UNIVERSITY OF CALIFORNIA SAN DIEGO

# Investigation of 3D Chromatin Modularity in Mouse Development

A thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Chemistry

by

Xiaofu Wei

Committee in charge:

Professor Wei Wang, Chair
Professor Galia Debelouchina
Professor Dong Wang

2023

The thesis of Xiaofu Wei is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

*To my dear parents*

*and Kolin*

*for their unconditional love, trust, and support.*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am grateful to have had the opportunity to meet Prof. Wei Wang during my second quarter at UCSD. His eye-opening experiences have been a great motivation for me to work in the field of computational biochemistry, which was entirely new to me. The knowledge and skills I acquired from him and other members of the group have provided me with a solid foundation in sequencing data analysis and genome-related research. Being a beginner in bioinformatics research, I am thankful to have joined Prof. Wang's lab and had the chance to work on unexplored problems. The interactions with Prof. Wang and the lab members have been invaluable and will benefit me throughout my career.

I am fortunate to have the support of many outstanding lab members. In particular, I would like to express my gratitude to Lina Zheng for helping me get started in the program and providing strong support in statistics for biology. I am also thankful to Cong Liu for assisting me with server usage. The guidance and support from Eunice Choi, Peiyao Wu, Yanmiao Du, and other lab members have been instrumental in expanding my research perspective and utilizing various software tools effectively.

I would also like to extend my appreciation to Prof. Galia and Dong for their valuable help and suggestions during the preparation of my thesis. Additionally, I am grateful to the exceptional professors at UCSD, including Prof. Brian and Prof. Kevin, who have imparted invaluable knowledge in various areas of biochemistry and computational tools.

This thesis includes coauthored unpublished material with Wei Wang, for which the thesis author serves as the primary investigator and author. The material presented in this thesis represents the contributions and findings of the thesis author.

VITA

| | |
|---|---|
| 2016-2020 | B. S. in Pharmaceutical Chemistry, University of California, Davis |
| 2021-2023 | M. S. in Biological Chemistry, University of California San Diego |

ABSTRACT OF THE THESIS

**Investigation of 3D Chromatin Modularity in Mouse Development**

by

Xiaofu Wei

Master of Science in Chemistry

University of California San Diego, 2023

Professor Wei Wang, Chair

Chromatin structure plays a crucial role in various genomic processes in eukaryotic cells, including genome replication, transcriptional silencing, and gene regulation. Extensive studies have focused on the three-dimensional organization of the genome, revealing the presence of topologically associating domains (TADs) and compartments, which are defined by spatial contacts identified through techniques such as Hi-C. However, understanding the direct role of histone modification in shaping the three-dimensional genome structure remains an ongoing challenge.

This thesis investigates changing patterns of regulation-associated modules (RAMs) in mouse development to understand the organization and function of RAMs and their boundaries.

RAMs, proposed in previous studies using human samples, offer insights into genome organization and regulation. However, comprehensive explanations for RAM formation, functions, and boundary factors are lacking.

Using the ”findRAM” tool, we have identified RAM regions and boundaries from a dataset of 72 mouse embryonic samples. Pairwise comparisons between tissues at specific time points and between subsequent times within the same tissue have revealed changes in RAMs. Through genome enrichment analysis of these regions, we have identified functional pathways, including cation binding, metal ion binding, and transcription-related pathways. Additionally, consensus RAM (cRAM) regions have been determined for each time point and tissue, highlighting regions that exhibit consistent patterns of RAMs and boundaries. Gene enrichment analysis has provided further support for some of the findings from pairwise comparisons, and these findings align with the potential mechanism of RAM boundary formation proposed in previous research on RAMs.

In conclusion, this thesis investigates 3D chromatin modularity through RAM analysis in mouse development data. We have identified pathways and genes potentially involved in RAM boundary formation through computational prediction and discussed improvements for the RAM identification model. These findings contribute to our understanding of the formation and functions of RAMs and boundaries, which are determined by histone modification marks. Ultimately, these findings highlight the connection between the structural and functional modularity of the 3D genome.

# Chapter 1

# Introduction

The organization of the genome in eukaryotes is complex and hierarchical, with the genome packaged inside the nucleus in a non-linear manner. The three-dimensional structure of the genome such as higher-order chromatin organization, which is linked to long-distance gene regulation that controls development and cell fate commitment [1], plays a crucial role in biological processes. Proper chromatin condensation and decondensation are important for accurate chromosome segregation during mitosis and meiosis. Besides, higher-order chromatin organization defects can cause developmental irregularities and illnesses[2]. Recent advancements in technology and analytical pipelines have revealed patterns associated with chromatin organization, including compartments[3] and topologically associated domains (TADs)[4]. TADs are regions of the genome where DNA sequences within that region interact more frequently with each other than with sequences outside that region, and the TAD boundaries are demarcated with CTCF sites or actively transcribed DNA sequences[5]. Compartments are regions of the genome with distinct patterns of chromatin accessibility and transcription activity, classified as compartment A (high levels of transcription activity and open chromatin) and compartment B (lower levels of transcriptional activity and more compact chromatin) [3]. Both findings are derived from Hi-C contact maps, which is a high-throughput genomic and epigenomic technique to capture

chromatin conformation [6]. Recently, some computational models have shown that histone modification signals are predictive of enhancer-promoter interactions [7], TAD boundaries[8], and compartments[9]. Proteins involved in transcriptional regulation, active promoters and enhancers, and transcriptional activity tend to form clusters in the nucleus, tightly associated with histone modifications [10][11].

Despite histone modifications reflecting chromatin activity in previous studies, the direct inference of the spatial modularity of the genome from histone modification patterns has not been explored. Unlike topologically associating domains (TADs) and compartments derived from Hi-C maps, the regulation-associated module (RAM) is a novel module that utilizes frequency profiles of H3K27ac histone modification peaks from chromatin immunoprecipitation sequencing (ChIP-seq) data [12] to generate a more comprehensive pattern across the entire genome in cells. H3K27ac modification involves acetylation of the lysine residue at position 27 of the histone H3 protein, and it is often considered a marker of active enhancers and promoters[13]. ChIP-seq is a powerful method for identifying genome-wide DNA binding sites for transcription factors and other proteins [14]. Histone modifications, such as H3K27ac and H3K4me3, play a critical role in determining chromatin structure and regulating gene expression. Active marks such as H3K27ac and H3K4me3 open chromatin to allow access to transcription factors (TFs) to promoters or enhancer. In contrast, repressive marks can condense chromatin and suppress gene expression [15]. Both active and repressive histone modifications contribute to the formation of euchromatin and heterochromatin, which differ in their level of compactness. These findings underscore the significance of histone modifications in shaping the three-dimensional structure of the genome at both regional and global levels. We chose H3K27ac peak density files due to their high correlation with other histone modification marks in previous RAM research on humans, and they can reflect signals from active enhancers and promoters.

For identifying the RAMs, we applied the computational tool called 'findRAM' that was previously used for predicting RAMs for humans. This method employs sliding window

strategies with a fixed flanking size of 500kbp and step size 250kbp to compute H3K27ac peak densities in the linear genome [12] for the data of both human and mouse to reach the maximum of shared boundaries across samples. In the previous study, they analyzed 93 normal samples and 19 cancer samples of humans. They provided evidence to demonstrate that RAMs are spatial modules, where enhancer-promoter interactions and ecDNA occur dominantly within RAMs, and RAMs are resistant to cohesion degradation. They also suggested that the RAM boundaries exhibit more insulating functions compared with topologically associating domains (TADs).

Furthermore, they not only showed the big differences between RAMs and other existing 3D chromatin modules but also proposed a mechanism for how the RAM forms. Based on many other studies of multivalent cations, calcium, magnesium, and manganese can reduce the electrostatic repulsion between the DNA chains and induce DNA condensation. These cations may bind to specific DNA sequences[16] and affect nucleosome positioning[17]. Therefore, a possible mechanism can be that genomic DNAs become densely packed around cations such as Ca2+, Mg2+, and Mn2+ to form RAM boundaries. Proteins such as calcium-binding proteins that carry many cations and their interacting partners may recognize specific DNA sequences such as those motifs enriched in cRAM boundaries to facilitate locus-specific localization of cations.

With numerous observations of RAMs (repetitive array motifs) in human data, we become increasingly curious about the reasons behind RAM formation, which exhibit both similar and distinct boundaries across samples. In this case, our objective is to understand how changes in time influence RAM alterations within the same tissue during developmental stages and how RAMs differ between tissues at identical time points. Additionally, we have attempted to investigate the functions of RAMs by examining the differences in gene expression levels within each changing region. Lastly, we aim to identify the genes associated with cation binding that contribute to the formation or undergo changes in RAM boundaries. However, the formation and patterning of RAMs in developmental data, as well as whether these patterns are shared across different species, remain unknown.

Understanding the patterns of RAMs across different species during development is crucial to comprehend the fundamental mechanisms of gene regulation. In this study, we applied the RAM calling modal 'findRAM'[12] to 72 samples of 12 different mouse embryonic tissue bulk ChIP-seq data of H3K27ac at eight variable time stages. By comparing results for different tissues at various time points and different consequent times at each tissue, we analyzed the expressed genes, differentially expressed genes, and regulatees, which are genes under the regulation of specific transcription factors in each variable RAM region.

We then analyzed pathways of genes under those variable regions using the g: profiler [18], a toolset widely used for finding biological categories enriched in gene lists, conversions between gene identifiers, and mappings to their orthologs. Interestingly, we found that many pathways shared a lot in most pairwise comparisons and are involved in various biological processes, including cell development and differentiation, tissue remodeling, and immune responses. To check the variability across all samples in each time point and in each tissue, we defined the consensus RAM (cRAM) and cRAM boundary regions in each time and tissue. After putting the cRAM boundaries at different time points and tissues together, we found that over 80% of cRAM boundaries are conserved in most time and tissue stages. We also observed a pattern where cRAM boundaries merged in some time or tissue stages and split into two or more in rest, which we referred to as MS regions. We further investigated the genes in these regions that have the function in the nucleus and identified some crucial genes for histone modification and chromatin organization, such as GATA3, ING3, and HDAC gene family. Some of those genes are correlated to cation and metal binding, which matched the hypothesis of the mechanism of how RAM forms mentioned in previous studies[19]. Validating the functions of these genes in the future with experiments may have significant implications in understanding the mechanisms of gene regulation and their potential roles in 3D chromatin organization and eventually leading to the improvement of disease studies.

Based on all our findings, we have realized that RAM patterns exhibit generic functions

in the majority of comparisons within variable and MS regions. However, interesting pathways and genes emerge when we examine differentially expressed genes (DEGs) and regulators in variable regions. Overall, our study provides insights into the patterns of RAMs in mouse embryonic development data and sheds light on the potential role of RAMs in gene regulation. Our findings can serve as a valuable resource for future studies aiming to explore further the molecular mechanisms underlying RAM formation and its potential impact on gene regulation in development and disease. Nonetheless, there are still some improvements required for the 'findRAM' model and additional analysis needed for the down-strain analysis. These aspects will be discussed in Chapter 4.

