Title:

Multi-omics analysis of chromatin accessibility and interactions with transcriptome by HiCAR

Author list:

Xiaolin Wei[1,2*], Yu Xiang[1,2*], Ruocheng Shan[3], Derek T. Peters[1,2], Tongyu Sun[1,2], Xin Lin[1,2], Wei Li[3], Yarui Diao[1,2,4,#]

Affiliation:

1. Department of Cell Biology, Duke University Medical Center, Durham, NC 27708
2. Regeneration Next Initiative, Duke University Medical Center, Durham, NC 27708
3. Center for Genetic Medicine Research, Center for Cancer and Immunology Research at Children's National Medical Center, Washington, D.C., 20010
4. Department of Orthopedic Surgery, Duke University Medical Center, Durham, NC 27708

* These authors contributed equally to this work.

[#] Corresponding author: yarui.diao@duke.edu

Abstract:

The long-range interactions of *cis*-regulatory elements (cREs) play a central role in regulating the spatial-temporal gene expression program of multi-cellular organism. cREs are characterized by the presence of accessible (or "open) chromatin, which can be identified at genome-wide scale with assays such as ATAC-seq, DHS-seq, and FAIRE-seq. However, it remains technically challenging to comprehensively identify the long-range physical interactions that occur between cREs, especially in a cost effective manner using low-input samples. Here, we report HiCAR (***Hi***gh-throughput ***C***hromosome conformation capture on ***A***ccessible DNA with m***R***NA-seq co-assay), a method that enables simultaneous assessment of *cis*-regulatory chromatin interactions and chromatin accessibility, as well as evaluation of the transcriptome, which represents the functional output of chromatin structure and accessibility. Unlike immunoprecipitation-based methods such as HiChIP, PLAC-seq, and ChIA-PET, HiCAR does not require target-specific antibodies and thus can comprehensively capture the *cis*-regulatory chromatin contacts anchored at accessible regulatory DNA regions and associated with diverse epigenetic modifications and transcription factor binding. Compared to Trac-looping, another method designed to capture interactions between accessible chromatin regions, HiCAR produced a 17-fold greater yield of informative long-range *cis*- reads at a similar sequencing depth and required 1,000-fold fewer cells as input. Applying HiCAR to H1 human embryonic stem cells (hESCs) revealed 46,792 *cis*-regulatory chromatin interactions at 5kb resolution. Interestingly, we found that epigenetically poised, bivalent, and repressed cREs exhibit comparable spatial interaction activity to those transcriptionally activated cREs. Using machine learning approaches, we predicated 22 epigenome features that are potentially important for the spatial interaction activity of cREs in H1 hESC. Lastly, we also identified long-range *cis*-regulatory chromatin interactions in GM12878 and mouse embryonic stem cells with HiCAR. Our results demonstrate that HiCAR is a robust and cost-effective multi-omics assay, which is broadly applicable for simultaneous analysis of genome architecture, chromatin accessibility, and the transcriptome using low-input samples.

Main Text:

Introduction

Cis-regulatory elements (cREs), such as enhancers, promoters, insulators and silencers, play a critical role in regulating spatial-temporal gene expression in development and diseases[1–3]. CREs are characterized by the presence of "open" or accessible chromatin that is depleted of packaging nucleosome particles, making way for the binding of Transcription Factors (TFs) and a variety of epigenetic remodelers. These accessible chromatin regions can be identified by Assay for Transposase-Accessible Chromatin using sequencing (ATAC-Seq)[4], DNase-Seq[5], and FAIRE-Seq[6] (Formaldehyde-Assisted Isolation of Regulatory Elements). cREs can form dynamic high-order chromatin interactions to precisely control the expression of distal target genes. The development of chromosome conformation capture (3C)-based technologies has greatly improved our understanding of the principles of high-order chromatin organization, and revealed how dynamic chromatin looping affects gene expression in a cell type specific manner. Among these technologies, Hi-C has been widely used to measure genome-wide chromatin architecture [7,8], but requires extremely deep sequencing depth (several billion reads) to resolve chromatin interactions at 5- to 10-kilobase resolution. To reduce the sequencing costs, alternative methods such as ChIA-PET, HiChIP, PLAC-seq and Capture-C have been developed [9–14]. However, these methods rely on ChIP-grade antibody (ChIA-PET, HiChIP and PLAC-seq) or pre-designed capture probes (Capture-C) to enrich a subset of chromatin interactions associated with specific proteins, histone modifications, or targeted genome regions. More recently, Trac-looping and Ocean-C have been developed to analyze interactions among accessible chromatin regions, independent of ChIP antibodies or capture probes [15,16]. Although these two methods do not require targeted immunoprecipitation or DNA pulldown, they require a large number of cells and yield a relatively low proportion of long-range *cis* reads, preventing their application to low input materials, such as clinical samples and primary tissues. Moreover, none of the methods described above enable simultaneous assessment of the transcriptome from the same biological sample, which is the key functional output of genome architecture and chromatin accessibility. Therefore, a robust, sensitive, and cost effective method is urgently needed to enable a comprehensive analysis of chromatin structure and function, including transcription output, using low-input materials.

Here, we introduce a new method called HiCAR, which allows genome-wide profiling of long-range *cis*-regulatory chromatin interactions, chromatin accessibility, and gene expression using the same input sample. By leveraging principles of *in situ* Hi-C, ATAC-seq, and SMART-seq2 methods, HiCAR requires only ~100,000 cells as input and avoids many potentially nucleic acid loss-prone steps, such as adaptor

ligation and biotin-pull down. With similar sequencing depth, HiCAR outperforms Trac-looping[15] by generating ~17-fold more (18.3% versus 1.1%) long-range (>20kb) *cis*- paired-end tags (*cis*-PET), even when starting from 1,000-fold fewer cells ($1x10^5$ versus $1x10^8$ million). As a multi-omics co-assay, HiCAR also yields high-quality chromatin accessibility and transcriptome data from the same low-input starting material. Applying HiCAR to H1 human embryonic stem cells (hESCs), we generated a comprehensive map of *cis*-regulatory chromatin contacts at 5kb resolution. Additionally, we provide a user-friendly data processing pipeline called HiCARTools (https://github.com/diao-lab/HiCARTools) for HiCAR data processing.

Results:

Principle of HiCAR

As a proof-of-principle, we performed HiCAR on H1 hESCs, because of the rich public genomic datasets available for this cell line that could be used to benchmark our approach (Table S1, list of public datasets used in this study) [2,17]. First, ~100,000 cross-linked H1 cells were treated with Tn5 transposase assembled with an engineered DNA adaptor (Table S2). The Tn5 adaptor contains a Mosaic End (ME) sequence for Tn5 recognition[18] as well as a single-stranded flanking sequence that can be ligated to the CviQI-digested DNA fragment with a splint oligo (Fig 1A, Table S2). Next, restriction enzyme digestion was performed using the 4-base cutter CviQI, followed by *in situ* proximity ligation to ligate Tn5 adaptor to the proximal genomic DNA. After *in situ* ligation, cross-links were reversed and the DNA was purified, digested by another 4-base cutter NlaIII, and circularized by re-ligation. The circularized DNA was used for PCR amplification to generate HiCAR DNA libraries for Next-Generation-Sequencing (NGS). Forward and reverse PCR primers (Table S2) were then used for library amplification, which anneal to the ME sequence and splint oligo sequence, respectively. Therefore, the resulting amplified chimeric DNA fragment contains one end derived from the CviQI digested genomic DNA (captured by Read 1 of each paired-end sequence, Fig 1A), and one end derived from the Tn5-tagmented open chromatin sequence (captured by Read 2 of each paired-end sequence, Fig 1A). Additionally, polyA RNAs from the cytoplasm and nucleoplasm were collected during the procedure (Fig 1A) and subjected to RNA-seq library preparation using a protocol modified from SMART-seq2[19] (detailed in materials and methods).

HiCAR libraries were made from 3 biological replicates of H1 hESC and each library was sequenced to a depth of ~300 million pair-end raw reads (Table S3). We first examined the enrichment of HiCAR reads around open chromatin regions defined by H1 hESC ATAC-seq data generated by the 4DN consortium[20]. We separately analyzed Read 1 (R1) and Read 2 (R2) of the HiCAR DNA library, and used the publicly available H1 hESC *in situ* Hi-C data from the 4DN consortium[20] (Table S1) as a reference

dataset without targeted enrichment. As expected, HiCAR R2 reads were highly enriched at the H1 hESC ATAC-seq peaks (Fig 1B), while the R1 reads and *in situ* Hi-C reads show no enrichment (Fig 1B). This result confirmed that HiCAR successfully captured and enriched the interactions between open chromatin regions (R2) and other genomic regions (R1). We refer to these interactions below as "open-to-all" interactions. This is different from Trac-looping[15], a different method capturing "open-to-open" interactions between pairs of open chromatin regions. Next, we compared the enrichment efficiency of HiCAR to that of Trac-looping and Ocean-C, two methods recently developed for mapping long-range interactions anchored at open chromatin regions [15,16]. Because HiCAR, Trac-looping and Ocean-C experiments were performed in different cell lines, we decided to assess open chromatin enrichment efficiency of each method by examining transcription start site (TSS) signal enrichment, a metric widely used as a quality control standard to compare signal-to-noise ratios of ATAC-seq data across different cell types[21]. We found that both HiCAR and Trac-looping reads show high TSS signal enrichment (Fig 1C, log2 fold change = 1.02 and 0.84, respectively, Wilcoxon test, both $p < 2.2e\text{-}16$), while Ocean-C reads show significant but much weaker enriched signal on TSS (Fig 1C, log2 fold change = 0.30, Wilcoxon test $p < 2.2e\text{-}16$). We carried out a similar analysis by comparing HiCAR data to the public DNase Hi-C data (Fig S1A) [12,13,20,22]. In the previous DNase Hi-C study, the authors concluded that DNase Hi-C does not introduce open chromatin bias into the chromatin contact matrix [22]. Consistent with their results, we found that the DNase Hi-C reads are indeed not enriched on TSS regions (Fig S1A, brown line).

We also performed a similar analysis to compare HiCAR data to the public HiChIP and PLAC-seq data (Fig S1A) [12,13,20,22]. As expected, we found that the signal enrichment of HiChIP and PLAC-seq at *cis*-regulatory sequences depends on the antibody used for chromatin immunoprecipitation (ChIP). For example, H3K4me3 modification is the mark of promoters [23], and the sequencing reads from H3K4me3 PLAC-seq data exhibited significant enrichment around TSS regions (Fig S1A, black line), whereas H3K4me1 (enhancer mark) HiChIP reads showed no enrichment on TSS (Fig S1A, purple line). Since open chromatin regions are bound by multiple TF and histone marks[24], we expected HiCAR reads could enrich comprehensive epigenome signatures associated with cis-regulatory sequences. Indeed, we found that the HiCAR R2 reads, but not R1 reads, are highly enriched on H1 hESC H3K27ac, H3K3me1, H3K4me3, H3K27me3, RAD21, CTCF, NANOG, SOX2, and POU5F1 ChIP-seq peaks (Fig S1B). Our results clearly illustrated that while HiChIP and PLAC-seq only enrich the reads that are bound by the specific ChIP antibody, HiCAR effectively enriches a broader array of readds anchored at open chromatin regions (Fig 1C) and associated with a spectrum of epigenetic modifications and transcription factor binding (Fig S1A).

Given the relative low TSS-enrichment efficiency of Ocean-C (Fig 1C), we excluded Ocean-C from the following analysis and only compared HiCAR data to the public Trac-looping[15] data. We included one *in situ* Hi-C library that was generated by the 4DN consortium[25] and sequenced at similar depth (Fig 1D, 373 million raw reads) as control data without targeted enrichment. Notably, HiCAR requires much less input material (100 thousand cells) than Trac-looping (100 million cells) and *in situ* Hi-C (2-5 million cells), while producing 4.15-fold more uniquely mapped PETs than Trac-looping (Fig 1D, 55.6% versus 13.4%). More importantly, compared to Trac-looping, HiCAR captured about 17-fold (18.3% versus 1.1%, blue bars in Fig 1E) more long-range (> 20kb) *cis*-PET, which are the informative reads to identify long-range chromatin interactions. Furthermore, we examined the genome-wide average contact frequency captured by HiCAR, *in situ* Hi-C, and Trac-looping. We found that HiCAR and *in situ* Hi-C show similar decay rate in capturing long-range chromatin interactions with increased linear genomic distance (Fig 1F), while Trac-looping captures more short-rage (less than 7kb) chromatin contacts but fewer long-range interactions (Fig 1F). Overall, we concluded that HiCAR outperforms Trac-looping and allows for efficient and comprehensive capture of *cis*-regulatory chromatin contacts independent of antibody immunoprecipitation using low-input cells.

HiCAR faithfully recapitulates the key features of high-order chromatin organization.

Next, we asked if HiCAR could identify the key features of genome architecture. To probe this question, we used the deeply sequenced (total of 6.2 billion raw reads, generated by 4DN consortium [20]) *in situ* Hi-C data generated from H1 hESCs as a “gold standard” in our analysis. We first visually examined the global chromatin contact matrix (sequencing depth normalized) of HiCAR and *in situ* Hi-C (Fig 2A). We found that HiCAR generated chromatin contact matrix highly similar to that of *in situ* Hi-C at chromosomes, compartments, topological associated domains (TADs), and 10kb-bin resolutions (Fig 2A, left to right). To further quantify the similarity of the HiCAR and Hi-C contact matrices, we used HiCRep[26] to compute the stratum-adjusted correlation coefficient (SCC) among three HiCAR replicates and the *in situ* Hi-C data[20]. At the genome-wide scale, we found that the three biological replicates of HiCAR library were highly reproducible (Fig S1C, SCC=0.98), and HiCAR captured a chromatin interaction pattern similar to the deeply sequenced *in situ* Hi-C dataset (Fig S1C, SCC = 0.90, 0.89, 0.89). Further analysis revealed that the A/B compartment PC1 score, insulation score, and directionality index calculated from the HiCAR and *in situ* Hi-C data are well correlated with each other (Fig 2B).

Notably, the HiCAR contact matrix, built from 488 million uniquely mapped PETs, revealed as much, if not greater, details on chromatin interactions compared to the deeply sequenced (2.53 billion uniquely mapped PETs) *in situ* Hi-C data (Fig 2A). Next, we asked whether HiCAR can enrich the long

range *cis*-PETs anchored on cREs. To probe this question, we collected the open chromatin peaks and ChIP-seq peaks of H1 hESC identified by ATAC-seq and ChIP-seq datasets (including CTCF, H3K27ac, H3K4me1, H3K4me3, and H3K27me3 ChIP-seq), and set these peaks as the center of the sub-chromatin contact matrix expanding +/- 250kb window from each peak center. Next, we aggregated the PET signal (sequencing depth normalized) from all the sub-chromatin contact matrices. Interestingly, we found that the aggregated HiCAR PET signal showed a clear stripe pattern extending from the peak centers of all the examined epigenetic features (Fig 2C, top tracks). By contrast, the stripe patterns of PET signal from the aggregated Hi-C contact matrices are much weaker (Fig 2C, bottom track). Compared to *in situ* Hi-C, we concluded that HiCAR can effectively enrich long-range *cis*-PETs anchored at *cis*-regulatory sequences and associated with diverse histone modification and TF binding.

HiCAR yields high-quality chromatin accessibility and transcriptome data from the same input biological sample.

In the HiCAR DNA library, the R2 reads are derived from the genomic sequences targeted by Tn5 tagmentation (Fig 1A). Therefore, the R2 reads can be treated as the single-end ATAC-seq reads to map genome-wide open chromatin regions. In a HiCAR experiment, the cytoplasm and nucleoplasm ployA-RNA can also be collected for RNA-seq library preparation (Fig 1A, detailed in material and methods). After deep sequencing, we confirmed that the HiCAR RNA-seq data and the DNA R2 reads were highly reproducible between biological replicates (Fig S1D, Pearson correlation coefficient = 0.95 for RNA and 0.87 for R2 reads). Next, we compared HiCAR RNA-seq to the public H1 hESC RNA-seq data (by ENCODE[27]), and the DNA library R2 reads to the ATAC-seq data (by the 4DN consortium[25]). As shown in Fig 2D, we observed very similar patterns of RNA and open chromatin signals on genome browser. At the genome-wide scale, the HiCAR RNA-seq data and the DNA R2 reads are highly correlated with the bulk RNA-seq and ATAC-seq datasets (Fig 2E, 2F, PCC = 0.91 and 0.77, respectively). We used MACS2[28] to call 1D open chromatin peaks from HiCAR R2 reads and compared to the ATAC-seq peaks. As shown in Fig 2G, we found that 57,069 (68.9% of total) HiCAR 1D peaks overlapped with ATAC-seq peaks. Further analysis revealed that the overlapping peaks are associated with more significant *p*-values (MACS2) in both ATAC-seq and HiCAR 1D peaks (Fig 2H). When we ranked the HiCAR 1D peaks based on their MACS2 *p*-value, we found that more than 82% of the high confidence 1D peaks (*p*-value < 10e-7) are validated by ATAC-seq peaks (Fig S1E). Taken together, HiCAR generated high-quality chromatin accessibility and transcriptome data using a single low-input sample.

Identification of long-range *cis*-regulatory chromatin interactions in H1 hESC with HiCAR.

