Amatya Pathak, Christian Raimondo, Briana Mullins, Caroline Martinez
December 5, 2021
BMIN-GA 1001 Bioinformatics: Topic 6

Assignment 5: Chromosome Conformation Capture

INTRODUCTION

As we work our way through increasingly macroscopic lenses of gene analysis, the next objective is to observe and analyze how eukaryotic chromosomes are structured, oriented and spatially organized, plus how their configurations affect gene expression.

All eukaryotic cells store their DNA in more condensed forms known as chromatin, held together by histone and chaperone proteins. Mammalian chromosomes in particular are organized in two compartments: “A” for less condensed, actively expressed genes, and “B” for more condensed and more repressed genes.

HiC is a method that couples the three-dimensional structure of a genome with it’s genomic sequence, derived from previous 3C chromosome capture technologies. Each of these technologies quantify the interactions between physical chromosome locations called loci, that might be significantly separated in a linearized genome, but almost adjacent when folded and condensed in a real three-dimensional space. Some of these 3D interactions may be random polymer looping collisions in DNA condensation, but more interestingly, others might be examples of a promoter-enhancer complex. Enhancers are DNA sequences that can “remotely” trigger the promoter sequence to begin gene transcription, as they do not have to be close to the expressed gene in the linear sequence.

The types of 3C technologies primarily differ in their scale. The original, eponymous version compares interactions between one pair of loci. A larger scale version is circular 3C, or 4C, which compares one locus to all loci in the genome. There is also 3C carbon copy (5C) which parallely compares hundreds of thousands of loci pairs, and finally, Hi-C, named for the high-throughput sequencing used to find the nucleotide sequences and compare all loci to each other. This assignment will implement the latter technology.

Hi-C works by sampling a short sequence from each fragment to be compared, and combining them together. For any combined, or ligated fragment, the two obtained sequences correspond to different restriction fragments. These can individually be aligned to the genome to determine their identity.

For this assignment, we analyzed mouse embryonic Hi-C data using Hi-C pipelines available on Big Purple. After aligning and filtering the raw fastqs, we were able to use contact matrices to then visualize the chromatin structure of the data and perform downstream compartment analyses.

METHODS

Before executing the Hi-C bench pipeline, two preprocessing steps were performed. First, a fastq directory was created and the Hi-C raw data available in BigPurple was linked for both samples. Next, a sample sheet was created to specify sample, group, fastq files, genome, enzyme, and cell-types for the pipeline analysis. For both samples, the group was set to 'ES_rep', the genome was set to 'mm10', and the enzyme used was 'Arima',

Next, the Hi-C-seq standard pipeline was performed on the inputted data. The mouse Hi-C and matched H3K27ac ChIP-seq fastq reads were aligned to the mouse genome using Burrows-Wheeler Aligners (BWA) and filtered on HiC-Bench using gtools ('01a-align'). The filtered reads were then assessed for quality control with HiC-Bench's '02b-filter-stats' step. After ensuring the .fastq files were good quality, Juicer, a loop and contact domain processing software, was used to generate contact matrices in binary format as a 'filtered.hic' file ('03a-tracks'). A software tool called 'Homer' was then used to analyze compartments (04a-compartments). These files were then visualized on Web Juicebox to identify A/B compartments. Normalized matrices were assessed for topologically associated domains (TADs), informally called loops. Then, the TAD boundary-scores, H3K27ac peak density, and housekeeping (HK) gene density were calculated.

For the downstream analysis, the bedtools module was used to cross reference the A/B compartments with the bedfiles including transcription start sites, HK genes, and H3k27ac peaks. The resulting '.tsv' files were uploaded into RStudio and the densities of shared markers were visualized for comparison between compartment A and compartment B.

RESULTS

QC Metrics

The aligned chimeric mouse fastqs showed similar quality when filtered (shown in Figure 1). The largest proportion of DNA for both samples were double-stranded accepted hits-intra by read counts and percentage. Very little DNA was found to be unmapped, unclassified, or unpaired.