# Chapter 2

# Methods

## 2.1 Methods for variable RAM and gene identification

### 2.1.1 Data Source

The 72 embryonic mouse bulk samples with processed narrow peaks ChIP-seq of H3K27ac in mm10 and the 72 embryonic mouse samples with processed gene quantification RNA-seq in mm10 were downloaded from ENCODE portal (https://www.encodeproject.org/)[20]. All the downloaded data met the ENCODE standards. Reference Table A.2 lists all the samples used in this study. Gene names and TSS regions are annotated using the vM21 annotation file for mouse downloaded from 'Gencode'[21].

### 2.1.2 RAM Identification in Individual Samples

After installing 'findRAM', we identified the H3K27ac narrow-peak density using a sliding window with a step size of 50kb, 100kb, 250kb, and 500kb, respectively, and a flanking size of 500kb for each window in every sample. The H3K27ac narrow-peak profiles were then smoothed using local polynomial regression fitting[22]. The RAM boundaries, identified as

valleys in the smoothing curves, and peaks, identified as summits in the smoothing curve, were detected using the ”findpeaks” function in the R package ”pracma” (R v4.1.2). RAM boundaries were also determined by any density peaks smaller than 0.1 proportional to the highest density peak in each chromosome for each sample.

### 2.1.3 Variable RAM Identification

For pairwise comparisons, we treated each tissue as a control group once and compared it with the remaining tissues at each time point to perform tissue comparisons. Additionally, for each tissue, we treated the early embryonic time point as a control group and the subsequent embryonic time point as the sample group. For example, in the case of the forebrain tissue, we compared embryonic day 10.5 as the control with embryonic day 11.5 as the sample, and then we compared embryonic day 11.5 as the control with embryonic day 12.5 as the sample. We recorded all regions that exhibited changes from boundaries in the control group to RAMs in the sample group. We recorded only those changing regions that showed a shift of at least 2 bin sizes between the control and sample groups. All analyses were performed using Python 3.8.

### 2.1.4 Expressed Genes Identification

We initially performed quantile normalization to align the two replicates in each RNA-seq sample, using the gene quantification files. Subsequently, we applied the TPM normalization method, taking into account the effective length and gene counts provided in the downloaded files. We assessed the distribution of log(TPM) for each gene in each sample by creating histograms. To mitigate the impact of noise, we set a cut-off at log(TPM) greater than 0. We then calculated the average TPM value for the two replicates and selected genes that exhibited TPM values greater than 1, expressing in over 60% of all samples. All analyses were performed using Python 3.8.

### 2.1.5 TSS Region Identification

In our annotation file, we lacked specific information regarding the promoter loci. However, it is widely recognized that the core promoter region, known as the transition starting site (TSS), is located in close proximity to the starting codon position, for which we possess detailed annotation data. Hence, we extended the annotation of the start codon by 1000bp in both the 5' and 3' directions for each gene and employed these extended loci as the TSS region. We utilized the same annotation file (vM21) that we employed for the gene loci. All analyses were performed using Python 3.8.

### 2.1.6 Regulatees Identification Using Taiji

After installing Taiji, we proceeded to run it using narrow-peak density files of H3K27ac ChIP-seq data and gene quantification RNA-seq data for each embryonic mouse sample. The data was obtained directly from ENCODE, as described in Chapter 2.1.1. Since we were working with bulk data, we utilized Taiji's EpiTensor functionality. Our analysis began by identifying the driver transcription factors (TFs) for each tissue and time point, employing the PageRank score as a measure. We first selected the top 12% of TFs with average ranking scores larger than 0.002 of all TF candidates across all samples, out of a pool of 880 TFs. Then, we selected those with a coefficient of variation (CV) value smaller than 0.3 for tissue comparisons and 0.4 for time comparisons. Subsequently, we applied a filtering process, selecting the top 700 regulatees under each driver TF based on the network score in each sample. Finally, we consolidated all the regulatees from all driver TFs in each sample and focused on the top expressed 4500 regulatees.

### 2.1.7 DEG Identification

We first calculated the sum of gene counts across all samples and replicates for each gene. We selected genes with a sum greater than 500 as candidates for further identification

of differentially expressed genes (DEGs) in each comparison. For the analysis, we utilized the 'DESEQ2' package (R v4.1.2) and selected DEGs with a logfold change value greater than 2.

### 2.1.8 cRAM Identification

To identify cRAMs, we analyzed the occurrence frequency of boundaries across all samples at different thresholds (30%, 40%, and 50%) for each time point and tissue. We observed that using a threshold of 30% resulted in a high consensus rate (over 93%) within each cRAM, while a threshold of 50% achieved an 80% consensus for shared boundaries. Based on this, we considered genome regions with an occurrence percentage above 50% as cRAM boundaries. We then merged boundaries within a distance of 250kb and imposed a minimum cRAM size requirement of 250kb. Additionally, we allowed for a one-bin size shift in the analysis. All of these analyses were performed using Python v3.8.

### 2.1.9 Merging and Splitting Region Identification

We merged all boundaries in cRAM files of each time or tissue stage separately using 'bedtools Merge'. Then, we intersected all the files to obtain the shared boundaries across all samples in each time or tissue stage using 'bedtools Intersect'. Next, we subtracted the intersecting boundaries from the merged regions using 'bedtools Subtract'. We then removed all regions that are smaller than 3 bin size. Finally, we removed those regions that occurred in less than 15% of the samples as boundaries or in more than 85% of the samples as boundaries. We utilized the gene loci in the annotation file (vM21) to generate genes located in the MS regions. All 'bedtools' functions were performed using v2.30.0, and other analyses were conducted using Python v3.8.

### 2.1.10 Gene Enrichment Analysis For Individual Genelist

Enrichment analyses were conducted using the R package 'g:profiler'. For individual analysis, we selected the top 30 significant pathways in four categories: MF (Molecular Function), BP (Biological Process), CC (Cellular Component), and KEGG(Kyoto Encyclopedia of Genes and Genomes). In each category, pathways were selected based on a p-adjusted value lower than 0.01. In enrichment analyses, the significance of pathways is determined by the p-adjusted value, where a lower value indicates higher significance.

# Chapter 3

# Results

## 3.1 Project Overview

This thesis presents the observation of the direct inference of the spatial modularity of the genome from histone modification patterns (H3K27ac) known as RAM (Regulation Associated Modules). We applied 72 mouse embryonic narrow-peak files of H3K27ac mark, and all samples are from various 12 tissue types and 8 time point. Detailed data descriptions are provided in Chapter 2.1.1 and table A.2. RAMs and boundaries were detected using the 'findRAM' method, and further analysis involved pairwise comparisons in both time and tissue scales. After identifying RAM and boundary (Figure 3.2) regions in all samples, we investigated the functionality of changing RAM regions through enrichment analysis of three categories of genes: expressed genes, differentially expressed genes(DEGs), and regulatees. Gene quantification files from the corresponding mouse stages, obtained from ENCODE, were used for gene expression levels. The methods for selecting expressed genes, DEG, and regulatees are described in Chapter 2.1.4-2.1.7, and examples were shown in figure 3.3 using integrative genomics viewer(IGV) as a presenter[23]. Overall, we conducted 60 different time comparisons and 606 tissue comparisons. For each comparison, we performed genome pathway analysis using g:profiler [18] for both time

and tissue comparisons

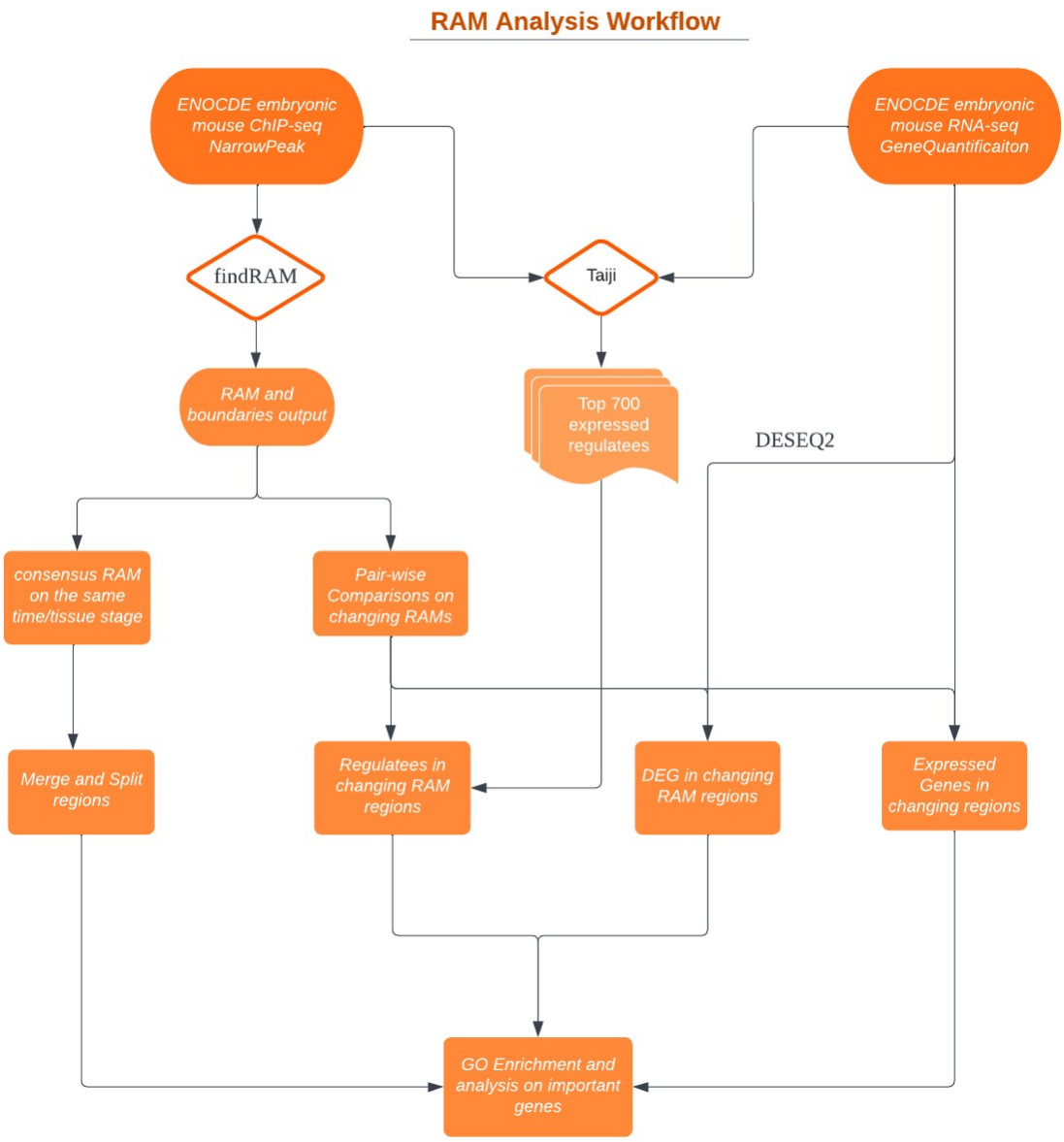



**Figure 3.1**: This is the overview of the entire thesis project. We utilized both H3K27ac ChIP-seq and RNA-seq for diverse analyses. The data for 72 mouse embryonic samples, obtained from various tissues and at different time points, were downloaded from the ENCODE website.

