CTCF mediated genome architecture regulates the dosage of mitotically stable mono-allelic expression of autosomal genes

Keerthivasan Raanin Chandradoss and Kuljeet Singh Sandhu*
Department of Biological Sciences
Indian Institute of Science Education and Research (IISER) – Mohali
Knowledge City, Sector 81, SAS Nagar 140306, India.

*To whom correspondence should be addressed
sandhuks@iisermohali.ac.in

Abstract
Mammalian genomes exhibit widespread mono-allelic expression of autosomal genes. However, the mechanistic insight that allows specific expression of one allele remains enigmatic. Here, we present evidence that the linear and the three dimensional architecture of the genome ascribes the appropriate framework that guides the mono-allelic expression of genes. We show that: 1) mono-allelically expressed genes are positioned in clusters that are insulated from bi-allelically expressed genes through CTCF mediated chromatin loops; 2) evolutionary and cell-type specific gain and loss of mono-allelic expression coincide respectively with the gain and loss of chromatin insulator sites; 3) dosage of mono-allelically expressed genes is more sensitive to loss of chromatin insulation associated with CTCF depletion as compared to bi-allelically expressed genes; 4) distinct susceptibility of mono- and bi-allelically expressed genes to CTCF depletion can be attributed to distinct functional roles of CTCF around these genes. Altogether, our observations highlight a general topological framework for the mono—allelic expression of genes, wherein the alleles are insulated from the spatial interference of chromatin and transcriptional states from neighbouring bi-allelic domains *via* CTCF mediated chromatin loops. The study also suggests that 3D genome organization might have evolved under the constraint to mitigate the fluctuations in the dosage of mono-allelically expressed genes, which otherwise are dosage sensitive.

Introduction
Though both copies of genes on autosomes have potential to be expressed, some genes escape the transcriptional activation of one of the alleles. These genes are known as mono-allelically expressed (MAE) genes. MAE genes can be genomically imprinted, i.e., the expressions of alleles are parentally fixed, or it can be random, i.e., any of the alleles can be expressed. A subset of random MAE genes has been shown to be mitotically stable in a clonal cell population, possibly through heritable epigenetic modifications of alleles[1]. It is intriguing that the random MAE genes have profound functional and evolutionary implications such as contributing to cellular diversity of immunoglobulins[2], interleukins[3, 4] and T-cell receptors[5]; ascribing choice of olfactory receptors in neurons[6] and increasing the evolvability of a locus[7]. Despite widespread presence of MAE genes in mammalian genomes, the underlying mechanisms that regulate allele-specific transcription remain poorly understood. Feedback mechanism in which expression of one allele induces the repression of other allele seems an interesting hypothesis. It has been shown that expression of a transgene odorant receptor in neurons leads to lack of endogenous expression of odorant receptor alleles, which is contrasting when compared to bi-allelically expressed genes which exhibit co-expression with the transgenes[8]. However, the mechanistic details of this phenomenon remain enigmatic. Directional switching of an upstream promoter can stochastically regulate the activity of downstream promoter of NK receptor gene in an allele-specific manner[9, 10]. Epigenetic mechanisms like differential chromatin states of alleles [11], non-coding RNA mediated repression[12, 13], distinct spatial localizations of alleles[14, 15] etc. have also been proposed and exemplified for certain MAE loci. While all these mechanisms are supported through experimental evidence, none of these could be generalized for most of the mono-allelic transcription of the genome. Moreover, any of the gene regulatory mechanisms implicated in regulating the two alleles distinctly would need a prerequisite of recognizing the MAE genes from the neighbouring BAE genes and one way this can be achieved is by insulating the MAE genes from BAE genes on either or both allelic loci. We, therefore, tested the hypothesis whether or not CTCF mediated insulation of chromatin domains implicate in guiding mono-allelic expression in the genome.

Results

We obtained experimentally identified mitotically stable MAE and BAE genes in human lymphoblastoid cell-line (hLCL), mouse lymphoblastoid cell-line (mLCL) and mouse embryonic stem cells (mESC) (Table 1). To further acquire the statistical robustness, we also included inferred MAE and BAE genes in hLCL, mLCL and mESC in the analysis (Table 1). To ensure that our analysis was not impacted by the already known properties, like clustering of imprinted genes, we removed the known imprinted loci from the present analysis. Known regulatory and functional differences of MAE and BAE genes prompted the hypothesis that the mono- and bi-allelic expression might be a domain property, of i.e., MAE genes might tend to segregate from BAE genes by clustering into chromosomal domains in a manner similar to imprinted genes. To test this, we calculated the density of MAE and BAE genes across chromosomes and identified regions enriched with MAE, BAE or both type of genes (Materials & Methods). We showed the gene clusters with either MAE or BAE genes were the majority as compared to ones with both, highlighting the preferred segregation of MAE and BAE domains (Figure 1a-b). Through randomizing the MAE and BAE labels of genes, we reported that the observed clustering of MAE and BAE genes was highly non-random (p-value=2.2e-16, Fisher's exact test, Materials & Methods). Further, to test if active and inactive alleles of MAE genes were also clustered separately, we obtained paternally active (PMAE) and maternally active MAE (MMAE) genes in hLCL (Table 1). Analysis of PMAE and MMAE suggested that active and inactive alleles did not exhibit random mixing and were mostly clustered separately (Figure S1a). These observations suggested that the linear gene order might have evolved under the evolutionary constraint to segregate MAE genes from BAE genes.

Given that the MAE genes taken in the analysis were mitotically stable, we further hypothesized that the clusters of MAE and BAE genes might be epigenetically insulated from each other. CTCF protein is presently the most popular candidate that serves as insulator between epigenetically distinct domains. We, therefore, tested the presence of CTCF binding at boundary of MAE and BAE domains. We obtained the CTCF binding sites that were associated with the insulator state in chromHMM annoations of hLCL and mESC genomes (Materials and Methods). We calculated Kronecker delta function ($\boldsymbol{\delta(x_i, x_{i+1})}$, where $\boldsymbol{x_i}$ is the allelic status of the gene $\boldsymbol{i}$) for consecutive genes around CTCF insulator sites. The function takes the value 1 if a MAE gene is followed by a MAE gene or a BAE gene followed by a BAE gene, otherwise the value remains zero. It was clear from the figure 1c that the allelic status of the gene changed after encountering a CTCF insulator site. This was also exemplified through numerous examples shown in the other panels of figure 1. Moreover, PMAE and MMAE genes also exhibited similar pattern as shown in figure S1b. We, therefore, conclude that the mono- or bi-allelic expression is the property of chromosomal domains, insulated by CTCF insulator sites, rather than individual genes.

CTCF orchestrates the genome in defined topological domains that are intervened by inter-domain or gap regions. To obtain further insight, we mapped the locations of MAE and BAE genes within gap-loop-gap architecture obtained from CTCF ChIA-PET data of hLCL and mESC. We observed that the MAE genes were preferably located inside the chromatin loop between insulator CTCF sites, while BAE genes exhibited lesser such preference (Figure 1d). This hinted that the insulator CTCF sites proximal to MAE genes might have been implicated and evolutionarily selected to maintain the mono-allelic expression. We, therefore, tested if the dynamics of allelic expression correlated with the gain and loss of insulator function of CTCF binding site in the proximity. Towards this, we compared the MAE and BAE genes in hLCL and K562 cell-lines. We first showed that the constitutive MAE genes (genes that maintained their mono-allelic status consistently in both cell-lines) exhibited consistently greater enrichment inside the chromatin loops mediated by insulator CTCF sites as compared to constitutive BAE genes in both the cell-lines (Figure 2a). Interestingly the genes that were MAE in hLCL, but had bi-allelic expression in K562 cell-line, had significant enrichment inside the chromatin loop in hLCL, but not in K562 cell-line (Figure 2b). This pattern reversed for the genes that were mono-allelic in K562 but had bi-allelic expression in hLCL (Figure 2b). The observed gain and loss of enrichment of MAE genes inside the chromatin loops mediated by CTCF insulator can be explained by the difference in CTCF mediated loops and the relative chromatin context of CTCF binding sites in the two cell-lines. This was illustrated through examples: 1) CTCF mediated loop around a locus that was MAE in hLCL, was not observed in K562 where the gene was expressed bi-allelically. By plotting RNAPII ChIA-PET data, we clearly observed lack of insulation and gain of abundant enhancer-

promoter and promoter-promoter interactions with the neighbouring regions in K562 cell-line (Figure 2c, left panel). 2) CTCF mediated loop remained intact in both the cell-lines, but the chromatin context of CTCF binding differed. In the cell-line where the gene was expressed bi-allelically (hLCL in this case), the promoter of the gene gained additional CTCF site, which was engaged with the other CTCF and non-CTCF sites associated with enhancer chromatin states (Figure 2c, right panel). It was interesting to note that the CTCF binding sites that function as insulator between MAE and BAE genes in one cell-line, can function as enhancer-linker in the other cell-line. Indeed, BAE genes were significantly associated with the enhancer-linking CTCF sites, as compared to MAE genes (Figure S2). The observed correlation between mono-allelic status of genes and their localization near CTCF insulator sites strongly supported the role of CTCF insulators in maintaining mono-allelic expression of genes. We also confirmed the above observations by comparing hLCL with the HMEC cell-line (Figure S3).

Further, to assess whether or not gain of CTCF insulator sites near MAE genes was evolutionarily selected, we compared the MAE and BAE genes of mLCL with that of hLCL. As shown in the figure 2d-e, the genes that maintained their mono-allelic expression in human and mouse LCLs were consistently associated with the CTCF insulator sites in the proximity. However, the evolutionary loss/gain of mono-allelic expression coincided with the loss/gain of insulator sequence in the proximity. These observations highlighted that the genetic and the epigenetic association with the CTCF insulator sites serve as a prerequisite for the mono-allelic expression of the genes. Since the loss of CTCF insulator function coincided with the gain of bi-allelic expression, it could also be inferred that CTCF insulator function was associated with the repressed alleles of MAE genes.

With the recent availability of genome-wide CTCF depletion datasets, it is now possible to explore if expression and insulation of certain predefined subset of genes is affected by the loss of CTCF function. We obtained CTCF depletion datasets for mLCL and mESC (Materials and Methods). We first showed that the CTCF depletion disrupted the insulation of MAE and BAE domains significantly (Figure 3a). More importantly, the disruption was more striking for MAE genes. We further showed that the dosage of MAE genes was strikingly more sensitive to CTCF depletion as compared to that of BAE genes (Figure 3b-c). It was also noticeable that the MAE genes that were upregulated after CTCF depletion outnumbered the ones that were downregulated when compared to BAE genes, again suggesting that the repressed allele was likely to be associated with insulator mediated chromatin loops (Figure S4). The up- and down-regulation of MAE genes coincided with the contrasting chromatin and transcriptional states in the neighbouring domains. The chromatin loops enclosing upregulated genes had greater enrichment of repressive marks within and active marks (active promoters and strong enhancers) in the neighbouring chromatin domains, while loops with downregulated genes had active marks within and repressive marks in the adjacent domains. Concomitantly, the presence of these marks correlated with the associated transcription levels of the genes, suggesting that the lack of insulation between neighbouring chromatin domains might have impacted the allelic status of MAE genes (Figure 3d-e).