HiCAR is designed to identify the long-range chromatin interactions anchored at cREs at high-resolution. To achieve this goal, we applied MAPS [29], a method recently developed for HiChIP and PLAC-seq data, to the HiCAR dataset. Using MAPS, we first removed the potential systemic biases from the contact matrix, including GC content, sequence mappability, 1D chromatin accessibility, and the density of restriction enzyme cutting [29] (detailed in material and methods). In total, we identified 46,792 significant (MAPS FDR < 0.01) chromatin interactions at 5kb resolution and anchored on H1 hESC open chromatin regions (Table S4). Next, we evaluated the sensitivity of HiCAR in detecting known chromatin interactions. Since there is no "gold standard" set of true positive interactions, we decided to compare HiCAR interactions to chromatin interactions defined by well-established methods such as *in situ* Hi-C, PLAC-seq, and HiChIP in matched cell types. Specifically, we used the public *in situ* Hi-C and H3K4m3 PLAC-seq data generated from H1 hESC by the 4DN consortium, as well as the CTCF HiChIP data generated from H9 hESC in a previous study [20,30]. Due to the lower sequencing depth of some public datasets, we decided to compare chromatin interactions at 10kb rather than 5kb resolution (Table S4). *In situ* Hi-C data was processed by HiCCUPS[31] while HiChIP and PLAC-seq data was processed by MAPS[29]. By visual examination of HiCCUPS loops and MAPS interactions in genome browser, we found that HiCAR interactions showed a similar pattern of loops and interactions identified by these well-established and widely used methods (Fig 3A). Interestingly, HiCCUPS loops (from *in situ* Hi-C data) and MAPS interactions (from H3K4me3 PLAC-seq and CTCF HiChIP data) represent a subset of the significant interactions identified by HiCAR (Fig 3A). To further quantify the sensitivity of HiCAR interactions, we filtered the *in situ* Hi-C loops and HiChIP/PLAC-seq interactions and only kept the "testable" loops and interactions with at least one anchor overlapping with ATAC-seq peaks for the following analysis. We found that HiCAR identified 92%, 81% and 69% of the "testable" loops and interactions identified by *in situ* Hi-C, H3K4me3 PLAC-seq, and CTCF HiChIP data, respectively (Fig 3B). These results indicate that HiCAR is a highly sensitive method in detecting "known" chromatin interactions identified by well-established methods.

Next, we assessed the precision of HiCAR-identified interactions. However, due to the lack of a complete list of "true interactions" in H1 hESCs, we instead asked whether HiCAR interactions recapitulate the known features of chromatin contacts. Based on the loop exclusion model, CTCF/Cohesin-associated loops have a preference for convergent CTCF motif orientations at loop anchors[9]. Thus, we examined the CTCF motif orientation of the HiCAR interactions identified by MAPS. We found that 62.8% of HiCAR interactions harbor convergent CTCF motifs on their anchors, and this ratio is comparable to that observed by PLAC-seq (Fig 3C, 60.3%). This result suggested that the precision of HiCAR in identifying interactions is comparable to PLAC-seq. Of note, there are more *in situ*

Hi-C loops (76.9%) anchored at the convergent CTCF motif (Fig 3C). We reasoned that such difference could be due the fact that HiCCUPS uses the local background model for loop calling, and therefore only identifies the most significant loop summits among a cluster of loops/interactions (Fig 3A). To further explore the regulatory role of HiCAR interactions on gene expression, we asked whether HiCAR interactions are enriched for expression quantitative trait loci (eQTL) and their associated genes (TSS) previously identified in human pluripotent stem cells (hPSC) [32]. We observed 5,368 human iPSC eQTL-TSS pairs overlapping with HiCAR loops, whereas only 3,228 eQTL-TSS pairs are expected to overlap with genomic region pairs which are randomly selected (shuffled 10,000 times) with linear distances matched to HiCAR interactions (Fig 3D, empirical *p*-value < 0.0001, detailed in material and Methods). The significantly enriched eQTL-TSS pairs at HiCAR interactions strongly suggest the regulatory role of HiCAR interactions on gene expression in human pluripotent stem cells.

Finally, to directly test the causal role of HiCAR interactions, we selected three putative *SOX2* enhancers for perturbation analysis. As shown in Fig 3E, two enhancers (#1 and #2) are located ~430kb from the *SOX2* TSS and enhancer #3 is located 788kb away from the *SOX2* TSS. All three candidate enhancers are open chromatin regions that form long-range interactions with the *SOX2* promoter as identified by HiCAR. We designed sgRNAs (Table S2) to specifically direct the epigenetic silencer dCas9-KRAB to the three candidate enhancers (Fig 3E). After introducing these CRISPR inhibition components into H1 hESCs to perturb these putative *SOX2* enhancers, we demonstrated significant down-regulation of *SOX2* mRNA expression by RT-qPCR (Fig 3F). Taken together, our results showed that HiCAR is a sensitive and accurate method to identify high-confidence *cis*-regulatory chromatin interactions at high-resolution. More importantly, HiCAR interactions likely reflect functional communication between *cis*-regulatory elements and their distal target genes.

The epigenetically poised, bivalent and repressed chromatin sequences exhibit extensive spatial activity comparable to the active chromatin regions.

Regulatory open chromatin sequences are associated with an array of diverse epigenome signatures. Therefore, we sought to determine whether the HiCAR interactions can enrich cRE-interactions anchored on different chromatin states. We took the 18-chromatin states annotation of H1 hESC defined by ChromHMM [2,17,33,34], and compared the enrichment fold of HiCAR interactions on each state to that of HiCCUPS loops identified by H1 hESC *in situ* Hi-C (Fig 4A). We found that HiCAR interactions showed higher enrichment fold across multiple chromatin states, including enhancers, promoters, and regions associated with active, poised, bivalent, and repressed states (Fig 4A, the chromatin states highlighted in blue text). Interestingly, compared to HiCCUPS loops, HiCAR interactions

277 are depleted at three chromatin states, namely Quiescence/low (Quies), ZNF genes & repeats
278 (ZNF/Rpts), and Heterochromatin (Het). We reasoned that the depletion of HiCAR interactions on these
279 three states could be due to the lack of open chromatin regions on those sequences, as the “Quies” state
280 lack any known marks associated with cRE, while the “ZNF/Rpts” and “Het” sequences are highly
281 enriched for the heterochromatin mark H3K9me3 [34]. Next, we examined how often one chromatin state
282 is interacting with all 18 chromatin states, and assessed whether the observed interaction frequency
283 between two chromatin states is over- or under-represented compared to the genome-wide background
284 (Table S5). Interestingly, we found that the chromatin regions associated with similar epigenome states
285 (epigenetically “active” states versus “inactive”” states, such as repressive/poised/repressed) tend to
286 interact with each other (Fig 4B, blue dots denote the “inactive-inactive” interaction”; red dots denote the
287 “active-active” interaction). On the contrary, the HiCAR interactions connecting the “active” versus
288 “inactive” chromatin states are significantly under-represented (Fig 4B, purple dots). Our results
289 suggested that the spatial proximity of cREs may play a role in facilitating the coordinated epigenomic
290 modification of cis-regulatory sequences.
291
292 Intrigued by the observation that both “active-to-active” and “inactive-to-inactive” interactions are
293 significantly enriched among the HiCAR interactions (Fig 4B), we decided to directly compare the
294 interactions anchored on the “active” versus “inactive” (poised/bivalent/repressed) chromatin states. In
295 ChromHMM, histone H3K27me3 modification is the common histone mark to annotate the poised,
296 bivalent, and repressed chromatin states, while the H3K27ac mark is used to denote transcriptionally
297 active chromatin regions[34]. We selected 14,845 and 10,287 HiCAR interactions with at least one anchor
298 overlapped with H1 hESC H3K27ac or H3K27me3 ChIP-seq peaks, respectively. The interactions
299 overlapped with both H3K27ac and H3K27me3 peaks were excluded from the following analysis.
300 Notably, using HiCAR, the two types of interactions were captured from one single assay independent of
301 antibody-specific ChIP enrichment, and therefore can be directly compared in terms of their numbers,
302 interaction strength/confidence, and transcriptional/enhancer activity. As expected, genes with promoters
303 located on H3K27ac anchors, had significantly higher mRNA expression levels compared with genes
304 with promoters located on H3K27me3 anchors (Fig 4C, Wilcoxon rank-sum, $p < 2.2e-16$). Interestingly,
305 when we compared the interaction strength quantified by -log10 FDR (output from MAPS) between the
306 two types of interactions, the H3K27me3-anchored interactions showed a similar distribution of FDR,
307 which are indistinguishable from the interactions anchored on H3K27ac peaks (Fig 4D, Wilcoxon rank-
308 sum, $p = 0.59$). We also found that the H3K27me3-anchored interactions showed significantly longer
309 linear genomic distance (median distance 145kb) than the H3K27ac-anchored interactions (median
310 distance 125 kb) (Fig 4E, Wilcoxon rank-sum, $p < 2.2e-16$). Furthermore, through gene ontology (GO)

311 analysis, we found that the genes with promoters located on the H3K27ac-anchored interactions are
312 enriched for GO terms related to transcription, metabolic, chromatin organization, and stem cell
313 proliferation/maintenance (Fig S2A), while genes associated with H3K27me3 anchors are enriched for
314 GO terms important for lineage specific tissue and organ differentiation/development (Fig S2B). This GO
315 enrichment analysis suggests that the two types of interactions may play different roles in regulating gene
316 expression in distinct biological processes. In summary, our results showed that the epigenetically
317 "inactive" (poised, bivalent, and repressed) cREs tend to form massive, long-range, and significant
318 chromatin interactions that are comparable to the interactions associated with "active" cREs.
319
320 **Identification of epigenome features important for the spatial interaction activity of *cis*-regulatory**
321 **sequences in H1 hESC**
322 Our high-resolution (5kb bin) cRE-contact map and the rich public epigenome datasets available
323 for H1 hESC (Table S1) gave us the opportunity to study the epigenome features important for the spatial
324 activity of cREs. To probe this question, we employed a method described previously [35,36] to calculate
325 the cumulative interactive score (sum of -log10 FDR) of each HiCAR interaction anchor (5kb bin) (Table
326 S6, detailed in material and methods). Interestingly, when we compare this cumulative interactive score
327 with gene expression (Fig S3A, mRNAs expressed from the gene promoters overlapped with anchors),
328 enhancer activity (Fig S3B, H3K27ac ChIP-seq signal on anchors), and chromatin accessibility (Fig S3C,
329 ATAC-seq signal on anchors), we found that the spatial interaction activity of cREs exhibit very weak
330 Pearson correlation coefficients with gene expression (PCC = 0.06), enhancer activity (PCC = 0.05) and
331 chromatin accessibility (PCC = 0.13). We then asked what are the chromatin epigenome features
332 important for the spatial activity of cREs. To address this question, we identified the cREs associated
333 with high-level chromatin interaction activity. We ranked all 42,463 anchors based on their cumulative
334 interactive score, and identified 2,096 anchors (Fig 5A, red dots) with extremely high-level spatial
335 interaction activity compared to other anchors (Table S6, detailed in material and methods). Consistent
336 with our observation that the spatial activity of cREs exhibit only weak, if any, correlation with
337 transcriptional activity (Fig S3A), we found that the mRNA levels of the genes with promoters located on
338 the 2,096 interaction hotspots are very similar to those of genes with promoters overlapped with regular
339 HiCAR anchors (Fig S3D, S3E, Wilcoxon rank-sum $p = 0.96$). Next, in order to determine the epigenome
340 features associated with these interaction hotspots, we analyzed the public ChIP-seq datasets generated
341 from H1 hESCs (Table S1) including 26 histone mark and 49 TF binding [2,17,27,37]. We identified 9 proteins
342 (KDM1A, HDAC2, RAD21, YY1, CTCF, CTBP2, RNF2, TCF12, and RNA Pol2) and 11 histone marks
343 (H2BK12ac, H2BK15, H2BK20ac, H2AK5ac, H2BK5ac, H3K4me1, H3K4m2, H3K4me3, H3K27me3,
344 H4K8ac, and H3K18ac) that are significantly enriched on the cRE-interaction hotspots (Fig 5B, red dots,

fold change > 1.2, FDR < 0.05; detailed in Table S7). 7 of these 20 enriched histone marks and TF binding signatures (RAD21, YY1, CTCF, RNF2, RNA Pol2, H3K4me1, and H3K27me3) were shown in previous studies to play important roles in regulating 3D chromatin [38–48], while the involvement of the other features in genome organization remains large unexplored. Interestingly, ZNF274, a transcriptional repressor important for the establishment and maintenance of the heterochromatin mark H3K9me3[49], is depleted on the open chromatin interaction hotspots compared to regular HiCAR anchors (Fig 5B, blue dot).

Finally, in order to gain a more comprehensive view of the epigenome features important for the spatial activity of chromatin, we used machine learning approaches to investigate the contribution of 26 histone modifications and the binding of 49 different TFs on chromatin spatial activity. We applied five regression methods [50,51] , namely Decision tree, Linear regression, XGBoost, Random forest, and Linear-kernel support vector machine (Linear SVM), to define the 15 top-ranked features from each model (Fig S4A, Table S8, detailed in material and methods). The five regression models have similar performance as indicated by comparable mean squared error (MES) and mean absolute error (MAE) (Fig S4B). In order to identify the high-confident epigenome features important to chromatin's spatial interactive activity, we required the positive features, defined as "union features", to be identified by at least two models independently. Using this approach, we predicted 22 "union features" as important for the spatial activity of chromatin (Fig 5C). Among these union features, Cohesin (RAD21), CTCF, and ZNF143 are the well-known regulators important for 3D genome organization [46–48]. We also identified additional features, such as pluripotency factor POU5F1, the PRC1 core component RNF2 (also known as RING1B), histone H3K27me3 modification, and transcription activation marks H3K36me3/H4K20me1/RNA Pol2, with known function in regulating high-order chromatin organization [38–44]. The identification of multiple union features with previously validated roles in regulating high-order chromatin organization (Fig 5C, highlighted in blue) suggests that our models are capable of accurately predicting regulators that are important for chromatin interaction activity.

Identification of long-range *cis*-regulatory chromatin interactions in GM12878 and mouse embryonic stem cells (mESCs) with HiCAR.

Lastly, In order to demonstrate the general applicability of HiCAR in other cell types, we applied HiCAR to human lymphoblastoid cell line GM12878 and mouse embryonic stem cells (mESCs). For each cell type, we used ~100,000 cells as input sample and generated high quality HiCAR DNA libraries (Table S3). Using the same approach described in Fig 3A-3C, we identified 42,459 and 91,809 significant (MAPS FDR < 0.01) high resolution (10kb bin) interactions in GM12878 and mESCs, respectively (Fig S5A, S5B;

Table S9, S10 for the full list of MAPS interactions and HiCCUPS loops identified in GM12878 and mESCs). Consistent with our analysis in H1 hESC, the GM12878 and mESC HiCAR interactions showed high sensitivity in detecting the "testable" HiCCUPS loops and MAPS interactions identified by *in situ* Hi-C, HiChIP, and PLAC-seq in GM12878 and mESCs (Fig S5C and Fig S5D). Importantly, 72.4% of GM12878 interactions and 63.7% mESC interactions identified by HiCAR harbor convergent CTCF motifs on their anchor regions. his ratio is comparable to that observed in GM12878 SMC1A HiChIP (75.8%), mESC CTCF PLAC-seq (62.7%), and mESC H3K4me3 PLAC-seq (55.7%), but lower than the ratio detected in HiCCUPS loops identified by *in situ* Hi-C in GM12878 (89.8%) and in mESC (86.7%) (Fig S5E, S5F). These results illustrate that the precision of HiCAR interaction called from GM12878 and mESC is comparable to that of PLAC-seq and HiChIP interactions. Successfully identification of these high-confident cis-regulatory chromatin interactions in GM12878 and mESCs clearly demonstrated the broadly applicability of HiCAR.

Discussion:

We applied HiCAR, a novel co-assay, in H1 hESC and identified 46,792 significant long-range chromatin interactions anchored on open chromatin regions at 5kb resolution. By integrating public epigenome datasets generated by the ENCODE, Epigenome Roadmap, and 4DN consortiums using the same H1 hESC line, we found that the epigenetically poised, bivalent, and repressed chromatin states can form massive, significant, and long-range chromatin interactions that are comparable to the interactions associated with active chromatin states. Consistent with the findings from recent H3K27me3 HiChIP and PRC2 ChIA-PET studies [38,52], the H3K27me3-anchored HiCAR interactions are enriched for genes that are silenced in pluripotency stem cells but important for tissue and organ development. Importantly, the high-resolution chromatin contact map generated by HiCAR provided the unique opportunity to compare the high-resolution cRE-anchored interactions associated with distinct epigenome modifications and chromatin states. Our analysis showed that the cREs with similar chromatin states ("active", or "inactive") tend to interact with each other more frequently, while the interactions between "active" versus "inactive" chromatin states are less frequent. These results suggest that the long-range chromatin interaction may play a role in coordinating epigenome modifications of cREs across linearly separated genomic loci.

Another interesting finding revealed by HiCAR analysis is that there appears to only be a weak correlation between cRE spatial interaction activity and transcriptional activity, enhancer activity, and chromatin accessibility. By integrating HiCAR data with public epigenome data, we identified 20 histone marks and TF binding interactions that are significantly enriched on cRE-anchored interactions hotspots.

We applied five machine learning approaches to predict 22 “union features” important for the spatial interaction activity of cREs in H1 hESC. Many of the epigenetic signatures which are enriched on HiCAR interaction hotspots or predicated by machine learning -- such as CTCF, Cohesin, ZNF143, POU5F1, RNF2, H3K27me3, H3K4me1 - as well as active transcription marks including H3K36me3, H4K20me1, RNA Pol2) are known regulators of 3D genome structure. In the future, it would be very interesting to explore the roles of these epigenome features in regulating genome architecture.

