Chromatin organization capture by Hi-C technology

Bioinformatics | Assignment 5 | Group 5

Sarnai Ganbold, Sabrina George, Aravind Rajagopalan, Yifan Li, Mingzhou Fu, Tania J. González Robles

Background

Chromatin organization can dictate the levels of gene transcription, enabling chromosomal associations that give rise to chromatin compartmentalization A or B, characteristic of active or silenced chromatin, respectively. Within the A and B compartments, we find a spectrum of intra- and inter-topologically-associated domains (TADs). TADs are described as megabase-sized self-interacting regions in the chromatin that is conserved across cell types and species. At the boundaries delimiting each TAD, insulator proteins such as CTCF, housekeeping (HK) genes, transfer RNAs and short interspaced retrotransposon elements are commonly enriched. These factors can be leveraged to estimate the extent to which a TAD belongs to an A or B compartment, corresponding to its active or inactive transcription, respectively. Three-dimensional (3-D) chromatin interactions are difficult to visualize under the microscope or through linear sequencing. In order to measure, such interactions researchers developed Chromosome Conformation Capture (3C) techniques.

The 3C techniques are a set of molecular biology tools used to analyze the spatial organization of chromatin in a cell. These methods quantify the number of interactions between genomic loci that are nearby in a 3-D space that may not be easily detected and quantified in the linear genome. Such interactions regulate biological functions governing promoter-enhancer activity in transcription, which can affect the gene transcription abundance or silencing. The interaction landscape can be analyzed directly or by converting the information into a distance matrix to reconstruct the chromatin 3-D structure. Among the several 3C techniques, Hi-C quantifies whole genome interactions between all possible chromosome pairs. To do so, the genomic material of a cell undergoes enzymatic fractionation followed by linkage, tag with biotin, and ligate together, followed by PCR amplification. The resulting fragments are then subjected to linear sequencing to capture the genomic reads comprising the contact regions of the chromosome, thus giving rise to the so-called contact matrix providing TADs information.

In our project, we used the Hi-C technique to determine the TADs in normal mouse Embryonic Stem (mES) cells. TADs are calculated based on contact matrices containing information from the whole genome. Because we are working with only two replicates of mES samples, we decided to report TADs information on the *Sox2* gene. *Sox2* is known to be involved in development and stem cell maintenance. Therefore, we anticipate this gene will be actively transcribed in mES cells. *Sox2* is found in chromosome 3 (chr3) starting at 181,711,925 bp

through the 181,714,436 bp. To visualize TADs within the *Sox2* genomic locus we used comparable pipelines, namely HiC-Bench and HiC-Explorer.

Methods

In this project, we used both HiC-bench and HiCExplorer to analyze the mouse embryonic stem cells. For HiC-bench, we followed the built-in pipeline. We started with alignment to the reference mm10 mouse genome using the restriction enzyme cocktail Arima which digested the restriction sequence at GATC, GAGTC, GACTC, GAATC and GATTC. Afterwards, we generated the filtered valid pairs. Then, we generated a HiC file to load into the juicer interface for contact matrix visualization with proper normalization.

The methods for HiCExplorer utilized a different set of tools. The paired-end alignment is performed with Burrows-Wheeler Aligner MEM (parameters used: A1 for matching score 1, B4 for mismatch penalty=4, E50 for gap extension penalty of 50, L0 for no clipping penalty) on NYU Langone High Performance Cluster. The local aligner was chosen to conform with the HiCExplorer algorithm for interaction frequency (IF)-based enrichment search. The HiC Explorer environment was set up with miniconda3 manager, python v.3.6, numpy v.1.15 and biopython v.1.77.

Next, the unsorted strand-specific pairs of ".bam" files were used to build four HiC contact matrices with imbedded HiCBuildMatrix tool (parameters used: bin size = 50000, threads 8, restriction sequences of Arima enzyme mix: GATC, GAGTC, GACTC, GAATC and GATTC, input buffer size = 100000). The four matrices for 2 replicates were merged into one using hicSumMatrices tool. The hicCorrectMatrix tool was employed to remove GC, open chromatin biases and to normalize the number of restriction sites per bin.