It was not entirely clear why the dosage of MAE genes was more sensitive to CTCF depletion. We suspected that the insulator function of CTCF might be more susceptible to CTCF depletion as compared to enhancer-linking. To test this, we compared the interaction frequencies of CTCF mediated enhancer-promoter loops with that of insulator anchored loops before and after CTCF depletion in mESC. We observed relatively lesser alteration in interaction frequencies of enhancer-linker sites after CTCF depletion as compared to the loops anchored to insulator sites (Figure 3f), suggesting that the minimal amount of CTCF was sufficient to link enhancer to their cognate promoter. Accordingly, the genes associated with insulator loops were more susceptible to CTCF depletion as compared to the ones associated with enhancer-linking loops (Figure 3g). Since MAE genes were flanked by insulator sites, while BAE genes were enriched near enhancer-linking CTCF sites, it could be concluded that distinct functional roles of CTCF around MAE and BAE genes might explain distinct susceptibility of MAE and BAE genes to CTCF depletion (Figure 3h and figure S5).

Discussion

CTCF mediated chromatin folding is known to implicate in maintaining allele-specific transcriptional states of H19-Igf2 imprinted locus. Loss of CTCF binding at H19-ICR leads to loss of maternal

repression of Igf2, which otherwise is insulated in a repressive loop mediated by CTCF[16]. Our observations suggested certain level of generality of insulation of inactive allele during mono-allelic expression. Indeed, we tested whether CTCF mediated chromatin loops were associated with the repressed alleles of MAE genes in a manner that was analogous to allele-specific repression of Igf2. Towards this, we obtained the haplotype resolved HiC data from Rao et al [25]. We first showed that the CTCF barrier loops associated with MAE genes exhibited greater variation between homologous chromosomes as compared to the ones associated with BAE genes, suggesting allele-specificity of chromatin conformation associated with MAE genes (Figure S6a). We inferred the maternally and paternally expressed alleles by making use of RNAPII ChIA-PET data (Materials and methods). By comparing the CTCF associated allele-specific chromatin interactions around maternally and paternally expressed genes, we showed that the inactive allele had relatively higher interaction frequency of CTCF mediated chromatin loop as compared to active allele, reconciling our proposal that the CTCF associated insulation was mostly associated with the inactive alleles (Figure S6b). However, due to subtle statistical difference seen in the analysis, we do not entirely deny the possibility that the upregulation of MAE genes upon CTCF depletion might not necessarily relate to repressive allele and instead the active allele might get further upregulated. Indeed, despite the fact that CTCF binding at H19-ICR is critical to establish and maintain proper H19-Igf2 imprinting, depletion of CTCF protein itself does not associate with the loss of mono-allelic expression of Igf2 and instead causes increased expression of the active allele itself, possibly *via* altering CTCF binding at other nearby sites[17]. This suggested that the minimal amount of CTCF is sufficient to maintain the imprinting at H19-Igf2 locus and that the dosage of imprinted gene can be susceptible to CTCF depletion in a non-allelic manner. We, therefore, largely restrict our claim to dosage sensitivity of MAE genes to the loss of CTCF mediated insulation around the locus.

Despite being relatively fewer in numbers, explanation to down-regulated genes was needed. It can be interpreted that the down-regulation was that of active allele. There can be following possibilities: 1) Association with the CTCF mediated chromatin loops, though statistically significant, might not always be related to repressed allele and at certain MAE loci, it might associate with active allele instead. Indeed, as shown in the figure 3e, the loops with downregulated MAE genes had significant enrichment of active promoters and strong enhancers within and repressive chromatin states in the neighbouring chromatin loops, which was in sharp contrast to loops with upregulated MAE genes, suggesting that the increased spatial interference with the neighbouring repressive domains might cause down-regulation of active alleles of MAE genes; 2) Gene regulatory network downstream to dysregulated transcription factors can also be a possibility. Regardless of the fact whether the MAE gene was upregulated or down-regulated or whether the change was allelic or non-allelic, significantly greater dosage sensitivity of MAE genes to CTCF depletion as compared to BAE genes is a novel and non-trivial observation, which has implication in understanding the constraints that shape the evolution of genome architecture. It has been hypothesized earlier that the MAE genes are likely to be dosage sensitive[18]. We tested this hypothesis using dosage sensitivity and copy number variation data in human. MAE genes had significantly greater overlap with the dosage sensitive genes obtained from ClinGen database (Materials & Methods, Figure S7a). Accordingly, BAE genes had greater overlap with the regions exhibiting common copy number variation in human (Figure S7b), suggesting their general insensitivity to dosage. Therefore, it can be concluded that evolutionary selection of appropriate dosage of MAE genes might have constrained the CTCF mediated chromatin insulator architecture of the genome.

Nora et al recently reported that the CTCF depletion impacts the transcriptional states but the chromatin states, as measured through H3K27me3, remains largely unaltered [19]. We also confirmed this in the context of the present study (Figure S8). Though the insulation and expression levels were significant altered, H3K27me3 levels largely remained unaltered after CTCF depletion. Accordingly we suggest that the loss of insulation of TAD boundaries might have introduced spatial interference of opposite chromatin and transcriptional states in the neighbourhood in the form of altered enhancer-promoter and promoter-promoter interactions across TAD boundaries, which might have impacted the expression of MAE genes. The BAE genes, on the other hand, were mostly flanked by the CTCF sites involved in enhancer-linking function, which surprisingly was relatively robust against CTCF depletion as compared to insulation. As a result, BAE genes exhibited comparatively lesser transcriptional dysregulation upon CTCF depletion.

Altogether, our analysis highlighted the significant dosage dependency of mitotically stable mono-allelically expressed genes, as compared to that of bi-allelically expressed genes, on the insulator function of CTCF protein. The correlation between evolutionary gain and loss of CTCF insulator sites with the gain and loss of mono-allelic expression together with the observation that the MAE genes are more dosage sensitive suggested that maintenance of mono-allelic expression of genes might have served as one of the potent evolutionary constraints that have shaped linearly and spatially compartmentalized genome organization. Availability of data on allelically biased expression of genes in other mammalian species would allow to infer the ancestral status of allelic expression in future, which would endow greater evolutionary insights to the phenomenon.

Materials and Methods

Datasets

Numbers and source of experimentally identified MAE and BAE genes and the ones inferred based on equal enrichment of H3K36me3 and H3K27me3 marks in human LCLs, mouse LCLs and mouse ESCs are given in the table 1. Chromatin insulator and other ChromHMM annotations for hLCLs and mESCs were taken from Ernst et al[20] and Yue et al [21] respectively. Histone modification data was obtained from ENCODE [22]. CTCF ChIA-PET datasets for hLCLs and mESCs were obtained from Tang et al[23] and Handoko et al[24] respectively. Haplotype resolved HiC dataset was obtained from Rao et al[25]. CTCF depletion datasets for mLCLs and mESCs were obtained from GSE98507 (unpublished) and Nora et al[19]. Dosage sensitive genes and CNV data were obtained from ClinGen resource (https://www.ncbi.nlm.nih.gov/projects/dbvar/clingen/) and Database of Genomic Variants (http://dgv.tcag.ca/dgv/app/home) respectively.

Analysis of genomic attributes

These genomic attributes like GC content, repeat elements, non-canonical DNA structures, conserved noncoding elements etc were mapped around +/- 50kb of MAE and BAE genes by taking 2kb bin-size. Repeat elements were downloaded from UCSC for hg19 and mm10 genome assemblies hg19/mm10. Non-B-DNA structures and conserved non-coding elements were taken from https://isp.ncifcrf.gov/isp/nonb_dwnld and Marcovitz et al., 2016 respectively. Heatmaps were plotted using the 'heatmap' function with default scaling in R-package.

Linear clustering of genes

hg19 and mm10 genomes were binned into 100kb windows. Density of MAE and BAE genes was calculated in those bins. It was then smoothed using 'running.mean' function with binwidth=3 from the "igraph" package and a linear regression line was called using 'lm' function in R. Bins exhibiting gene density above the linear regression line were defined as gene-clusters. Bins having only MAE (or BAE) were called as MAE (or BAE) bins. Bins with both MAE and BAE genes were called as 'mixed' bins. For re-sampling, 9332 and 10012 genes were picked randomly and termed as MAE and BAE genes respectively. The above procedure was followed to classify bins as MAE, BAE and mixed bins. P-value was calculated by Fisher's exact test for original and randomized dataset.

Kronecker delta calculation for CTCF insulation

Five genes upstream and five genes downstream were mapped around each barrier CTCF site. Kronecker delta was calculated for the pair of consecutive genes moving from upstream to downstream direction. If two genes were having same allelic status, value "1" was given. If one was MAE and other was BAE, value "0" was given. It was done for all insulator CTCF sites. Then the values were scaled between 0 to 1 and the average value of Kronecker delta was plotted.

CTCF gap-loop-gap analysis

CTCF ChIA-PET loops which were having at least three PET pairs and spanned up to 1mb, were taken as per recommendations from Fullwood et al[26] and Li et al[27]. CTCF ChIA-PET loops that overlapped with insulator CTCF sites were taken for the analysis. Relative enrichment of TSS sites of MAE and BAE genes in the loops and the flanking regions, of same length as that of loops, was calculated. Average aggregation values for all loops were normalized with total gene count

accordingly. These final values were then scaled using (x-minimum/maximum-minimum) and average values were plotted.

Analysis of conserved and variable allelic status of genes

To compare the allelic status between cell-lines, we obtained the insulator sites for K562 cell-line (breast epithelial) and compared with that of hLCL. MAE and BAE genes that maintained their allelic status in both cell-lines and the ones that switched from mono-to-biallelic expression and vice-versa in two cell-lines were assessed for the presence of insulator binding in the proximity (<20kb). For human-mouse comparison, orthologous gene information was taken from Ensembl. Orthologous insulator CTCF sites of human (hg19) in mouse (mm10) were obtained using UCSC liftover (with minimum ratio of bases that must remap as 0.1). Proximal presence of insulator sequence was assessed for MAE and BAE genes that maintained their allelic status in both human and mouse LCL cell-line and the ones that switched from mono-to-bi-allelic expression and vice-versa in two species (<20kb). . P-values were calculated by Fisher's exact test.

Analysis of allele-specific chromatin loops

MAE and BAE genes from human LCL cell-line were assessed for the difference in the chromatin interaction frequency of insulator occupied sites. Due to the lack of public availability of haplotype-resolved CTCF ChIA-PET data, we overlaid CTCF ChIA-PET loops for GM12878 from Tang et al [23] onto haplotype-resolved Hi-C data for GM12878 (resolution: 5kb, normalization: VC) from Rao et al[25]. These CTCF ChIA-PET loops were having at least three PET pairs and were up to 1Mb in length [26, 27]. Chromatin loops with only MAE's and only BAE's gene TSSs were classified as MAE and BAE loops respectively. Squared difference between maternal and paternal loci was calculated and then divided by the maximum of their interaction frequencies for normalization. P-value was calculated by one tailed Mann Whitney U test.

