysis of CORs indicates that different sets of TF motifs are enriched by

CORs of HPV positive and negative cell lines (Fig. 6D, Table S7), suggesting that the HPV infection may change the gene expression regulation of cervix cancer cells by changing TF-binding profiles. The comparison of CORs also reveals that there is some cell-specific CORs in two different HPVs-infected cervix cancer cells (Fig. 6E). The motif analysis reveals that these CORs possesses different TF motifs (Figure S4). The gene annotation of differential CORs reveals that the GO terms enriched by HPV positive cancer cells are all closely related with cancer process and cell metabolic activity (Fig. 6E).

## Identification of relationship between COS and mutation hotspot

Hotspot mutations exist in multiple types of cancers, and many cancer diagnosis and therapy methods were based on detections of these hotspot mutations. In order to investigate whether there is relationship between these hotspot mutations and COS, the cancer genes collected in MSK-IMPACT™ panel were analyzed. Normalized to a background generated by all human gene promoters, the COS of these cancer genes shows high diversity among different cancer cell lines (Fig. 7A). In different cell lines, the promoters of about 60% of these cancer genes show high COL (Fig. 7B). Using the mutation data from COSMIC, the mutation positions of these cancer genes were determined. The COL was calculated for each base from TSS to TES of a cancer gene. The COLs of mutation positions of a gene were then normalized with the average COL of the gene. The results indicate that there are cell-specific genes with high-COL mutations among all cancer types (Fig. 7C). The mutations with high COL shows high cell specificity; however, this kind of high-COL mutations occupy low percentage of all analyzed mutations (Fig. 7D). In order to further validate the cell-specific genes with high COL, the mutations associated with cell-specific genes collected by COSMIC of each tissue were analyzed and the frequency in each tissues was calculated. The results showed that the mutation frequency of cell-specific genes in the cell origin tissue is higher than other tissues (Figure S5). These results demonstrate that the hotspot mutations are related to the COS around the loci.

## Characterization of COS of mouse liver tissue

In order to investigate the COS difference between fresh tissue cells and cancer cell line, the COS of healthy BALB/c mouse liver and mouse liver cancer cell line (Hepa1-6) was detected with SALP-sEq. The whole-genome distribution of reads and CORs indicates that there is great difference between two samples (Fig. 8A). Some regions show great difference (such as chromosome X and 7) (Fig. 8A). Hepa1-6 possesses much more COR than healthy liver (Fig. 8A and B). Large percentage of CORs is located in 300 kb downstream region of genes in both healthy and cancer samples (Fig. 8C). However, the healthy mouse liver shows higher COL in proximal promoters than Hepa1-6 (Fig. 8C). More CORs are located in 1 kb regions around TSS in healthy liver than in Hepa1-6 (Fig. 8D). The de novo motif analysis reveals two different sets of TF motifs were enriched by the differential CORs of two samples (Fig. 8E, Table S8), suggesting differential gene regulation between normal tissue cell and cancer cell line. The gene annotation reveals that the GO terms are enriched by the genes assigned to Hepa1-6-specifc CORs (Fig. 8F). These GO terms are often r closely related to cancer process. Similar to GO terms enriched in

human normal live cell line, several GO terms related to cell metabolic processes are enriched by the mouse healthy liver-specific CORs (Fig. 8F).

# Discussion

Developments in modern genomics techniques has led to rapid progress in our understanding of the genetic basis of cancer, especially about the mutations in cancers. Almost all of detected tumors carry the acquired somatic mutations, but these mutations are not sufficient to cause cancer [3]. The epigenetic abnormal in cancer was verified recent years [28, 29]. For these reasons, epigenetic study of cancer has great potential to bridge the gap in understanding of the tumorigenic process.

The comparison of chromatin state between human normal cell lines and cancer cell lines shows significant difference. Totally 2 normal cell lines and 9 cancer cell lines were compared comprehensively, indicating the high COL in cancer cell lines (Fig. 2B), consistent with the results derived from cancer and para-carcinoma tissue [30]. As important regulation regions, CORs can provide binding sites for TFs. The difference of functional TFs in different cell lines is caused by a variety of COSs in different cell lines. For cancer-specific CORs, the enriched motifs are significantly different (Figure S1), demonstrating the cancer specificity. There are some regions with high COL among all cancer cell lines, which serves as binding sites for TFs in all cancer types. TFs, which are bound with common CORs, serve regulation function in various cancer types. This type of TFs has great potential in pan-cancer therapy. The TFs enriched from cell-specific CORs are closely related with the feature of origin tissues. For example, The motif of Fra-1 is enriched in A549 cell, this TF can enhance the rate of proliferation, motility, and invasion of A549 [31, 32] (Figure S1). The motif of STAT2 is enriched in HepG2 cell, this TF plays important function in inflammation and immunity during the liver cancer process (Figure S1) [33]. This cancer type-specific CORs and TFs provide potential targets for cancer type-specific diagnosis and therapy.

The process of tissue transformation from healthy state to cancer is dependent on not only the accumulation of mutations, but also the change of chromatin, such as the change of DNA methylation status [34, 35], histone modification [36, 37] and 3D structure of chromatin [38, 39]. The change of COS plays key roles in initiation and development of cancer. The comparison of COS of HepG2 and HL7702 cells demonstrates that the COL increases in canceration (Fig. 4A). Compared with normal liver cell lines, HepG2 shows high COL around TSSs (Fig. 4B), suggesting high gene activity in hepatocarcinoma. Additionally, more CORs far away from TSSs appear in HepG2 (Fig. 4C), suggesting that distal regulation plays important roles in hepatocarcinoma. In these differential CORs, different sets of TF motifs were found. The HNF4α motif is enriched from HL7702-specific CORs. This TF can regulate the differentiation of normal liver and plays an important role in maintainment of normal liver function [40]. The expression of this TF decreases in liver cancer cells and its transgenic over-expression was used to treat hepatocarcinoma by differentiation therapy [41]. As a result, no HNF4α motif was enriched in HepG2-specific CORs (Fig. 4D). However, the motifs of some hepatocarcinoma-related TFs were enriched in HepG2-specific CORs (Fig. 4D). These results demonstrate that the COS change provides a different TF binding profile in liver hepatocarcinoma, which promotes expression of a new set of genes helpful for

liver canceration (Fig. 4E). Therefore, the identified HepG2-specific CORs and TFs provide potential targets for hepatocarcinoma diagnosis and therapy.

Fibroblast was reprogrammed into to pluripotent stem cells (iPSCs) by introducing 4 transcription factors (TFs), Oct4, Sox2, Klf4 and c-Myc [42–44]. Since then, fibroblast was the most widely used cells to produce iPSCs. Despite its facile availability, the underlying reason may be more closely related to its high inducing efficiency. The high inducing efficiency should be dependent on its more open COS, which provides high accessibility of iPSC-inducing TFs to their target genes. This study reveals that the fibroblast MRC-5 possesses much more CORs than a normal liver cell HL7702. These CORs provide much more binding sites to three of OSKM TFs used to induce iPSCs, especially to the most important TF Oct4 (Fig. 5D). These results shed new insight on the mechanism why fibroblast was widely selected to induce iPSCs with OSKM factors. The mechanism may be used to further improve the induction efficiency of iPSCs.