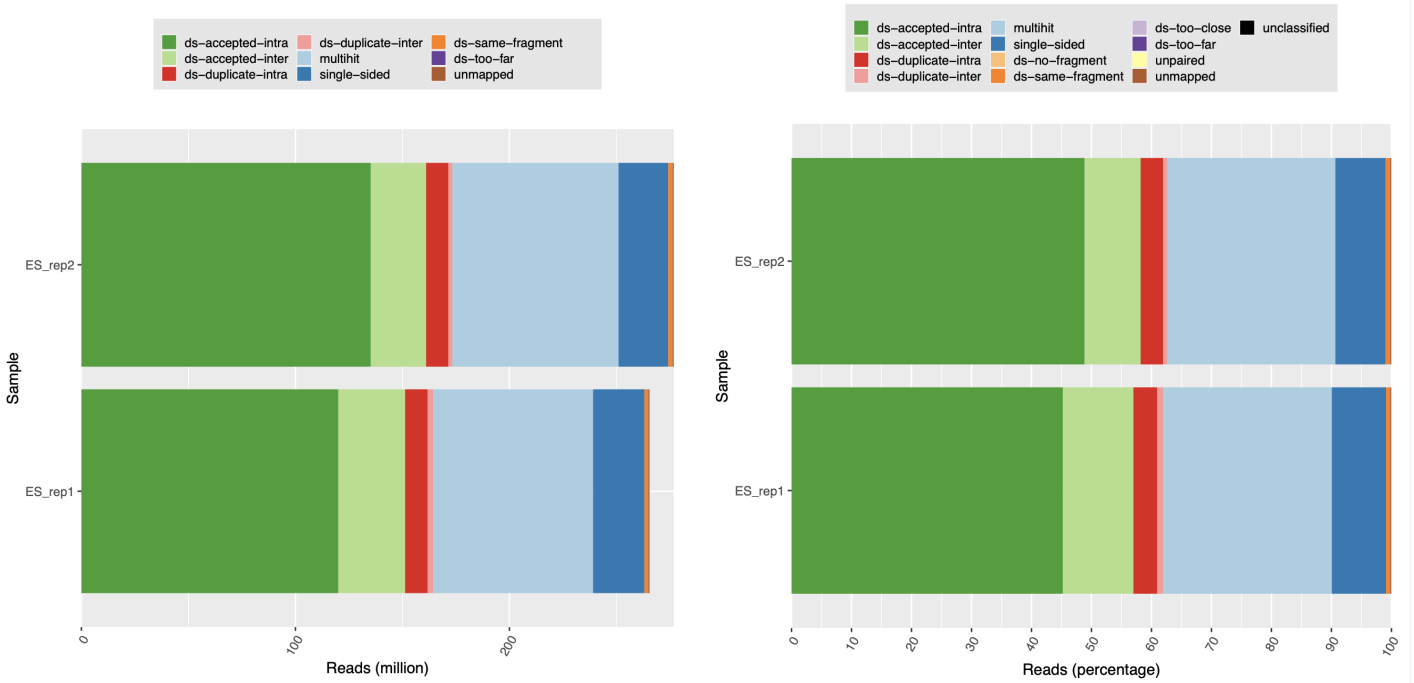

Figure 1. Quality Control statistics of aligned and filtered sample reads showing count and percentage of double-stranded, single-stranded, multi-hit, unmapped DNA etc. The left image shows counts per million per sample read and the right image shows percentage per sample read.

JuiceBox Visualization

Following the Juicer and Homer analyses, these two mouse samples were then able to be uploaded into JuiceBox for further visualization of Compartment A/B, housekeeping (HK) genes, transcription start site genes (TSS), and Peaks. We chose to focus our JuiceBox analysis on Chromosome 5 of the mouse genome. **Figure 2** shows the data using 25 kb resolution. Compartment scores appear to be positive in compartment A and negative in compartment B. Housekeeping genes can be seen to fall within Compartment A. Additionally, most transcription site genes visually appear to fall within Compartment A, and almost all HK327ac Peaks can be seen to fall within Compartment A. Numerous topologically associated domains (TADs), as represented by red triangles along the diagonal, can be seen at this resolution (both intra- and inter-TADs). Their frequency appears to decrease in Compartment B. Loops are difficult to locate at this resolution. **Figure 3** shows the same Chromosome, however the resolution is at 10 kb. At this resolution we see only the Compartment A loci. Additionally, TADs are more visible as well as Loops (represented by red dots). There appears to be a lack of a TAD which corresponds to a large space in histone peaks.