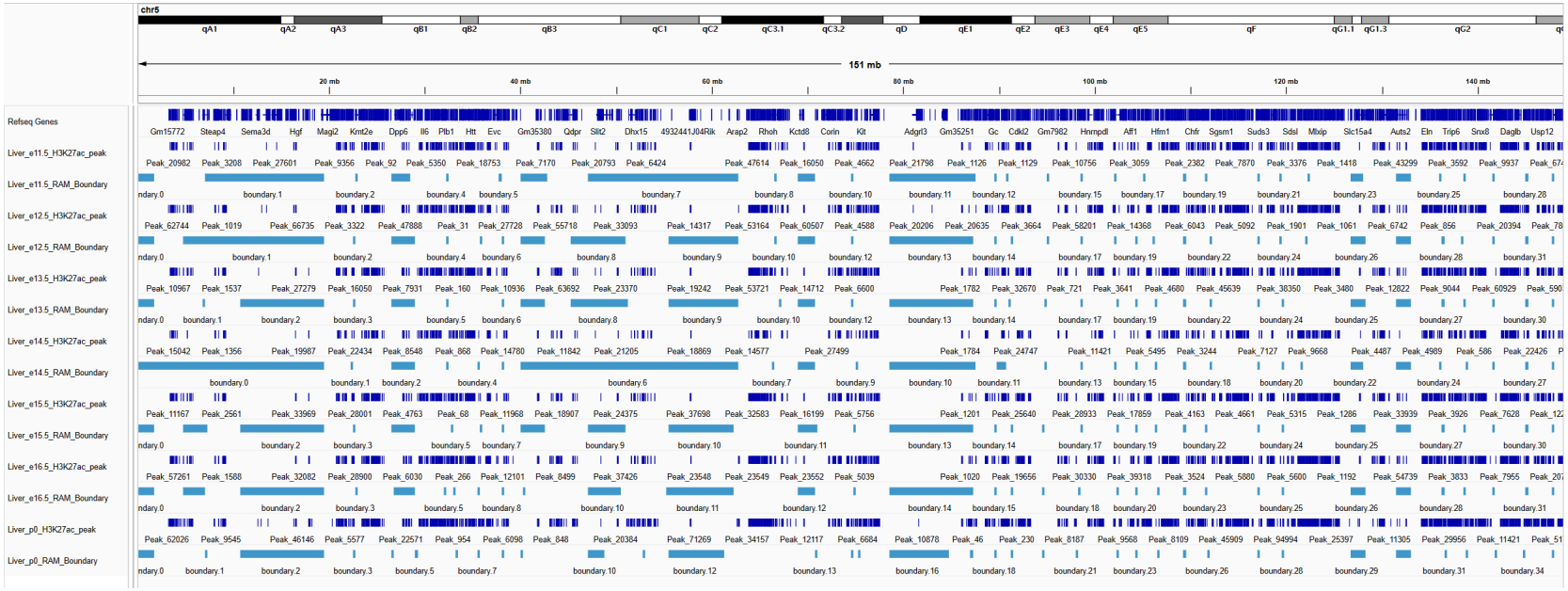



**Figure 3.2**: This is the IGV plot of histone modifications of H3K27ac in liver narrow-peaks and RAM (Reproductive Accessible Memory) boundaries at embryonic stages 11.5, 12.5, 13.5, 14.5, 15.5, 16.5, and postnatal day 0 (p0) on chromosome 5

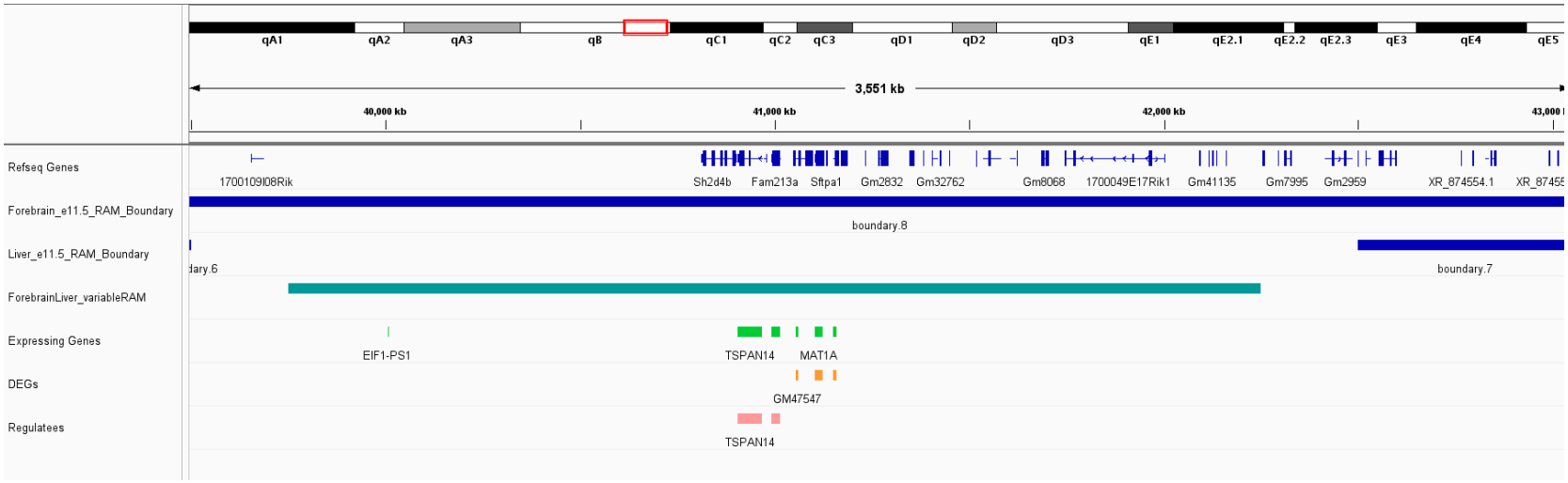



**Figure 3.3**: This is the IGV plot comparing the liver and forebrain at the embryonic time point of 11.5 days, focusing on a partial segment of chromosome 14. The forebrain serves as the control group, while the liver represents the sample group. The original RAM boundaries of the liver and forebrain are depicted in dark blue, while the RAM variable regions transitioning from boundaries to non-boundaries are shown in teal. Expressed genes within the changing RAM regions are marked in green, differentially expressed genes (DEGs) are highlighted in orange, and the regulatees are indicated in pink.

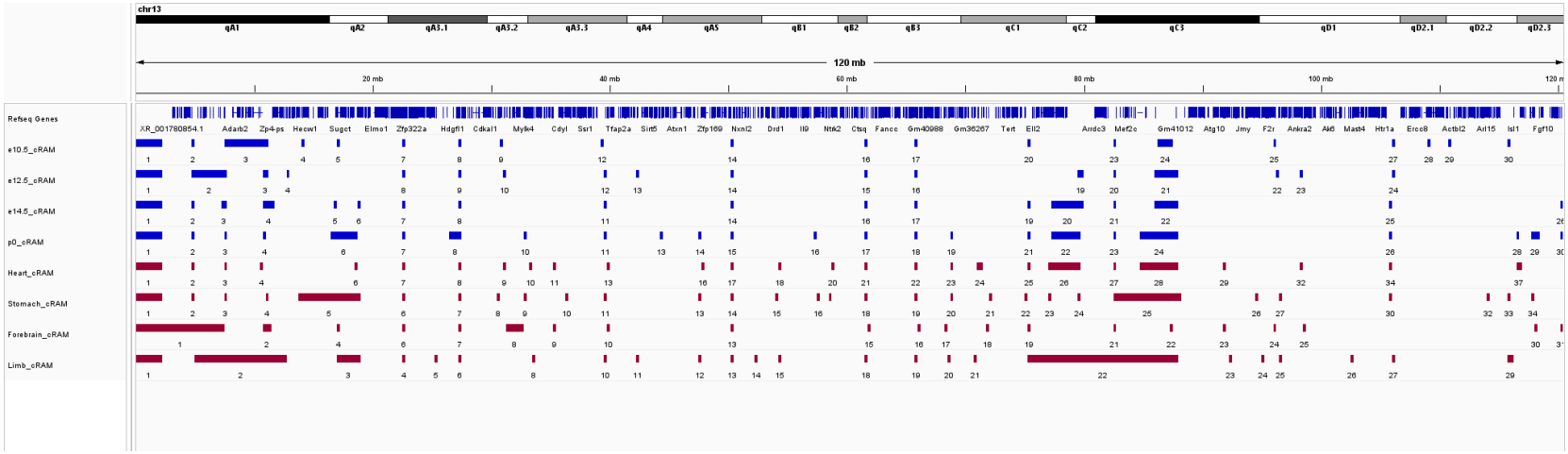



**Figure 3.4**: This is an example of cRAM results shown in IGV. The colored regions represent boundaries, while the uncolored regions represent RAMs. Here, we present cRAM results for four different time points (in blue) and four tissues (in red) on chromosome 13.

After analyzing the enriched pathways in pairwise comparisons, we proceeded to assess the variability of the data at each time and tissue stage. Consensus RAM (cRAM) and boundary regions were defined for each time point based on all tissues, and for each tissue based on all time points. We then examined the pattern changes for all cRAMs in tissue and time comparisons (Figure 3.4).Our focus was on the boundary regions that merged in some cases but split in others, which we referred to as merging and splitting (MS) regions. Pathway analysis was performed on these regions, and we conducted a gene search within the interesting pathways to identify genes correlated with histone modification and chromatin organization, specifically located in the MS regions. These genes were found to be associated with cation and metal ion binding.

## 3.2 RAM Identification

The identification of RAM involved using the 'findRAM' pipeline on a dataset consisting of 72 different samples of peak density files of H3K27ac. To initiate the analysis, we conducted preliminary tests using various sliding window sizes, including 50kb, 100kb, 250kb, and 500kb, with a 500kb flanking size for each window in mouse samples. We observed that increasing the step size resulted in larger RAM sizes and a higher percentage of shared RAMs among the mouse samples. Specifically, we determined that the 250kb step size allowed us to identify the maximum boundaries of RAMs for mouse embryonic samples, similar to the findings observed for human data.