With HiCAR data, we identified 2,096 open chromatin-anchored interaction hotspots in H1 hESCs. In previous studies, other groups carried out similar analyses with *in situ* Hi-C and PLAC-seq data, and discovered frequently interacting regions (FIREs)[35] and super-interactive promoters (SIPs)[36] in the human genome. Like FIREs and SIPs, HiCAR interaction hotspots exhibit unusually high chromatin interaction activity compared to other genomic loci. Notably, FIREs are enriched for super-enhancers and are near genes that are tissue-specifically expressed in 21 primary human tissues and cell types. HiCAR interaction hotspots, however, are not enriched for the super-enhancer mark H3K27ac. Our GO enrichment analysis found that GO terms overrepresented on HiCAR interaction hotspots predominantly related to cell proliferation, chromatin organization, as well as neuronal, cardiovascular, blood vessel, and skeletal system differentiation. (Table S6). Unexpectedly, we did not find pluripotency genes or pluripotency related GO terms enriched on HiCAR interaction hotspots. In contrast, SIPs are enriched for lineage-specific genes in human brain cells. We hypothesize that these differences between HiCAR interaction hotspots, FIREs, and SIPs may be due to two potential phenomena: (1) the genome organization of hESCs is intrinsically different from that of terminally differentiated cells found in human adult tissues; or (2) *in situ* Hi-C, PLAC-seq, and HiCAR each capture a subset of the “true” interactions present in the 3D genome. Therefore, FIREs (by Hi-C), SIPs (by H3K4me3 PLAC-seq), and HiCAR interaction hotspots may represent the top ranked interaction hotspots or hubs that are sampled from different types of chromatin interactions. To test this hypothesis, in the future, it would be interesting to carry out a systematic analysis with well controlled samples, experimental methods, computational pipelines, and potentially with new approaches independent of 3C[53].

Most importantly, we showed that HiCAR is a robust, sensitive, and cost-effective method that can be used to simultaneously study genome architecture, chromatin accessibility, and the transcriptome from the same low-input samples. Compared to existing methods, the technical advantages of HiCAR are multifold. HiCAR requires substantially less sequencing depth than *in situ* Hi-C to identify high-resolution, significant, long-range chromatin interactions anchored on cREs. Second, compared with HiChIP and PLAC-seq, HiCAR does not rely on ChIP-grade antibody-mediated immunoprecipitation to

pull down chromatin interactions bound by a specific protein or histone modification. Thus, HiCAR enables comprehensive analysis of open chromatin-anchored interactions associated with an array of diverse histone mark, TF binding, and chromatin states. Third, compared to state-of-the-art methods such as Trac-looping, with similar sequencing depth, HiCAR generates ~17-fold more informative long-range cis-PETs despite starting from 1,000-fold lower input cell number. Fourth, by applying HiCAR in GM12878 and mESCs, we showed that HiCAR is a sensitive and robust assay which is broadly applicable in multiple cell types with low input samples. Taken together, our results clearly demonstrated the technical advancement and general applicability of HiCAR, which can be used for multimodal analysis of low-input materials.

Accession Codes and Data Availability:

Sequencing data have been deposited to the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE162819. Additional materials, data, code, and associated protocols are available upon request.

Acknowledgements:

We thank Drs. Kenneth Poss (Duke University), Brigid Hogan (Duke University) and David Gorkin (Emory University) for feedback on previous versions of the manuscript. This work is supported by Duke Whitehead Scholar (to Y.D.), startup fund from Duke School of Medicine and Regeneration Next Initiative (to Y.D.), and National Institute Health (NIH) 4D Nucleome Consortium Phase 2 project U01HL156064 (to Y.D.). Y.X. is supported by the post-doctoral fellowship from Regeneration Next Initiative (RNI) at Duke University.

Author contributions:

X.W. and Y.D. conceived the idea for HiCAR; X.W. performed the experiments with help from T.S., and X.L.; Y. X. conducted data analysis with help from X.W., R.S., and W.L.; X.W., Y.X., and Y.D. wrote the paper.

Figure legends

Figure 1. Overview of HiCAR experimental design and HiCAR data quality control. **(A)** In a HiCAR experiment, the nuclei are isolated from cross-linked cells and treated by Tn5 transposase loaded with engineered DNA adaptors, followed by restriction enzyme digestion with 4 base cutter CviQI and in situ ligation. The engineered Tn5 adaptors can be ligated to the proximal genomic DNA digested by CviQI. After in situ ligation, the genomic DNA are purified after reverse crosslinking, and subjected to a second restriction enzyme digestion by another 4 base cutter NlaIII. Then resulting DNA fragments are circularized and PCR amplified for deep sequencing. The DNA sequences amplified from the splint oligo sequence and the Tn5/ME region are defined as R1 reads and R2 reads, respectively. The cytoplasm and nuclei RNA fractions are collected and pooled together for RNA-seq analysis **(B)** The aggregated signals of HiCAR R2 reads (red), R1 reads (blue), and in situ Hi-C (black) within +/- 3kb window centered at H1 hESC ATAC-seq peaks. The HiCAR R1, R2 and Hi-C reads are normalized against sequence depten (counts per million). Signal coverage (y-axis) was calculated as sequencing read depth per base within +/- 2kb window of peak center. **(C)** The aggregated signals of HiCAR R2 reads (red), Trac-looping reads (green), Ocean-C reads (orange), and in situ Hi-C reads (blue) within +/- 2kb window centered at TSS. Enrichment was calculated by comparing the normalized reads signal on peak center against the signal at +/- 2kb region. **(D)** The number of input cells and sequencing outputs of three methods. **(E)** Percentage of uniquely mapped short range (<20kb) cis, long range (>=20kb) cis, and the trans (inter-chromosomal) reads from HiCAR, in situ Hi-C and Trac-looping data. **(F)** Contact frequency as a function of distance measured by HiCAR, in situ Hi-C and Trac-looping data.

Figure 2. HiCAR captures the key features of chromatin organization, chromatin accessibility and transcriptome.
(A) The contact matrices of H1 hESC obtained from HiCAR (top right, above the diagonal) and in situ Hi-C (bottom left, below the diagonal) data at successive zoom-in views. The H1 hESC in situ Hi-C data was obtained from 4DN data portal. The color represents sequence depth normalized reads signal (counts per million mapped reads). **(B)** Scatter plots show the global correlation of compartment scores (top panel), TAD insulation score (middle panel) and TAD directionality index (bottom panel) computed from HiCAR and in situ Hi-C, respectively. The R value: Pearson correlation coefficient. **(C)** Aggregated HiCAR (top) and in situ Hi-C (bottom) contact matrix (10kb bin) within +/- 250kb window centered on the indicated peak regions of H1 hESC. **(D)** A representative genome browser view showing the signals of HiCAR RNA-seq (pink) and HiCAR 1D open chromatin profile (light blue). The red track indicates the H1 hESC bulk RNA-seq and the dark blue track indicates ATAC data, downloaded from ENCODE and 4DN data portal, respectively. **(E-F)**. Scatter plots showing the correlation of **(E)** HiCAR RNA-seq vs. bulk RNA-seq dataset**,** and **(F)** HiCAR R2 reads v.s. ATAC-seq reads**. (G)** Venn diagram showing open chromatin peaks identified by HiCAR R2 reads (1D open chromatin peaks) and ATAC-seq in H1 hESC. MACS2 was used for peak calling. **(H)** We compared the open chromatin peaks identified by HiCAR R2 reads and ATAC-seq. The overlapping open chromatin peaks and the non-overlapping peaks are separated. Boxplot showing the distribution of the MACS2 P-value of the peaks. Wilcoxon rank-sum test was used for statistical analysis to compute P value.

Figure 3. Identify long-range cis-regulatory chromaitn interactions with HiCAR.
(A) Genome browser screenshot showing ChIP-seq (NANOG, SOX2, CTCF, H3K4me1, H3K4me3), RNA-seq, ATAC-seq of H1 hESC, as well as the chromatin loops and interactions identified by HiCAR, CTCF HiChIP, H3K4me3 PLAC-seq and in situ Hi-C data with H1 or H9 hESCs. **(B)** The chromatin loops and interactions with at least one anchor overlapping with ATAC-seq peaks are defined as "testable" loops/interactions. We calculated the proportion of the "testable" loops/interactions that can be discovered by HiCAR interaction to estimate the sensitivity of HiCAR interaction calling. **(C)** We examined the orientation of CTCF motif located on the pairwise anchors of each chromatin loop and interactions. The length of the color bar indicates the proportion of convergent, tandem and divergent CTCF motif pairs among tested HiCCUPS loops and MAPS interactions. **(D)** The TSS-eQTL pairs identified in human pluripotent stem cells are significantly enriched on HiCAR interactions. Red line: the number of observed eQTL-TSS pairs overlapping with HiCAR interactions. The histogram represents the distribution of the number of eQTL-TSS pairs overlapped with randomly sampled (10,000 times shuffling) pairwise DNA regions with matched linear genomic distance to HiCAR interactions. (Empirical p-value < 0.0001). **(E)** Genome browser screenshot showing H1 hESC ATAC-seq track and HiCAR interactions near *SOX2* locus. The three arrowheads point to the three candidate SOX2 enhancers (highlighted in light blue). **(F)** The sgRNAs were designed to specifically target the SOX2 candidate enhancers showing in (E). The H1 hESC were infected by lentiviral vectors expressing dCas9-KRAB together with control sgRNA or the sgRNAs targeting enhancer regions. After lentiviral infection, the hESCs were selected by Puromycin for 3-days, then cultured for another 7-days without Puromycin. The total RNA was extracted and subjected to RT-qPCR analysis. The mRNA level of SOX2 was normalized against housekeeping gene GAPDH. The data was collected from three biological replicates. P values are calculated by two-tailed Student's t test.

Figure 4. The poised, bivalent, and repressed chromatin regions form massive, long-range, and significant chromatin interactions comparable to the active chromatin states.
(A) We took the anchor (5kb bin) sequences of all interactions identified by HiCAR, and calculated the "observed" number of anchors overlapped with each individual chromatin state defined by chromHMM. Based on the genome-wide distribution of each chromHMM state, we also calculated the "expected" number of anchors overlapped with each state. The fold change (y-axis) of HiCAR interaction for each chromHMM state was calculated as "observed/expected". The fold change of Hi-C loops for each chromHMM state was calculated in the same way. The 18-states ChromHMM annotation: TssA: Active TSS, TssFlnk: Flanking TSS, TssFlnkU: Flanking TSS Upstream, TssFlnkD: Flanking TSS Downstream, Tx: Strong transcription, TxWk: Weak transcription, EnhG1: Genic enhancer1, EnhG2: Genic enhancer2, EnhA1: Active Enhancer 1, EnhA2: Active Enhancer 2, EnhWk: Weak Enhancer, ZNF: Rpts ZNF genes & repeats, Het: Heterochromatin, TssBiv: Bivalent/Poised TSS, EnhBiv: Bivalent Enhancer, ReprPC: Repressed PolyComb, ReprPCWk: Weak Repressed PolyComb, Quies: Quiescent. **(B)** Based on HiCAR interaction, we first computed the "observed" interaction frequency of pairwise chromatin states (total 18 states determined by ChromHMM). Next, based on the genome-wide distribution of each chromHMM state, we computed the "expected" interaction frequency between any two states. The fold change of pairwise interaction frequency and P-value were calculated using the "annotateInteractions" function from Homer. X-axis: log2 (fold change) of "observed" interaction frequency over "expected" interaction frequency. Y-axis: -log10(FDR), the FDR is the output from HOMER. Red dots: the interactions between "active" chromatin states; Blue dots: the interactions between "inactive" states, including

bivalent/repressed/poised chromatin states; Purple dots: the interactions between "active" versus "inactive" states. **(C-D)** We selected 14,845 and 10,287 HiCAR interactions with at least one anchor overlapped with H3K37ac and H3K27me3 peaks, respectively. For these two types of interactions (H3K27ac v.s. H3K27me3), we compared **(C)** the mRNA level of genes expressed from the promoters located on anchors; **(D)** interaction strength quantified by -log10 FDR, the FDR is output from MAPS; and **(E)** the linear genomic distance between anchors of interactions. Boxplot: P value is calculated from Wilcoxon rank-sum test.

Figure 5. The epigenome features important for chromatin spatial interactive activity.
(A) The 5kb anchors of HiCAR interactions are ranked along the x-axis based on their cumulative interactive score (sum of -log10 FDR, y-axis). FDR is the output of MAPS of each significant interaction. Total 2,096 anchors were identified as interaction hotspots associated with abnormal high level interactive score (red dots, detailed in methods). **(B)** Scatterplot showing the significantly enriched (red dots) or depleted (blue dot, ZNF274) histone mark and TF binding on interaction hotspots versus regular interaction anchors. Total 75 public ChIP-seq data listed in Table S1 was used for signal enrichment analysis. **(C)** We employed five machine learning algorithms, including Decision tree, Linear regression, XGBoost, Random forest, and Linear-kernel support vector machine, to predict the top ranked epigenome features that are potentially important for the spatial interactive activity of cREs. The "union features" are defined as the features predicted by at least two algorithms. The features highlighted in blue color are the features with known function in regulating 3D chromatin interactions.

577 **Materials and methods**:

578 **Cell culture and crosslink.**

579 H1 hESCs (WiCell, WA01) were cultured in Matrigel (corning, 354230) coated plates with Stabilized
580 feeder-free maintenance medium mTeSR™ Plus (STEMCELL, #05825). mTeSR™ Plus was changed
581 every other day. For crosslinking, cells were washed once by PBS, then treated by accutase (biolegend,
582 #423201) for 10mins at 37°C. After removing the accutase, cells were resuspended by DMEM.
583 Formaldehyde was added to the final concentration of 1%, incubated at room temperature for 10mins.
584 Glycine was added to the final concentration of 0.2M, incubated at room temperature for 10 mins to
585 quench formaldehyde. Fixed cells were pelleted by centrifugation for 5 min at 4°C and washed with ice-
586 cold PBS once
587
588 **Tn5 Purification**
589 Briefly, Rosetta DE3 cells transformed with Tn5 expression plasmid pTXB1-Tn5 (Addgene #60240) were
590 cultured in 500ml LB and incubated at 16°C overnight for protein induction. The bacteria were collected
591 by centrifuge and resuspend by pre-cooled HEGX (40mM Hepes-KOH pH 7.2, 1.6M NaCl, 2 mM EDTA,
592 20% Glycerol, 0.4% Triton-X100, Roche Complete Protease Inhibitor), sonicated to release the protein.
593 PEI (10% PEI, 4.44% HCl, 800mM NaCl, 20mM Hepes, 0.3mM EDTA, 0.2% Triton X-100, pH 7.2) were
594 then added to the lysate in dropwise to precipitate the E. coli DNA. The lysate was then centrifuged and
595 supernatant was loaded to Chitin column (BIO-RAD, #7372522). The column was rotated at 4°C for 2-
596 3h then washed by HEGX buffer. 15ml HEGX buffer containing 100mM DTT was added to elute the
597 protein. The column was incubated for another 24 hr at 4°C. The elution fraction was collected and
598 concentrated to about 1ml by Amicon Ultracel 30K (Millipore, #UFC903024), then dialyzed twice by 1L
599 dialysis buffer (100 HEPES-KOH pH 7.2, 0.2 M NaCl, 0.2 mM EDTA, 2 mM DTT, 0.2% Triton X-100,
600 20% glycerol) for 24h using dialysis membrane tube (Spectra, D1614-11). Then the protein was added
601 80% glycerol to a final concentration of 50%.
602
603 **Tn5 transposase assembly**
604 To assemble Tn5, 50ul of 200mM ME-rev and 50ul of 200mM BfaI-truseqR1-pmeI-nextera7 (Table S2)
605 were annealed by the following program: 95°C 5min, cool to 14°C with a slow ramp 1°C /min. The
606 annealed adaptor was mixed with Tn5 Transposase in 1: 1.5 molar ratio, the mixture was mixed by pipette
607 and incubated at room temperature for 30mins.
608
609 **HiCAR protocol**
610 Step1. Nuclei preparation and tagmentation:

100,000 crosslinked cells were treated by 1ml NPB (PBS containing 5% BSA, 1mM DTT, 0.2% IGEPAL, Roche Complete Protease Inhibitor) at 4°C for 15min to isolate the nuclei. After centrifugation, the supernatant containing cytoplasm RNA was saved for future RNA-seq analysis. The isolated nuclei were resuspended in 350ul 2X TB buffer (66mM Tris-AC pH 7.8, 132mM K-AC, 20mM Mg-AC, 32% DMF), 335ul water and 15ul assembled Tn5 transposome. The oligos used for Tn adaptors are listed in Table S2).

Step 2. CviQI digestion and in situ ligation

After tagmentation, the nuclei were permeabilized by 2% SDS at 62ºC for 10 minutes. After centrifugation at 850g for 5min, the supernatant containing nuclei RNA was collected for future RNA-seq library construction. The nuclei were then digested in 90ul 1.1X NEBuffer 3.1 containing 100U CviQI (NEB, #R0639L) After digestion, we added 48ul 10X T4 ligation buffer, and 2ul T4 DNA ligase (400U/ul, NEB, #M0202S) for in situ ligation with TruseqR1 oligo (Table S2) at room temperature for 4h.

Step 3. Reverse crosslink and DNA purification

After centrifugation, the supernatant was discarded. The nuclei were resuspended in 200ul of 10mM Tris-HCl (pH 8.0), 5ul Proteinase K (Thermofisher, #AM2546), 10ul 20% SDS, incubated at 60°C for 30min. Next, we added 22ul 5M NaCl to the buffer and incubated the nuclei at 68°C for at least 1h to reverse crosslink. The DNA was purified by Phenol:Chloroform:Isoamyl Alcohol (25:24:1, v/v, SPECTRUM, #136112-00-0) treatment followed by ethanol precipitation. The DNA was dissolved by 21ul 10mM Tris-HCl (pH8.0).