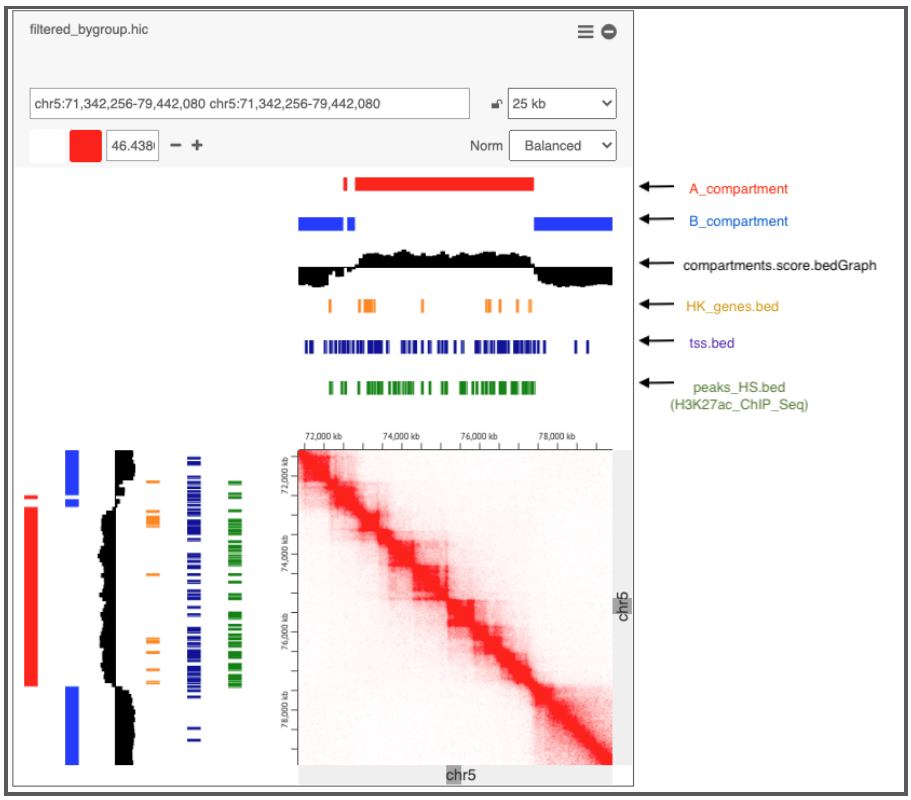

Figure 2. JuiceBox visualization of Mouse Chromosome 5 HiC data mapped alongside housekeeping genes, transcription start sites, and matched CHIP-Seq H3K27ac histone markers, at a resolution of 25 kilobases (kb).