Subsequently, we proceeded with the pipeline using a 250kb bin size and processed the narrow-peak files specific to the mouse version mm10. This process yielded density peak files and boundary loci information for each chromosome, as depicted in Figure 3.5. On average, approximately 20% of the regions across all mouse genomes (chromosomes 1-19 and X) were identified as boundary regions. Each sample contained an average of 609 boundaries. For a more comprehensive overview of the specific number of boundaries, please refer to the appendix table

A1.

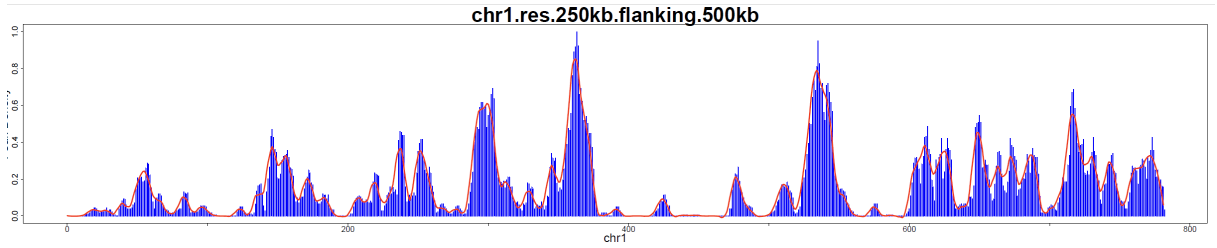



**Figure 3.5**: This is an example of the density peak file for chromosome 1 of the mouse embryonic facial prominence at e10.5 day. The x-axis represents the bin position, with each bin measuring 250kb in length. The y-axis indicates the proportion of the density peak counts after applying the sliding window algorithm. In the visualization, the RAM boundaries are indicated by the valleys or minima on the smoothing curves, represented by the red line. Similarly, the peaks or maxima on the smoothing curve of the red line indicate the RAM peaks.

We also observed that the number of RAM boundaries in mouse developmental data was slightly smaller compared to humans when using the same bin size in 'findRAM'. This discrepancy could potentially be attributed to the shorter genome length in mice, which is approximately 14% shorter than that of humans.

### 3.2.1 Variable RAM Identification

Following the generation of RAM boundaries, our next step involved conducting pairwise comparisons between these boundaries. In order to perform these comparisons, we designated each tissue as a control group once and compared it with the remaining tissues at each time point. Additionally, for each tissue, we treated the early time point as the control group and the subsequent time point as the sample group. During these comparisons, we recorded all regions where boundaries in the control group transitioned to RAMs in the sample group. However, we only considered regions that exhibited a minimum shift of 2 bin sizes between the control and sample groups. We specifically investigated pathways that displayed a one-bin size shift and found that they often led to problematic results in differential expression genes (DEGs), such as strong head development in comparisons involving the liver and limbs. This occurrence

could potentially be attributed to noise in the ChIP-sequence data, as no laboratory work or data processing is entirely flawless. On average, approximately 5% of the genome was identified as variable RAMs in tissue comparisons, while 4% of the genome exhibited variable RAMs in time comparisons, as illustrated in Figure 3.8.

## 3.3 Expressing Genes, DEGs, and Regulatees Identification in Variable RAM for Pairwise Comparisons

### 3.3.1 Expressing Genes in Variable RAMs

Before choosing the cutoff for expressed genes, we checked the quality of the processed data from all gene quantification files provided by ENCODE. We processed 10 samples, starting from raw data, and generated gene counts using the STAR [24] and RSEM [25] methods through a pipeline called RNA-seq nf-core [26]. We observed that the counts generated by the nf-core pipeline were similar to the counts provided by ENCODE. Next, we analyzed the distribution of gene counts for all expressed genes (non-zero counts) using the TPM (transcripts per million) normalization method. We created histograms of log(TPM) values for each sample. Across all samples, we identified significant noise when log(TPM) values were smaller than 0, indicating TPM values below 1. In Figure 3.6, we present an example plot of embryonic facial prominence replicate 1 at e13.5 day, which demonstrates the observed pattern. Similar patterns were observed in the remaining samples. Based on this analysis, we selected genes with log(TPM) values greater than 0 in over 60% of the samples as expressed genes. In total, we identified 16,438 genes as expressed genes.

To identify the relevant genes located within the variable RAMs, we performed an overlap analysis between the loci of expressed genes and the variable RAMs in each pairwise comparison. Our objective was to pinpoint the gene functions associated with the variable RAMs. During this

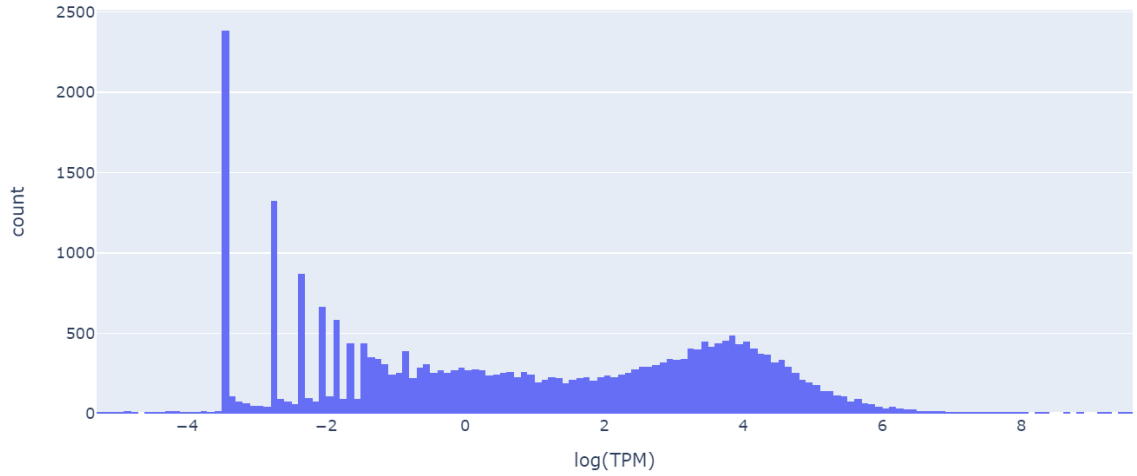



**Figure 3.6**: This graph is an example of a histogram displaying the log(TPM) gene counts. It represents the embryonic facial prominence replicate 1 at e13.5 day.

analysis, we also explored the differences between utilizing gene body annotation and annotated transcription start site (TSS) loci for gene enrichment results. The TSS corresponds to the core promoter region bound by RNA Polymerase II (RNA Pol II)[27], the primary enzyme responsible for transcription. Notably, the annotation of TSS regions had fewer annotations compared to gene body annotations. Specifically, we had annotations for 22k TSS regions and 45k gene bodies. Therefore, by utilizing TSS annotation, we lost approximately 17% of genome information concerning expressed genes. After conducting pathway analyses for the expressed genes located within the variable RAMs, we observed minimal disparities in the identification of significant pathways between using TSS and gene body annotations. To avoid losing valuable genomic information, we decided to prioritize gene body annotation as the criterion for assessing overlapping genes within the variable RAMs during pairwise comparisons, as well as for merging and splitting cRAM regions later on. Further details regarding these comparisons will be elaborated in Chapter 3.4.1 and Chapter 3.5.1.

### 3.3.2 DEGs in Variable RAMs

To identify differentially expressed genes (DEGs), we initially filtered out genes with low counts across all samples, resulting in 17k genes that were considered for further analysis. The filtering details are mentioned in chapter 2.1.7. We employed the DESEQ2 method to generate p-adjusted values and log2 fold-change values for each pairwise comparison. Since p-adjusted values are typically used when there are more than 5 replicates, we disregarded them in our analysis due to the limited number of replicates (only two). Instead, we focused on selecting DEGs based on a log2 fold-change value greater than 2, which corresponds to regions transitioning from boundaries to RAMs. In each tissue, for every time point except the last one, we treated the preceding time point as the control group, and the subsequent time point as the sample group. For instance, if e10.5 day was considered the control, then e11.5 would be the sample. At each time point, every tissue would become the control, and the rest of the tissues would become samples to determine the DEGs for tissue comparisons. We then examined the DEGs located within the variable RAMs for both time and tissue comparisons. More detailed results can be found in Chapter 3.3.2 and Chapter 4.3.2.

### 3.3.3 Regulatees in Variable RAMs

To study the regulatory network underlying 3D chromatin patterns, we employed a pipeline called Taiji[28] to identify key genes under regulation. Here, we planned to check whether these regulators and their regulatees are associated with variable RAM. Thus, we used Taiji to first identify the driver TFs for each sample and eventually chose the top regulatees of each driver TF. Taiji is a comprehensive system that utilizes various genomics information to construct transcriptional regulatory networks by predicting regulatory interactions between transcription factors (TFs) and genes. The PageRank score assigned to each TF within the network was used to assess its genome-wide influence, which is reflected in gene expression patterns.

After running Taiji, we initially obtained the PageRank scores for 887 different TFs in each sample. To ensure the reliability of our results, we examined the correlation of these scores across all samples. Additionally, we performed validation analyses before identifying the driving TFs and the most important regulatees. One of the validation methods involved employing unsupervised clustering, specifically hierarchical clustering, to distinguish different tissues and time stages. The results revealed that nervous-system-related tissues formed a distinct cluster, while other tissues exhibited similar clustering patterns (Figure 3.7). Furthermore, we applied Pearson correlation analysis to assess the correlation between samples. We observed a strong correlation across tissues at different stages, with tissues displaying stronger correlations at each time point compared to different time points. These validation procedures provided evidence supporting the validity and interpretability of our results for subsequent analysis.

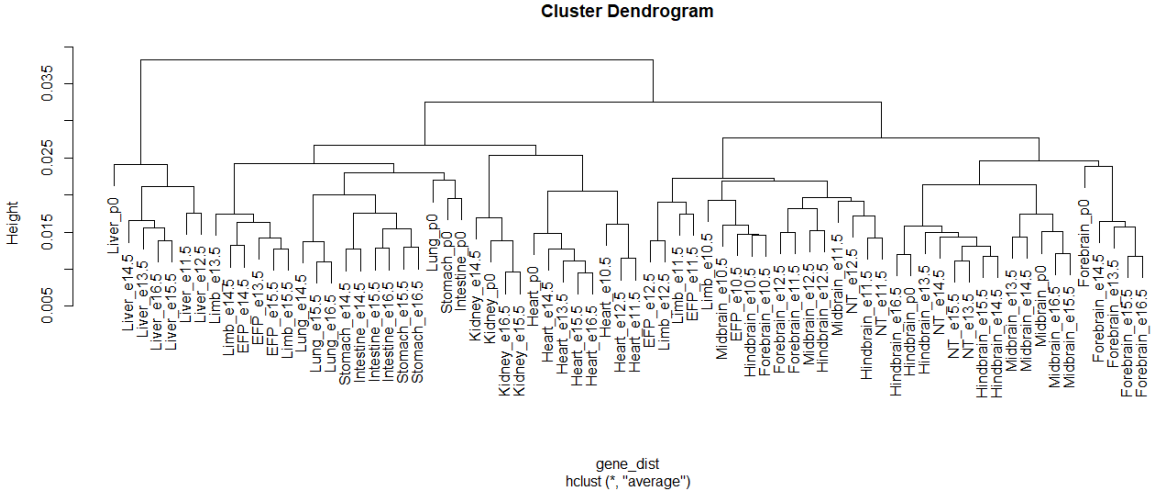



**Figure 3.7**: This graph shows Taiji hierarchical clustering with 72 results for PageRank scores of regulatees Validation.