The cervical cancer is the second largest cause of cancer deaths in women [27]. The HPV infection is the main cause of cervical cancer, especially the infection of two high risk HPV type, HPV16 and HPV18 [45]. This study reveals that there is great difference of CORs between the two HPV-infected and one HPV-free cervical cancer cell lines (Fig. 6A), indicating that HPV infection exerts great effect on COS of cervical epithelial cell and thus resulting in development-related gene expression profile that is closely related to cancer (Fig. 6E). Moreover, this study finds the distinct CORs and TF motifs resulted from infections of two different HPV subtypes (Fig. 6A, Figure S4). These results shed new insights on the molecular mechanisms of cervical canceration and provide new potential targets for treatment of this cancer.

Many cancers are closely related to some hotspot mutations of a particular set of genes. Detection of these mutation has already become essential for the clinical diagnosis and targeting therapy of these cancers. However, why these cancer-related genes become hotspot mutations in human genome remains unclear. This study explores the issue at the point view of COS because it is supposed that the opened chromatin provides the damaging factors such as radicals and deaminases more chance to contact DNA. The results reveal that each cancer cell line has a set of specific cancer genes with high-COL mutations (Fig. 7C). Additionally, the bases with high COL are positively related to their mutation frequency in COSMIC. This provides new insights on the mechanism of the formation of hotspot mutations.

Analyzing the COS of cancer tissues is important for the dissection of potential epigenetic causes of cancer. Therefore, 410 tissue samples derived form 23 types of tumors were analyzed with ATAC-seq for identifying the cancer-specific CORs of primary human cancers [7]. These CORs are helpful for cancer diagnosis and classification. This study characterizes the CORs of healthy BALB/c mouse liver tissue and compares them with those of mouse hepatocarcinoma cell Hepa1-6. The results reveal that the liver cancer cell harbors more CORs than normal liver cells (Fig. 8B), similar to the CORs distribution in human liver cancer and normal cell lines (Fig. 4A). These new cancer-derived CORs and their related TFs and genes shed new insights on the epigenetic causes of liver tumorigenesis, providing new potential markers and targets for liver cancer diagnosis and therapy.

The SALP-seq technique has recently been developed as a new method for constructing NGS library by using a new kind of single strand adaptor [9]. This technique has been successfully applied to constructing NGS library of DNA fragments sheared by any forms including sonication, enzyme digestion and tagmentation [9], and high degraded blood cell-free DNA [10]. By using barcoded Tn5 adaptors, SALP-seq is beneficial for the comparative characterization of COS of various cells in high-throughput format [9]. The comparative study of COS of various cells and tissues is helpful for dissecting the epigenetic causes of various biological and pathological processes such as development and tumorigenesis. This study comparatively characterized the CORs of two normal human cell lines (MRC-5 and HL7702), nine human cancer cell lines (A549, HepG2, HT-29, PANC-1, SKOV-3, GM12878, HeLa, SiHa and C-33A), one mouse cancer cell line (Hepa1-6), and healthy mouse liver tissue with SALP-sEq. This study thus made a set of systematic and useful COR profiles of human and mouse cells and tissues, which can be used to explore some interesting life science issues such as the potential epigenetic causes of various human cancers, inducing iPSCs with fibroblast, and cancer-related hotspot mutations. These comparative COR profiles shed new insights on the potential molecular mechanisms under many important life processes and provide new potential markers or targets for cancer diagnosis and therapy.

# Conclusions

In this study, the COS of multiple cell lines and mouse liver tissue were analyzed using SALP-sEq. The obtained comparative COR profiles shed new insights on the potential epigenetic mechanisms underlying several important biological processes, including (i) cancer cell lines possess higher COL than normal cell lines, indicating that more chromatin regions are opened in tumorigenesis; (ii) fibroblast has higher COL than other normal cell lines, allowing more effective iPSCs induction than other cells; (iii) HPV subtypes differently affect on the COS in cervix cancer, suggesting their differential pathogenesis; (iv) the cancer-related mutation hotspots are closely related to high CORs, indicating the contribution of COS to gene mutation.

# Abbreviations

ATAC-seq: transposase-accessible chromatin using sequencing; BTAs: barcoded Tn5 adaptors; COL: chromatin open level; COR: chromatin open region; COS: chromatin open state; COSMIC: catalogue of somatic mutations in cancer; FDA: Food and Drug Administration; GO: gene ontology; HPV: human papillomavirus; iPSCs: induced pluripotent stem cells; ME: mosaic end; OSKM: Oct4, Sox2, Klf4 and c-Myc; SALP: single strand adaptor library preparation; SSAs: single strand adaptors; TF: transcription factors; TPS: transposome assemble buffer; TSS: transcription start site.

# Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable

### Availability of data and materials

The raw reads data from SALP-seq are available at NCBI GEO with the accession number: GSE GSE 136391. Supplementary Data are available at Online.

### Competing interests

The authors declare that they have no competing interests.


### Funding

This work was supported by the National Natural Science Foundation of China (61971122).


### Authors' contributions

JW conceived the study. JLG, TL and WD performed cell culture. JW and SYZ performed SALP-seq experiemnts. JW performed all bioinformatics analysis and prepared figures. JW, JHL and JKW drafted the manuscript. JKW provided overall supervision of the study. JW, JLG, SYZ, TL, WD, and JHL contributed to the discussion. All author had read and approved the final manuscript.


### Acknowledgements

Not applicable.


# References


1.
Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
2.
Flavahan WA, Gaskell E, Bernstein BE. Epigenetic plasticity and the hallmarks of cancer. Science. 2017;357(6348):eaal2380.
3.
Taipale J. The chromatin of cancer. Science. 2018;362(6413):401–2.
4.
Zhu P, Fan Z. Cancer stem cells and tumorigenesis. Biophys Rep. 2018;4(4):178–88.
5.
Wang GG, Allis CD, Chi P. Chromatin remodeling and cancer, Part I: Covalent histone modifications. Trends Mol Med. 2007;13(9):363–72.
6.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Meth. 2013;10(12):1213.
7.
Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, Silva TC, Groeneveld C, Wong CK, Cho SW, et al. The chromatin accessibility landscape of primary human cancers. Science. 2018;362(6413):eaav1898.
8.
Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, Silva TC, Groeneveld C, Wong CK, Cho SW, et al. The chromatin accessibility landscape of primary human cancers. Science. 2018; 362(6413).
9.
Wu J, Dai W, Wu L, Wang J. SALP, a new single-stranded DNA library preparation method especially useful for the high-throughput characterization of chromatin openness states. BMC Genom. 2018;19(1):143.
10.
Wu J, Dai W, Wu L, Li W, Xia X, Wang J. Decoding genetic and epigenetic information embedded in cell free DNA with adapted SALP-sEq. Int J Cancer. 2019;145(9):2395–406.
11.
Sambrook J, Russell DW. Purification of nucleic acids by extraction with phenol: chloroform. Cold Spring Harb Protoc. 2006; 2006:pii: pdb.prot4455..
12.
Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.
13.
Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38(4):576–89.
14.
Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
15.
Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
16.
Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44.
17.
Yu G, Wang L-G, He Q-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics. 2015;31(14):2382–3.
18.