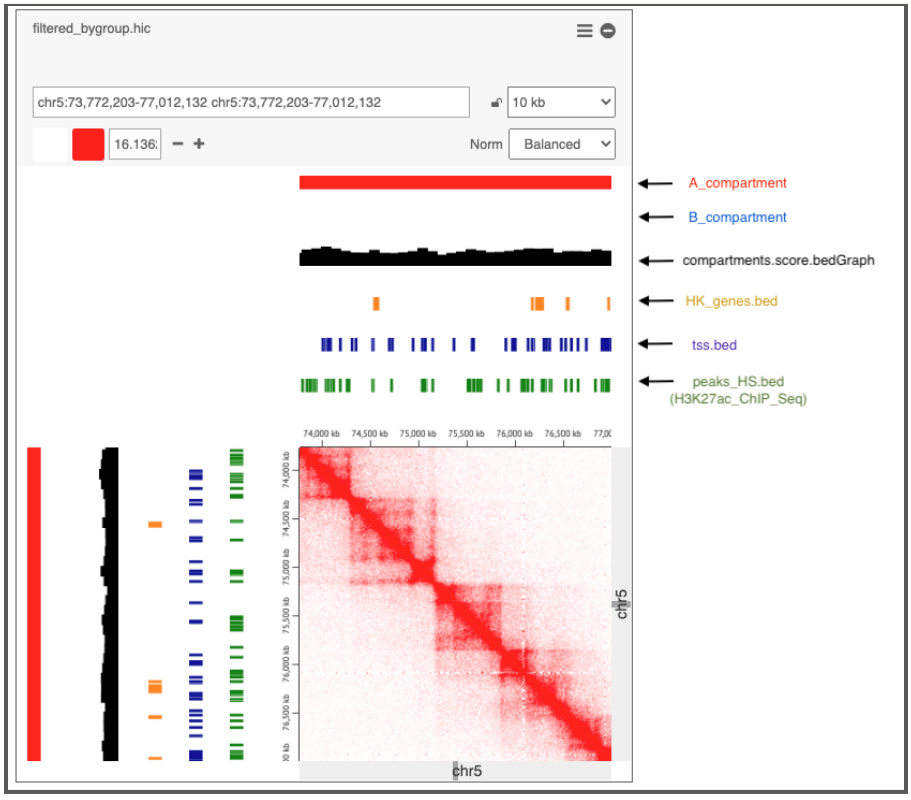

Figure 3. JuiceBox visualization of Mouse Chromosome 5 HiC data mapped alongside housekeeping genes, transcription start sites, and matched CHIP-Seq H3K27ac histone markers, at a resolution of 10 kilobases (kb).