HiCFindTADS computes a TAD-separation score based on a z-score matrix for all bins.
Then bins with local minimum of the TAD-separation score are compared to the surrounding bins and p-values are assigned. Finally, an FDR cutoff of 0.05 is applied to select the bins more likely to be TAD boundaries. TAD detection code produced several files: (1) a TAD-separation score file, (2) a z-score matrix, (3) a .bed file with boundary locations, (4) a .bed file with domains, and (5) a .bedgraph file with the TAD-score that can be visualized in a genome browser. A and B compartments for the chromosome of interest were detected with PC1 and PC2 method. Using the hicPCA tool, Pearson correlation method was used to calculate covariance matrix, and eigenvectors 1 and 2 were used to determine principal components 1 and 2 to separate the space between compartments A and B. As PC1 can capture the separation for two compartments in a more accurate way, the first eigenvector of the correlation matrix was used to define the compartment score.

Finally, we took an in-depth look at chromosomal compartment A which is a more active transcribed region and chromosomal compartment B which is a less active transcribed region with respect to gene density, housekeeping gene density and H3K27ac peak density. We then

visualized the bed files from compartmentalization step into IGV for visualization along with the compartment-score bed graph. Finally, bedtools intersect was used to calculate the amount of genes in each compartment and a ratio for comparison between the two compartments.

Results

Hi-C Bench

Hi-C paired-end sequencing data processing and Quality Control

The raw Hi-C sequencing data (.fastq files) were provided by the instructor and accessed on Big Purple. The data included pair-end sequencing of two mouse Embryonic Stem (ES) cell replicates, previously digested with the commercially available Arima enzyme cocktail. We used this data as input for our chromatin organization analysis. First, we assess the provided dataset to ensure the minimal number of reads necessary to calculate TADs. The quality control showed that just under half of all mappable reads were optimal for downstream analyses (Figure 1).

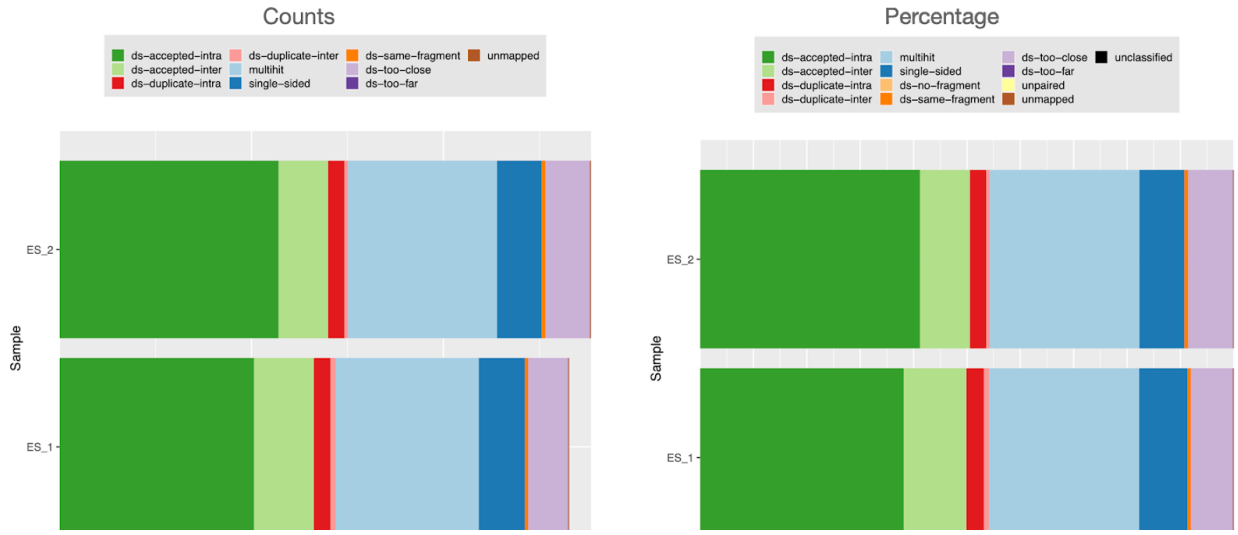

Figure 1. HiC-Bench quality control with both raw counts (left) and percentages (right).