de la Torre-Ubieta L, Stein JL, Won H, Opland CK, Liang D, Lu D, Geschwind DH. The dynamic landscape of open chromatin during human cortical neurogenesis. Cell. 2018;172(1–2):289–304. e18.
19.
Mas G, Blanco E, Ballaré C, Sansó M, Spill YG, Hu D, Aoi Y, Le Dily F, Shilatifard A, Marti-Renom MA. Promoter bivalency favors an open chromatin architecture in embryonic stem cells. Nat Genet. 2018;50(10):1452–62.
20.
Gurtner A, Fuschi P, Martelli F, Manni I, Artuso S, Simonte G, Ambrosino V, Antonini A, Folgiero V, Falcioni R. Transcription factor NF-Y induces apoptosis in cells expressing wild-type p53 through E2F1 upregulation and p53 activation. Cancer Res. 2010;70(23):9711–20.
21.
Aguilar A. Kidney cancer: OCT2 demethylation cracks open oxaliplatin resistance. Nature Reviews Nephrology. 2016;12(10):581.
22.
Behrens A, Sibilia M, David JP, Möhle-Steinlein U, Tronche F, Schütz G, Wagner EF. Impaired postnatal hepatocyte proliferation and liver regeneration in mice lacking c-jun in the liver. The EMBO journal. 2002;21(7):1782–90.
23.
Si-Tayeb K, Lemaigre FP, Duncan SA. Organogenesis and development of the liver. Dev Cell. 2010;18(2):175–89.
24.
Lee D-H, Park JO, Kim T-S, Kim S-K, Kim T-h, Kim M-c, Park GS, Kim J-H, Kuninaka S, Olson EN. LATS-YAP/TAZ controls lineage specification by regulating TGFβ signaling and Hnf4α expression during liver development. Nat Commun. 2016;7:11961.
25.
Min L, Ji Y, Bakiri L, Qiu Z, Cen J, Chen X, Chen L, Scheuch H, Zheng H, Qin L. Liver cancer initiation is controlled by AP-1 through SIRT6-dependent inhibition of survivin. Nat Cell Biol. 2012;14(11):1203–11.
26.
Raab S, Klingenstein M, Liebau S, Linta L. A comparative view on human somatic cell sources for iPSC generation. Stem cells international. 2014; 2014:768391.
27.
Roden R, Wu T-C. How will HPV vaccines affect cervical cancer? Nat Rev Cancer. 2006;6(10):753.
28.
Jones PA, Issa J-PJ, Baylin S. Targeting the cancer epigenome for therapy. Nat Rev Genet. 2016;17(10):630.
29.
Bennett RL, Licht JD. Targeting epigenetics in cancer. Annu Rev Pharmacol Toxicol. 2018;58:187–207.
30.
Britton E, Rogerson C, Mehta S, Li Y, Li X, Fitzgerald RC, Ang YS, Sharrocks AD. Open chromatin profiling identifies AP1 as a transcriptional regulator in oesophageal adenocarcinoma. PLoS Genet.

2017;13(8):e1006879.
31.
Rajasekaran S, Vaz M, Reddy SP. Fra-1/AP-1 transcription factor negatively regulates pulmonary fibrosis in vivo. PLoS One. 2012;7(7):e41611.
32.
Adiseshaiah P, Lindner DJ, Kalvakolanu DV, Reddy SP. FRA-1 proto-oncogene induces lung epithelial cell invasion and anchorage-independent growth in vitro, but is insufficient to promote tumor growth in vivo. Cancer Res. 2007;67(13):6204–11.
33.
Yu H, Pardoll D, Jove R. STATs in cancer inflammation and immunity: a leading role for STAT3. Nat Rev Cancer. 2009;9(11):798–809.
34.
Klutstein M, Nejman D, Greenfield R, Cedar H. DNA methylation in cancer and aging. Cancer Res. 2016;76(12):3446–50.
35.
Teschendorff AE, Gao Y, Jones A, Ruebner M, Beckmann MW, Wachter DL, Fasching PA, Widschwendter M. DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. Nat Commun. 2016;7:10478.
36.
Audia JE, Campbell RM. Histone modifications and cancer. Cold Spring Harb Perspect Biol. 2016;8(4):a019521.
37.
Kim KH, Roberts CW. Targeting EZH2 in cancer. Nat Med. 2016;22(2):128–34.
38.
Kim K, Jang K, Yang W, Choi E-Y, Park S-M, Bae M, Kim Y-J, Choi JK. Chromatin structure–based prediction of recurrent noncoding mutations in cancer. Nat Genet. 2016;48(11):1321–6.
39.
Morgan MA, Shilatifard A. Chromatin signatures of cancer. Genes Dev. 2015;29(3):238–49.
40.
Parviz F, Matullo C, Garrison WD, Savatski L, Adamson JW, Ning G, Kaestner KH, Rossi JM, Zaret KS, Duncan SA. Hepatocyte nuclear factor 4α controls the development of a hepatic epithelium and liver morphogenesis. Nat Genet. 2003;34(3):292–6.
41.
Cheng Z, He Z, Cai Y, Zhang C, Fu G, Li H, Sun W, Liu C, Cui X, Ning B. Conversion of hepatoma cells to hepatocyte-like cells by defined hepatocyte nuclear factors. Cell Res. 2019;29(2):124.
42.
Zhao R, Daley GQ. From fibroblasts to iPS cells: induced pluripotency by defined factors. J Cell Biochem. 2008;105(4):949–55.
43.

Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell. 2007;131(5):861–72.
44.
Mai T, Markov GJ, Brady JJ, Palla A, Zeng H, Sebastiano V, Blau HM. NKX3-1 is required for induced pluripotent stem cell reprogramming and can replace OCT4 in mouse and human iPSC induction. Nat Cell Biol. 2018;20(8):900–8.
45.
Liu P, Chen M, Liu Y, Qi LS, Ding S. CRISPR-based chromatin remodeling of the endogenous Oct4 or Sox2 locus enables reprogramming to pluripotency. Cell Stem Cell. 2018;22(2):252–61. e4.