Since allele-specific expression data for the same clone of LCL cell-line was not available in Rao et al's article, we used RNAPII ChIA-PET data[23] to infer active and inactive alleles in Rao et al's HiC data[25]. RNA-pol2 ChIA-PET loops for GM12878 were overlaid onto maternal and paternal Hi-C datasets. Maternal-to-paternal ratio of HiC interaction frequencies for RNAPII ChIA-PET loops was calculated for each TSS site. The upper and lower quartiles of the ratio were then taken as maternally and paternally biased genes. For these M-biased and P-biased genes, allele-specific interaction frequencies of CTCF mediated loops were obtained from the HiC data. Maternal-to-paternal ratio of interaction frequencies of CTCF mediated loops was calculated and viewed as boxplots. P-value was calculated using one tailed Mann-Whitney U test.

CTCF depletion analysis

Hi-C data (.cool format, resolution: 20kb, mm9, untreated and 2 days auxin treated) were taken from Nora et al., 2017[19]. Dip prominence scores was used as provided by the authors. Higher dip prominence signified highly insulating boundaries. Dip prominence was mapped to 20Kb to the TSSs of MAE and BAE genes. Hi-C data (.cool format) was extracted using Cooler (https://github.com/mirnylab/cooler). 'Heatmap' function of R-package was used to plot TAD domains.

To analyse the transcriptome, SRA files for LPS induced CTCF depleted B-cells were downloaded from GSE98507 and were converted into fastq format using fastq-dump (SRA Toolkit). Pilot-run fastq files were not used for further analysis. fastq files were mapped onto mouse reference genome (mm10) and calculated differential gene expression between control and CTCF depleted samples using tophat and cufflinks without new gene/transcript discovery[28]. For mESC, RPKM values for control (untreated) and 2 days auxin-treated CTCF depleted cells were downloaded from GSE98671 [19]. Fold change was calculated with respect to control cells. Distributions of fold change (FC) of up-regulated (log2 FC > 0) and down-regulated (log2 FC < 0) MAE and BAE genes were plotted as boxplots. P-values were calculated by Mann-Whitney U test. For the analysis of expression in the neighbouring domains MAE genes with at-least 1.5 fold up and downregulation were used The nearest neighbour was taken on both sides and their RPKM values before CTCF depletion were plotted as boxplots.

Table legends

Table 1. Overall statistics of MAE and BAE genes taken for the analysis

Figure legends

Figure 1. Genomic compartmentalization of MAE and BAE genes. (a) Pie charts showing the distribution of clustered MAE, BAE and mixed (MAE/BAE) genes in different cell-lines. hLCL: human lymphoblastic cell-line, mLCL: mouse lymphoblastic cell-line, mESC: mouse embryonic stem cells. (b) Examples illustrating the linear compartmentalization of MAE and BAE genes. Chromosome coordinates are given for hg19 assembly. (c) Scaled average aggregation plot of Kronecker delta function over five consecutive genes upstream and downstream to insulator CTCF sites. (d) Normalized enrichment of MAE and BAE genes inside and around chromatin loops mediated by insulator CTCF sites. (e) Example snapshots from WashU EpiGenome Browser showing MAE genes (orange bars), BAE genes(blue bars), H3K36me3 (red), H3K27me3 (green), CTCF ChIA-PET loops (black arcs) and TAD domains (heatmaps, 25/50 kb resolution). Shown regions are chr4:79.6-81mb, chr8:67.3-67.625mb, chr5:131.25-131.875mb and chr2:24.8-25.7mb in hLCL (hg19).

Figure 2. Genetic and epigenetic association between allelic status of genes and their proximity to insulator CTCF sites. (a) Enrichment of MAE and BAE genes that maintained their allelic status in hLCL and K562 cell-lines, inside and around insulator loops (b) Enrichment of MAE and BAE genes that switched their allelic status in hLCL and K562 cell-lines, inside and around insulator loops. (c) Examples illustrating switch in allelic status of gene between two cell-lines. Shown are the RefSeq genes, chromHMM, CTCF ChIP-Seq, RNAPII ChIA-PET and CTCF ChIA-PET tracks for hLCL and K562 cell-lines. In the left panel, ITPKB gene was MAE in hLCL and BAE in K562 cell-line. In the right panel, GNPDA1 gene was BAE in hLCL and MAE in K562 cell-line. (d) Proportion of MAE and BAE genes that maintained their allelic status in human and mouse lymphoblastoid cell-lines and were having at-least one insulator CTCF site within 20kb to TSS. (e) Proportion of MAE and BAE genes that switched their allelic status in human and mouse lymphoblastoid cell-lines and were having at-least one insulator CTCF site within 20kb to TSS. P-values were calculated using Mann Whitney U test.

Figure 3. Impact of CTCF depletion on MAE and BAE genes. (a) Boxplots of dip prominence (insulation) of MAE and BAE genes in the control and CTCF depleted mESC. (b-c) Log2 fold change of expression of MAE and BAE genes after CTCF depletion in (b) mESC, and (c) mLCL. P-values were calculated by Mann-Whitney U test. (d) Median expression of genes which were located in the neighbouring chromatin domain of upregulated and downregulated MAE genes. P-value was calculated Mann Whitney U test (e) Relative enrichment of active (active promoters, strong enhancers) and inactive (H3K27me3 repressed) chromatin states inside and around the chromatin loops enclosing up and downregulated MAE genes. Ratio of enrichment for upregulated genes to that of downregulated genes was plotted. (f) Distribution of chromatin interaction frequencies of genes enclosed within chromatin loops mediated by insulator CTCF sites and the genes with their promoters linked to enhancers via enhancer-linking CTCF sites. (g) Expression of genes associated with insulator loops and ehancer linking loops before after CTCF depletion in mESC. (h) An example illustrating chromatin insulation of MAE genes in mESC and lack thereof after CTCF depletion. Shown are MAE, BAE genes, CTCF ChIA-PET loops and the heatmaps for HiC data before and after CTCF depletion and difference thereof.

Supplementary Figure Legends

Figure S1. (a) Pie chart showing the percentage of bins with only MMAE, only PMAE and both mixed within gene-clusters. (b) Scaled average aggregation plot of Kronecker delta function for MMAE and PMAE genes.

Figure S2. Normalized enrichment of MAE and BAE genes inside and around chromatin loops mediated by enhancer-linker CTCF sites in hLCL.

Figure S3. (a) Proportion of MAE and BAE genes that maintained their allelic status in hLCL and HMEC cell-lines, having at least one insulator CTCF site within 20kb. (b) Proportion of MAE and BAE genes that switched their allelic status in hLCL and HMEC cell-lines, having at least one insulator CTCF site within 20kb. P-values were calculated using Mann Whitney U test. (c) Examples illustrating the switch in allelic status of genes between two cell-lines. Shown are the chromHMM tracks around: i) MYB gene, which was bi-allelic in hLCL but had mono-allelic expression in HMEC cell-line; and ii) NRP2 gene that had mono-allelic expression in hLCL but was expressed bi-allelically in HMEC. Solid and dashed green arcs are experimentally identified RNAPII ChIA-PET and CTCF ChIA-PET loops in hLCL respectively. Chromatin loops for HMEC were not available.

Figure S4. Ratio of up-regulated to down-regulated genes after CTCF depletion in mLCL and mESC. P-values were calculated by Fisher's exact test.

Figure S5. An example illustrating the alterations in the contact frequencies (TAD disruption) after CTCF depletion in mESC at locus chr11:78680000-79960000 (mm9).

Figure S6. (a) Density plot of log10 normalized squared difference in the CTCF loop interaction frequencies between the maternal and the paternal loci of MAE and BAE genes. (b) Box-plot of the maternal-to-paternal ratio of CTCF loop interaction frequencies of maternally and paternally biased MAE genes. Dashed line is the median of all MAE genes. All the p-values were calculated by Mann-Whitney U test (one tail).

Figure S7. Proportion of MAE and BAE genes overlapping with (a) dosage sensitive genes and (b) common CNVs, as annotated by ClinGen resource. P-values were calculated using Fisher's exact test.

Figure S8. Normalized enrichment of H3K27me3 before and after CTCF depletion in the chromatin loops (and flanking regions) enclosing (a) upregulated and (b) downregualted genes.