Detecting TADs and A/B compartment calling

The Hi-C files generated from the alignment and filtering steps were imported into juicebox for visualization. Juicebox is an online tool, maintained by the Aiden Lab, that allows for interactive visualization of Hi-C contact matrices. For closer visualization, we focused on the *Sox2* gene, which encodes a transcription factor that is crucial in cell development and stem cell maintenance (Figure 2).

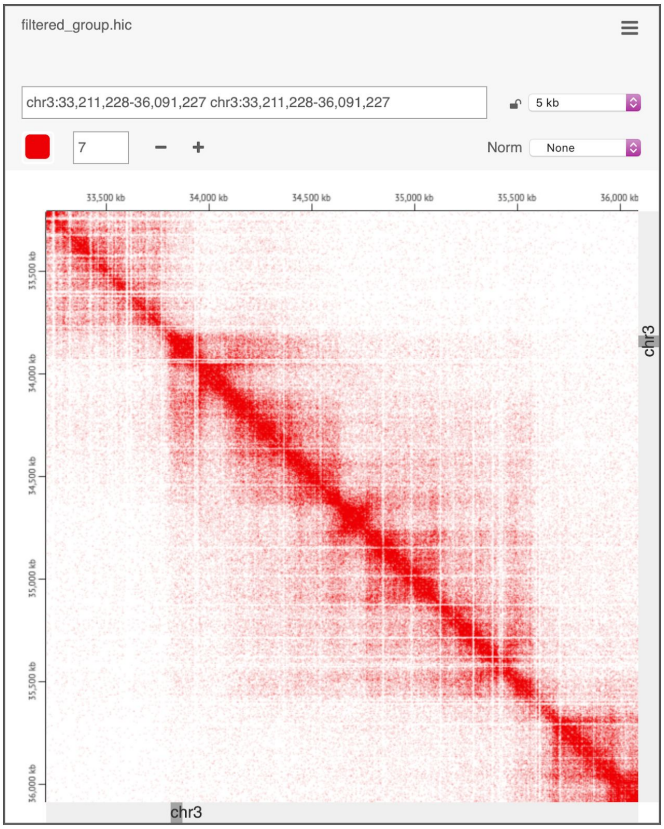

Figure 2. Visualization of the *Sox2* gene using Juicebox. *Sox2* locus contained within chromosome 3 between bases 33,211,228 - 36,091,227.

The ChIP-seq H3K27ac peaks were also provided by the instructor and accessed on Big Purple, which had 28875 2kb regions. The ChIP-seq H3K27ac peaks were used for compartment calling. Upon completion of this step, the .bedgraph file containing the compartment scores was uploaded into IGV for visualization. The housekeeping genes (HK_genes.bed) and the full list of annotated transcription start site (TSS) genes (genes-tss_annot.bed) were also given by the instructor and uploaded into IGV as references for downstream analysis. The .bedgraph and .bed files were all referenced against the mm10 genome. In accordance with the previous *Sox2* visualization in juicebox, we looked at the distribution of compartments, housekeeping genes, and all annotated genes in chromosome 3 (Figure 3), and a similar zoomed-in 8.9 Mb locus at the *Sox2* gene site (Figure 4). The matrices in Figures 2, 3, and 4 were all normalized. Other visualizations of individual interest were included in the supplemental github repository.