# Figures

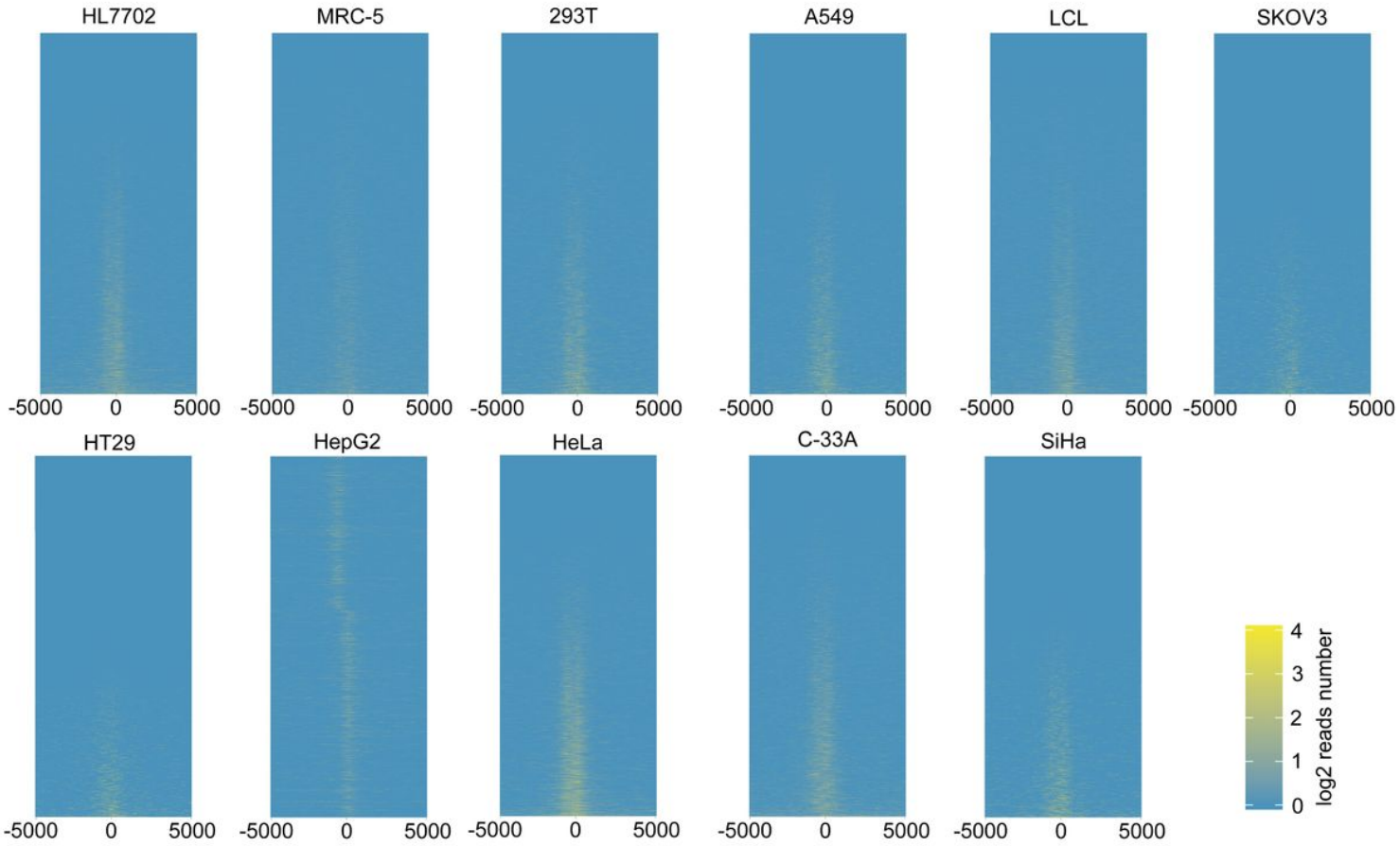
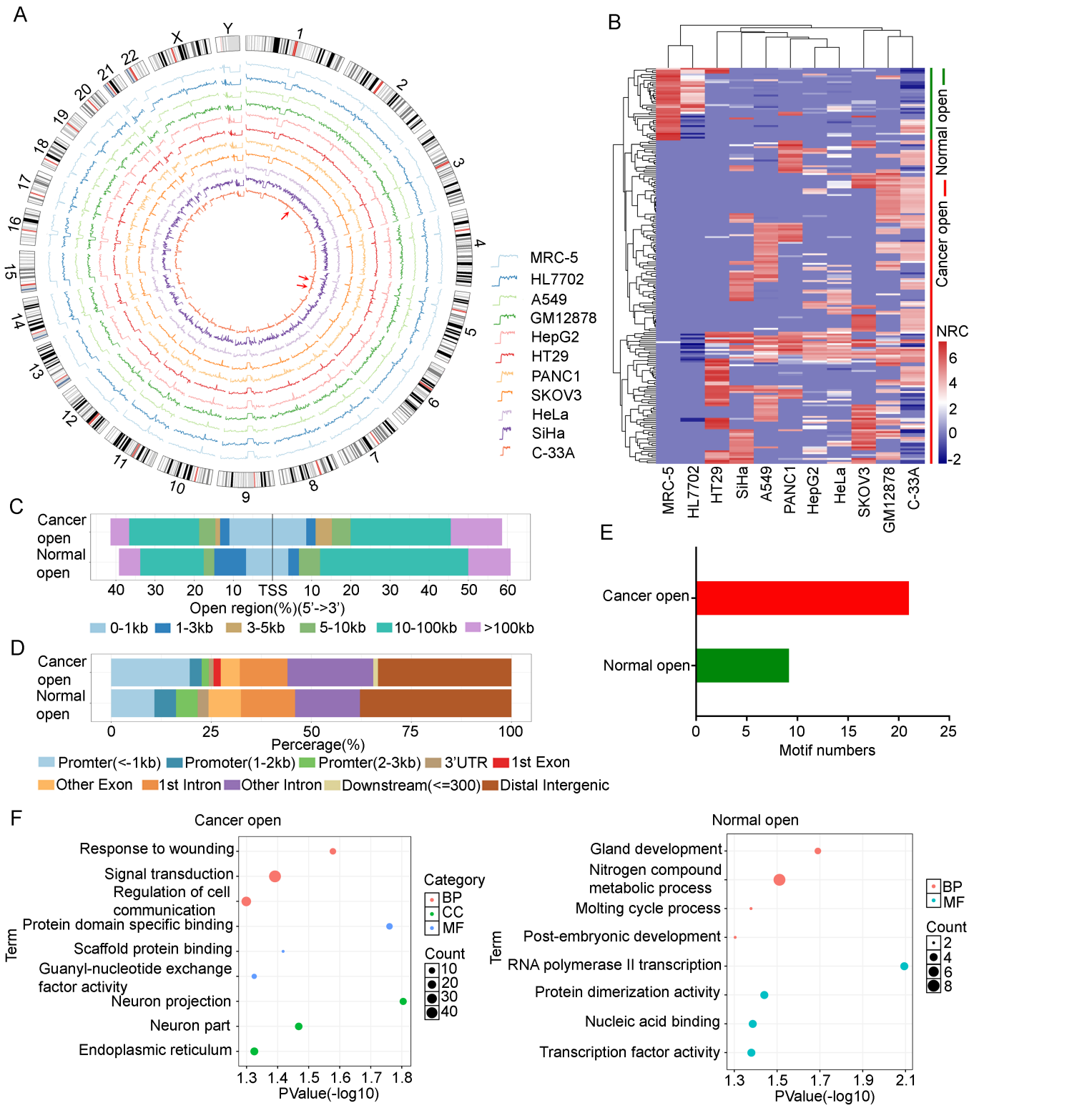