Step 4. NlaIII digestion, circularization, and DNA library amplification by PCR

The purified DNA was incubated with 4ul 10mM dNTP, 5ul 10X Cutsmart buffer 1.5ul T4 DNA polymerase (NEB, # M0203L) and 20.5ul H2O at room temperature for 30min to repair the Tn5 transposition gap. Next, the reaction was incubated at 75°C for 20min to inactivate T4 DNA polymerase. After that, 43ul water, 5ul 10X CutSmart buffer, and 2ul NlaIII (NEB, # R0125L) were added into the sample followed by incubation at 37°C for 1h. The digested DNA was purified by 0.9X (90ul) volume SPRI beads (BECKMAN, # B23319), and dissolved in 80ul 10mM Tris-HCl (pH8.0) buffer. Next, the DNA was diluted to 0.6ng/ul and circulated in T4 Ligation Buffer by T4 DNA ligase (400U/ul, NEB, #M0202S). The sample is mixed and incubated at room temperature for at least 2h. The DNA was purified by DNA clean & concentrator kit (Zymo, #D4013) and eluted in 20ul water. The PCR library amplification was performed using the following program (step 1: 72 °C 5 min, 98 °C 30 s; step 2: 98 °C 10 s, 59 °C 30 s, 72 °C 45s, repeating step 2 for an additional 11 cycles; step 3: 72°C 5 min and 4°C forever). After PCR, the DNA product

between 400-600bp was purified by gel extraction using DNA recovery kit (Zymo, #D4002) for deep sequencing.

Step 5. HiCAR RNA libraries construction

The cytoplasmic and nuclei RNA fraction was combined. We added 20% SDS to the pooled RNA fraction to make the final concentration of SDS as 1%. The sample was mixed and incubated at 60°C for 30min. After incubation, we added 1/9 volume of 5M NaCl to make the final concentration of NaCl 500mM, the sample was incubated at 68°C for at least 1.5h for reverse crosslinking. Next, the RNA was purified by Phenol:Chloroform:Isoamyl Alcohol (25:24:1, v/v, SPECTRUM, #136112-00-0) extraction and ethanol precipitation. The sample was dissolved in 21ul 10mM Tris-HCl (pH8.0). Then the sample was treated by 0.5ul DNaseI at 37°C for 30min to remove DNA in solution. The RNA was purified by 2X volume of SPRI beads, dissolved RNA by 20ul 10mM Tris-HCl (pH8.0). Then take out 2.3ul RNA to make an RNAseq library using smartseq2 protocol[19].

HiCAR data processing

HiCAR datasets were processed following the distiller pipeline (https://github.com/mirnylab/distiller-nf). Briefly, reads were aligned to hg38 reference genome using bwa mem with flags -SP. Alignments were parsed, and paired end tags (PET) were generated using the pairtools (https://github.com/mirnylab/pairtools). PET with low mapping quality (MAPQ < 10) were filtered out. PET with the same coordinate on the genome or mapped to the same digestion fragment were removed. Uniquely mapped PETs were flipped as side 1 with the lower genomic coordinate and aggregated into contact matrices in the cooler format using the cooler tools[54] at delimited resolution (5kb, 10kb, 50kb, 100kb, 250kb, 500Kb, 1Mb, 25MB, 50MB,100MB). The dense matrix data were extracted from cooler files and visualized using HiGlass[55]. The R1 and R2 reads signal around TSS or peaks were calculated with EnrichedHeatmap[56] before PET flipping.

Hi-C matrix correlation SCC (stratum-adjusted correlation coefficient)

The similarity between different Hi-C datasets were measured by HiCRep[26]. The stratum adjusted correlation coefficient (SCC) is calculated on a per chromosome basis using HiCRep on 100 kb resolution data with a max distance of 5 Mb. The SCC was calculated as a weighted average of stratum-specific Pearson's correlation coefficients.

Compartments A and B, directionality and Insulation score

Compartmentalization, directionality index and insulation score was assessed using cooltools (https://github.com/mirnylab/cooltools). Briefly, eigenvector decomposition was performed on cis contact maps at 100-kb resolution. The first three eigenvectors and eigenvalues were calculated, and the eigenvector associated with the largest absolute eigenvalue was chosen. An identically binned track of GC content was used to orient the eigenvectors. The insulation score and directionality Index were computed by cooltools using 'find_insulating_boundaries' and 'directionality' function, respectively.

Contact probability decaying curve

The curves of contact probability as a function of genomic separation were generated by pairsqc following the 4DN pipeline (https://github.com/4dn-dcic/pairsqc). Briefly, the genome is binned at log10 scale at interval of 0.1. For each bin, contact probability is computed as number of reads/number of possible reads/bin size.

HiCAR RNA profile processing

Reads were aligned to hg38 genome with Hisat2[57] using hg38 genome_tran index obtained from Hisat2 website (http://daehwankimlab.github.io/hisat2/download/). Raw reads for each gene were quantified using featureCounts[58].

HiCAR 1D open chromatin peak processing

Unique mapped HiCAR DNA library R2 reads were extracted before PET flipping. R2 reads from long range (>20kb) and the inter-chromosome trans-PETs were combined and processed to be compatible as MACS2[28] input BED files. R2 reads from the short-range cis-PETs were discarded to avoid the potential bias due to proximity to CviQI enzyme cut sites[59]. . MACS2[28] was used to identify ATAC peaks following the ENCODE pipeline (https://github.com/ENCODE-DCC/atac-seq-pipeline) with the following parameters: "-q 0.01 --shift 150 --extsize -75--nomodel -B --SPMR --keep-dup all " .

CTCF motif orientation analysis

CTCF ChIP-seq peak list of H1 was downloaded from ENCODE (accession No. ENCFF821AQO) and searched for CTCF sequence motifs using gimme[60] and CTCF motif (MA0139.1) from the JASPAR database[61]. We then selected a subset of interactions with both ends containing either a single CTCF motif or multiple CTCF motifs in the same direction. The frequency of all possible directionality of CTCF motif pairs, convergent, tandem and divergent, are evaluated.

Chromatin interaction calling

For HiCAR, PLAC-seq and HiChIP datasets, we used the MAPS[29] to call the significant chromatin interactions. First, paired-end tags were extracted from cooler datasets at 5KB or 10Kb resolution using the "cooler dump" function with parameters: "-t pixels -H --join". The interaction anchor bins were defined by the ATAC peaks or corresponding ChIP-seq peaks called using MACS2[28]. MAPS applied a positive Poisson regression-based approach to normalize systematic biases from restriction enzyme cut sites, GC content, sequence mappability, and 1D signal enrichment. We grouped interactions that were located within 15 kb of each other at both ends into clusters and classified all other interactions as singletons. We retained only interactions with 6 or more and normalized contact frequency (raw read counts/expected read counts) >= 2 and the significant interactions were defined by FDR < 0.01 for clusters and FDR < 0.0001 for singletons. For in situ Hi-C dataset, the .hic file is downloaded from 4DN data portal (accession No. 4DNES2M5JIGV) and HiCCUPS[31] is applied to call interactions at 10Kb resolution with the following parameters: "-r 10000 -k KR -f .1,.1 -p 4,2 -i 7,5 -t 0.02,1.5,1.75,2 -d 20000,20000".

Chromatin states enrichment analysis at chromatin interaction anchors

Chromatin state calls using a 18-state model for H1 cell line were obtained from the Roadmap Epigenomics Mapping Consortium. To determine which pairs of chromatin states were enriched at interaction anchors at a statistically significant level, we examined the distribution of chromatin states at interaction anchors using HOMER and assess if a connection between the feature is over or under represented given the general enrichment for each chromatin states at the interaction anchors. We used the HOMER "annotateInteractions" function to obtain the p value and enrichment fold ratio for all pairs of chromatin states. The FDR adjusted p values were obtained using the p.adjust function from the R package, with option method="fdr".

Comparison between eQTL-TSS association and HiCAR interaction

To test the enrichment for HiCAR identified interactions in significant eQTL-TSS association, we first obtain the eQTL-TSS associations in H1 hESC from the previous study[32]. To assess the significance of the enrichment, we generated a null distribution by creating a simulated interaction datasets by resampling the same number of interactions at random from distance-matched interactions (with 10,000 repeats). The empirical P-value was computed by comparing the observed overlapping number with the null distribution.

Machine learning approaches to identify features associated with interaction activity

We next collected epigenetic features from the public ENCODE consortium from H1 hESC lines. There are 75 ChIP-seq datasets collected for the H1 cell line, including 26 histone mark datasets and 49 transcription factors (redundant datasets from different labs are removed). Average bigWig signals on each 5kb anchor are computed using the bigWigAverageOverBed command from UCSC. We used regression-based machine learning. For regression, we used a sigmoid function to scale the chromatin interaction score into a [0,1] range:

$$f(x) = \frac{1}{1 + e^{-c1(x-c2)}}$$

We set $c1 = 0.05$ and $c2 = 20$ empirically, such that the bins with stronger interactions have a value closer to 1 after sigmoid conversion. We used the regression methods in the scikit-learn Python package[50] for regression analysis, including linear regression, decision tree, xbgboost, random forest and linear-kernel support vector machine (SVM). The XGBoost Python package[51] was used for XGBoost regression analysis.

Gene Ontology enrichment analysis

We used Clusterprofile[62] to examine whether particular gene sets were enriched in certain gene lists. GO categories with “BH” adjusted p value < 0.05 were considered as significant.

Reference

1. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).
2. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
3. Diao, Y. *et al.* A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods* **14**, 629–635 (2017).
4. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
5. Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* **2010**, db.prot5384 (2010).
6. Simon, J. M., Giresi, P. G., Davis, I. J. & Lieb, J. D. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nat. Protoc.* **7**, 256–267 (2012).
7. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
8. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
9. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
10. Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557–572.e24 (2017).
11. Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
12. Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).
13. Fang, R. *et al.* Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res.* **26**, 1345–1348 (2016).
14. Davies, J. O. J. *et al.* Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat. Methods* **13**, 74–80 (2016).
15. Lai, B. *et al.* Trac-looping measures genome structure and chromatin accessibility. *Nat. Methods* **15**, 741–747 (2018).

16. Li, T., Jia, L., Cao, Y., Chen, Q. & Li, C. OCEAN-C: mapping hubs of open chromatin interactions across the genome reveals gene regulatory networks. *Genome Biol.* **19**, 54 (2018).
17. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
18. Reznikoff, W. S. Tn5 as a model for understanding DNA transposition. *Mol. Microbiol.* **47**, 1199–1206 (2003).
19. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
20. Krietenstein, N. *et al.* Ultrastructural Details of Mammalian Chromosome Architecture. *Mol. Cell* **78**, 554–565.e7 (2020).
21. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).
22. Ma, W. *et al.* Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nat. Methods* **12**, 71–78 (2015).
23. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
24. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019).
25. Dekker, J. *et al.* The 4D nucleome project. *Nature* **549**, 219–226 (2017).
26. Yang, T. *et al.* HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.* **27**, 1939–1949 (2017).
27. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
28. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
29. Juric, I. *et al.* MAPS: Model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. *PLoS Comput. Biol.* **15**, e1006982 (2019).
30. Lyu, X., Rowley, M. J. & Corces, V. G. Architectural Proteins and Pluripotency Factors Cooperate to Orchestrate the Transcriptional Response of hESCs to Temperature Stress. *Mol. Cell* **71**, 940–955.e7 (2018).
31. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98 (2016).
32. DeBoever, C. *et al.* Large-Scale Profiling Reveals the Influence of Genetic Variation on Gene Expression in Human Induced Pluripotent Stem Cells. *Cell Stem Cell* **20**, 533–546.e7 (2017).
33. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat.*

Methods **9**, 215–216 (2012).

34. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* **12**, 2478–2492 (2017).
35. Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep.* **17**, 2042–2059 (2016).
36. Song, M. *et al.* Cell-type-specific 3D epigenomes in the developing human cortex. *Nature* (2020) doi:10.1038/s41586-020-2825-4.
37. Zheng, R. *et al.* Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res.* **47**, D729–D735 (2019).
38. Kraft, K. *et al.* Polycomb-mediated Genome Architecture Enables Long-range Spreading of H3K27 methylation. *bioRxiv* (2020).
39. Cai, Y. *et al.* H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. doi:10.1101/684712.
40. van Steensel, B. & Furlong, E. E. M. The role of transcription in shaping the spatial organization of the genome. *Nat. Rev. Mol. Cell Biol.* **20**, 327–337 (2019).
41. Abboud, N. *et al.* A cohesin-OCT4 complex mediates Sox enhancers to prime an early embryonic lineage. *Nat. Commun.* **6**, 6749 (2015).
42. Donohoe, M. E., Silva, S. S., Pinter, S. F., Xu, N. & Lee, J. T. The pluripotency factor Oct4 interacts with Ctcf and also controls X-chromosome pairing and counting. *Nature* **460**, 128–132 (2009).
43. McLaughlin, K. *et al.* DNA Methylation Directs Polycomb-Dependent 3D Genome Re-organization in Naive Pluripotency. *Cell Rep.* **29**, 1974–1985.e6 (2019).
44. Boyle, S. *et al.* A central role for canonical PRC1 in shaping the 3D nuclear landscape. *Genes Dev.* **34**, 931–949 (2020).
45. Yan, J. *et al.* Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers. *Cell Res.* **28**, 387 (2018).
46. Yu, M. & Ren, B. The Three-Dimensional Organization of Mammalian Genomes. *Annu. Rev. Cell Dev. Biol.* **33**, 265–289 (2017).
47. Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163**, 1611–1627 (2015).
48. Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat. Rev. Genet.* **19**, 789–800 (2018).
49. Frietze, S., O'Geen, H., Blahnik, K. R., Jin, V. X. & Farnham, P. J. ZNF274 recruits the histone methyltransferase SETDB1 to the 3' ends of ZNF genes. *PLoS One* **5**, e15082 (2010).

50. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).
51. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *arXiv [cs.LG]* (2016).
52. Ngan, C. Y. *et al.* Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development. *Nat. Genet.* **52**, 264–272 (2020).
53. Dekker, J. The three 'C' s of chromosome conformation capture: controls, controls, controls. *Nat. Methods* **3**, 17–21 (2006).
54. Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36**, 311–316 (2020).
55. Kerpedjiev, P. *et al.* HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol.* **19**, 125 (2018).
56. Gu, Z., Eils, R., Schlesner, M. & Ishaque, N. EnrichedHeatmap: an R/Bioconductor package for comprehensive visualization of genomic signal associations. *BMC Genomics* **19**, 234 (2018).
57. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
58. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
59. Lareau, C. A. & Aryee, M. J. hichipper: a preprocessing pipeline for calling DNA loops from HiChIP data. *Nature methods* vol. 15 155–156 (2018).
60. van Heeringen, S. J. & Veenstra, G. J. GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments. *Bioinformatics* **27**, 270–271 (2011).
61. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).
62. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

Supplementary Information

Including Supplementary figures 1-5 with figure legends

The title of Supplementary Table 1-10

Supplementary Figure 1. HiCAR library enrichment analysis and data quality control.

(A) The aggregated signals of HiCAR R2 reads (red), R1 reads (blue), and in situ Hi-C (black) reads within +/- 3kb window of indicated peak regions of H1 hESC. The HiCAR R1, R2 and Hi-C reads are normalized against sequence depth (counts per million). Signal coverage (y-axis) was calculated as sequencing read depth per base within +/- 2kb window of peak center. **(B)** The aggregated signals of HiCAR R2 reads (red), R1 reads (blue), H3K4me1 HiChIP (purple), H3K4me3 PLAC-seq (black), and DNase Hi-C (brown) within +/- 2kb window centered at TSS. Enrichment fold was calculated by comparing the reads coverage on peak center against the reads coverage at +/- 2kb region. **(C)** We used HiCrep to compute the similarity of chromatin contact matrice including three HiCAR biological replicates and 4DN in situ Hi-C data. The number is the SCC value computed from HiCrep. **(D)** Scatter plots with PCC of the reads counts from two biological replicates of HiCAR RNA-seq library (left) and HiCAR DNA library R2 reads (right panel). **(E)** The HiCAR 1D open chromatin peaks are called by MACS2. The peaks are ranked along x-axis based on their MACS P value (-log10). At a given P value, the y-axis indicates the proportion of the HiCAR 1D peaks that can be validated by H1 hESC ATAC-seq peaks.

Supplementary Figure 2. Gene Ontology terms associated with H3K27ac- and H3K27m3-anchored HiCAR interactions

We selected the genes whose promoters are overlapped with HiCAR interaction anchors for Gene Ontology enrichment analysis. **(A)** GO terms enriched on H3K27ac-anchored interactions and **(B)** GO terms enriched on H3K27me3-anchored interactions.

Supplementary Figure 3. The spatial interactive activity of cis-regulatory sequence shows very weak correlation with its transcriptional activity, enhancer activity, or chromatin accessibility.

(A-C) Scatter plots showing the cumulative interactive score (sum of -log10FDR) of HiCAR interaction anchor on y-axis, against x-axis showing: **(A)** mRNA level (log2 FPKM) of the genes expressed from the promoters overlapped with anchors; (B) H3K27ac ChIP-seq signal of anchors indicating their enhancer activity mark; and **(C)** chromatin accessibility of anchors measured by ATAC-seq signal. PCC: Pearson correlation coefficient. **(D)** Histogram and **(E)** boxplot showing the distribution of mRNA levels expressed from the gene promoters overlap with HiCAR interaction hotspots or regular anchors. The P value (0.96) was calculated by Wilcoxon rank-sum test in (D).

Supplementary Figure 4. Prediction of histone mark and TF binding important for cRE's spatial interactive activity using machine learning.

(A) The top ranked 15 features predicted by five machine learning algorithms, including Decision tree, Linear regression, XGBoost, Random forest, and Linear-kernel support vector machine (Linear SVM). **(B)** Mean absolute error and Mean squared error of each regression method.

Supplementary Figure 5. Identify long-range cis-regulatory chromaitn interaction in GM12878 and mESCs with HiCAR.