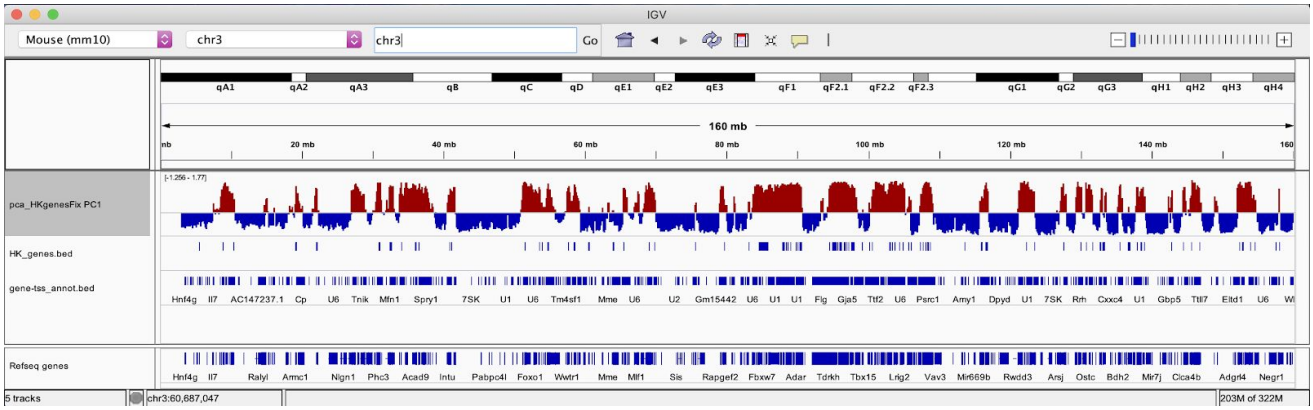

Figure 3. Visualization of Chromosome 3 centered around the *Sox2* gene. Consistent with prior observations, HK_genes were enriched in A compartments and sparse in B compartments.

Figure 4. Visualization of a zoomed-in 8.9Mb locus centered at *Sox2* gene.

Hi-C Explorer

Hi-C paired-end sequencing data processing and Quality Control

Initially, the four Hi-C fastq files for 2 technical replicates of a mouse (mus musculus) were aligned to the reference genome (UCSC, mm10).

Mapping percentages: replicate_1_1_R1=95.99%, replicate_1_1_R2=94.68%, replicate_1_2_R1=96.13%, replicate_1_2_R2=94.37%, replicate_2_1_R1=96.22%, replicate_2_1_R2=94.77%, replicate_2_2_R1=96.42%, replicate_2_2_R2=94.90%.

Mus musculus has 21 haploid chromosomes (chr1-19, chrX, chrY) shown in Fig 1b. Generated histograms of interaction frequency (IF) counts per bin per chromosome shown in Fig 1a. A diagnostic plot of a binomial distribution of average normalized counts shown in Fig 1c. The subsequent outliers region filtering (-2 Z score as the lower threshold and +2 Z score as the upper threshold) shown with red dashed lines in Fig 1c.