**Figure 1**

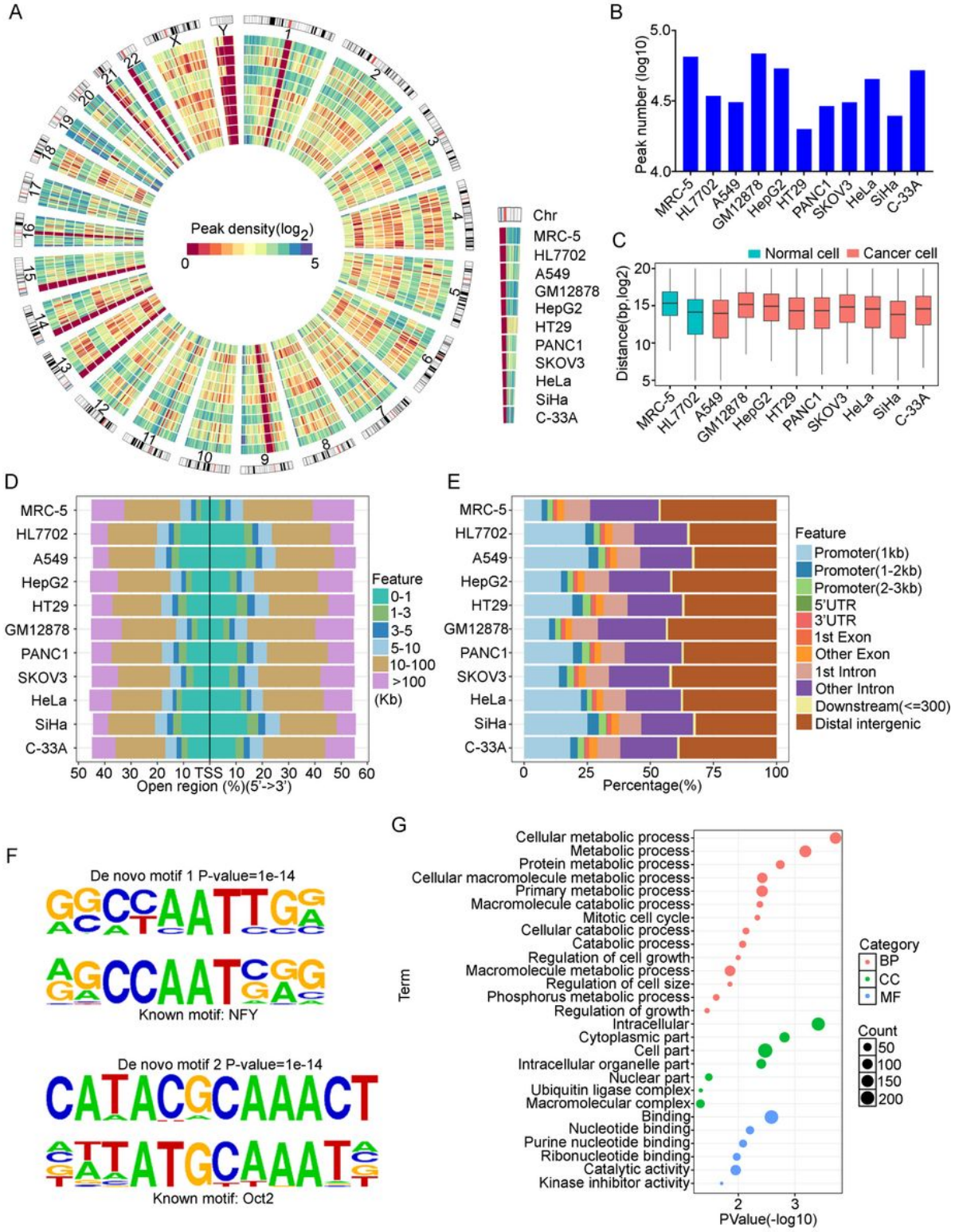
The reads distribution around TSS. The reads density of ± 5 kb regions around TSS were calculated and normalized using 100-bp windows.




## Figure 2

The COS comparison between human normal and cancer cell lines. (A) The reads distribution of different cell lines in the whole genome. The reads density was calculated with 1-Mb window through the whole genome. (B) The chromatin regions with significantly different open state between normal and cancer cell lines. NRC: normalized reads count. (C) The distribution of distance between TSS and the specific CORs of normal and cancer cell lines. (D) The genomic feature of normal and cancer cell line-specific CORs. (E)

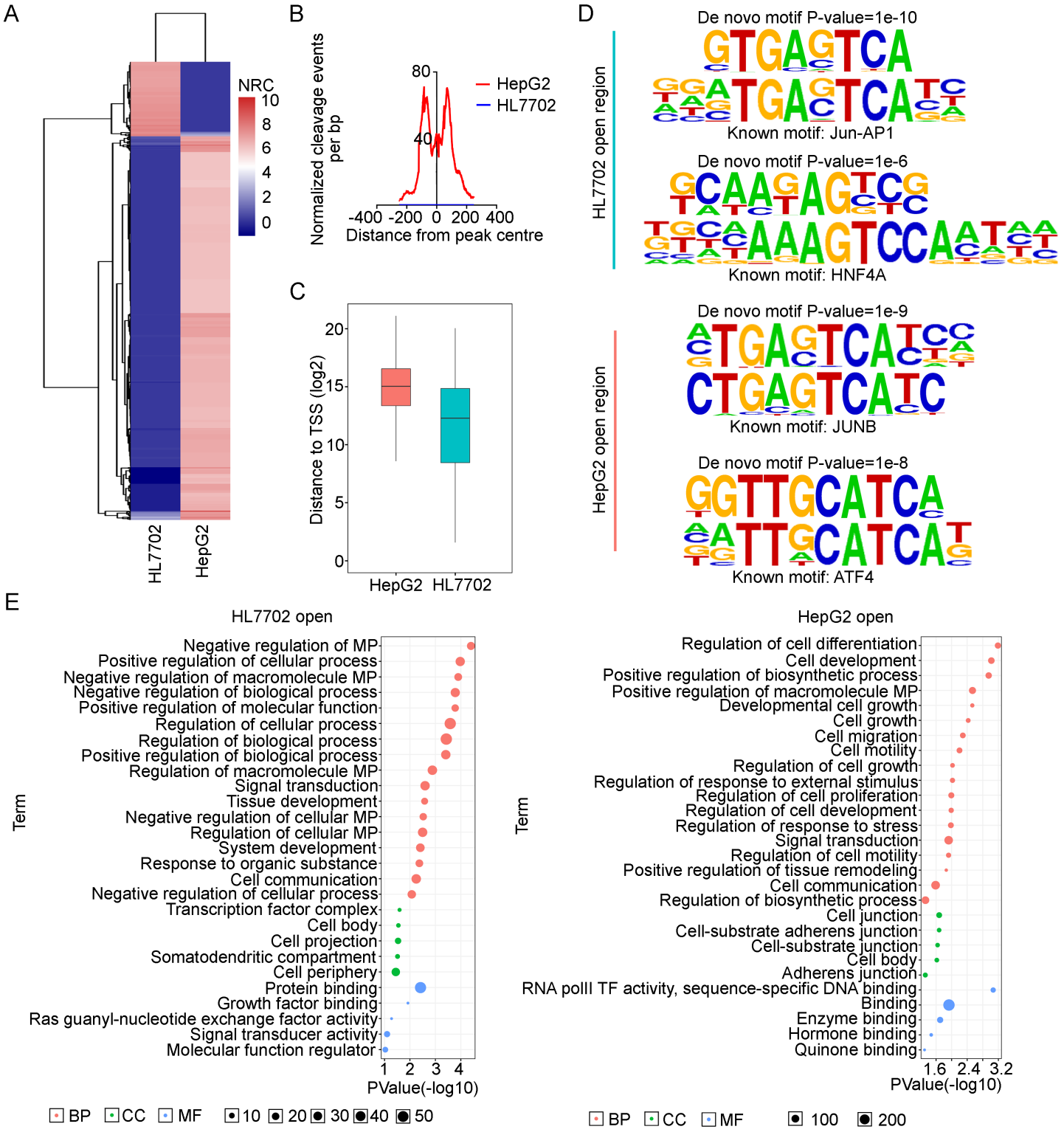
The change of TFs motif numbers. The lost or acquired TF numbers caused by the COS change were calculated and compared. (F) The GO term analysis of genes related to the normal or cancer-specific CORs. BP: biological process, CC: cellular component, MF: molecular function.