After validating the Taiji results, we proceeded to identify the driver transcription factors (TFs) for each tissue or time point. We initially selected TFs with average ranking scores greater than 0.002, which corresponded to the top 12% of all TFs. Next, we retained TFs with coefficients of variation (CV) less than 0.3 for each tissue and 0.4 for each time. The coefficient of variation measures the relative dispersion of data points around the mean, with lower CV values indicating

less variability and a higher likelihood of being a driver TF across different samples.

As a result, we obtained an average of 45 driver TFs for each tissue and 30 driver TFs for each time point, as shown in Table A.3. Notably, brain-related tissues such as the forebrain (19), midbrain (20), and hindbrain (29) exhibited a lower number of driver TFs compared to tissues like the kidney (86) and stomach (80). This discrepancy may be attributed to the availability of more time point data for the brain and heart tissues. Additionally, during the early stages, there were more driver TFs compared to the later stages. At e10.5 days, there were 73 driver TFs, but this number significantly decreased in subsequent days, with only 19 driver TFs present at p0 day. This suggests less differentiation among all tissues in the early stages, as heart and brain-related tissues show significant similarities in driver TF identifications. However, this trend could also potentially be due to the limited availability of data during the early stages. Tissues such as the stomach, intestine, and lungs did not have any data available during the early stages.

Once the driver TFs were identified for each time and tissue point, we proceeded to identify the top 700 regulatees for each driver TF in each comparison, based on the network score provided in the Taiji results. The network score takes into account various elements such as gene expression level, motif binding, and peak intensity. Subsequently, we combined regulatees across all tissues for each time point and regulatees across all time points for each tissue. Finally, we selected the top 4500 genes based on the sum of transcripts per million (TPM) values across all available tissues or time points.

## 3.4 Pairwise Comparison of Tissue Comparison

**Summary** We conducted an analysis of the pathways in the Gene Ontology (GO) [29] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [30] enrichment of 606 varied RAM boundaries between different tissues at each time point using g:profiler. GO provides a comprehensive framework and standardized vocabulary for describing the functions of gene products across

all organisms. Our analysis focused on all three major categories: molecular function (MF), biological process (BP), and cellular component (CC). KEGG, on the other hand, serves as a valuable resource for understanding the functions and interactions of biological systems, leveraging molecular-level information derived from genome sequencing and other high-throughput experimental technologies.

In the subsequent sections, we compared and presented selected detailed findings derived from the GO and KEGG analyses. Our selection criteria were designed to ensure a high level of statistical significance, with all results in these categories being chosen based on a threshold of a p-adjusted value below 0.01. It is important to note that a lower p-adjusted value indicates a higher degree of statistical significance.

### 3.4.1 Pairwise Comparison of Expressed Genes in Different Tissues at Each Time Point

After analyzing the results for each sample, we began by comparing the number of expressed genes and the percentage of variable regions. We observed that the number of expressed genes within variable RAM regions, relative to all expressed genes, was approximately the same as the percentage of variable RAM regions across the entire genome length. This finding suggests that expressed genes are not specifically enriched or avoided in variable regions. Next, we examined the detailed pathways associated with each comparison. Our findings revealed numerous generic pathways in biological processes (BP), molecular functions (MF), and cellular components (CC). However, in the KEGG results, we did not observe any pathways that exhibited a level of significance comparable to the other three categories.

Here, we presented two comparison samples (Figure 3.9-3.10). The analysis revealed interesting similarities in the driver gene ontology (GO) terms between comparisons involving different tissues and time points. Despite the differences, common driver GO terms were identified. For example, in terms of molecular function (MF), the most important pathways were related to

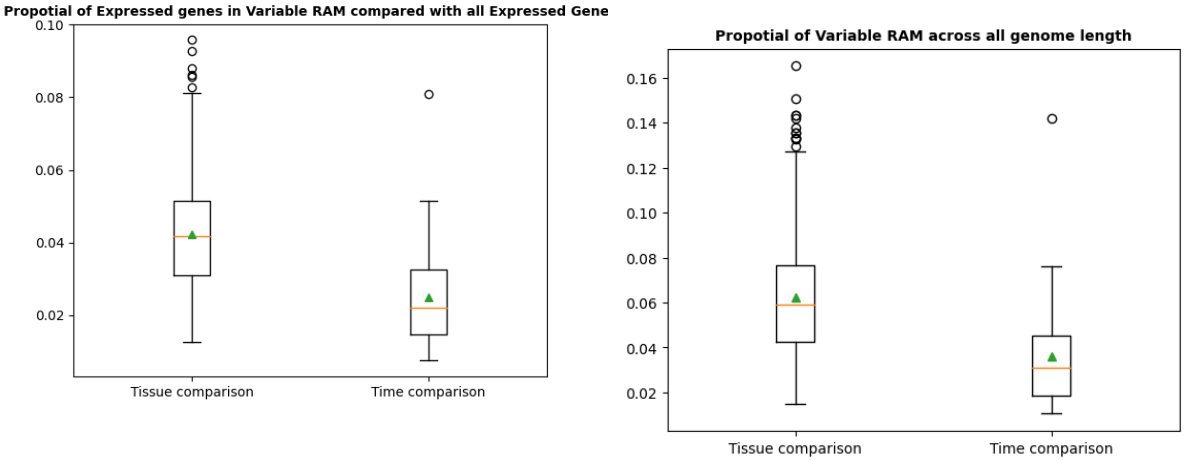
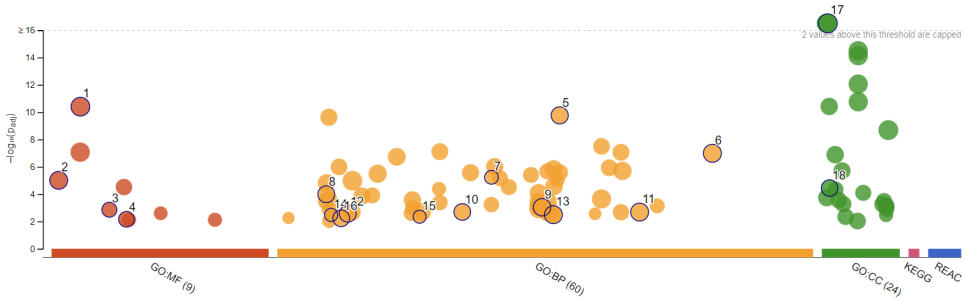
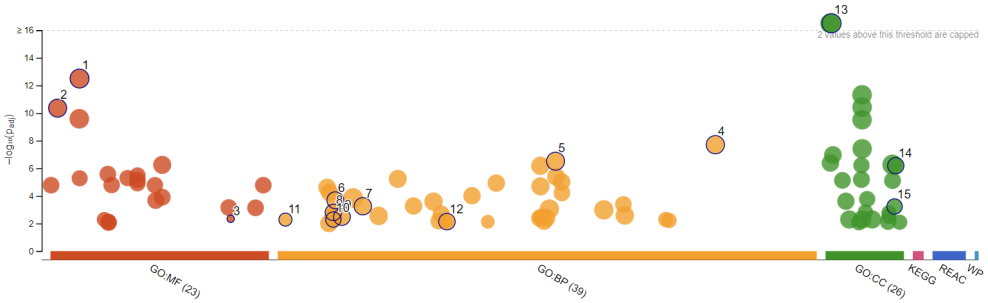



**Figure 3.8**: The left box plot represents the comparison of expressed genes located in variable RAMs out of all selected 15,438 expressed genes across all tissue (606 samples) and time (60 samples) comparisons. The right box plot illustrates the comparison of variable RAMs region, encompassing all tissue and time comparisons, out of the total genome lengths across Chromosome 1-19 and Chromosome X for the mouse.




| ID | Source | Term ID | Term Name | $p_{adj}$ (query_1) |
|---|---|---|---|---|
| 1 | GO:MF | GO:0005515 | protein binding | $4.003\times10^{-11}$ |
| 2 | GO:MF | GO:0003824 | catalytic activity | $9.845\times10^{-6}$ |
| 3 | GO:MF | GO:0016788 | hydrolase activity, acting on ester bonds | $1.392\times10^{-3}$ |
| 4 | GO:MF | GO:0030234 | enzyme regulator activity | $6.950\times10^{-3}$ |
| 5 | GO:BP | GO:0051641 | cellular localization | $1.729\times10^{-10}$ |
| 6 | GO:BP | GO:1901564 | organonitrogen compound metabolic process | $1.030\times10^{-7}$ |
| 7 | GO:BP | GO:0043087 | regulation of GTPase activity | $5.905\times10^{-6}$ |
| 8 | GO:BP | GO:0006793 | phosphorus metabolic process | $1.025\times10^{-4}$ |
| 9 | GO:BP | GO:0048731 | system development | $9.221\times10^{-4}$ |
| 10 | GO:BP | GO:0035556 | intracellular signal transduction | $2.032\times10^{-3}$ |
| 11 | GO:BP | GO:0080090 | regulation of primary metabolic process | $2.080\times10^{-3}$ |
| 12 | GO:BP | GO:0009653 | anatomical structure morphogenesis | $2.743\times10^{-3}$ |
| 13 | GO:BP | GO:0051171 | regulation of nitrogen compound metabolic proce... | $3.235\times10^{-3}$ |
| 14 | GO:BP | GO:0007160 | cell-matrix adhesion | $3.335\times10^{-3}$ |
| 15 | GO:BP | GO:0031032 | actomyosin structure organization | $4.532\times10^{-3}$ |
| 16 | GO:BP | GO:0009056 | catabolic process | $5.915\times10^{-3}$ |
| 17 | GO:CC | GO:0005737 | cytoplasm | $8.304\times10^{-30}$ |
| 18 | GO:CC | GO:0005856 | cytoskeleton | $3.604\times10^{-5}$ |

**Figure 3.9**: This graph displays the driver Gene Ontology (GO) pathways associated with variable RAM regions of the embryonic facial prominence, specifically compared with the forebrain at e10.5 days. The analysis includes three categories: molecular function (MF), biological process (BP), and cellular component (CC), with a p-adjusted value smaller than 0.01.