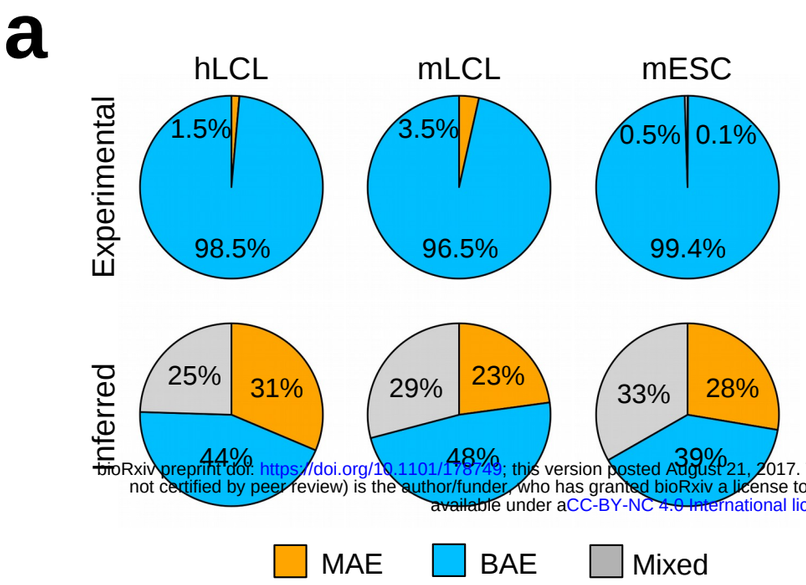
References

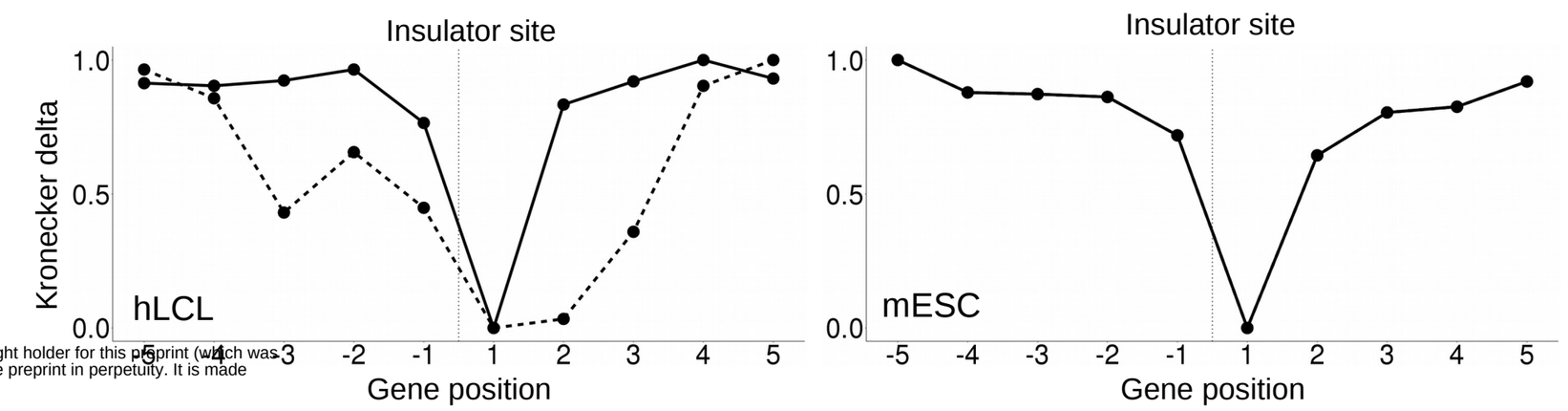
1. Gimelbrant A, Hutchinson JN, Thompson BR, Chess A: **Widespread monoallelic expression on human autosomes.** *Science* 2007, **318:**1136-1140.
2. Pernis B, Chiappino G, Kelus AS, Gell PG: **Cellular localization of immunoglobulins with different allotypic specificities in rabbit lymphoid tissues.** *J Exp Med* 1965, **122:**853-876.
3. Hollander GA, Zuklys S, Morel C, Mizoguchi E, Mobisson K, Simpson S, Terhorst C, Wishart W, Golan DE, Bhan AK, Burakoff SJ: **Monoallelic expression of the interleukin-2 locus.** *Science* 1998, **279:**2118-2121.
4. Rhoades KL, Singh N, Simon I, Glidden B, Cedar H, Chess A: **Allele-specific expression patterns of interleukin-2 and Pax-5 revealed by a sensitive single-cell RT-PCR analysis.** *Curr Biol* 2000, **10:**789-792.
5. Brady BL, Steinel NC, Bassing CH: **Antigen receptor allelic exclusion: an update and reappraisal.** *J Immunol* 2010, **185:**3801-3808.
6. Chess A, Simon I, Cedar H, Axel R: **Allelic inactivation regulates olfactory receptor gene expression.** *Cell* 1994, **78:**823-834.
7. Chess A: **Mechanisms and consequences of widespread random monoallelic expression.** *Nat Rev Genet* 2012, **13:**421-428.
8. Lewcock JW, Reed RR: **A feedback mechanism regulates monoallelic odorant receptor expression.** *Proc Natl Acad Sci U S A* 2004, **101:**1069-1074.
9. Saleh A, Davies GE, Pascal V, Wright PW, Hodge DL, Cho EH, Lockett SJ, Abshari M, Anderson SK: **Identification of probabilistic transcriptional switches in the Ly49 gene cluster: a eukaryotic mechanism for selective gene activation.** *Immunity* 2004, **21:**55-66.
10. Meaburn EL, Schalkwyk LC, Mill J: **Allele-specific methylation in the human genome: implications for genetic studies of complex disease.** *Epigenetics* 2010, **5:**578-582.
11. Shoemaker R, Deng J, Wang W, Zhang K: **Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome.** *Genome Res* 2010, **20:**883-889.
12. Mancini-Dinardo D, Steele SJ, Levorse JM, Ingram RS, Tilghman SM: **Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes.** *Genes Dev* 2006, **20:**1268-1282.
13. Sleutels F, Zwart R, Barlow DP: **The non-coding Air RNA is required for silencing autosomal imprinted genes.** *Nature* 2002, **415:**810-813.
14. Takizawa T, Gudla PR, Guo L, Lockett S, Misteli T: **Allele-specific nuclear positioning of the monoallelically expressed astrocyte marker GFAP.** *Genes Dev* 2008, **22:**489-498.
15. Armelin-Correa LM, Gutiyama LM, Brandt DY, Malnic B: **Nuclear compartmentalization of odorant receptor genes.** *Proc Natl Acad Sci U S A* 2014, **111:**2782-2787.
16. Kurukuti S, Tiwari VK, Tavoosidana G, Pugacheva E, Murrell A, Zhao Z, Lobanenkov V, Reik W, Ohlsson R: **CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2.** *Proc Natl Acad Sci U S A* 2006, **103:**10684-10689.
17. Lin S, Ferguson-Smith AC, Schultz RM, Bartolomei MS: **Nonallelic transcriptional roles of CTCF and cohesins at imprinted loci.** *Mol Cell Biol* 2011, **31:**3094-3104.
18. Jeffries AR, Collier DA, Vassos E, Curran S, Ogilvie CM, Price J: **Random or stochastic monoallelic expressed genes are enriched for neurodevelopmental disorder candidate genes.** *PLoS One* 2013, **8:**e85093.
19. Nora EP, Goloborodko A, Valton AL, Gibcus JH, Uebersohn A, Abdennur N, Dekker J, Mirny LA, Bruneau BG: **Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization.** *Cell* 2017, **169:**930-944 e922.
20. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al: **Mapping and analysis of chromatin state dynamics in nine human cell types.** *Nature* 2011, **473:**43-49.
21. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, et al: **A comparative encyclopedia of DNA elements in the mouse genome.** *Nature* 2014, **515:**355-364.
22. Consortium EP: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57-74.
23. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J, Ruszczycki B, et al: **CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription.** *Cell* 2015, **163:**1611-1627.
24. Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CW, Ye C, Ping JL, Mulawadi F, et al: **CTCF-mediated functional chromatin interactome in pluripotent cells.** *Nat Genet* 2011, **43:**630-638.
25. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell* 2014, **159:**1665-1680.
26. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, et al: **An oestrogen-receptor-alpha-bound human chromatin interactome.** *Nature* 2009, **462:**58-64.

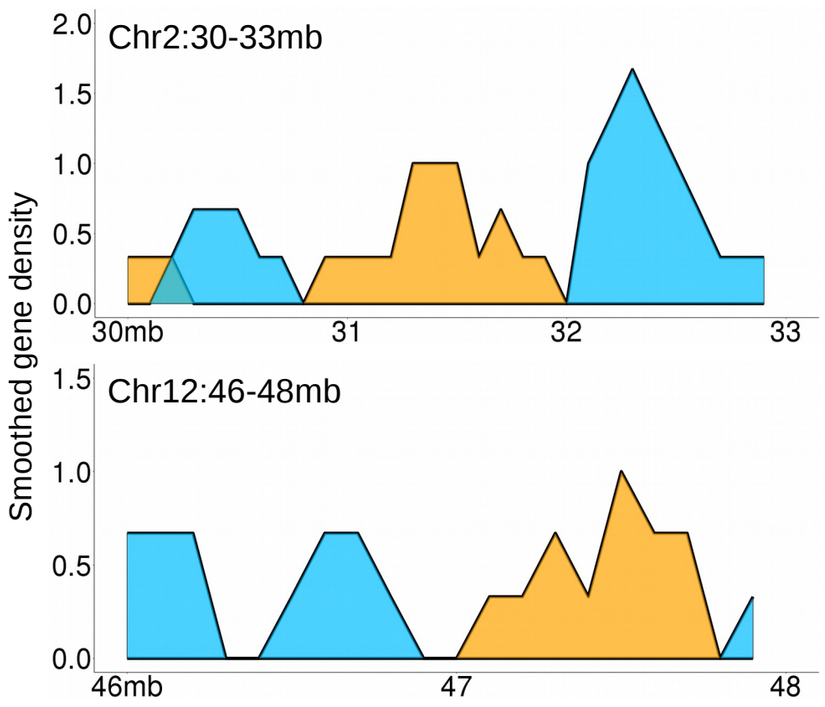
27. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, et al: **Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation.** *Cell* 2012, **148:**84-98.
28. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.** *Nat Protoc* 2012, **7:**562-578.
29. Zwemer LM, Zak A, Thompson BR, Kirby A, Daly MJ, Chess A, Gimelbrant AA: **Autosomal monoallelic expression in the mouse.** *Genome Biol* 2012, **13:**R10.
30. Gendrel AV, Attia M, Chen CJ, Diabangouaya P, Servant N, Barillot E, Heard E: **Developmental dynamics and disease potential of random monoallelic gene expression.** *Dev Cell* 2014, **28:**366-380.
31. Nag A, Savova V, Fung HL, Miron A, Yuan GC, Zhang K, Gimelbrant AA: **Chromatin signature of widespread monoallelic expression.** *Elife* 2013, **2:**e01256.
32. Nag A, Vigneau S, Savova V, Zwemer LM, Gimelbrant AA: **Chromatin Signature Identifies Monoallelic Gene Expression Across Mammalian Cell Types.** *G3 (Bethesda)* 2015, **5:**1713-1720.
33. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjornson R, Kong Y, Kitabayashi N, et al: **AlleleSeq: analysis of allele-specific expression and binding in a network framework.** *Mol Syst Biol* 2011, **7:**522.