## Figure 3

The COS comparison between different cancer cell lines. (A) The reads distribution of different cancer cell lines in the whole genome. The reads density was calculated with 1-Mb window through the whole

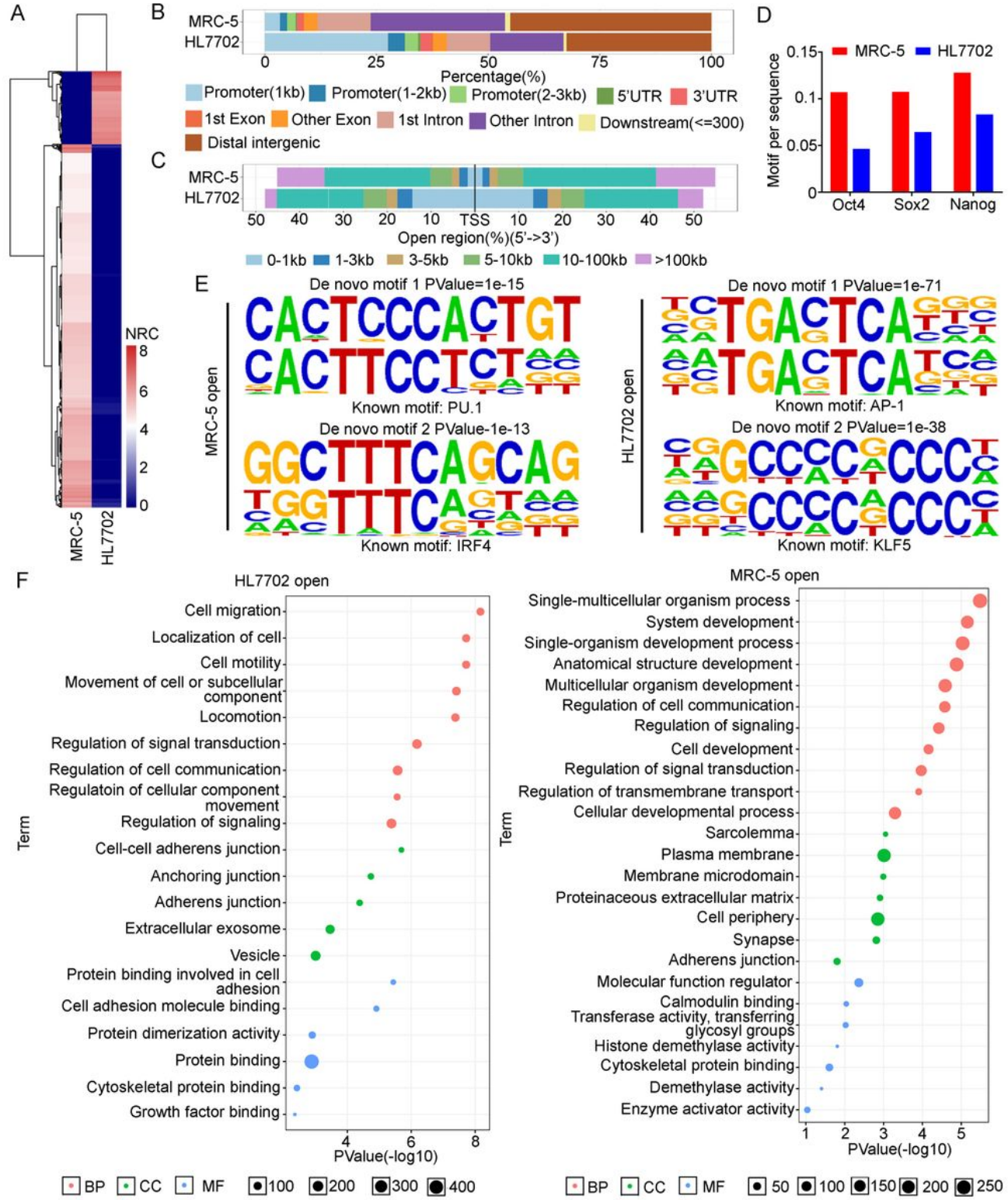
genome. (B) The number of CORs in different cancer cell lines. 106 reads were random selected from each cell lines and CORs were enriched and compared. (C) The comparison of distance between CORs and TSS in each cell line. (D) The distribution of distance between TSS and cell type specific CORs. The distance between TSS and CORs in each cell lines were calculated and compared. (E) The genomic feature of cell type-specific CORs. The cell type-specific CORs were annotated with genomic features and the percentage of each feature was calculated. (F) The top 2 de novo enriched motifs from cancer common CORs. The known motifs with high similarity were also shown. (G) The GO term analysis of genes related to common CORs in cancer cell lines. BP: biological process, CC: cellular component, MF: molecular function.