| ID | Source | Term ID | Term Name | $p_{adj}$ (query_1) |
|---|---|---|---|---|
| 1 | GO:MF | GO:0005515 | protein binding | $3.197\times10^{-13}$ |
| 2 | GO:MF | GO:0003824 | catalytic activity | $4.425\times10^{-11}$ |
| 3 | GO:MF | GO:0099038 | ceramide floppase activity | $4.589\times10^{-3}$ |
| 4 | GO:BP | GO:1901564 | organonitrogen compound metabolic process | $1.984\times10^{-8}$ |
| 5 | GO:BP | GO:0051179 | localization | $3.095\times10^{-7}$ |
| 6 | GO:BP | GO:0007399 | nervous system development | $2.090\times10^{-4}$ |
| 7 | GO:BP | GO:0010646 | regulation of cell communication | $5.650\times10^{-4}$ |
| 8 | GO:BP | GO:0007267 | cell-cell signaling | $1.580\times10^{-3}$ |
| 9 | GO:BP | GO:0009056 | catabolic process | $3.370\times10^{-3}$ |
| 10 | GO:BP | GO:0007268 | chemical synaptic transmission | $5.185\times10^{-3}$ |
| 11 | GO:BP | GO:0001678 | intracellular glucose homeostasis | $5.227\times10^{-3}$ |
| 12 | GO:BP | GO:0033554 | cellular response to stress | $7.365\times10^{-3}$ |
| 13 | GO:CC | GO:0005737 | cytoplasm | $7.954\times10^{-22}$ |
| 14 | GO:CC | GO:1902494 | catalytic complex | $6.536\times10^{-7}$ |
| 15 | GO:CC | GO:0140535 | intracellular protein-containing complex | $5.881\times10^{-4}$ |

**Figure 3.10**: This graph displays the driver Gene Ontology (GO) pathways associated with variable RAM regions of the heart, specifically compared with the liver at e16.5 days. The analysis includes three categories: molecular function (MF), biological process (BP), and cellular component (CC) with a p-adjusted value smaller than 0.01.

protein binding and catalytic binding. Additionally, the driver GO term for cellular component (CC) indicated a significant pathway associated with the cytoplasm. Furthermore, we conducted a thorough check of all pathways listed in both the GO and the KEGG database. We found that not only did the driver pathways share high similarity, but almost all pathways were generic and similar, such as enzyme binding, ion binding, and hydrolytic activity in MF. Moreover, the majority of pathways overlapped and were very generic in the context of embryonic development. No KEGG results were obtained.

In order to investigate the potential implications of utilizing gene body loci instead of enhancer and promoter regions for gene enrichment analysis, we conducted a comparative analysis between the two approaches. One approach utilized gene bodies as annotated loci, while the other approach focused on promoter loci. It is important to note that H3K27ac histone modification is associated with both enhancer and promoter signals. However, we deliberately

chose not to employ enhancer loci as indicators for identifying expressed genes. This decision was motivated by the complex nature of enhancer regulation on genes and the limited availability of comprehensive annotation for enhancers. Enhancers have the ability to regulate multiple genes, and conversely, a single gene can be influenced by multiple enhancers through enhancer-promoter interactions. Additionally, different enhancers may exhibit diverse modes of regulation, and the competition between enhancers in regulating a particular gene further complicates accurate predictions. To address these challenges, we opted to use transcription start site (TSS) loci as indicators for promoter regions. The details of TSS identification are described in Chapter 2.1.5.

To provide a summary of the results obtained from these comparisons, we have generated bar plots. These plots depict the differences in pathway enrichment between the two loci annotation approaches. They serve as a comprehensive overview of the findings and provide a reference for further analysis and interpretation.

Based on the analysis, we discovered a total of 16,438 expressed genes across all samples. However, when considering only the genes with annotated transcription start site (TSS) regions, the count decreased to 13,452 genes. This reduction suggests that approximately 18% of genome information was lost when relying solely on TSS loci annotation. Additionally, we investigated the disparity in gene identification between TSS loci and gene loci specifically within the variable RAMs. On average, the gene list derived from TSS loci within variable RAMs was approximately 20% smaller compared to the gene list obtained from gene loci. This finding indicates a proportional decrease in the number of genes when utilizing TSS loci annotation, which is consistent with the loss of annotation information.

After conducting a thorough analysis of all comparisons using TSS and gene loci, we have compiled a comprehensive summary of consistently identified pathways across all samples. These pathways have been categorized into three distinct categories: molecular function (MF), biological process (BP), and cellular component (CC). To present the results clearly and concisely, we have created tables displaying the identified pathways. In each comparison, we focused on

the top 30 significant pathways within each category, highlighting the most noteworthy findings while minimizing potential noise. This selection process was necessary due to the presence of pathways in the database with limited officially recorded genes, which could yield significant findings based on the provided gene lists. Moreover, considering the database includes predictions for the functions of numerous genes, it is crucial to ensure that the identified pathways genuinely reflect the role of variable RAM regions.

Upon comparing the pathways identified using TSS loci and gene loci, it was observed that more than 90% of the pathways were consistently identified in both categories across all tissue comparisons. The tables presented in the molecular function (MF) (Figure 3.11), biological process (BP) (Figure 3.12), and cellular component (CC) (Figure B.2) categories clearly highlight the striking similarity in the enrichment analysis between TSS loci and gene body loci in variable RAMs. Minor variations were observed in the number of significant pathways across all samples, but the most shared pathways generally appeared in a similar order, with only slight rearrangements of the most significant pathways. These findings provide strong evidence of the high degree of consistency and agreement in the pathway analysis results obtained from TSS and gene loci annotations. Consequently, these results bolster confidence in the reliability and validity of the enrichment analysis conducted solely using gene loci.

It is noteworthy that, in the majority of pathways identified using TSS loci in variable RAM regions, the number of samples was only slightly lower compared to gene loci. This observation can be attributed to the close proximity of TSS regions to the genome. Consequently, when a gene is located within a variable RAM region and its size is much smaller than the encompassing variable region, it is highly likely that the TSS region of this expressed gene is also situated within the same variable region. In order to minimize the loss of genome information, the decision was made to utilize gene body loci instead of TSS loci. Fortunately, the pathways identified using gene body loci exhibited a striking similarity to those identified using TSS loci. This implies that the inclusion of gene body loci would not introduce a significant number of

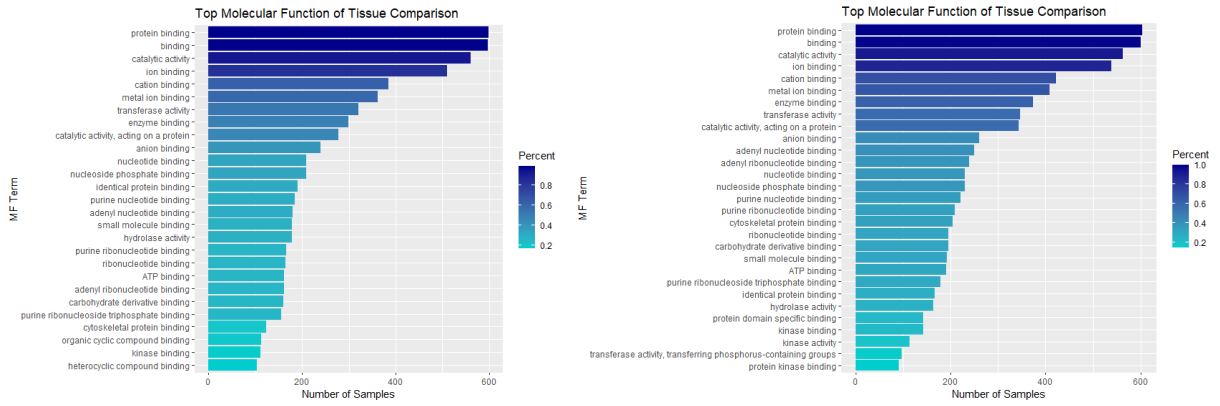



**Figure 3.11**: Here are all the pathways selected based on a p-adjusted value ¡ 0.01 and appearing in 15% of all tissue comparisons. The left panel represents the pathways identified using transcription start site (TSS) loci located in variable RAMs, while the right panel represents the pathways identified using gene body loci located in variable RAMs.

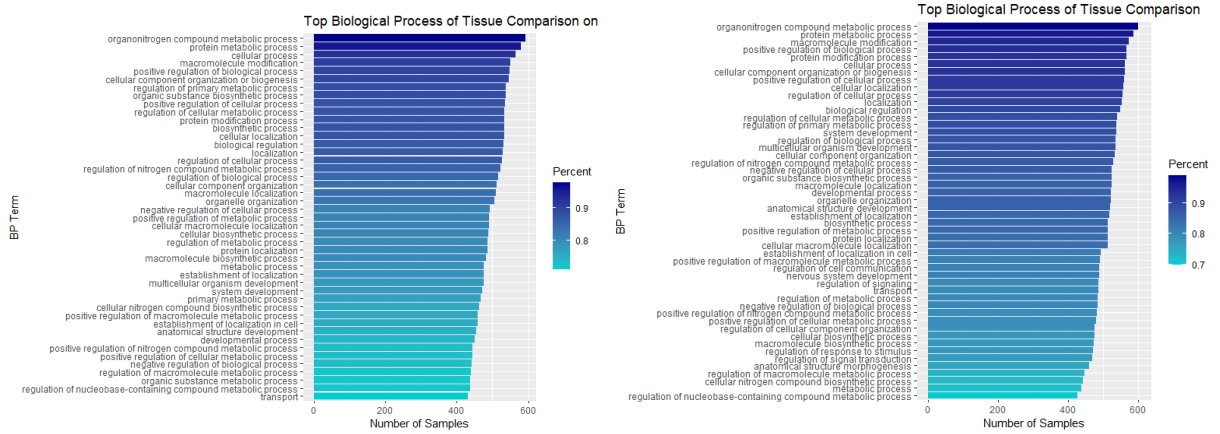



**Figure 3.12**: Here are all the pathways that were selected based on a p-adjusted value ¡ 0.01 and appeared in 65% of all tissue comparisons. The left panel represents the pathways identified using transcription start site (TSS) loci located in variable RAMs, while the right panel represents the pathways identified using gene body loci located in variable RAMs.

additional genes that could potentially disrupt the final results.

From the analysis of the MF (Molecular Function) results, it was observed that pathways related to protein binding, binding, enzyme or catalytic binding, and ion binding consistently ranked among the top significant pathways across all comparisons. These pathways involved a substantial number of genes and were consistently present in the majority of comparisons. Notably,

during the early stages of development, when samples included embryonic facial prominence, forebrain, midbrain, hindbrain, neural tube, limb, and heart, pathways related to cation binding and metal ion binding were particularly significant. In the comparison involving lung samples, pathways associated with metal and cation ion binding were also observed. However, it is important to note that in some comparisons where cation and metal binding pathways were present, they exhibited lower significance compared to the protein and ion binding pathways.

Interestingly, pathways associated with nervous system development were not only observed in variable RAMs of brain-related tissues but were also present in more than 80% of all tissue comparisons, albeit with varying levels of significance. In the case of the neural tube, it consistently ranked among the top 5 significant pathways compared to other tissues at any given time point. However, it is interesting to observe that the significance of nervous system development pathways in brain-related tissues was comparable to that of the heart, facial prominence, and liver. Therefore, the presence of nervous system pathways in the pathway analysis cannot solely be attributed to a specific tissue when utilizing expressed genes in variable RAMs.