(A) Genome browser screenshot showing CTCF ChIP-Seq, DNase hypersensitive (DHS), and the HiCCUPS loops and MAPS interactions identified by HiCAR, in situ Hi-C and SMC1A HiChIP in GM12878 cells. **(B)** Genome browser screenshot showing H3K27ac ChIP-seq and the HiCCUPS loops and MAPS interactions identified by HiCAR, in situ Hi-C, CTCF PLAC-seq, H3K4me3 PLAC-seq in mESC cells. **(C, D)** The chromatin loops and interactions with at least one anchor overlapping with ATAC-seq peaks are defined as "testable" loops/interactions. We calculated the proportion of the "testable" loops/interactions that can be discovered by HiCAR interaction to estimate the sensitivity of HiCAR interaction calling in GM12878 and mESCs. **(C)** In GM12878 cells, HiCAR discovered 79% and 62% of "testable" loops/interactions identified by in situ Hi-C and SMC1A HiChIP, respectively. **(D)** In mESC, HiCAR discovered 74%, 70% and 85% of "testable" loops and interactions identified by in situ Hi-C, H3K4me3 PLAC-seq and CTCF PLAC-seq, respectively. **(E, F)** We examined the motif orientation of CTCF on the anchors of chromatin loop and interactions. The length of the bars indicating the proportion of chromatin loops/interactions harbor convergent, tandem and divergent CTCF motif on their anchors. (E) In GM12878 cells, 72.4%, 75.8%, and 89.8% HiCAR interactions, SMC1A HiChIP interactions, and in situ Hi-C loops harbor convergent CTCF motif on their anchors. **(F)** In mESC, 63.7%, 62.7%, and 55.7% of HiCAR interactions, CTCF PLAC-seq interactions, and H3K4me3 PLAC-seq interactions harbor convergent CTCF motif on their anchors.

Supplementary Table 1

The list of public datasets used in this study.

Supplementary Table 2

Oligo and DNA sequences used in this study.

Supplementary Table 3

Summary of all a total of seven HiCAR DNA libraries generated with H1 hESC, GM12878 and mESCs.

Supplementary Table 4

The full list of chromatin loops and interactions in H1 identified by HICCUPS and MAPS from in situ HiC, HiChIP, PLAC-seq and HiCAR data.

Supplementary Table 5

Statistical analysis of pairwise chromHMM states interaction frequency.

Supplementary Table 6

HiCAR anchor cumulative interactive score and GO term enrichment on interaction hotspots.

Supplementary Table 7

Statistical analysis of ChIP-seq signals enrichment on HiCAR interaction hotspots versus regular anchors.

Supplementary Table 8

The full list of top-ranked important features predicted by five regression models.

Supplementary Table 9

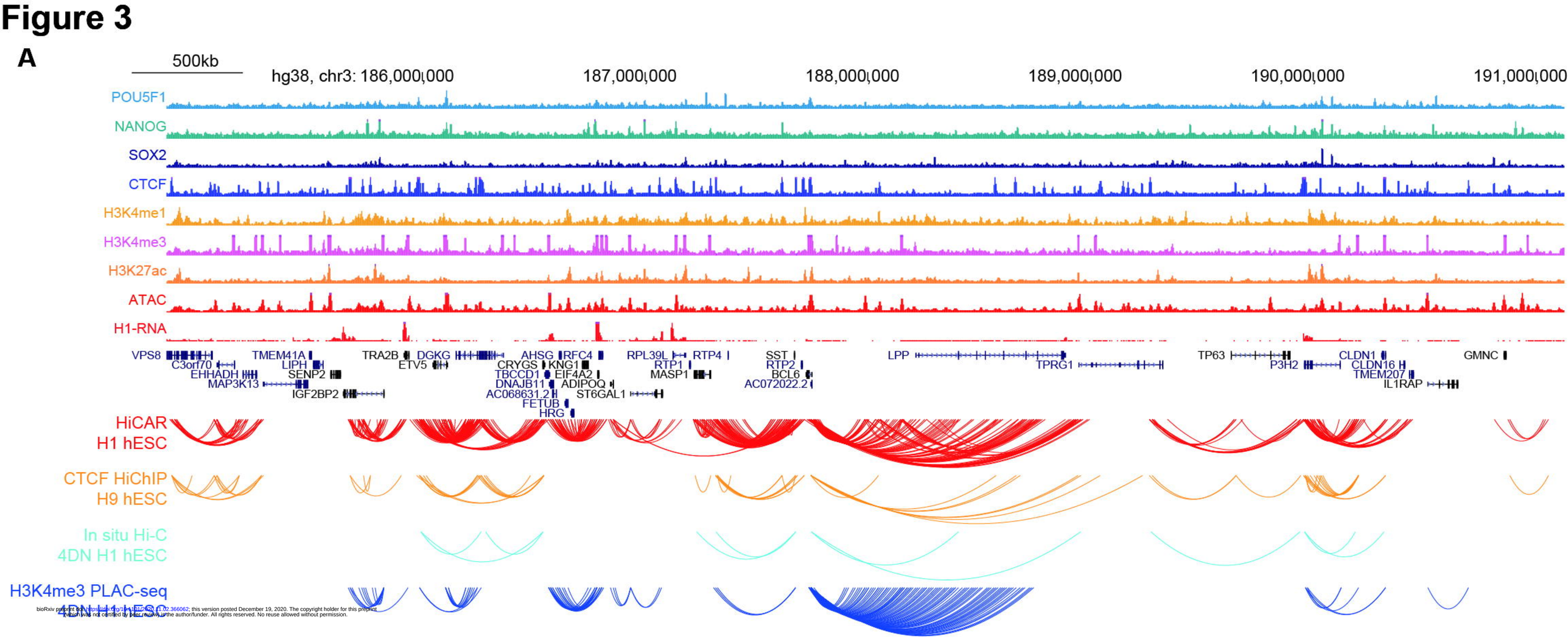
The full list of mESC HiCCUPPS loops and MAPS interactions identified by *in situ* Hi-C, PLAC-seq and HiCAR datasets.

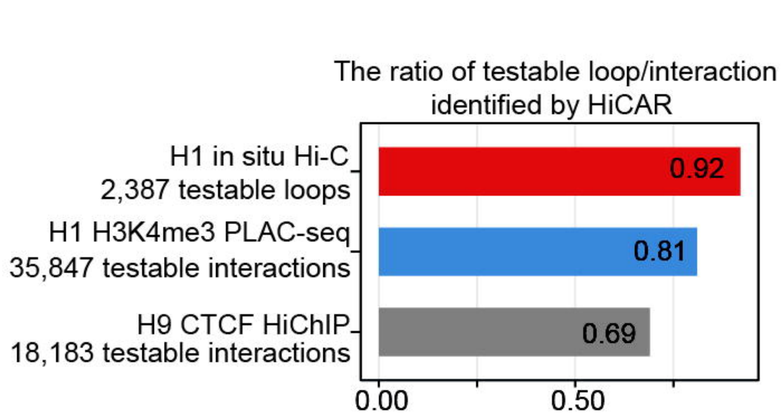
Supplementary Table 10

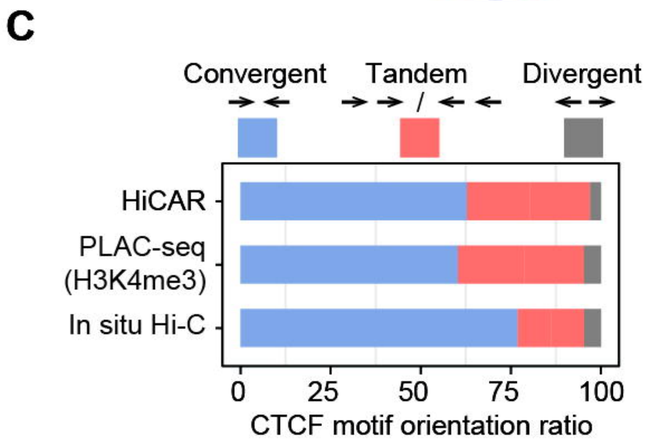
The full list of GM12878 HiCCUPPS loops and MAPS interactions identified by *in situ* Hi-C, HiChiP, and HiCAR datasets.

Figure 2

Figure 3







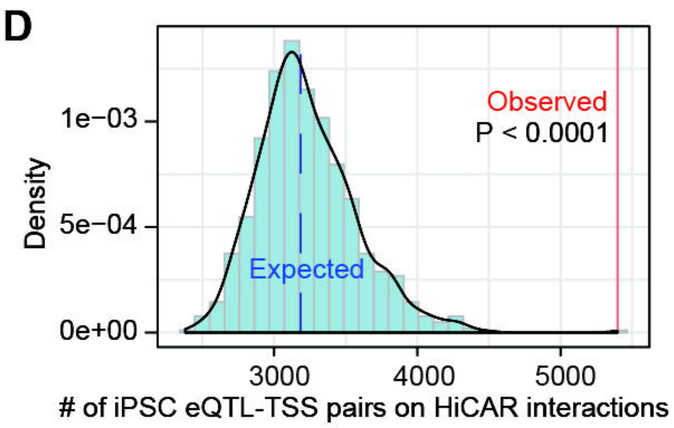
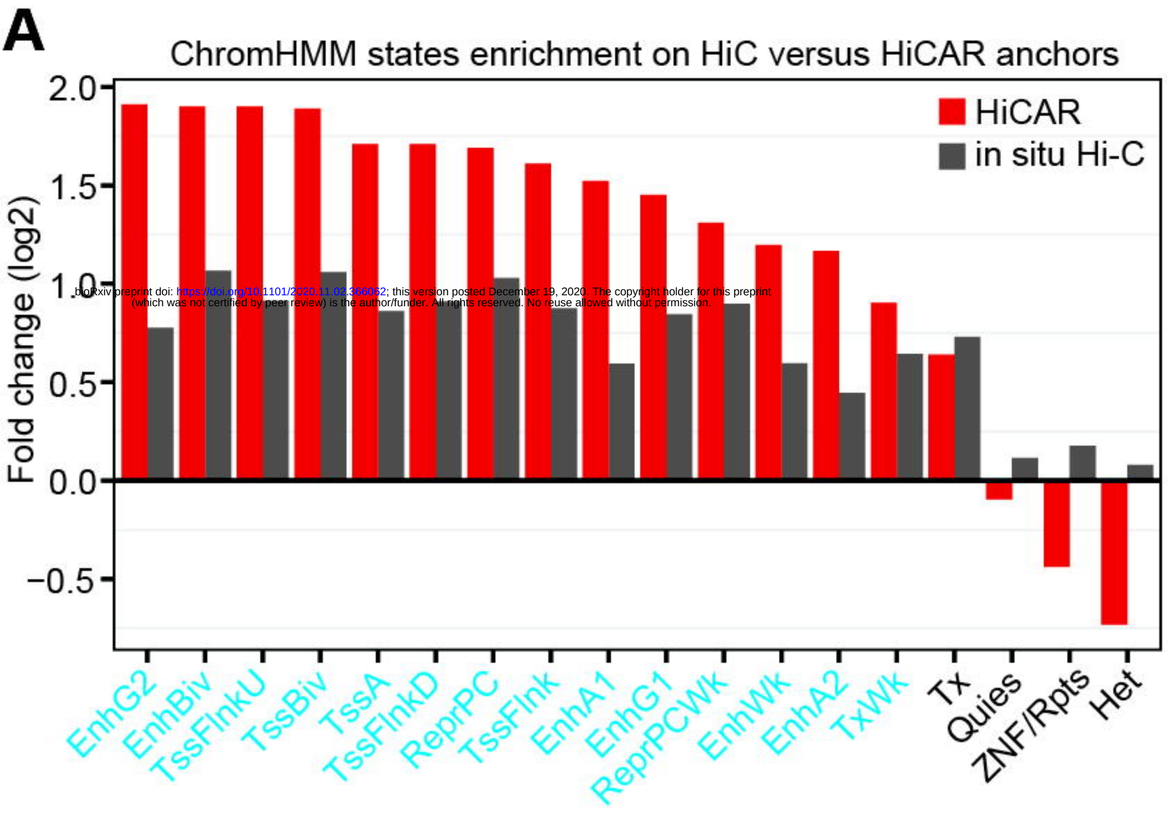
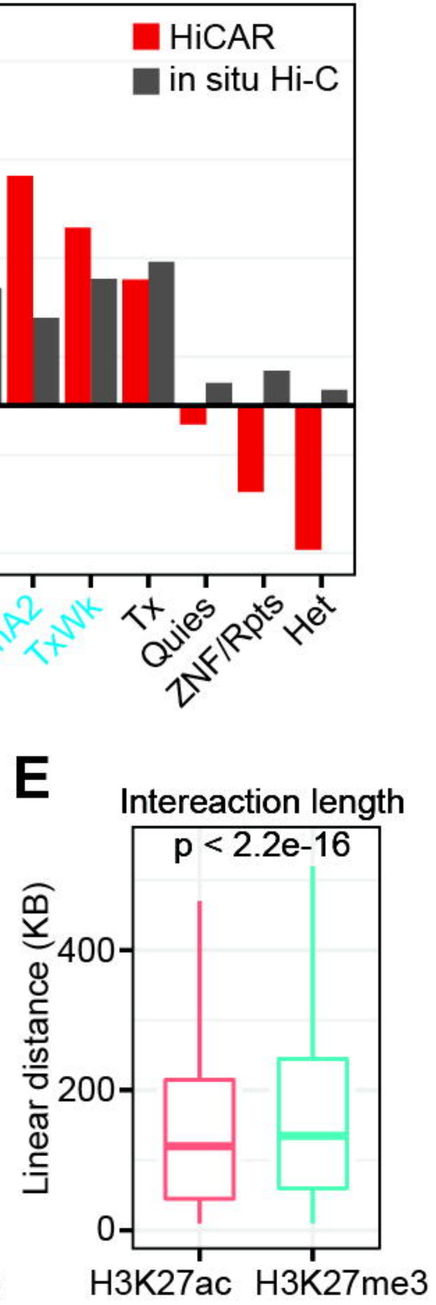