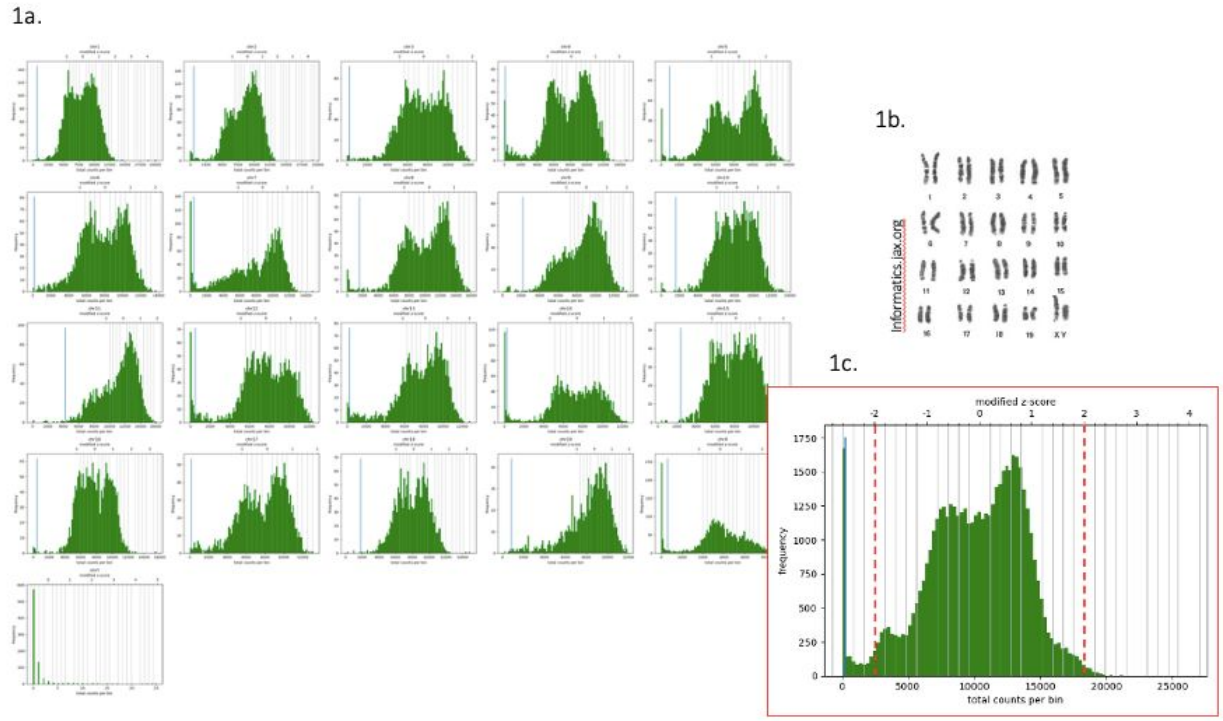

Figure 1. Histograms of normalized interaction frequencies per chromosomes before filtering out outliers.

Detecting TADs and A/B Compartment Calling

HiCExplorer computed a TAD-separation score based on a z-score matrix for all bins. Bins with local minimum of the TAD-separation score were compared to the surrounding bins and p-values were assigned. An FDR cutoff was applied to select bins which were more likely to be TAD boundaries. The corrected contact matrix for chromosome 3 is plotted with the hicPlotMatrix tool (Figure 2, 3). The *Sox2* gene was a gene of interest due to its importance in stem cell development and maintenance (Figure 4).