## Figure 4

The COS comparison between human normal and cancer liver cell lines. (A) The differential CORs between HL7702 and HepG2 cell line. NRC: normalized reads count. (B) The comparison of Tn5 cleavage events in different cell lines. (C) The distance between CORs and TSS in each cell lines. (D) Selected motifs enriched from cell specific CORs. The known motifs with high similarity were also shown. (E) The

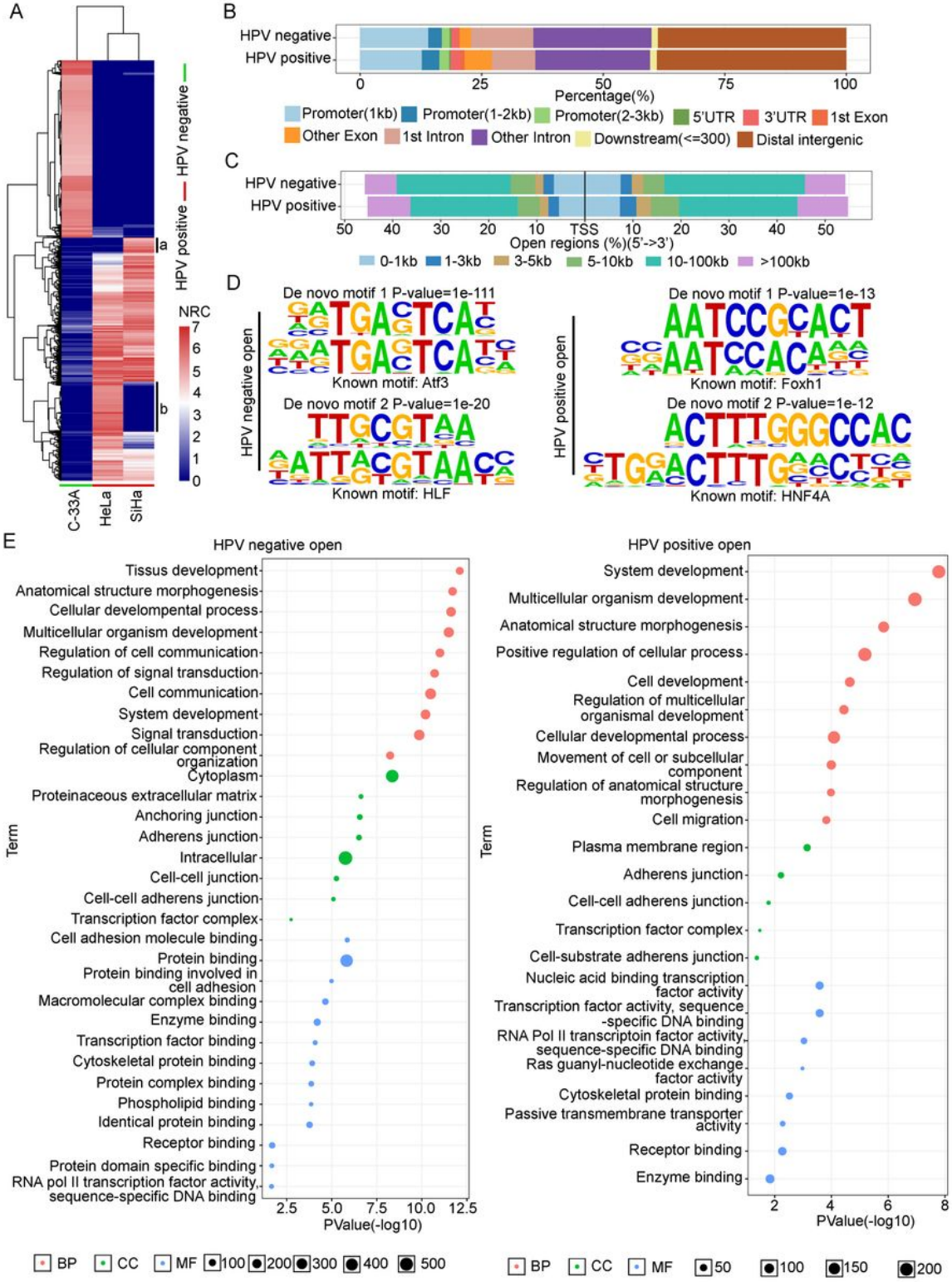
GO term analysis of genes related to cell specific CORs. BP: biological process, CC: cellular component, MF: molecular function.




## Figure 5

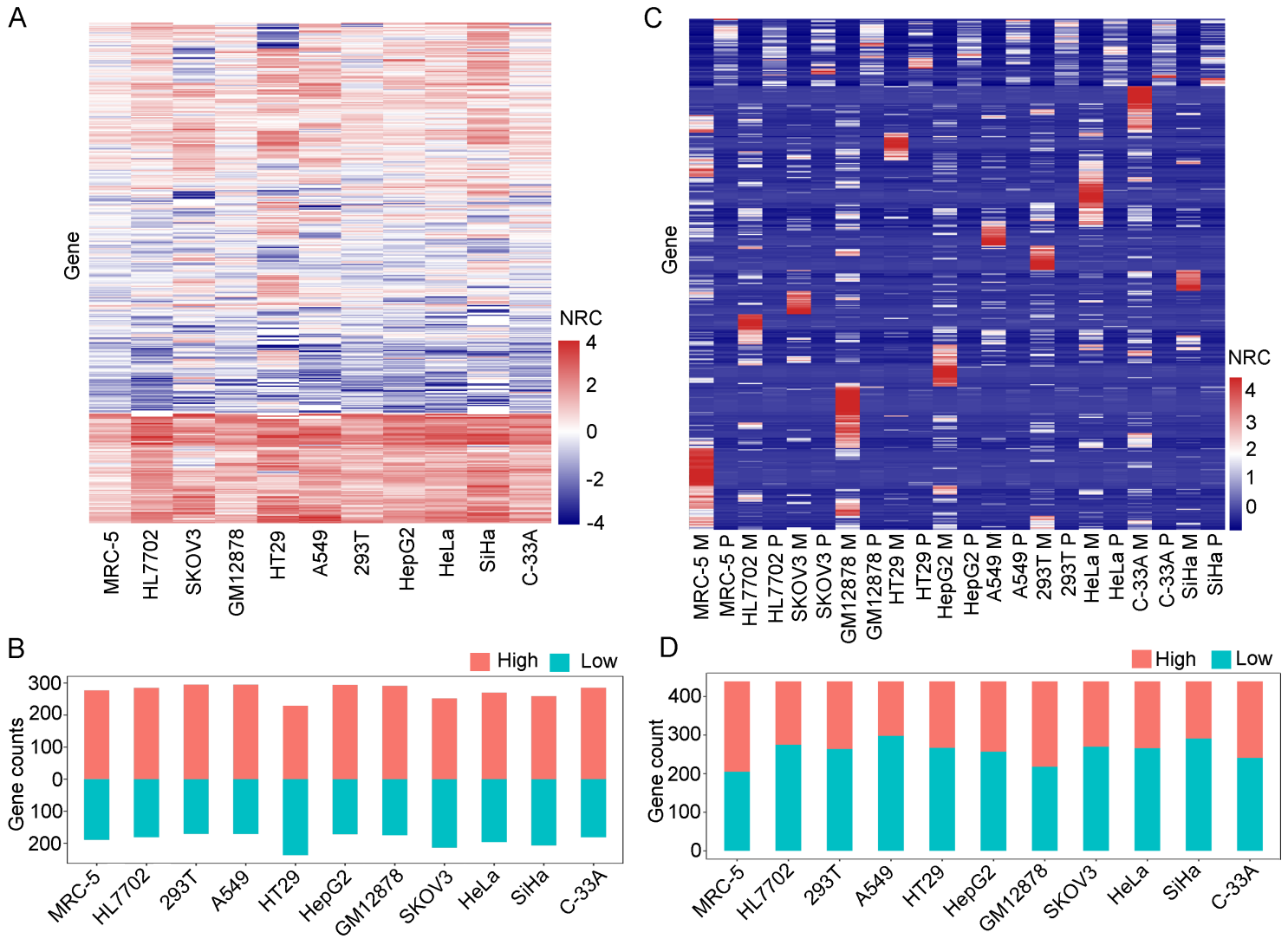
The COS comparison between fibroblast and normal liver cell lines. (A) The differential CORs between HL7702 and MRC-5 cell line. NRC: normalized reads count. (B) The genomic feature of CORs of two cell lines. The CORs were annotated with genomic features and the percentage of each feature was

calculated. (C) The distribution of distance between TSS and CORs in two cell lines. The distance between TSS and CORs in each cell lines was calculated and compared. (D) The frequency of three TF motifs located in the MRC-5- and HL7702-specific CORs was compared. (E) The top 2 motifs enriched from cell specific CORs. The known motifs with high similarity were also shown. (F) The GO term analysis of genes related to cell specific CORs. BP: biological process, CC: cellular component, MF: molecular function.