### 3.4.2 Pairwise Comparison of Differentially Expressed Genes in Different Tissues at Each Time Point

We discovered that the number of differentially expressed genes (DEGs) identified within each sample was limited, posing challenges in identifying significant pathways across most categories. To address this limitation and focus on tissue-specific comparisons, we decided to consolidate DEGs from comparisons involving the same control and sample groups across all time points. Subsequently, we conducted enrichment analysis using g:profiler, following the same methodology as before. Overall, our analysis included a total of 132 comparisons. Consistent with our observations for expressed genes, we found that DEGs were not preferentially enriched or avoided in variable RAM regions. The number of DEGs within variable RAM regions displayed a nearly proportional relationship to the total number of DEGs, with a ratio of approximately 1:1

(Figure 3.13).

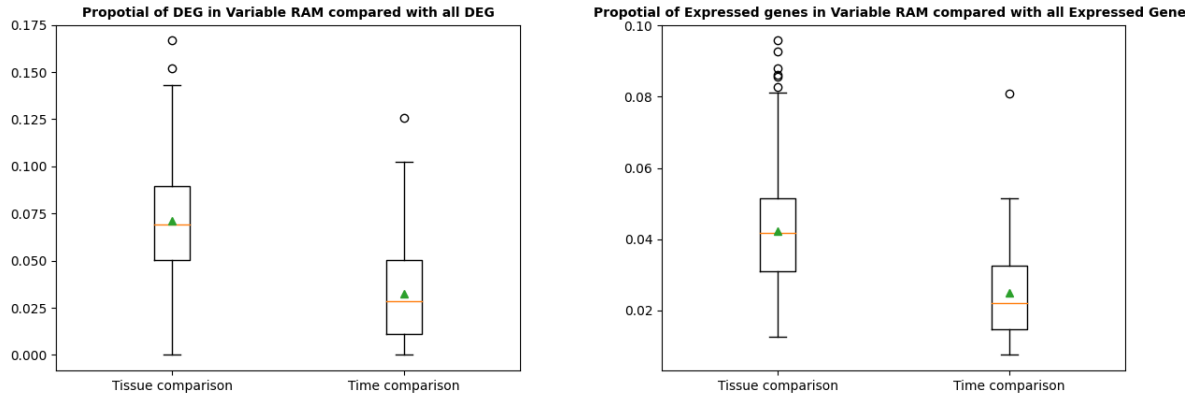



**Figure 3.13**: The left box plot represents the comparison of differentially expressed genes (DEGs) located in variable RAMs out of all DEGs per sample across all tissue (606 samples) and time (60 samples) comparisons. The right box plot illustrates the comparison of variable RAMs region, encompassing all tissue and time comparisons, out of the total genome lengths across Chromosome 1-19 and Chromosome X for the mouse.

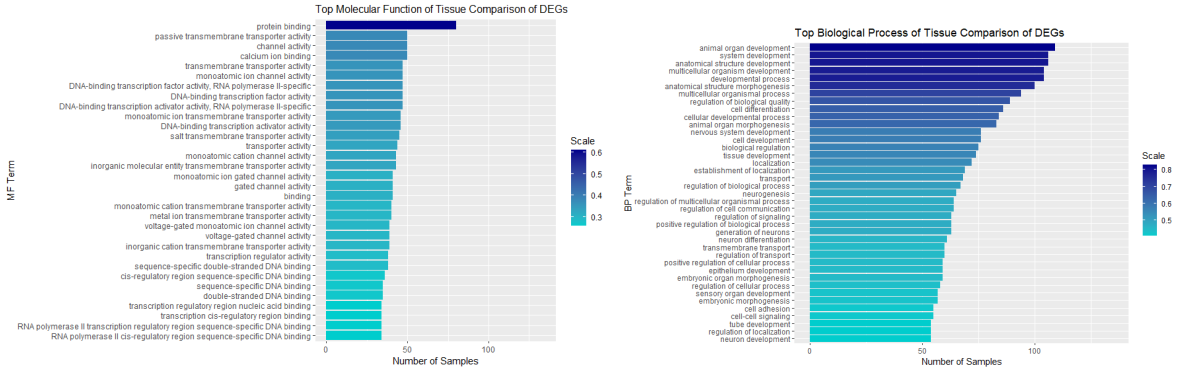



**Figure 3.14**: The left bar plot illustrates the selected MF pathways, while the right bar plot displays the selected BP pathways. These pathways were chosen based on a p-adjusted value smaller than 0.01 and were observed in over 15% for MF and over 30% for BP across all comparisons.

However, in contrast to the high consensus observed among expressed genes across all samples in each Gene Ontology (GO) category, the results obtained for DEGs exhibited significant variations across all samples. Notably, the analysis of DEGs revealed a larger number of KEGG

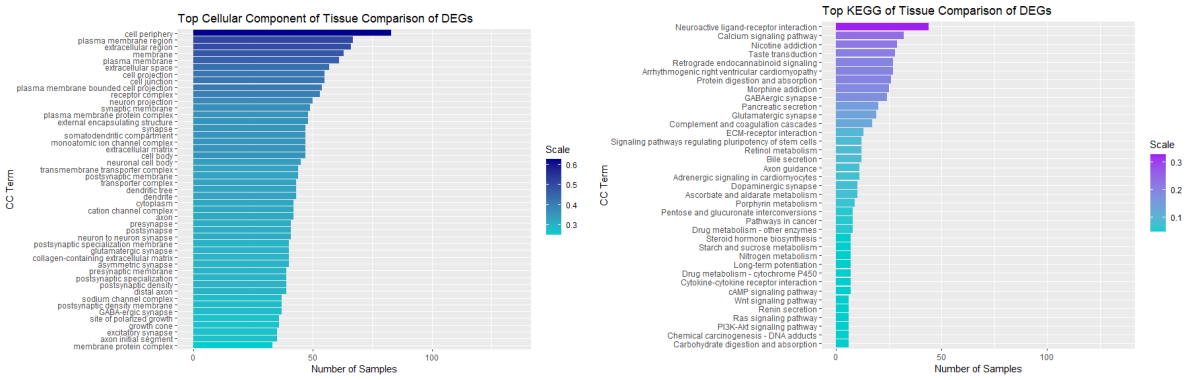



**Figure 3.15**: The bar plot on the left showcases the selected CC pathways, while the one on the right presents the selected KEGG pathways. These pathways were chosen based on a p-adjusted value smaller than 0.01 and were observed in over 25% for CC and over 30% for KEGG across all comparisons.

terms compared to the analysis of expressed genes. To provide a comprehensive overview of the findings, we have summarized the top terms in the Molecular Function (MF), Biological Process (BP), and Cellular Component (CC) categories in the following bar plots (Figures 3.14-3.15). These plots serve to highlight the notable differences and variations observed in the enrichment analysis results of DEGs.

The top significant pathway in the MF category for DEGs is protein binding, which aligns with previous findings. However, it is observed in only 65% of the comparisons, indicating a higher variability in DEGs located in variable RAMs across samples. The increased number of total MF pathways shared over 25% across all comparisons suggests diverse functional roles for DEGs in variable RAMs. Both findings indicate an increased variability of DEGs located in variable RAMs, which is consistent with the normal features of DEGs across different tissues.

One interesting finding is the high significance of calcium binding in multiple comparisons. This finding aligns with previous research that suggests calcium plays diverse roles in cellular development across various tissues. Calcium acts as an essential intracellular signaling mediator and is involved in processes such as neurodegeneration [31]. Furthermore, studies have demonstrated the dependence of embryonic morphological development and DNA synthesis on

the concentration of $Ca^{2+}$ in the growth medium [32].

Another reason for the focus on calcium ion binding proteins is their potential involvement in the formation of RAM boundaries, as mentioned in a previous RAM paper. Studies have suggested that cations such as calcium play a critical role in maintaining the structural integrity of chromosomes, particularly during the condensation of mitotic chromosomes following nuclear envelope breakdown (NEB) and the compaction of chromatin fibers [19]. Cation ion binding proteins might have the ability to transport cations such as $Ca^{2+}$ and $Mg^{2+}$ along with partner proteins into the nucleus, where they can bind to specific positions on DNA. This binding can ultimately lead to chromatin condensation and the formation of RAM boundaries. Therefore, the significance of calcium ion binding proteins in our analysis may indicate their involvement in these processes [12].

After conducting a thorough analysis, we found that certain comparisons, such as those involving the heart, limb, and embryonic facial prominence as samples, exhibited a particularly strong association with calcium ion binding. Additionally, when the stomach was used as a control group, brain-related tissues also showed significant calcium ion binding. This finding aligns with previous research indicating the diverse importance of calcium-binding proteins, including parvalbumin, calbindin, and calretinin, in the central nervous system. Experimental studies involving the knockdown or overexpression of these genes in vivo and in vitro have demonstrated their role in determining neuronal survival in different locations [33]. To further investigate whether calcium ion binding is specifically related to DNA condensation in tissues, we examined the cellular component (CC) categories associated with genes correlated with calcium ion binding.

After analyzing the cellular component (CC) pathways, we observed that genes related to calcium ion binding were not primarily located in the nucleus, where chromosomes are situated. However, it is worth noting that some genes exhibited dual localization in both the cytoplasm and nucleus. Unfortunately, our analysis did not reveal a substantial number of genes associated with

histone modification functions and chromatin modification among the DEG candidates. However, our investigation yielded significant findings related to important genes as expressed genes and regulatees in pairwise comparisons within variable RAMs and cRAM, as discussed in Chapter 3.6.

Even when excluding the most common embryonic cell development mouse genes, our analysis of pathways still revealed a strong signal associated with organ growth, system development, and tissue development in biological processes. This finding suggests that these processes play a significant role in the formation and development of tissues, extending beyond the context of embryonic cell development.

### 3.4.3 Pairwise Comparison on Different Tissues for Each Time with Regulatees

In the analysis of regulatees in variable RAMs, we observed that they do not have as many genes as expressed genes in each tissue comparison per time. To address this limitation, a decision was made to merge all time points for each tissue comparison and conduct enrichment analysis. This approach resulted in a total of 132 different comparisons, providing a more comprehensive understanding of the enriched pathways associated with regulatees in variable RAMs.

The analysis presented in Figure 3.16 demonstrated that most pathways exhibit similarities across all three categories. However, a notable difference was observed in the molecular function (MF) category, where mRNA binding and transcript coregulator activity were prominently present in the regulatees analysis, but seldom observed in the expressed genes. This observation suggests that many highly expressed regulatees are involved in regulatory functions that affect gene expression level.