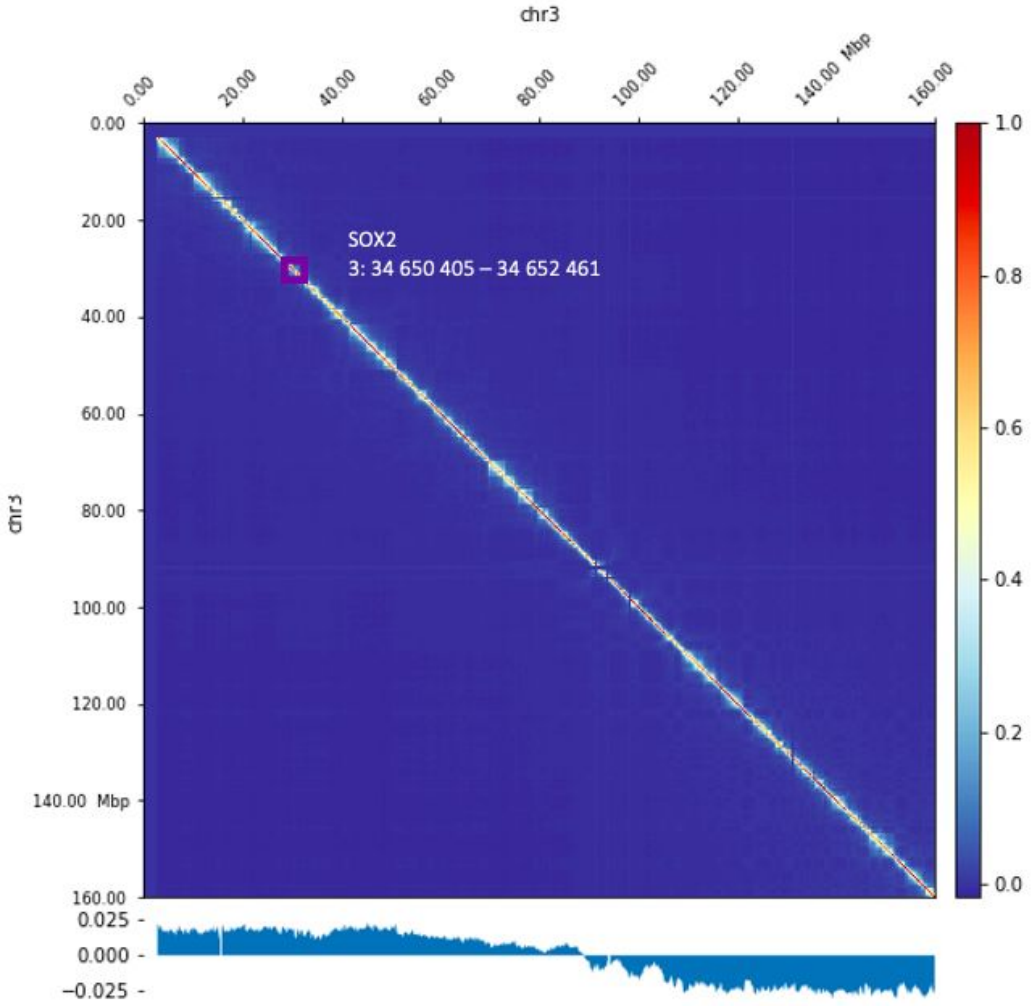

Figure 2. Visualization of HiC contact matrix for the entire chromosome 3 with the *Sox2* gene locus labeled.

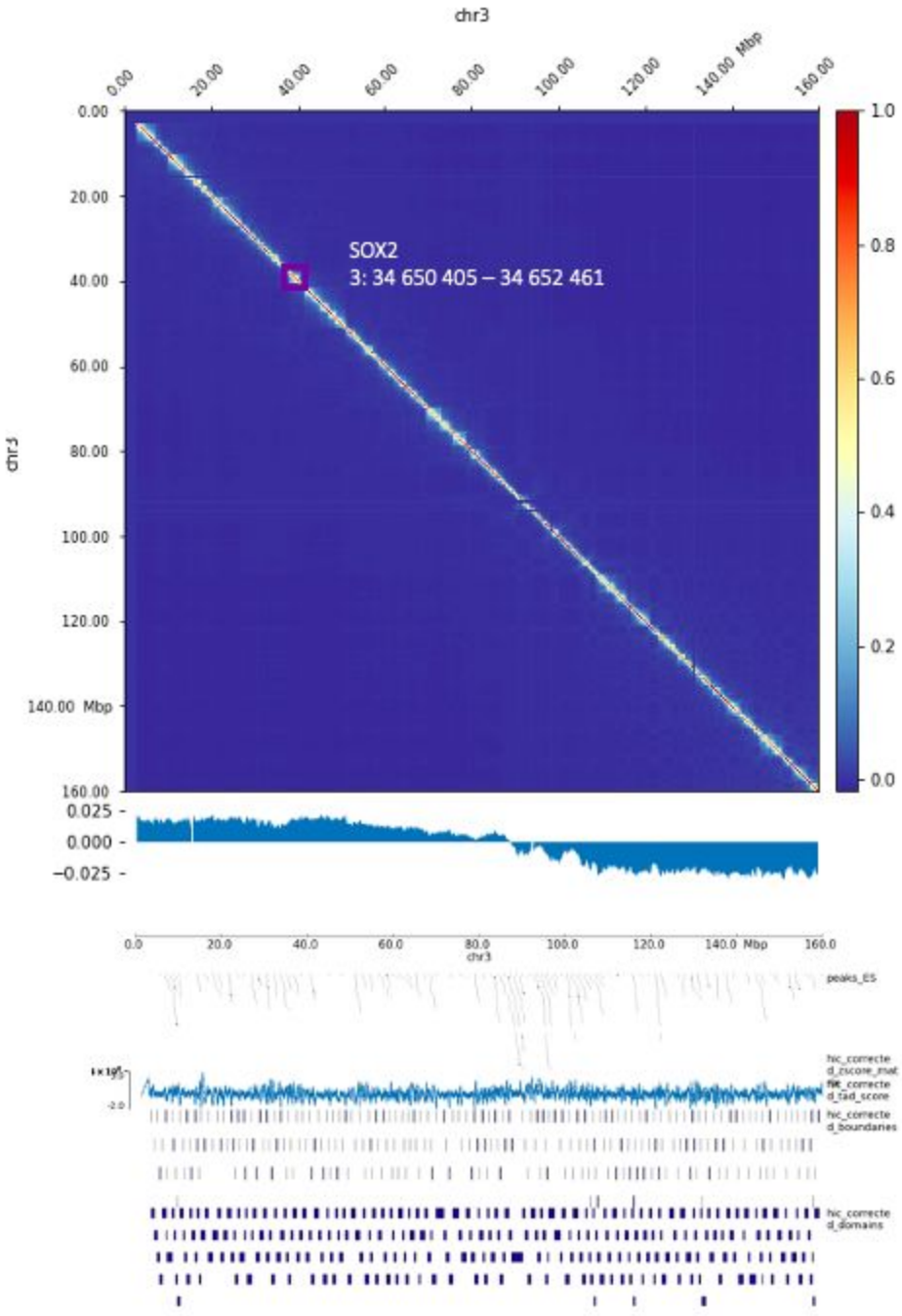

Figure 3. HiC-Explorer visualization of the same region as Figure 2 but with additional peak scores, hic-corrected z-scores, boundaries, and domains.