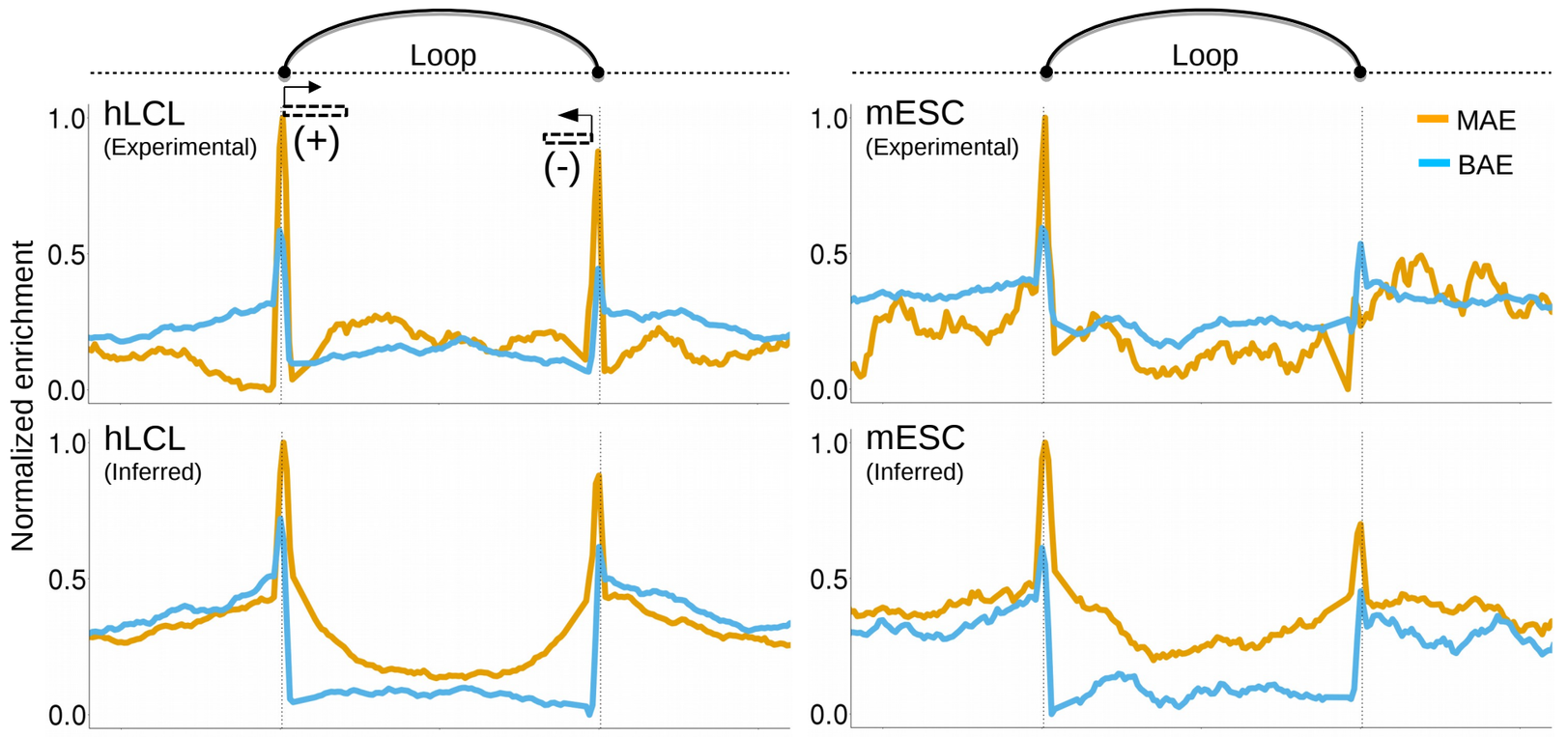
Table 1

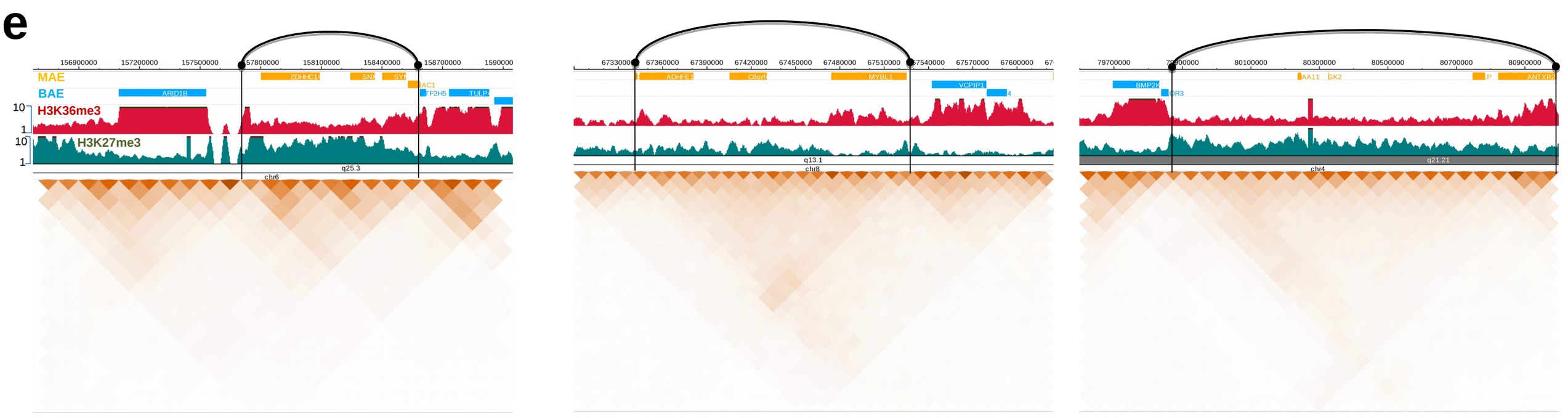
Method	Species	Cell-line	MAE	BAE	Source
Experimentally identified	Human	Lymphoblastoid	399	3045	Gimelbrant et al [1]
	Mouse	Pre-B	294	1101	Zwemer et al [29]
		ESC	629	9827	Gendrel et al [30]
Inferred	Human	Lymphoblastoid	9469	10085	Nag et al[31]
		HMEC	6529	12880	Nag et al [31]
		K562	8574	10455	Nag et al [31]
	Mouse	ESC	10427	11955	Nag et al [32]
		Lymphoblastoid	8967	11647	Nag et al[32]
Experimentally identified	Human	Lymphoblastoid	**MMAE**	**PMAE**	
			484	422	Rozowsky et al [33]