Compartment A/B Comparison

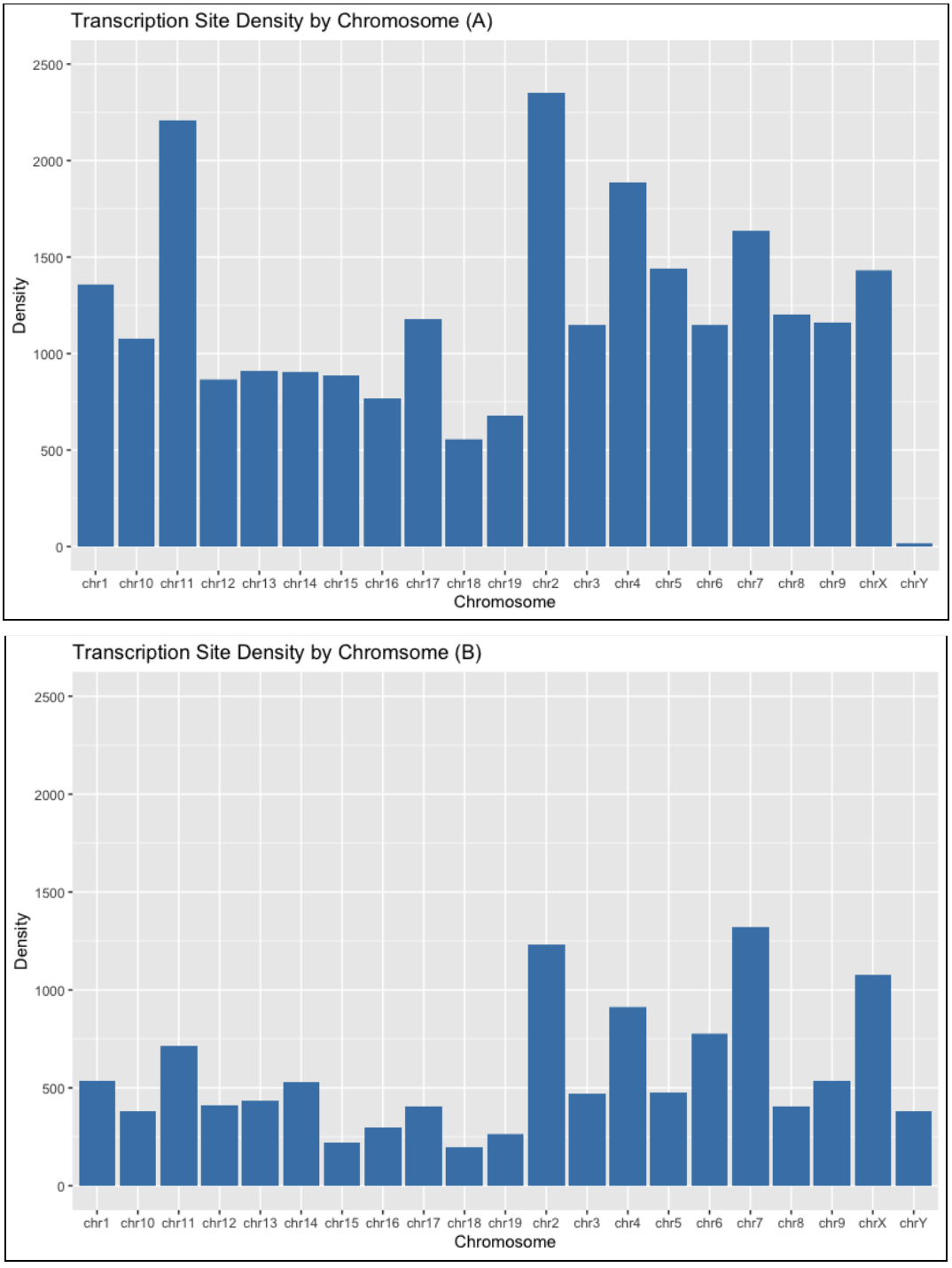

Figure 4. Gene densities of Transcription Start Sites displayed by chromosome for compartment A (top) and compartment B (bottom).