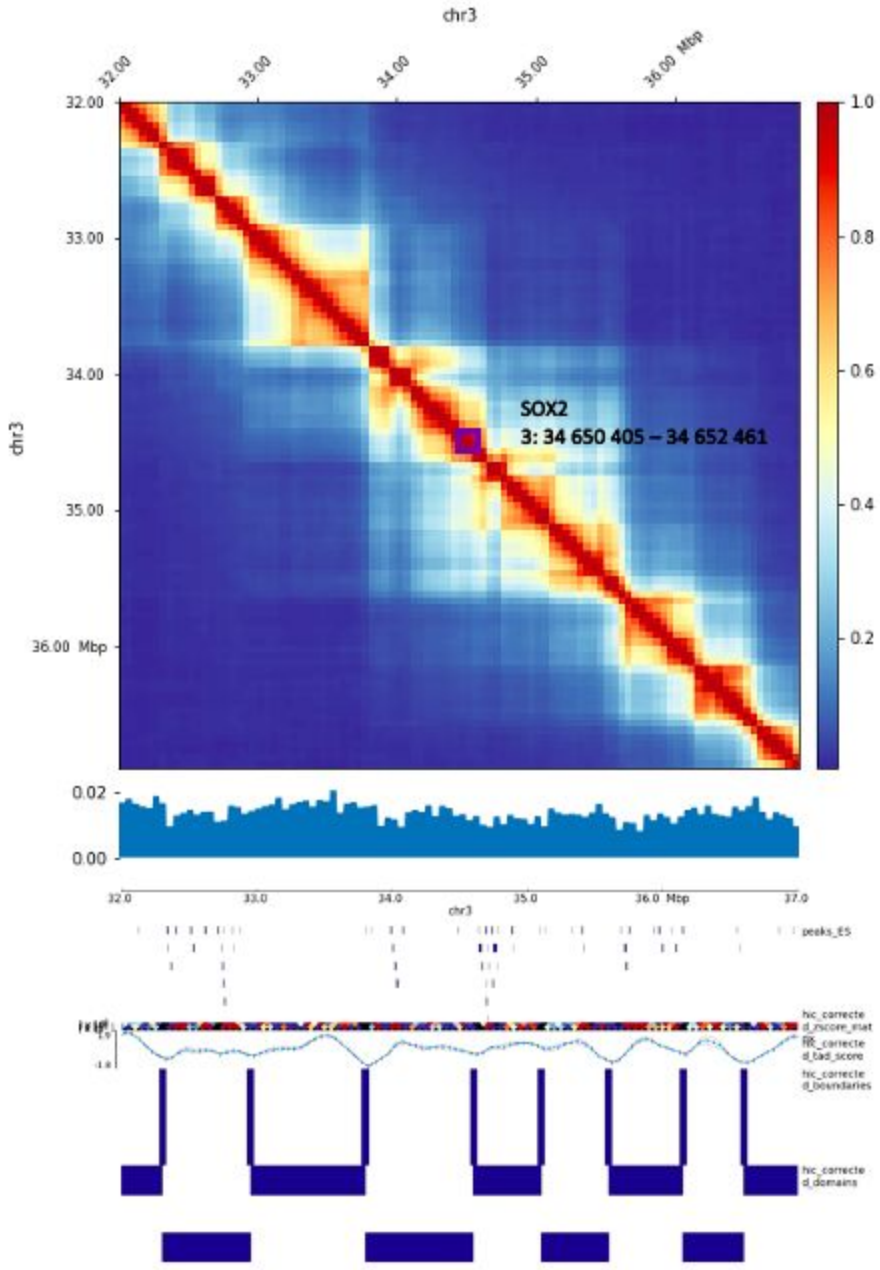

Figure 4. Corrected contact matrices for the chromosome 3 region of interest where *Sox2* is located.