Figure 4





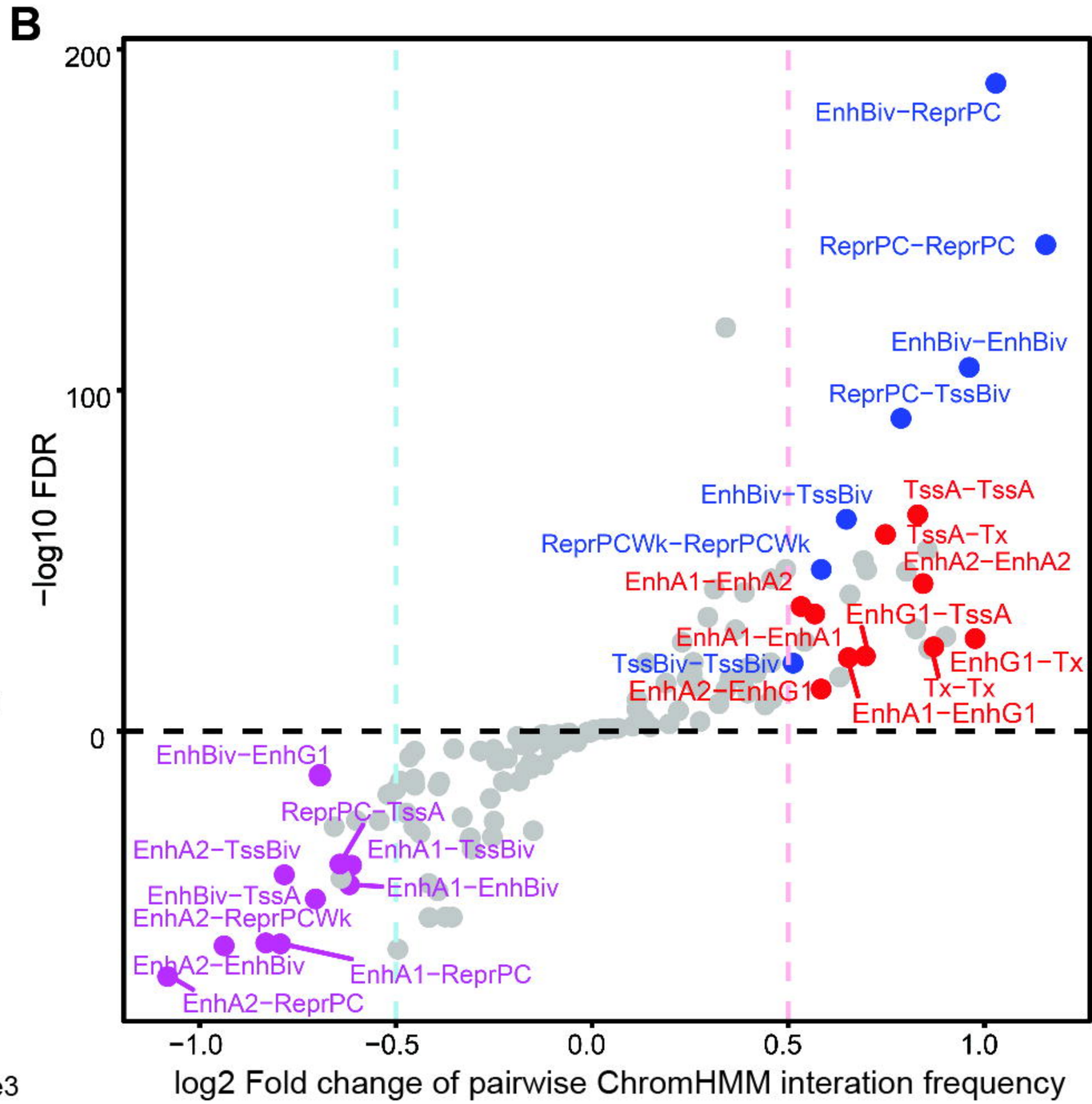
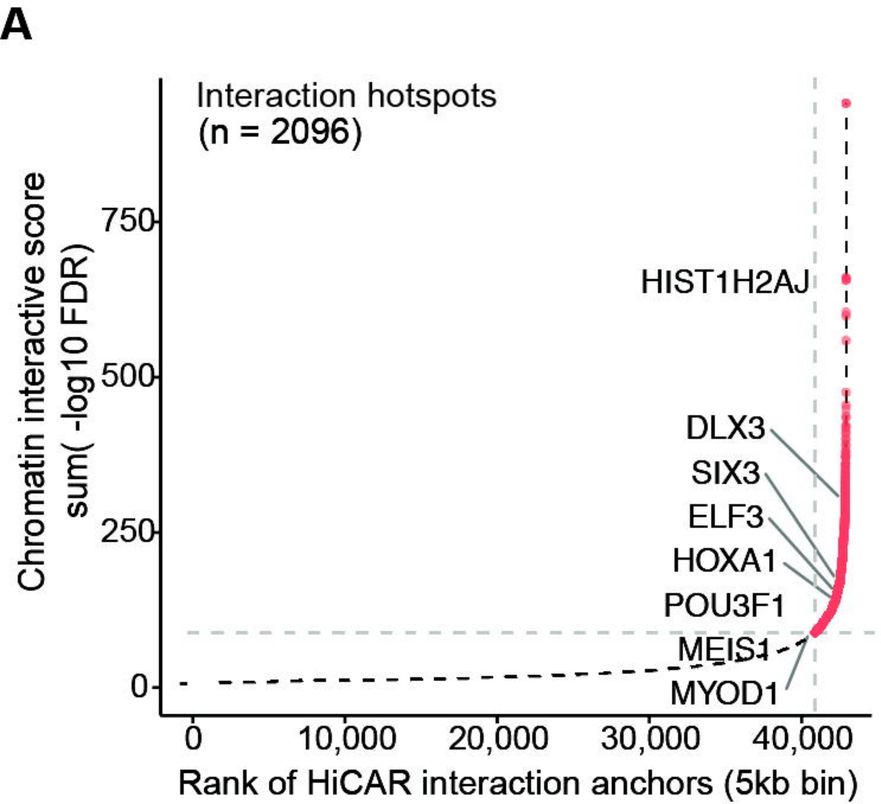
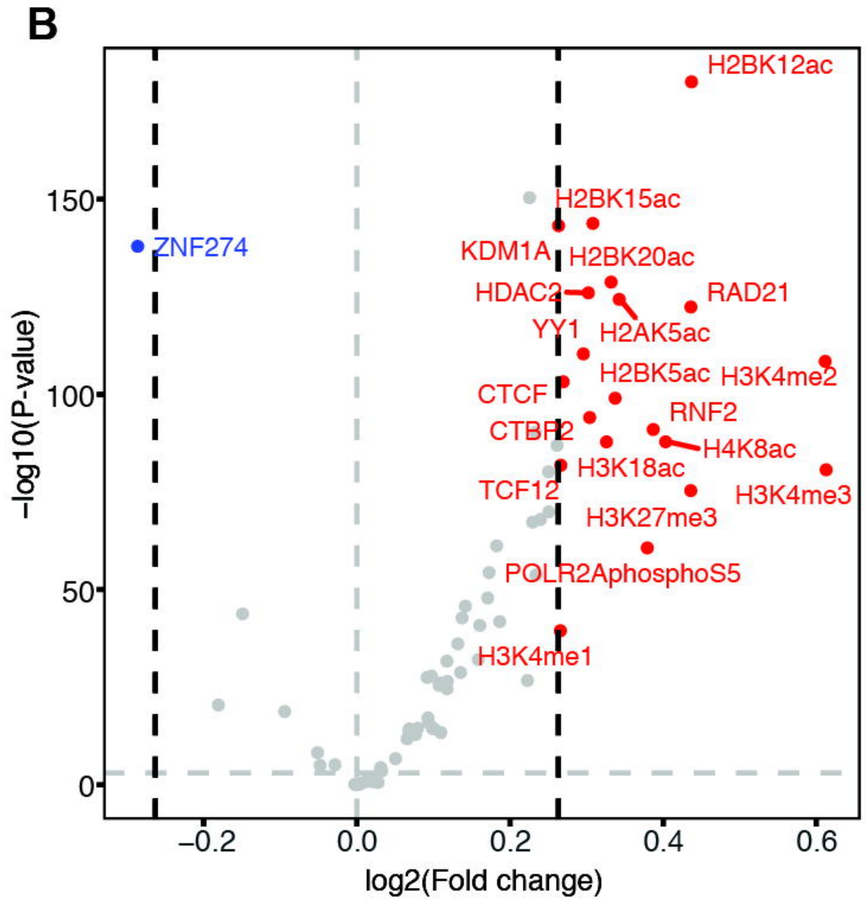
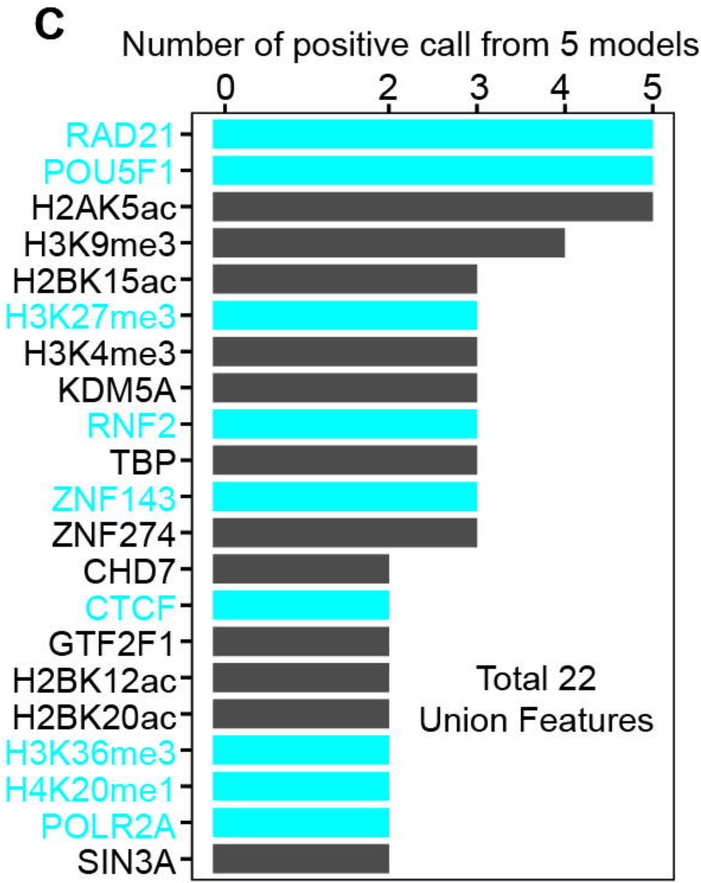