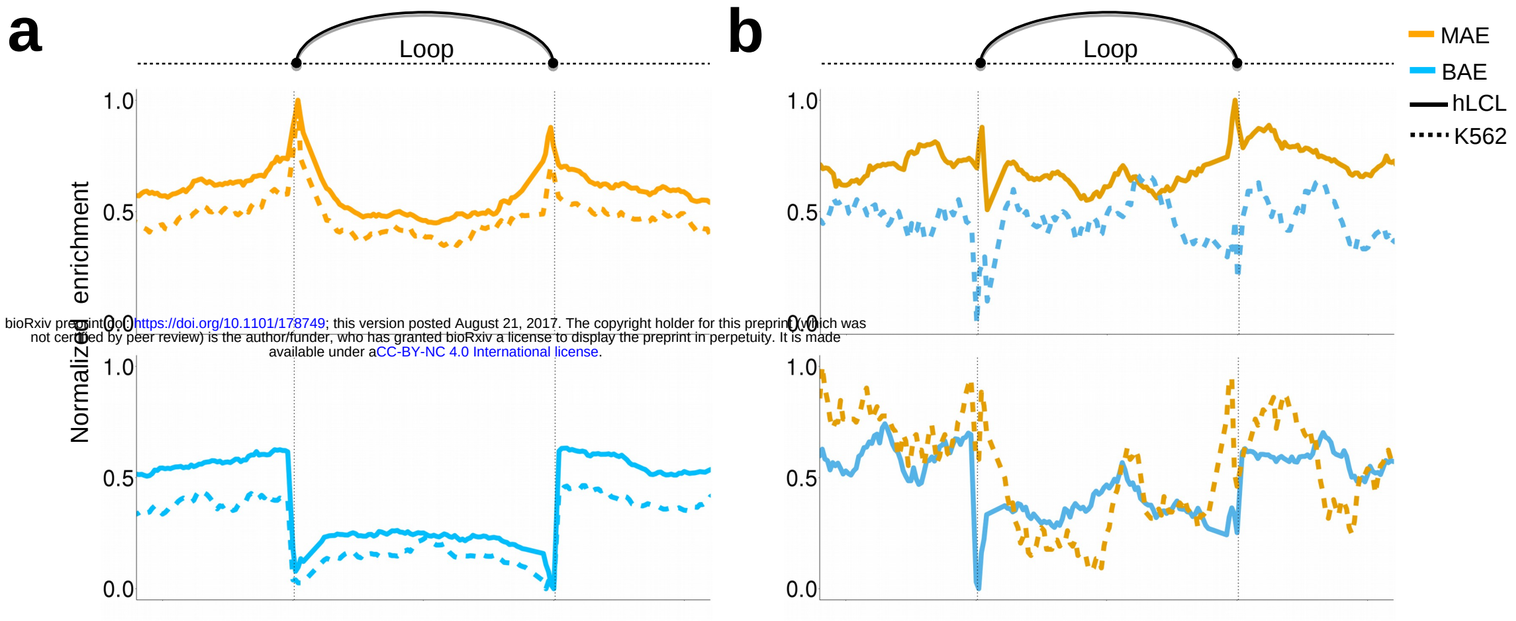


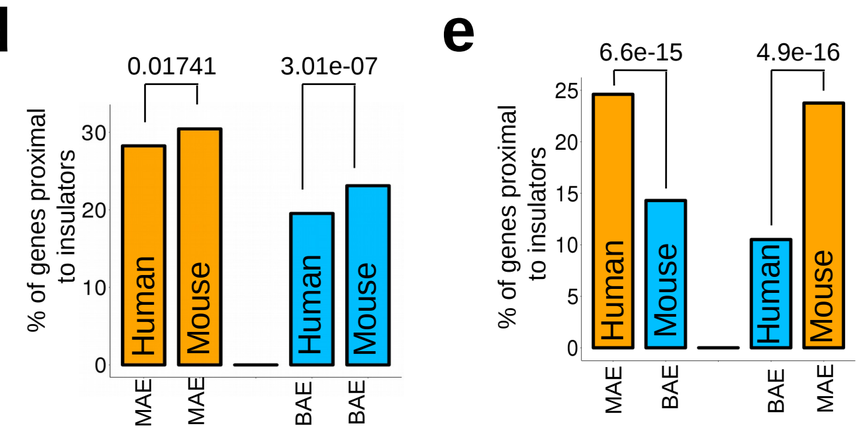


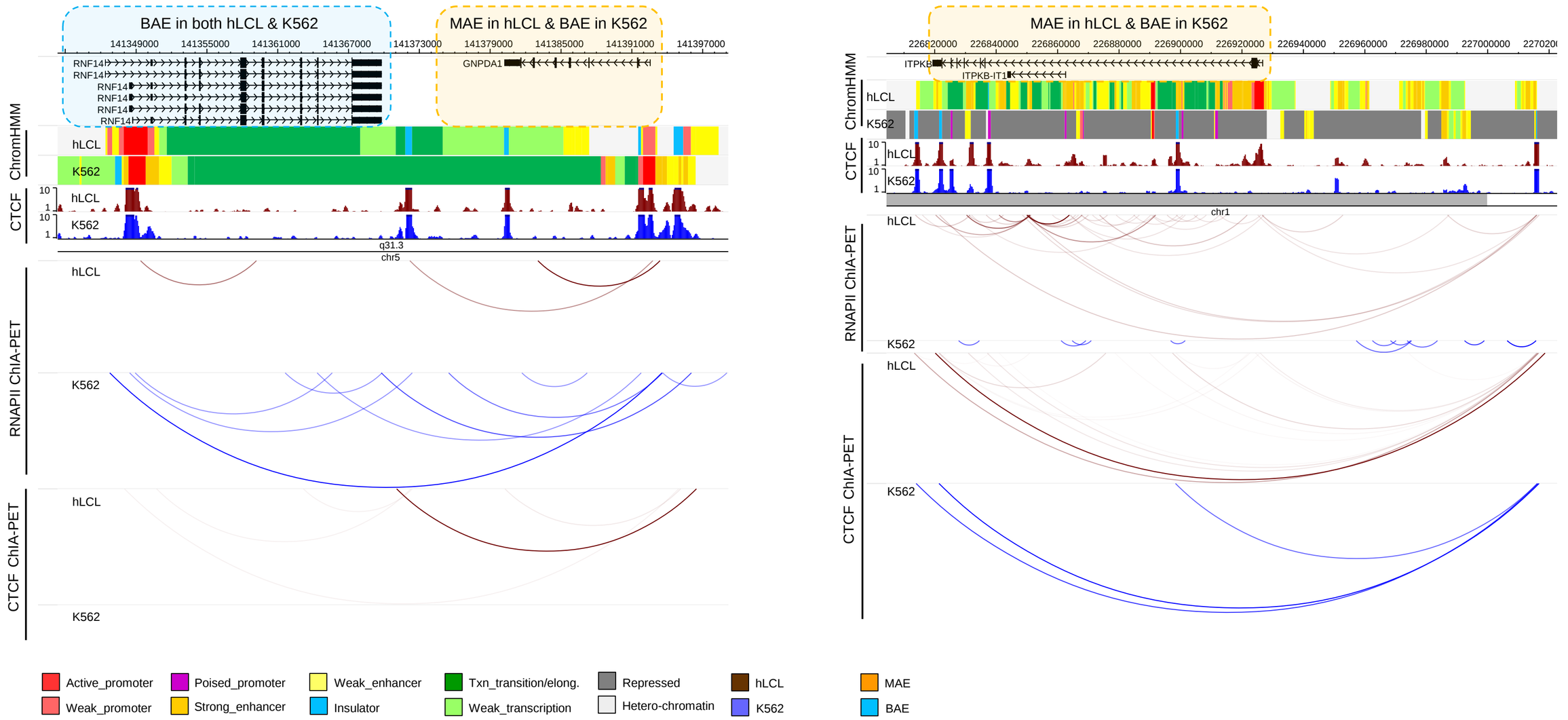












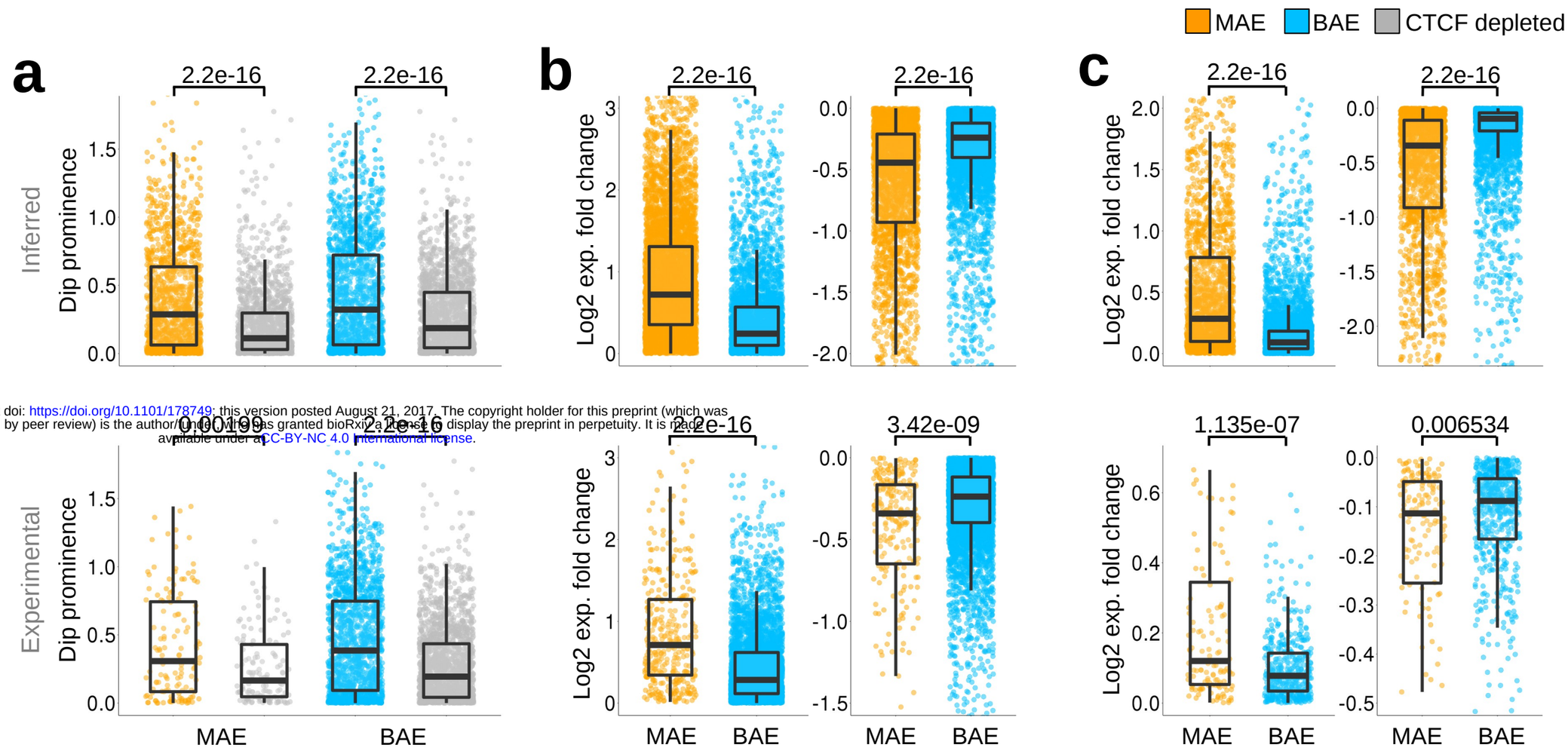


Figure S1:

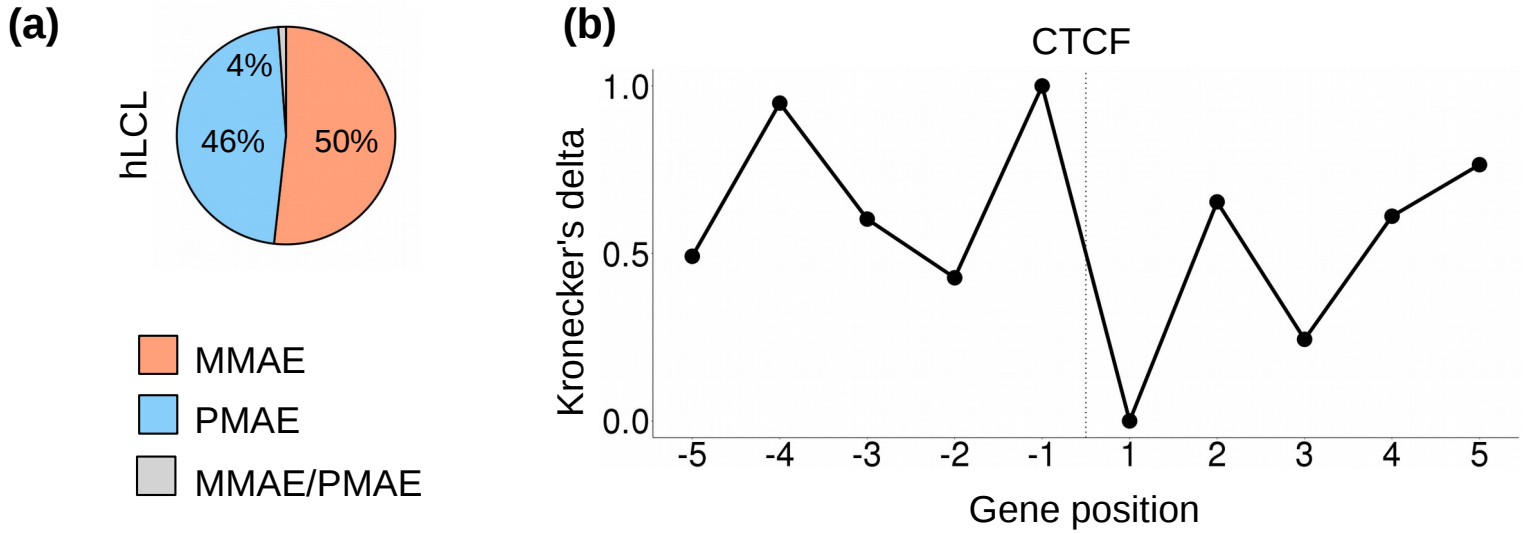

(a)
hLCL
4%
46%
50%
MMAE
PMAE
MMAE/PMAE
(b)
CTCF
Kronecker's delta
1.0
0.5
0.0
-5
-4
-3
-2
-1
1
2
3
4
5
Gene position

Figure S2:

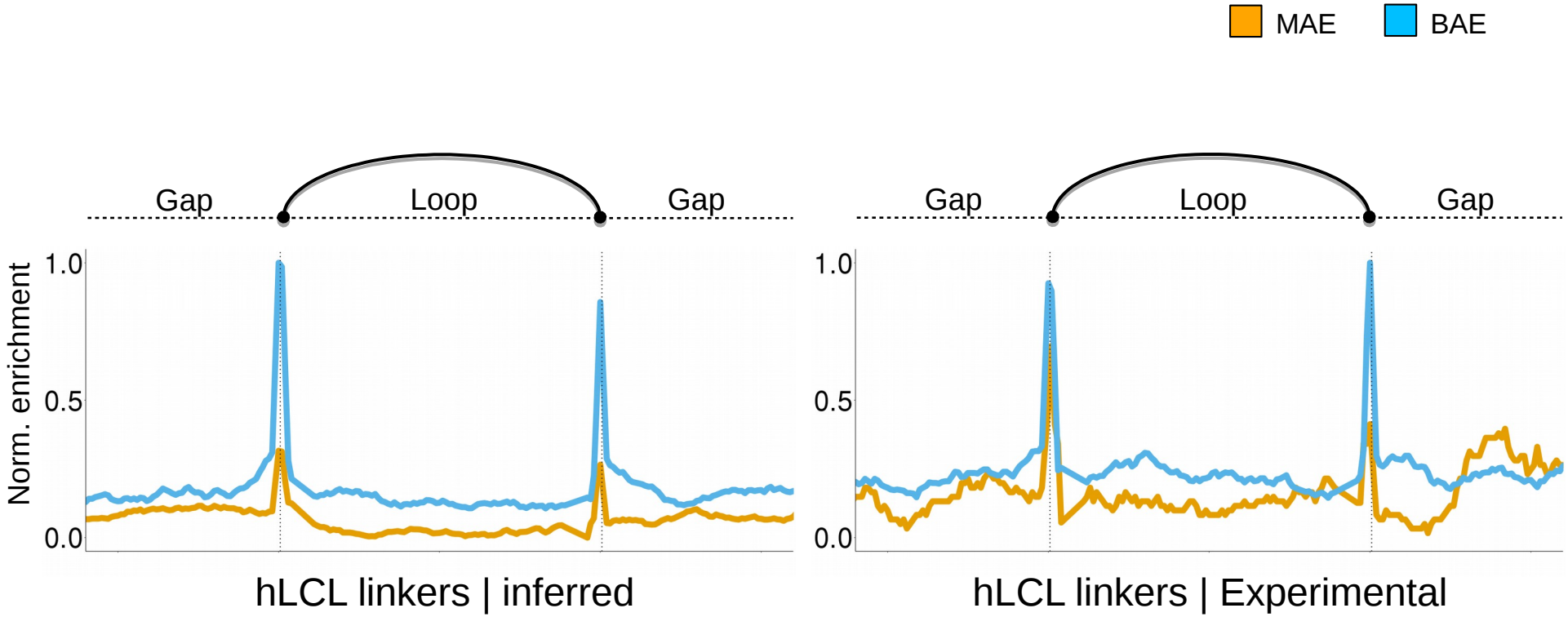
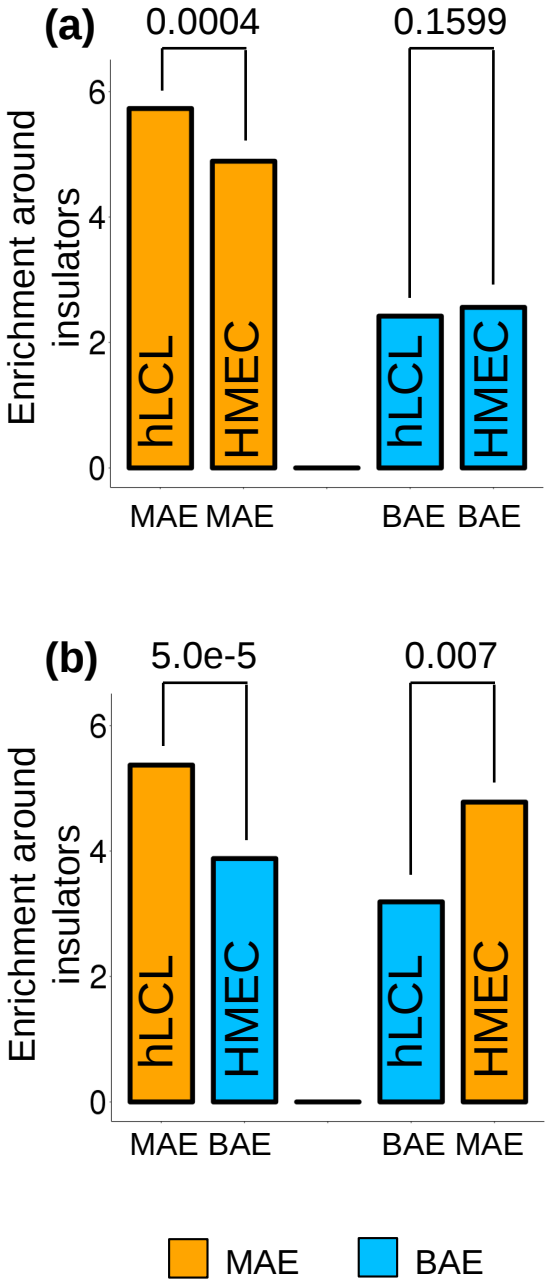