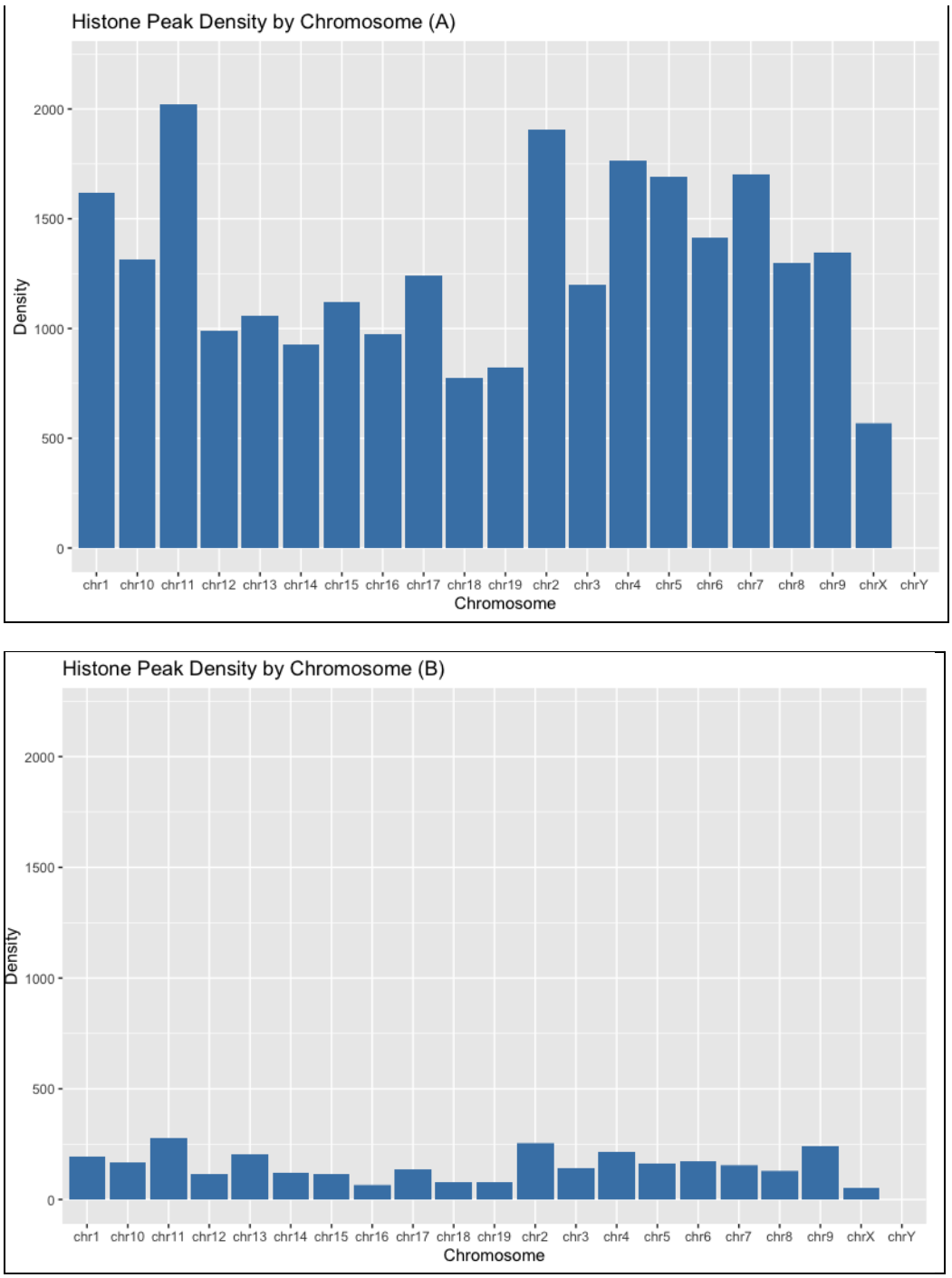

Figure 5. Gene densities of histone markers (H3K27ac Peaks) displayed by chromosome for Compartment A (top) and Compartment B (bottom).