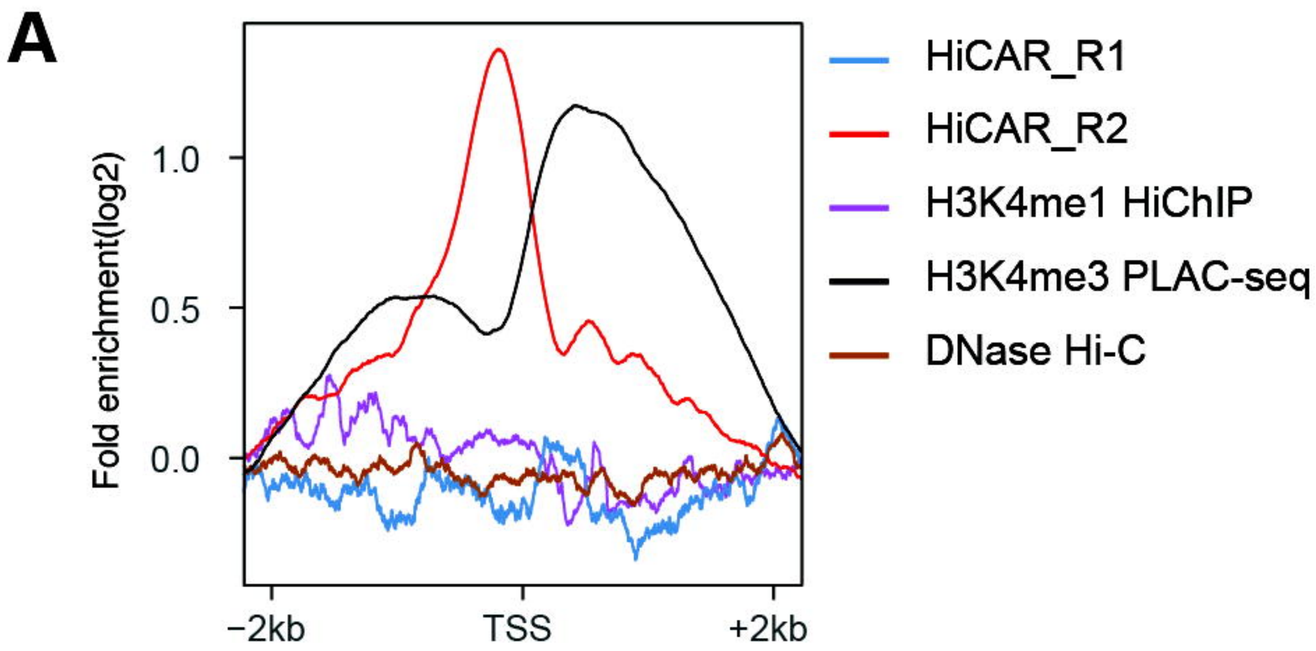
Figure 5

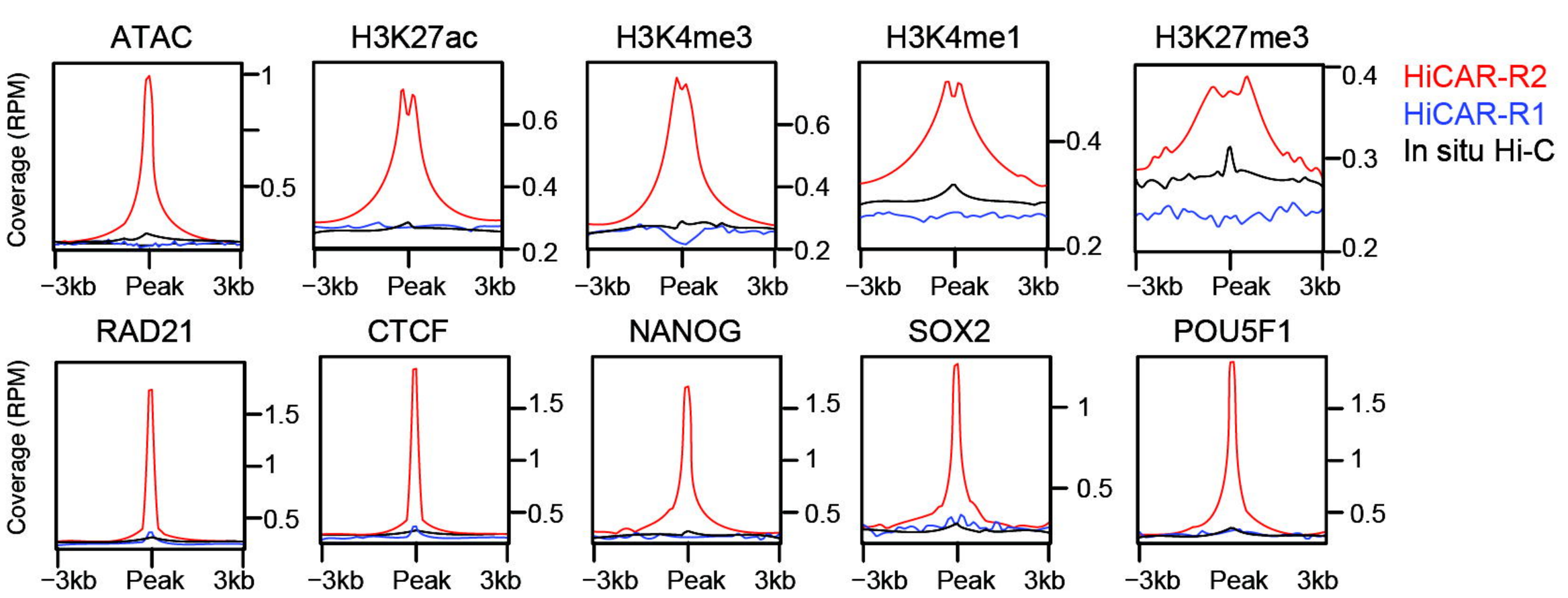


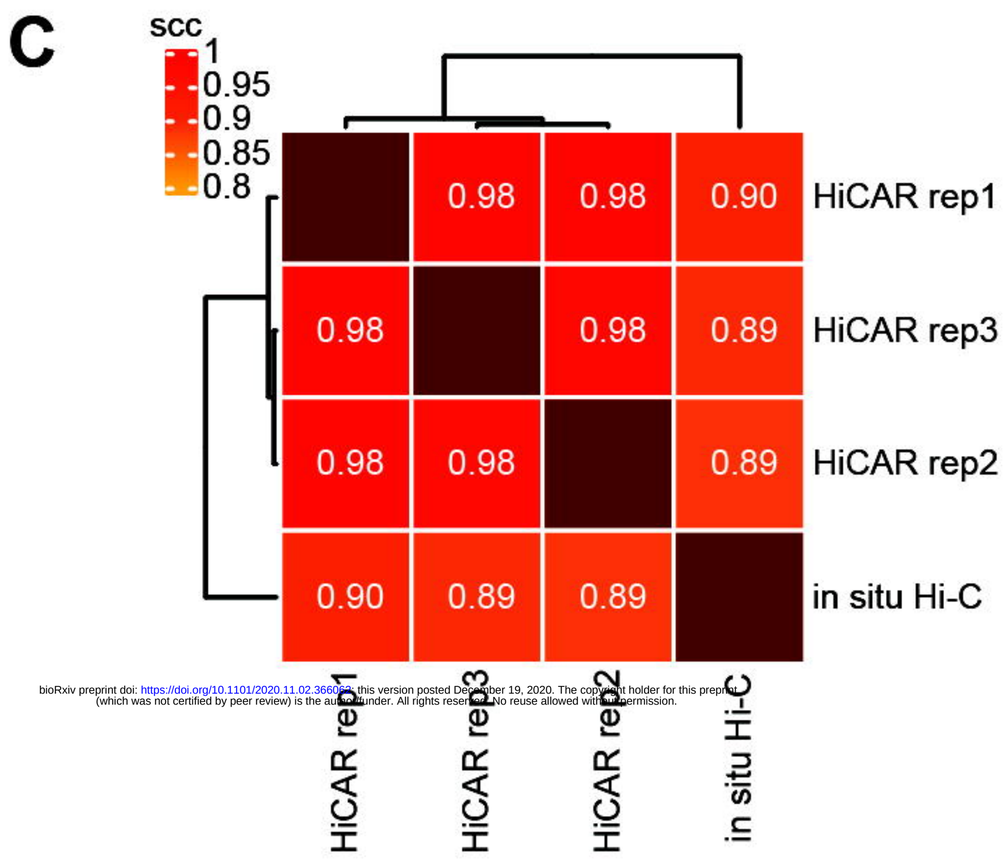


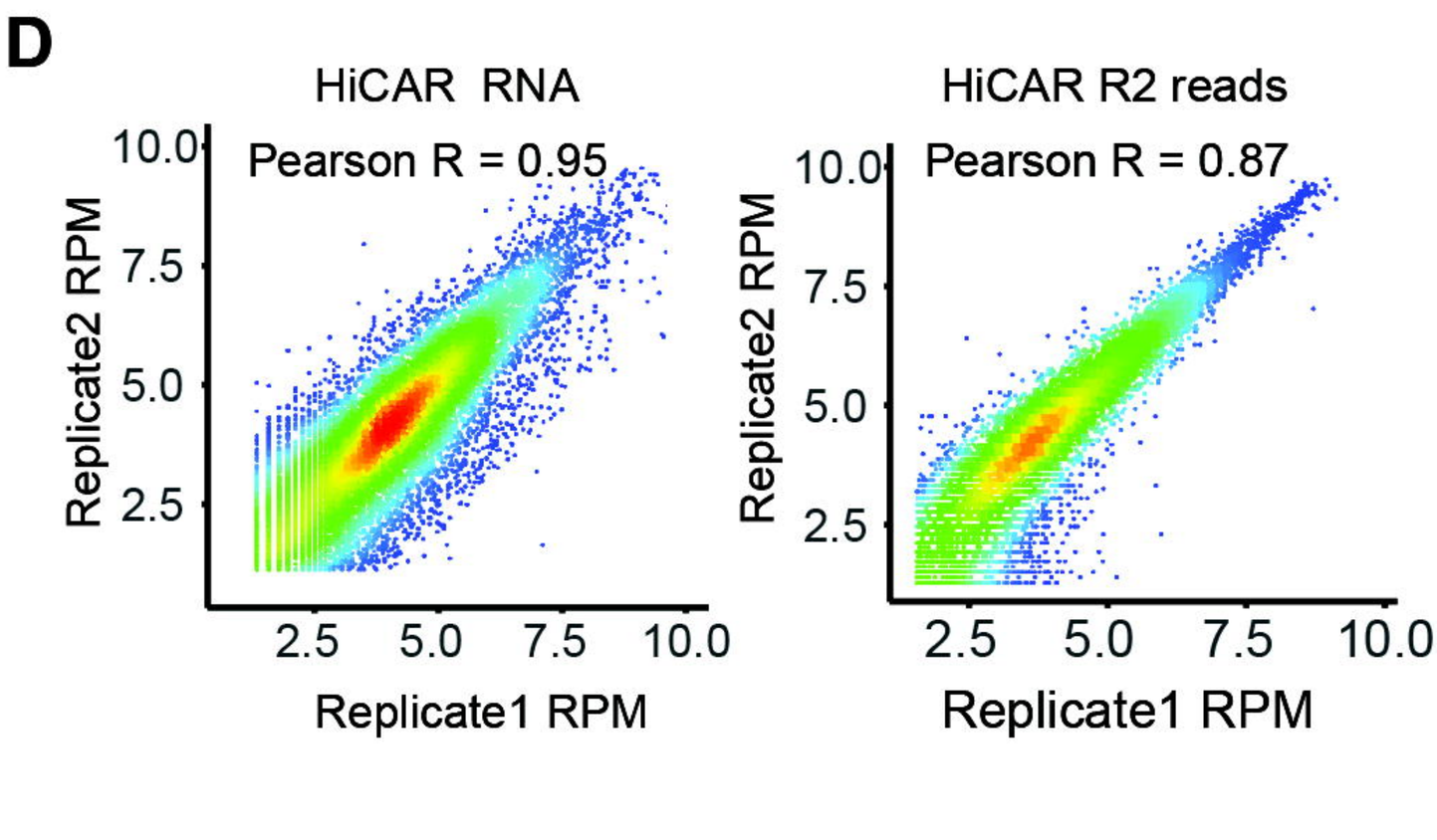


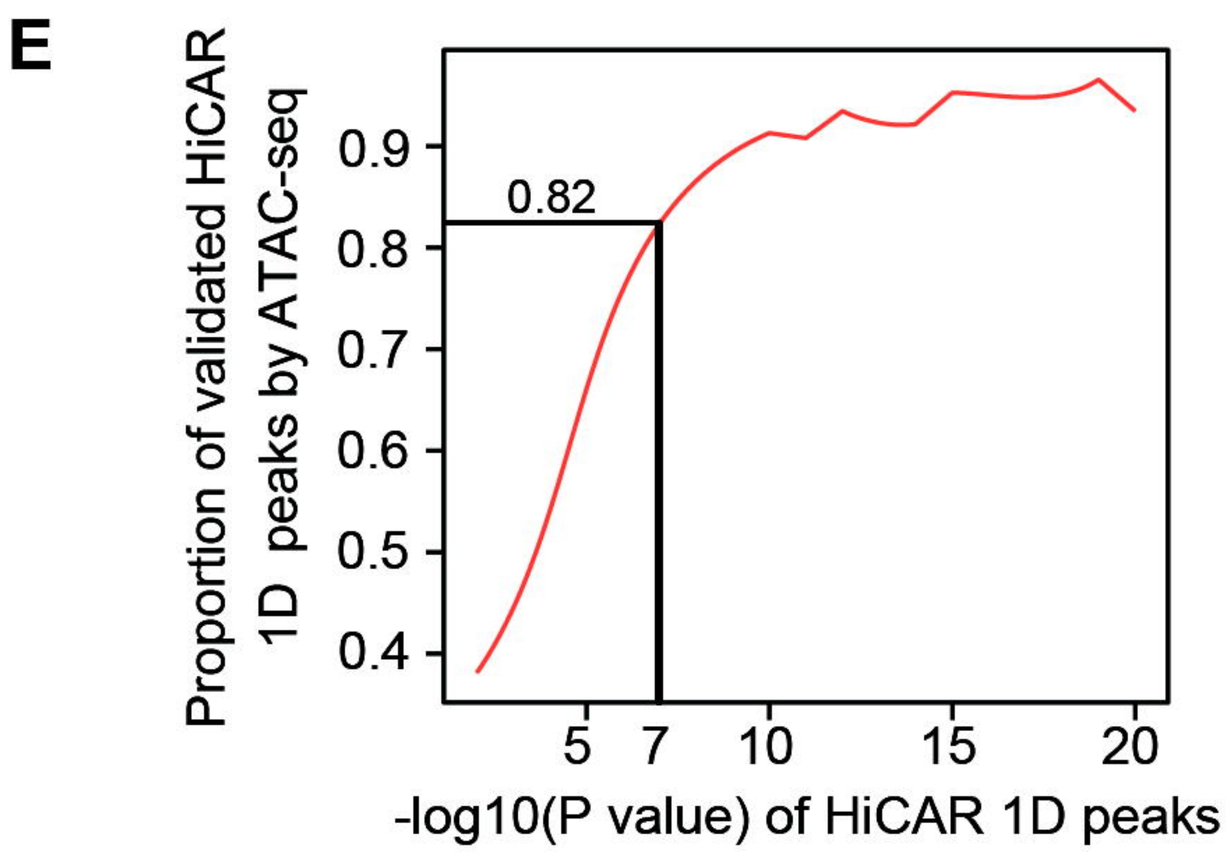
Supplementary Figure 1



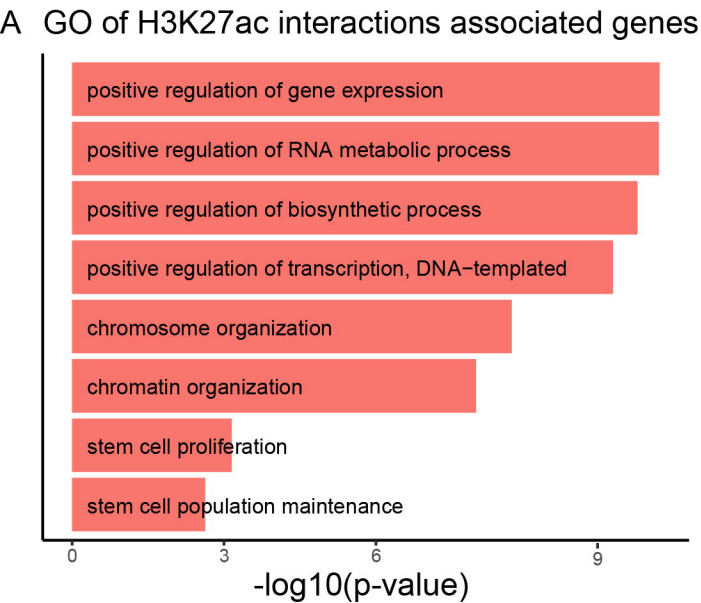


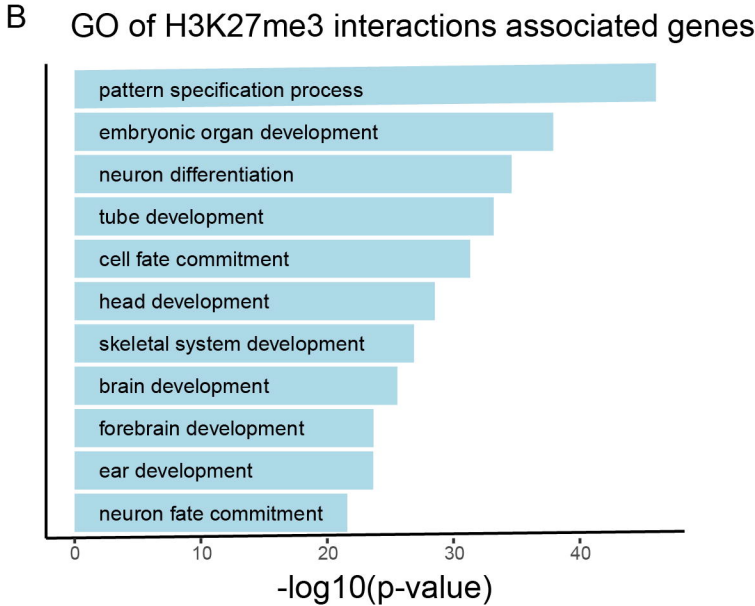




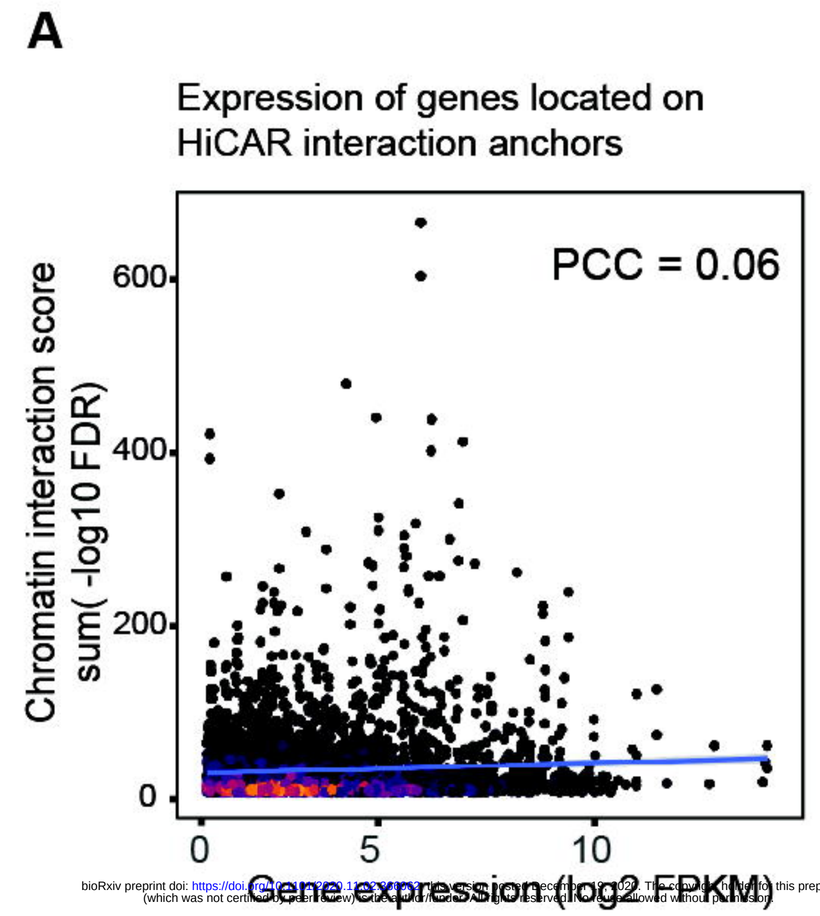


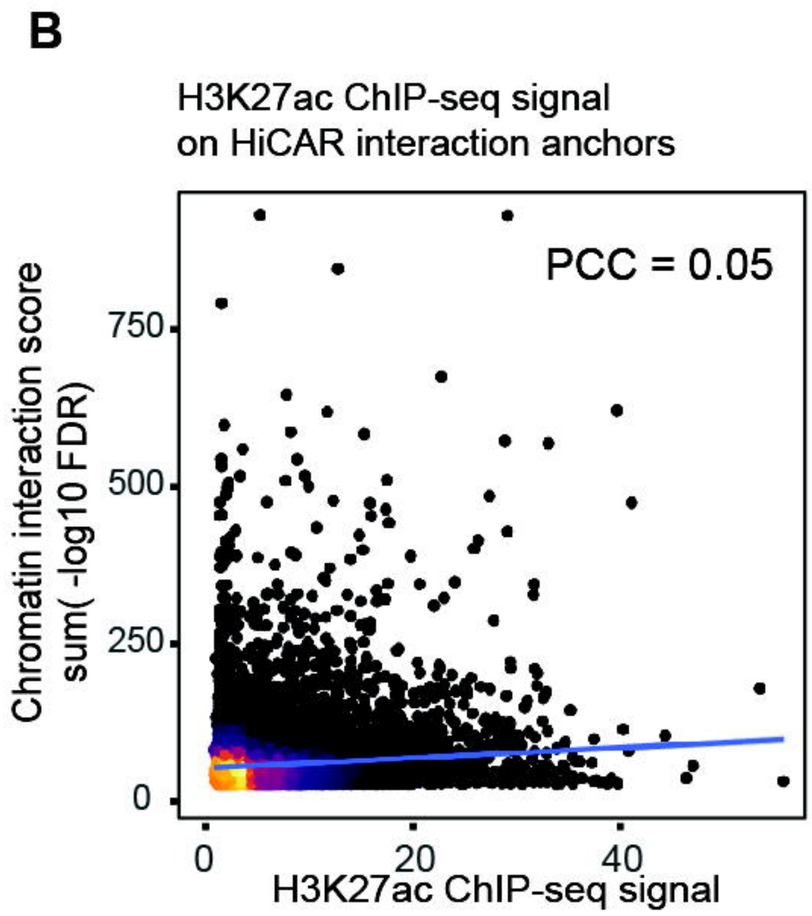
Supplementary Figure 2

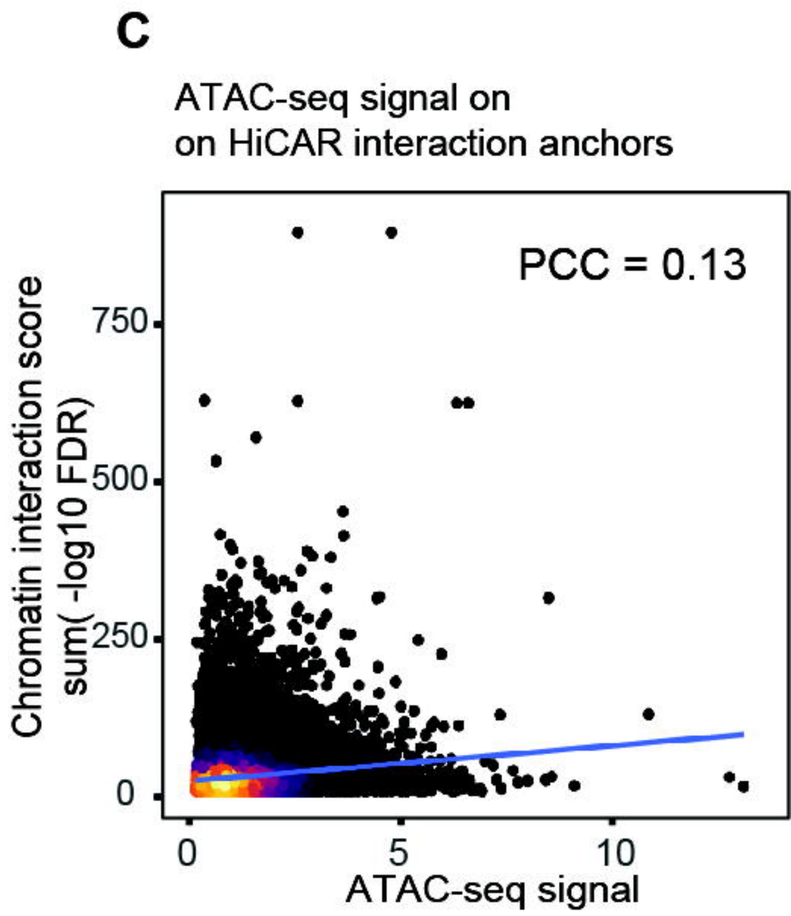