MAE
BAE
Gap
Loop
Gap
Norm. enrichment
1.0
0.5
0.0
hLCL linkers | inferred
Gap
Loop
Gap
1.0
0.5
0.0
hLCL linkers | Experimental

Figure S3:


(a)
0.0004
0.1599
Enrichment around insulators
hLCL
HMEC
hLCL
HMEC
MAE MAE
BAE BAE
(b)
5.0e-5
0.007
Enrichment around insulators
hLCL
HMEC
hLCL
HMEC
MAE BAE
BAE MAE
MAE
BAE

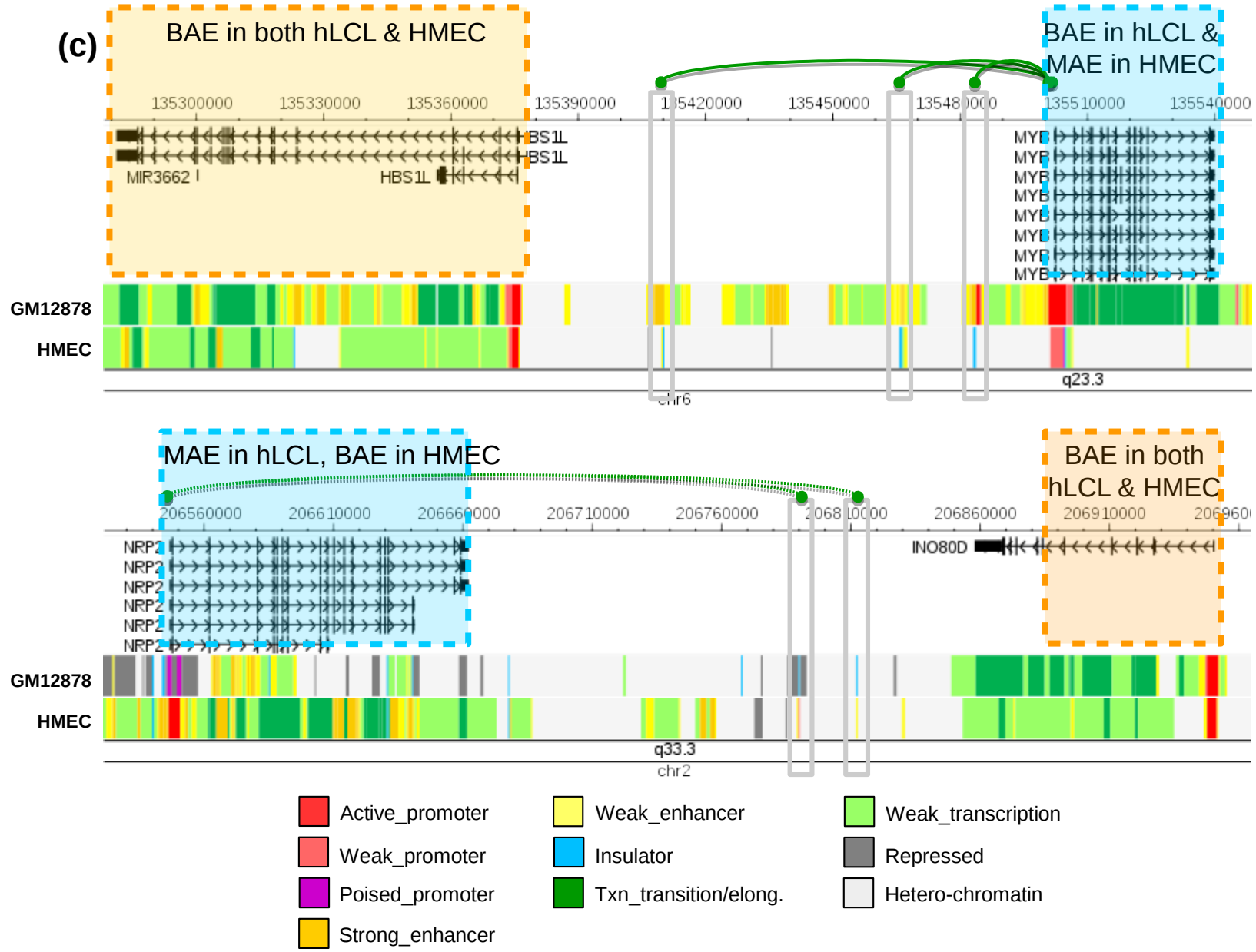

(c)
BAE in both hLCL & HMEC
BAE in hLCL & MAE in HMEC
MAE in hLCL, BAE in HMEC
BAE in both hLCL & HMEC
GM12878
HMEC
q23.3
chr6
q33.3
chr2
Active_promoter
Weak_promoter
Poised_promoter
Strong_enhancer
Weak_enhancer
Insulator
Txn_transition/elong.
Weak_transcription
Repressed
Hetero-chromatin

Figure S4:

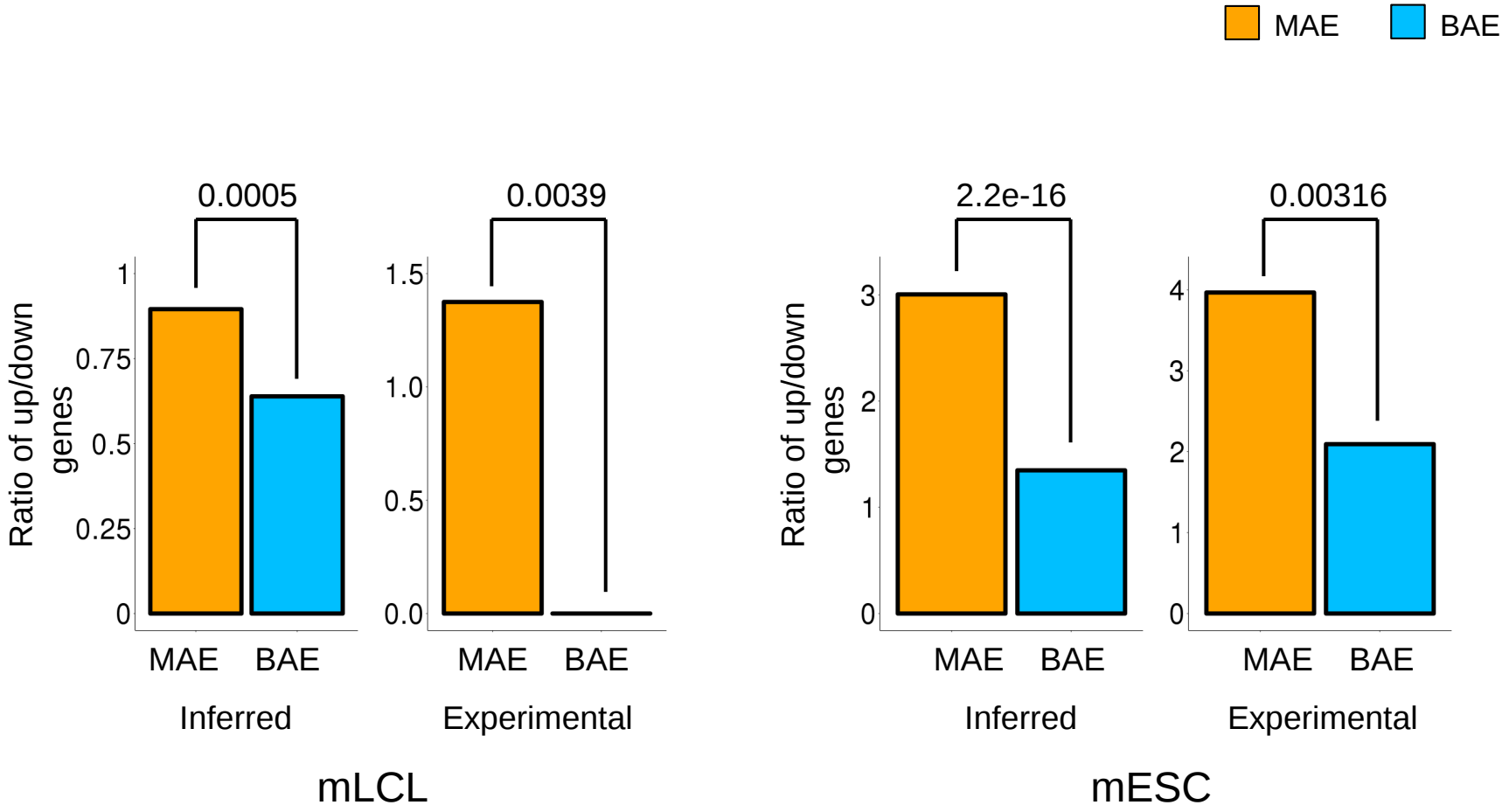

MAE
BAE
0.0005
0.0039
2.2e-16
0.00316
Ratio of up/down genes
1
0.75
0.5
0.25
0
1.5
1.0
0.5
0.0
3
2
1
0
4
3
2
1
0
MAE
BAE
Inferred
Experimental
mLCL
mESC

Figure S5:

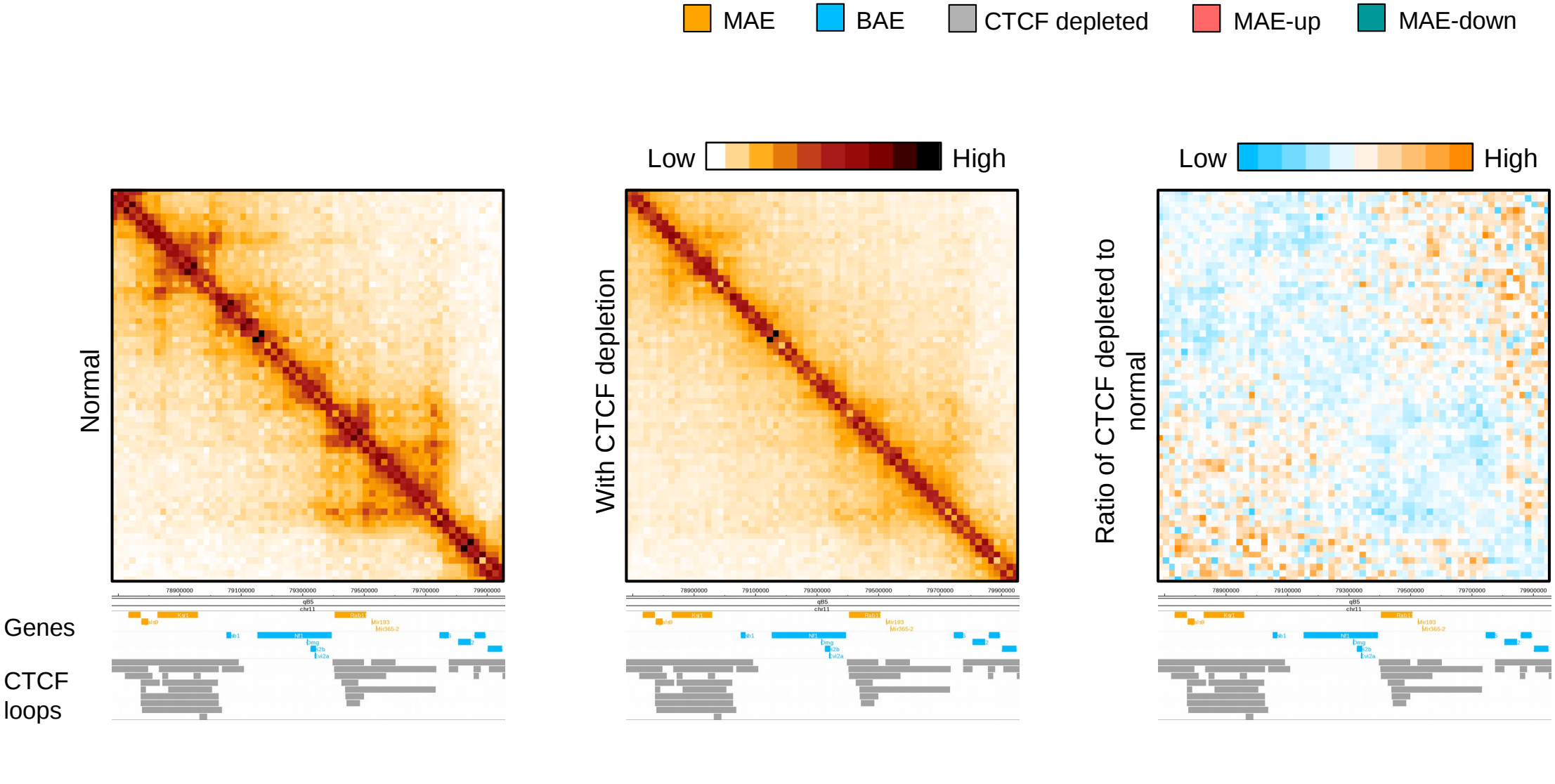

MAE
BAE
CTCF depleted
MAE-up
MAE-down
Low
High
Low
High
Normal
With CTCF depletion
Ratio of CTCF depleted to normal
Genes
CTCF loops

Figure S6:

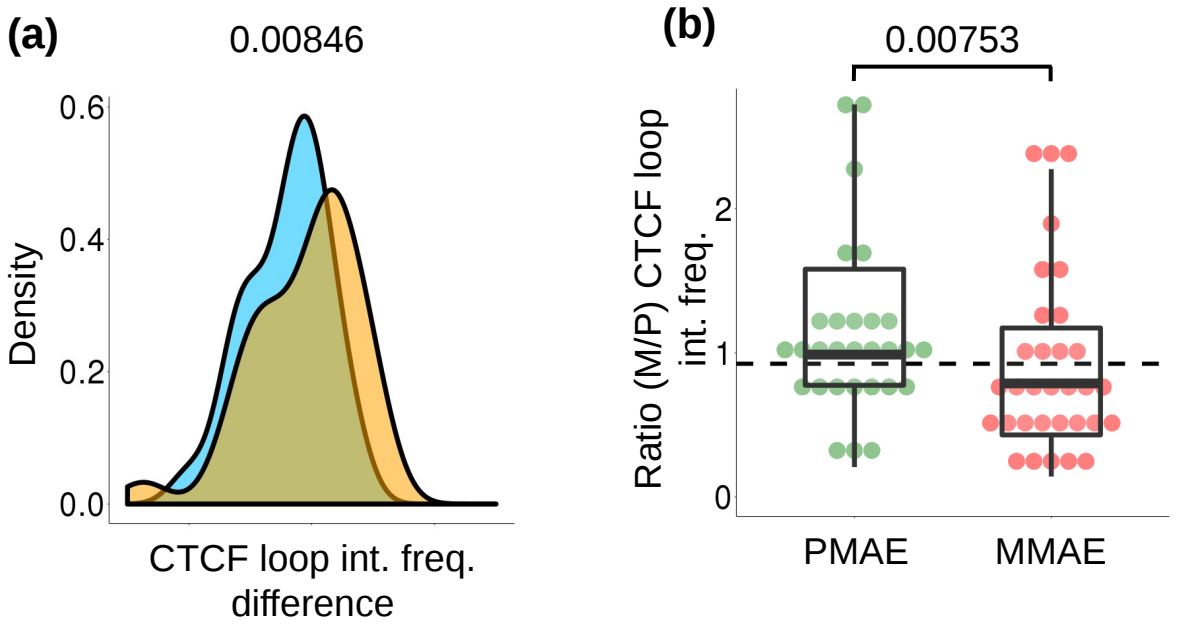

(a)
0.00846
0.6
0.4
0.2
0.0
Density
CTCF loop int. freq. difference
(b)
0.00753
Ratio (M/P) CTCF loop int. freq.
2
1
0
PMAE
MMAE

Figure S7:

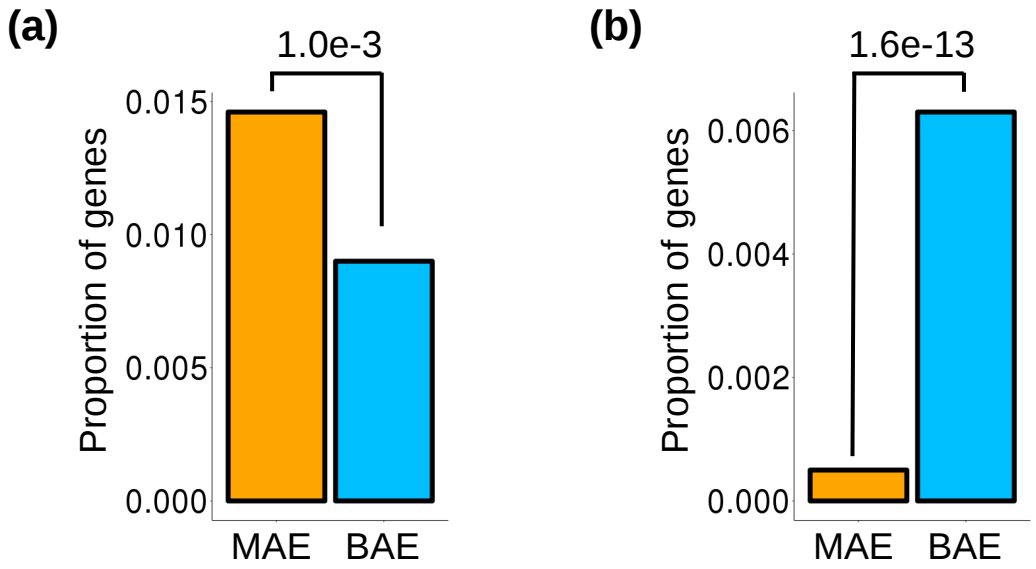
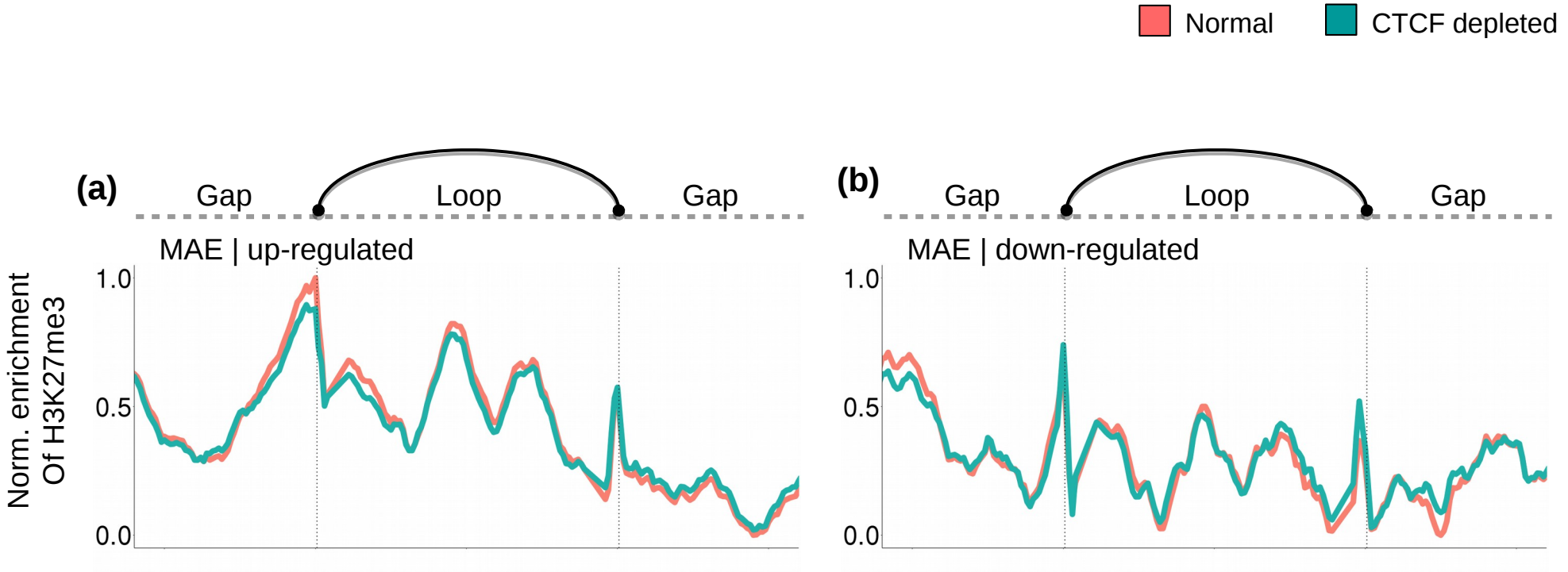

(a)
1.0e-3
Proportion of genes
0.015
0.010
0.005
0.000
MAE
BAE
(b)
1.6e-13
Proportion of genes
0.006
0.004
0.002
0.000
MAE
BAE

Figure S8:


Normal
CTCF depleted
(a)
Gap
Loop
Gap
MAE | up-regulated
Norm. enrichment
Of H3K27me3
1.0
0.5
0.0
(b)
Gap
Loop
Gap
MAE | down-regulated
1.0
0.5
0.0