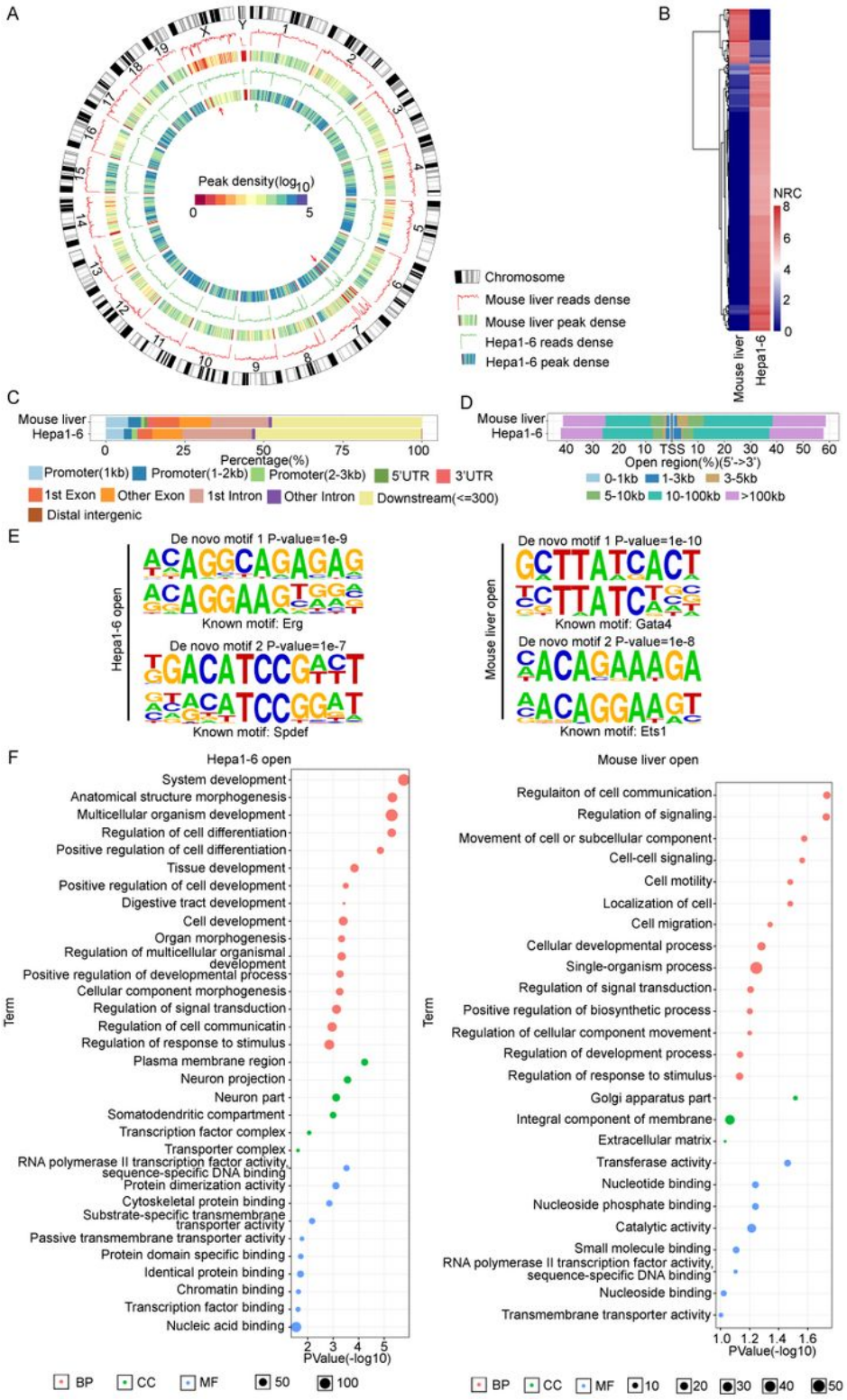

## Figure 6

The COS comparison between different cervix cancer cell lines. (A) The differential CORs between the HPV positive and negative cervix cancer cell lines. a: SiHa-specific CORs; b: HeLa-specific CORs; NRC: normalized reads count. (B) The genomic feature comparison of CORs of HPV positive and negative cell lines. (C) The distribution of distance between TSS and CORs in the HPV positive and negative cell lines. (D) The top 2 motifs enriched from the HPV positive and negative cell lines-specific CORs. (E) The GO term analysis of genes related to HPV infection status specific CORs. BP: biological process, CC: cellular component, MF: molecular function.




## Figure 7

The COS analysis of mutation hotspot. (A) The COS of regions ± 50 kb around the TSSs of genes collected by the MSK-IMPACT™ panel. NRC: normalized reads count. (B) The comparison of chromatin state between regions ± 50 kb around TSSs and background. High: COL in defined regions higher than background; Low: COL in defined regions lower than background; (C) The COS of mutation loci. The COS of gene mutation loci collected by COSMIC was analyzed and compared with background. M: reads count in mutation base; P: average reads count per base; NRC: normalized reads count. (D) The change of COS in the mutations of genes collected by the MSK-IMPACT™ panel. High: the COL in mutation base higher than background; Low: the COL in mutation base lower than background.




## Figure 8

The COS comparison between mouse healthy liver tissue and liver cancer cell line. (A) The comparison of reads density and COR distribution between healthy liver tissue and Hepa1-6. Red arrow: the regions with different reads distribution trend between healthy liver tissue and Hepa1-6; Green arrow: the regions with same reads distribution trend between healthy liver tissue and Hepa1-6. (B) The differential CORs between healthy liver tissue and Hepa1-6. NRC: normalized reads count. (C) The genomic feature

comparison of CORs of healthy liver tissue and Hepa1-6. (D) The distribution of distance between TSS and CORs of healthy liver tissue and Hepa1-6. (E) The top 2 motifs enriched from healthy liver tissue and liver cancer cell line specific CORs. (F) The GO term analysis of genes related to healthy liver tissue- and Hepa1-6-specific CORs. BP: biological process, CC: cellular component, MF: molecular function.

## Supplementary Files

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

- SupplementaryFilesBMCGenomics.pdf