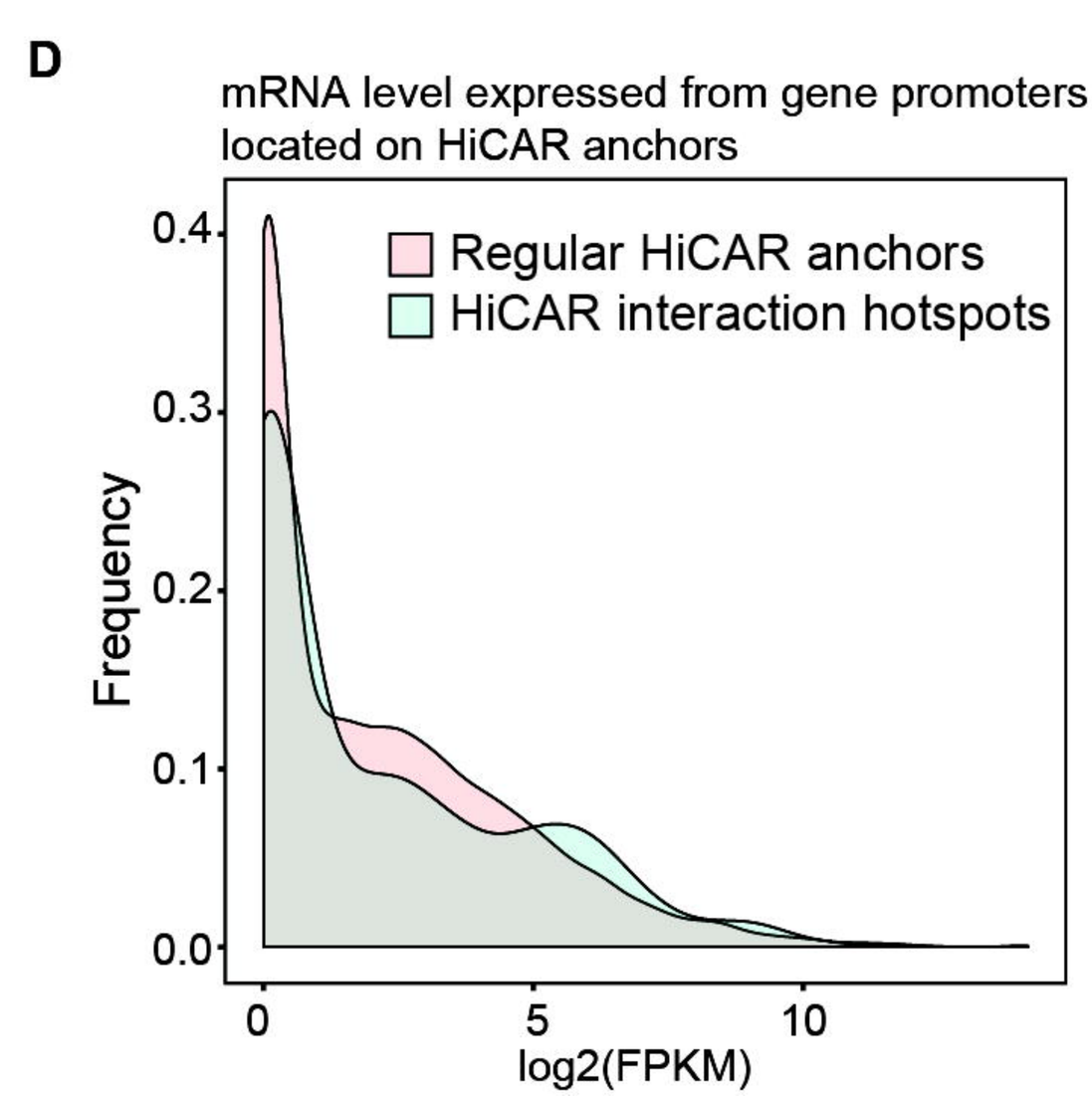


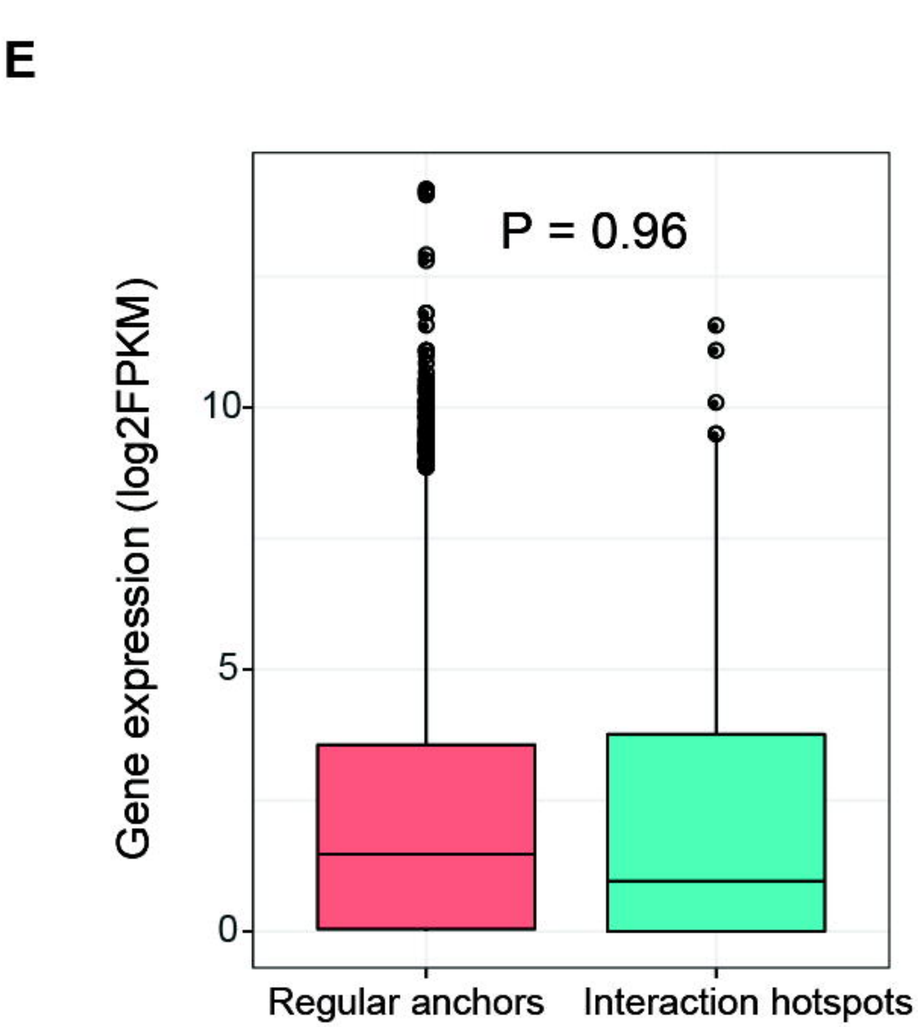
Supplementary Figure 3





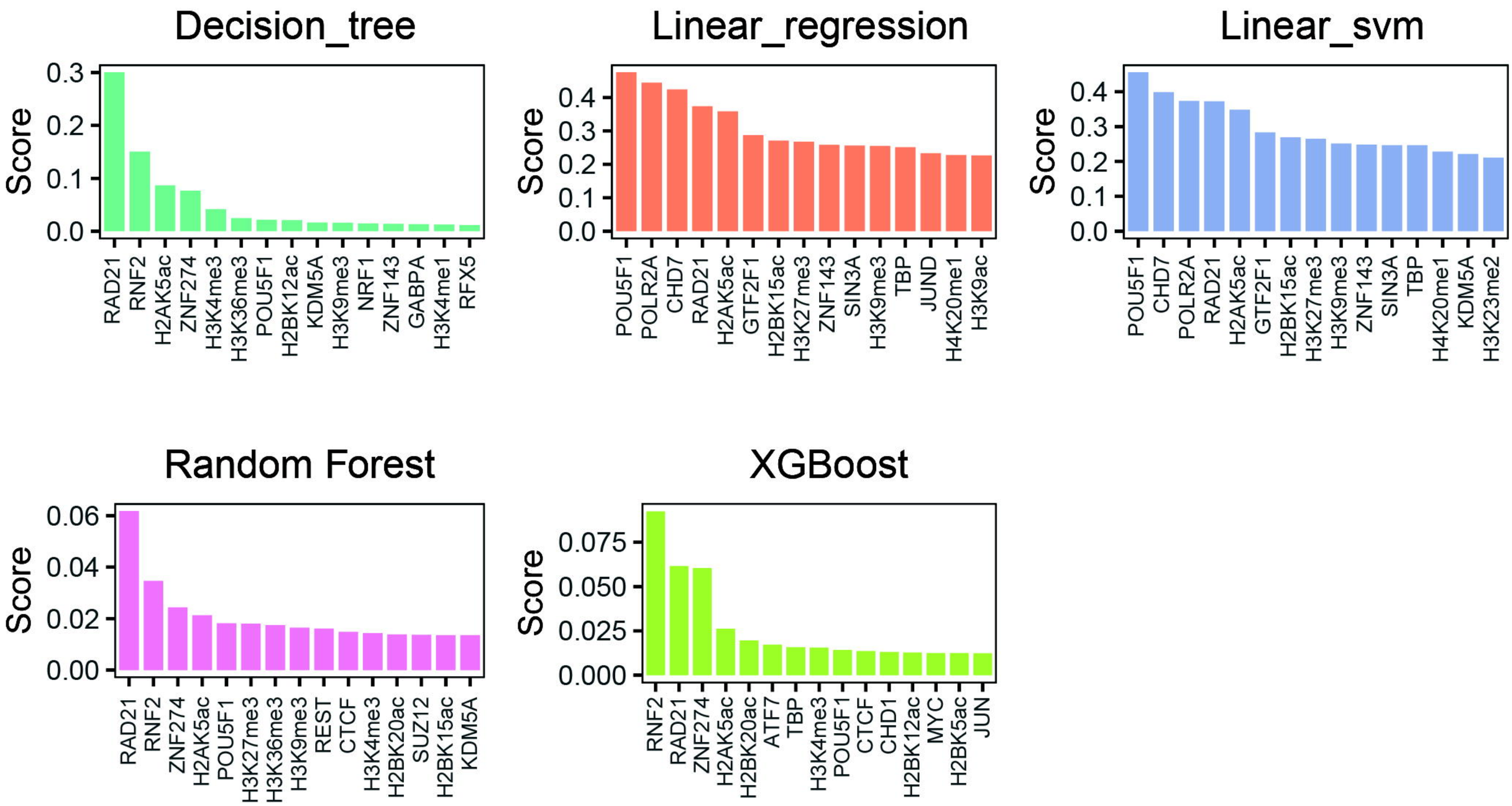




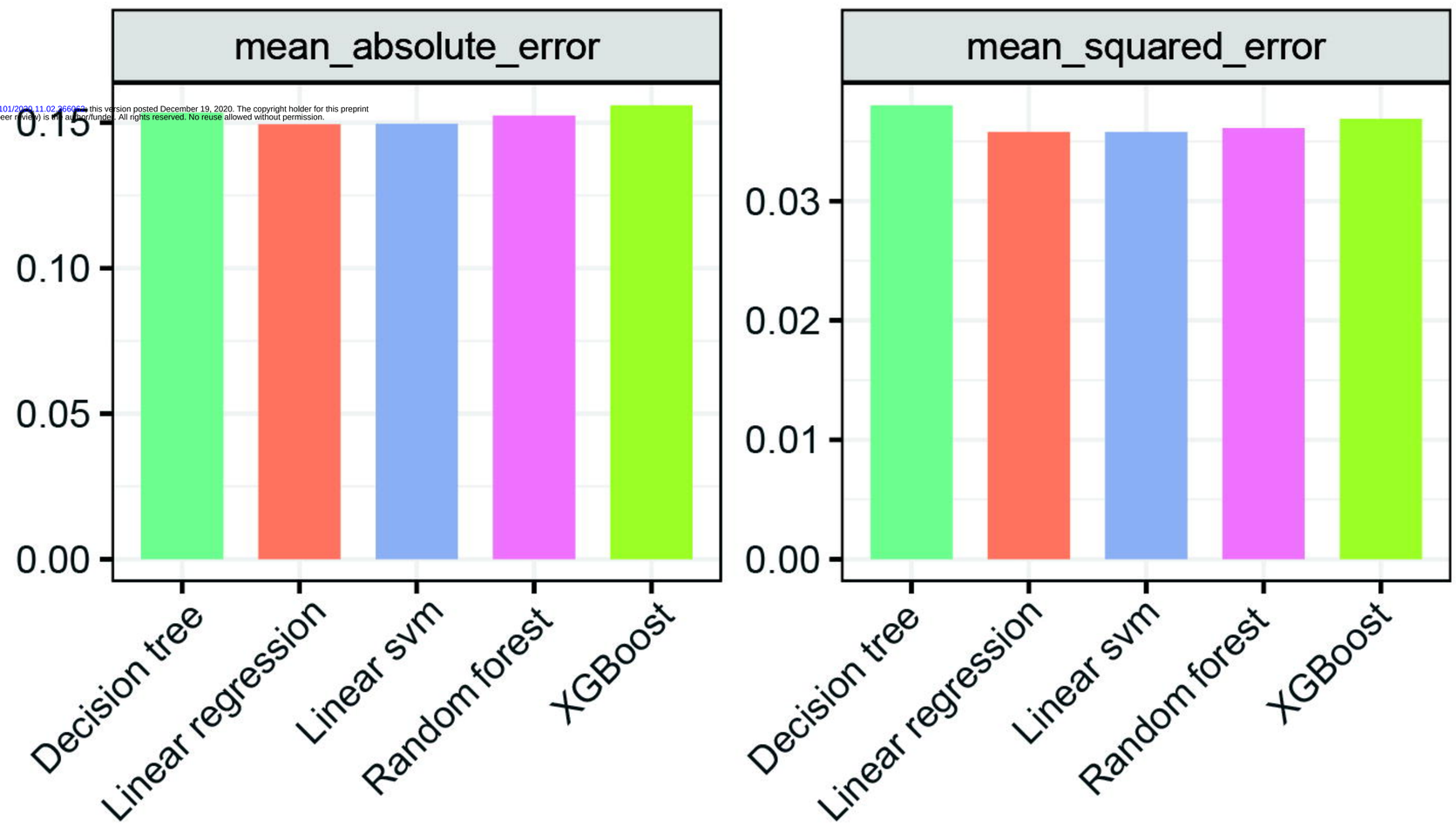


Supplementary Figure 4

A



B



Supplementary Figure 5

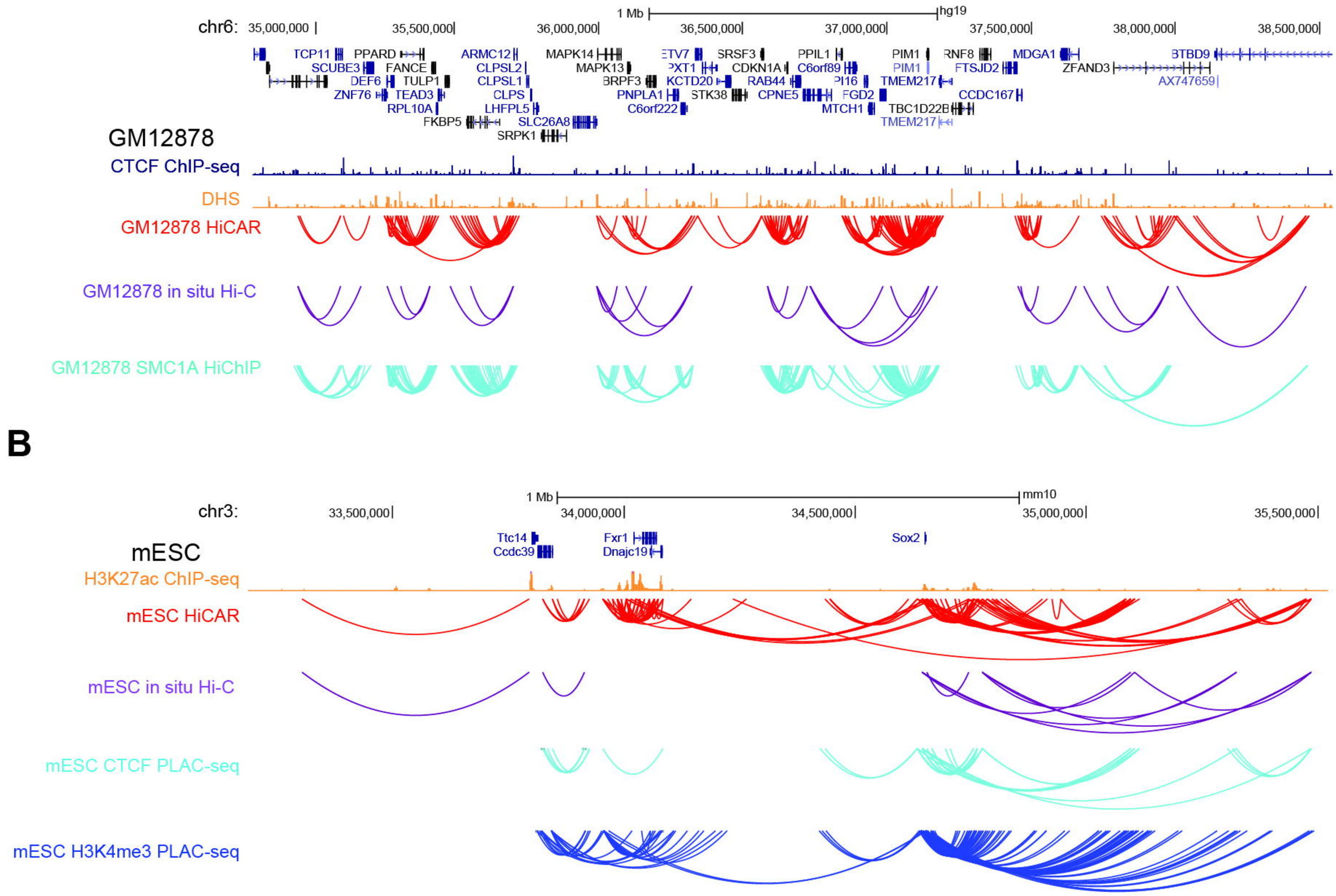
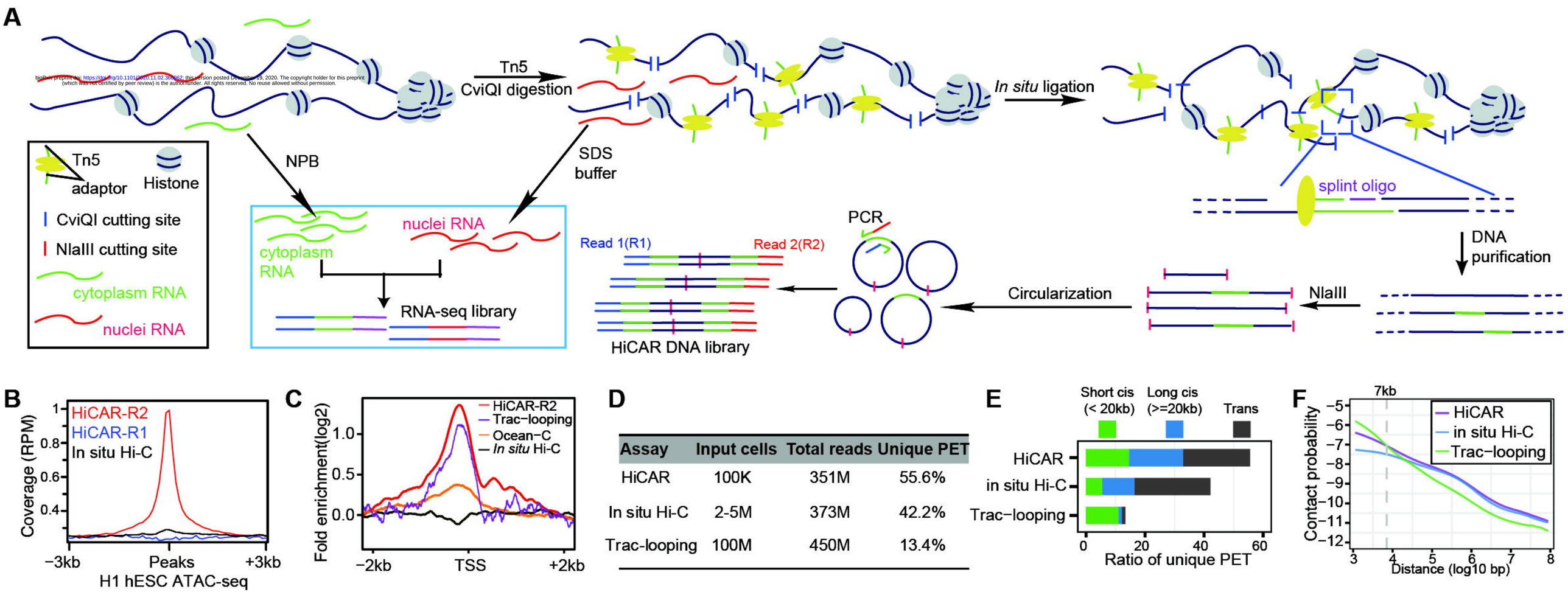


Figure 1



Assay	Input cells	Total reads	Unique PET
HiCAR	100K	351M	55.6%
In situ Hi-C	2-5M	373M	42.2%
Trac-looping	100M	450M	13.4%