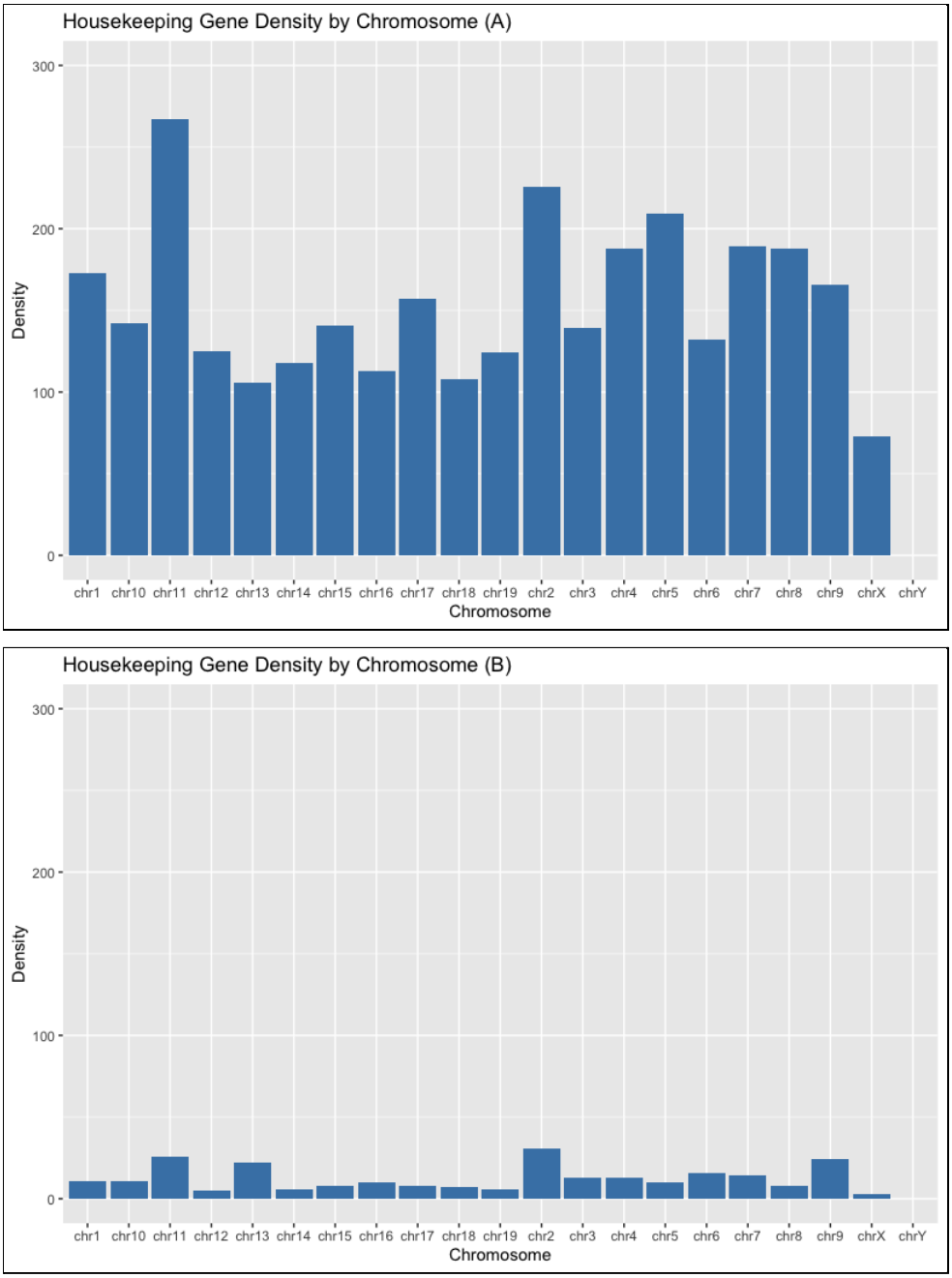

Figure 6. Gene densities of Housekeeping Genes displayed by chromosome for Compartment A (top) and Compartment B (bottom).

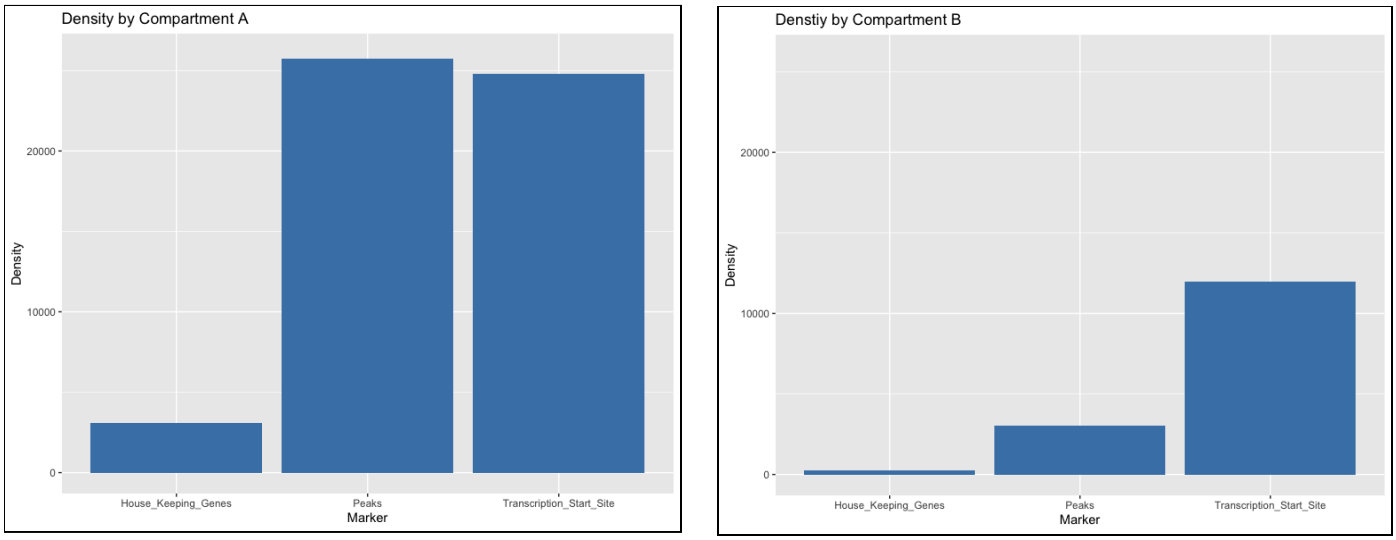

Figure 7. Sum of Gene densities of housekeeping genes and transcription site genes as well as density of histone markers (H3K27ac Peaks) by Compartment A (left) and Compartment B (right) across all chromosomes.