Compartment A vs. B comparison

Finally, the bedtool intersect features was used to calculate the sizes and relative ratio of the compartments (Table 1). There were significantly more acetylated genes in the A compartment vs. B compartment. Housekeeping genes and all genes (annotated transcription start sites) were also more present in A compartments, but their relative ratios were much lower compared

to acetylated peaks. Between these two groups, there were relatively more housekeeping genes compared to all genes in compartment A vs. B, as shown by a slightly higher ratio of 2.48.

Feature	Compartment	Size Normalized per megabases	A/B Ratio
K27AC	A	23.67	112.55
K27AC	B	0.2103	112.55
HK	A	94.89	2.48
HK	B	38.203	2.48
TSS	A	0.02103	1.33
TSS	B	0.01583	1.33

Table 1. A/B compartment comparison in normalized size per Mb and A/B ratios. Bash Script and R Script for ratio calculation.

Discussion

In this experiment we used both the Hi-C Bench and Hi-C Explorer pipelines to analyze two biological replicates of mouse embryonic stem cells. Both pipelines yielded similar results, and Hi-C Explorer allowed additional visualization of peaks, TAD scores, and domain boundaries overlayed with the contact matrices. Our samples were taken from mouse embryonic stem cells and aligned against the mm10 reference genome. Alignment of sample data to a reference genome is an important step in inferring which genes are expressed. It was expected that only a fraction of the reads generated could be used for downstream analysis due to possible duplicate reads, reads mapping to multiple sites, or overall unmappable reads. The counts and percentages from the alignment and filtering quality control step were consistent with this expectation, and the mappable reads were used for the downstream analyses. We then visualized our filtered .hiC files on Juicebox. Highly red areas indicate the presence of certain chromosome structures such as TADs, compartments and 3D chromatin loops. The contact matrices showed that many of these structures of interest were present within the same chromosome. Although they are still being actively studied, many of these structures are thought to interact and regulate gene expression regulation, and their proximity to one another makes them more likely to interact. After step 7 of the pipeline, we locally copied a bedGraph file containing compartment scores for A and B compartments. A compartments represent euchromatin, portions of chromatin that are acetylated and therefore open to transcription because they are no longer tightly wrapped around the histone. In contrast, B compartments represent heterochromatin, which is densely packed and has overall lower gene expression. The IGV visualizations compared A and B compartments and levels of expression against housekeeping genes and all annotated gene transcription start sites. Housekeeping genes are expected to be actively transcribed within stem cells, are they are crucial for many biological processes involved in both stem cell maintenance and development. Other genes were not necessarily expected to be actively transcribed, as they could be dependent on external

conditions/factors/stimuli, important for stem cell differentiation, or any number of other functions. The results confirmed that the housekeeping genes were more highly correlated with A compartments and therefore more highly expressed, whereas the rest of the genes were split more evenly between A and B compartments. To further support this idea, we focused on the expression of the *Sox2* gene in our visualizations. As *Sox2* is a transcription factor vital to maintaining pluripotency in embryonic and neural stem cells, we expected its locus to fall in the A compartment--this was confirmed with our local IGV visualizations.

This experiment helped us understand the organization of chromatin within the genome of a mouse embryonic stem cell. Future analyses could be conducted to rigorously compare the two pipelines, further explore regions of interest, and better understand self-interacting and associating structures within chromatin. This has implications for understanding various biological dysfunctions such as oncogene activations and disruptions in developmental processes.

Self-Assessment

All members contributed equally to the execution of the experiment, discussion of results, and drafting of the report. Sarnai, in particular, completed the experiment with the HiCExplorer software and collated her results into this report. The remaining group members conducted simultaneous analyses using the HiC-bench pipeline. To avoid duplication, one set of HiC-bench results has been included with the report.

Note: Many of us also looked at other genes that are important in the maintenance of embryonic stem cells or for which we had a personal research interest. The images are in the links.

Sarnai - 5/5
Sabrina - 5/5
Tania - 5/5
Aravind -5/5
Yifan - 5/5
Mingzhou - 5/5