The frequency of housekeeping genes, transcription start sites, and acetylated histone peaks (H3K27ac Peaks) can be visualized in **Figures 4-6** per chromosome, where we cross reference their densities in the mouse samples with reference data . The transcription start sites appear to have a greater density in compartment A (**Figure 4**) and also appear to have increased density on chromosome 11,2, 4,7, and X. The acetylated histone peaks have a much greater density in compartment A compared to compartment B (**Figure 5**). Their spread across chromosomes is much more even, however there is a greater density of peaks on chromosome 11 and 2 (Compartment A). Housekeeping genes also show increased density across all chromosomes in compartment A compared to compartment B (**Figure 6**), with increased densities among chromosomes 2 and 11.

The frequency of housekeeping genes, transcription start sites, and acetylated histone peaks (H3K27ac Peaks) are summed in **Figure 7**. All three markers of gene transcription enhancement occur at greater frequency in Compartment A compared to compartment B. Transcription start sites and histone peaks occur at much greater density than housekeeping genes.

DISCUSSION

Using the Hi-C Bench pipeline on BigPurple, we were able to align, filter, and visualize high dimensional markers of chromatin structure for mouse fastq sample reads. We were able to generate a contact matrix to define areas of high chromosome contact and assign compartment scores based on topological features and genetic sequence proximity. These compartments (A and B) were investigated further using JuiceBox and density analyses of gene enhancement.

In JuiceBox, we visualized Compartment A and B on mouse Chromosome 5 at both 25kb and 10 kb, cross-referencing the compartments with mapped housekeeping genes, transcription start sites, and histone markers (H3K27ac peaks). From this visualization we could clearly see that compartment A had positive compartment scores, while compartment B had negative compartment scores. Additionally, housekeeping genes, transcription start sites, and histone marker peaks were clearly associated with Compartment A. At 10kb resolution we were able to zoom in to the Compartment A loci and better visualize TADs and Loops.

When comparing the presence of housekeeping genes, transcription start sites, and histone marker peaks between compartment A and B (using reference data), it became obvious that these genes and transcription site enhancers are at a higher frequency in Compartment A than Compartment B (see **Figures 4-6** for density by chromosome, **Figure 7** for summed chromosome densities).

Both the JuiceBox visualization and our own visualization of transcription site enhancer frequencies indicate that areas of greater transcription occur in Compartment A. It is known that TADs regulate transcription by restricting enhancers to genes within that TAD and therefore preventing enhancement of genes outside of this TAD. It is no surprise then that we see an increase in TADs and histone peaks within Compartment A.

Housekeeping genes are frequently expressed in almost all cell types, and encode fundamental cellular functions including but not limited to respiration, mitosis, and ubiquitination. These genes must be constantly transcribed and therefore would be presumed to occur in areas where their transcription would be enhanced. This background information matches the higher abundance of housekeeping genes visible in Compartment A, both via JuiceBox visualization and counted frequencies. The samples analyzed are specifically mouse embryonic stem cells which would require timely and coordinated transcription for differentiation into other cell types. Therefore the regulation of these cell types at the chromatin level is crucial. Our results suggest that these cell types are highly regulated via TADs, loops, and histone acetylations.

Although the information from Hi-C is vast, unlike other sequencing technologies which have single cell granularity, we cannot extrapolate our results to single cells. A potential improvement could combine single cell sequencing with higher dimensional chromatin methods like HiC for more scalable results, which would be advantageous in the future.

CONTRIBUTIONS:

Christian and Briana ran the BigPurple pipeline and created the R scripts to visualize the gene density charts, while Amatya and Caroline compiled the written report.