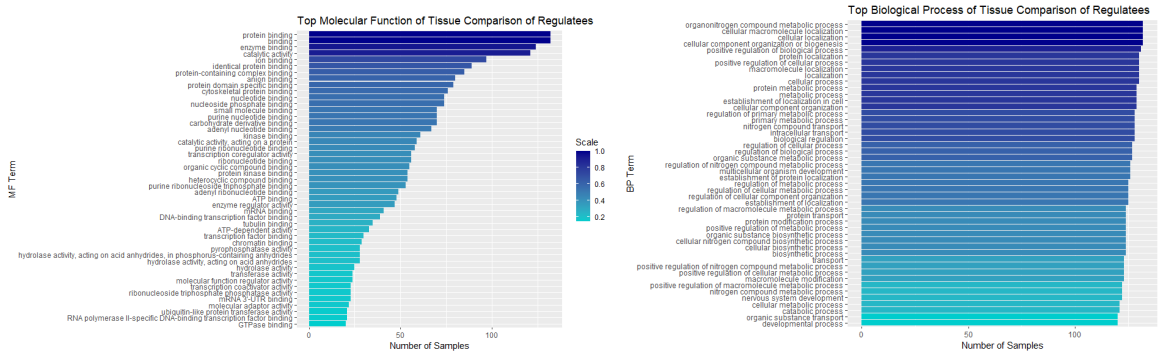



**Figure 3.16**: The bar plot on the left showcases the selected MF pathways, while the one on the right presents the selected BP pathways. These pathways were chosen based on a p-adjusted value smaller than 0.01 and were observed in over 15% for MF and over 90% for BP across all comparisons.

## 3.5 Pairwise Comparison of Time Comparison

### 3.5.1 Pairwise Comparison of Subsequent Time Points for Each Tissue with Expressed Genes

Similar to the tissue comparison, we also conducted pathway analysis between TSS and expressed genes located in variable RAMs for subsequent time points. Consistent with our previous findings,the differences between these two sets of data are very small (Figure 3.17). Therefore, we proceeded with our analysis using expressed genes using their loci.

The consistent patterns observed in all GO terms between time and tissue comparisons indicate a high degree of similarity across different comparisons. Despite the lower number of genes in time comparisons compared to tissue comparisons, the top terms in the molecular function (MF), biological process (BP), and cellular component (CC) categories exhibit remarkable similarity. Processes such as protein binding, catalytic binding, ion binding, metal binding, and cation binding are frequently observed in both types of comparisons. These processes are fundamental to cell development and survival and predominantly occur in the cytoplasm. These findings suggest that these essential cellular processes and molecular interactions play a crucial

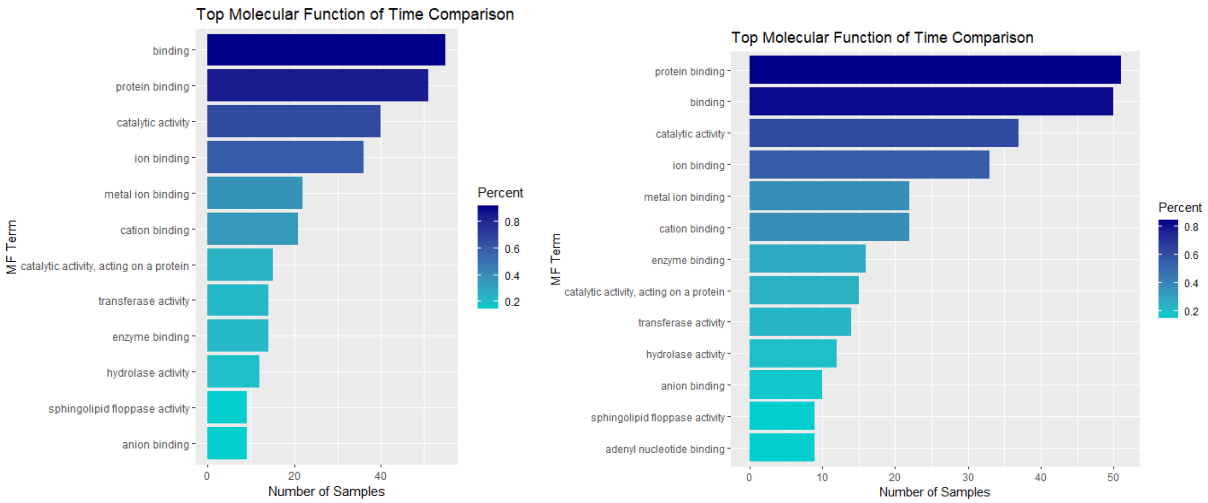



**Figure 3.17**: The bar plots illustrate the selected molecular function (MF) pathways that were present in over 15% of the genes located in variable RAM regions. The left bar plot corresponds to the results obtained using TSS loci, while the right bar plot represents the results obtained using gene body loci.

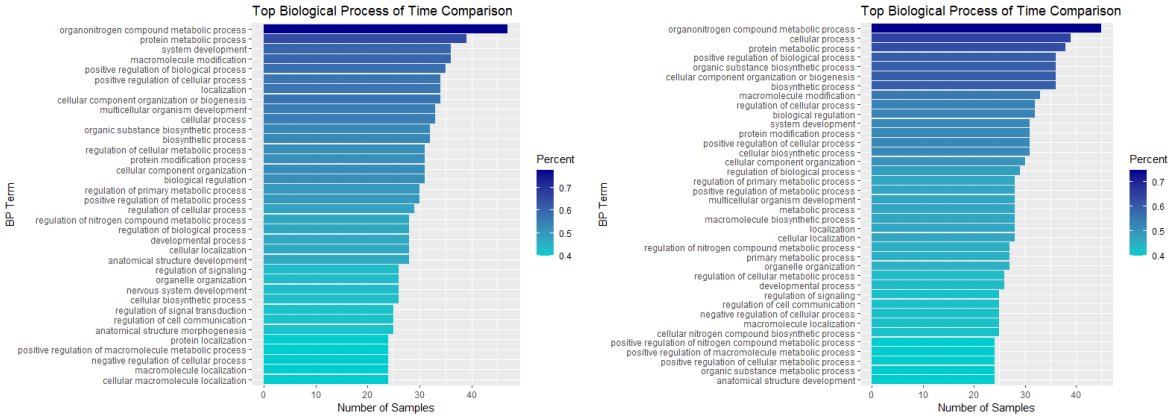



**Figure 3.18**: The bar plots illustrate the selected biological process (BP) pathways that were present in over 45% of the genes located in variable RAM regions. The left bar plot corresponds to the results obtained using TSS loci, while the right bar plot represents the results obtained using gene body loci.

role across various conditions and stages of development in variable RAM regions.

### 3.5.2 Pairwise Comparison of Subsequent Time Points for Each Tissue with Differentially Expressed Genes

Similar to the approach used for tissue comparisons, we consolidated all DEGs located in variable RAMs for all time comparisons within each tissue. This resulted in a total of 12 samples for pathway analysis in time comparisons across 12 tissues.

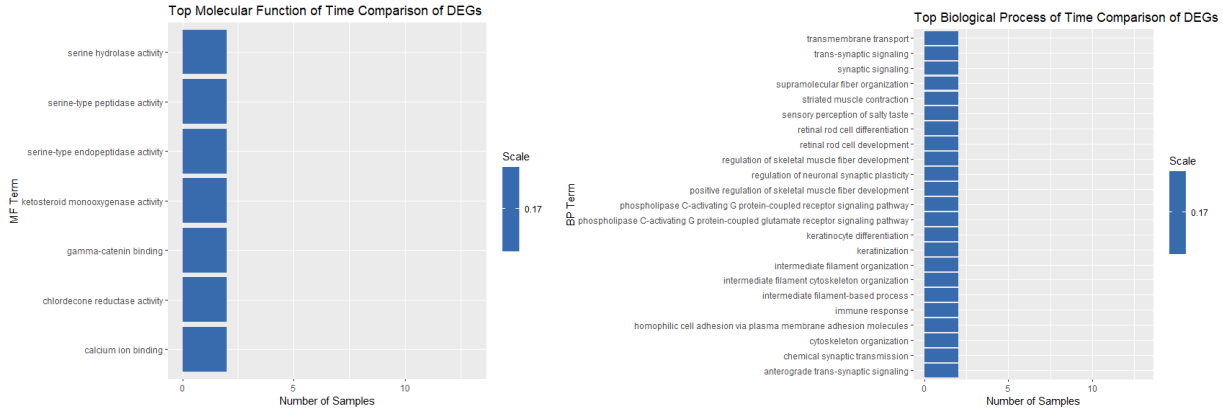



**Figure 3.19**: The bar plot on the left showcases the selected molecular function (MF) pathways, while the one on the right presents the selected biological process (BP) pathways. These pathways were chosen based on a p-adjusted value smaller than 0.02 and were observed in over 10% of the comparisons for MF and over 10% for BP across all comparisons.

The comparison of MF and BP pathways shared between tissue comparisons for DEGs in time comparisons for expressed genes reveals a significant difference (Figure 3.19). The overlap of pathways in MF and BP is remarkably low, with a maximum of only 2 shared pathways. This finding is surprising, considering the analysis was conducted on three distinct brain regions and the neural tube, which showed considerable similarities in previous analyses. The observed variability in MF and BP suggests distinct changes in RAMs within the brain regions over time, emphasizing the dynamic nature of brain development and the unique regulatory processes involved.

The absence of pathways related to protein binding, enzyme binding, and ion binding in the Molecular Function (MF) category, as well as the significant variability in the top Cellular Component (CC) terms, indicates distinct differences between time point comparisons and

tissue comparisons for DEGs located in variable RAMs. Unlike the results obtained from tissue comparisons and expressed genes in time comparisons, generic pathways such as cytoplasm, organelle, and cytosol are not prominently featured in the time point comparisons. This disparity can be attributed to the limited number of DEGs available for the time point comparisons, which is a consequence of the high degree of similarity among tissues even at different developmental stages. In the case of certain tissues, such as the heart, there is a relatively small proportion of DEGs located in variable RAMs, potentially because their early development begins prior to the available data [34]. Similarly, tissues like the neural tube also undergo early development stages that precede the available online data [35], resulting in limited information on differential gene expression at those specific time points.

Moreover, certain tissues, such as the stomach and intestine, have data available for only four distinct time points, resulting in a smaller number of DEGs and consequently fewer pathways available for enrichment analysis. This is in contrast to tissues like the forebrain, which have data available for eight different time points, leading to a larger number of DEGs and more comprehensive pathway results. Therefore, the variability and limited pathway outcomes observed in the time point comparisons of DEGs located in variable RAMs can be attributed to the restricted number of DEGs and the developmental disparities among tissues at specific time points.

### 3.5.3 Pairwise Comparison of Subsequent Time Points for Each Tissue with Regulatees

We also conducted time comparisons for regulatees in each tissue. Following a similar approach as with DEGs in variable RAMs, we consolidated the regulatees from the same tissue, resulting in 12 distinct comparisons. Our analysis of the time comparisons for regulatees revealed a high frequency of molecular function (MF) terms related to protein binding and binding.

We observed a greater number of significant pathways in the molecular function (MF),

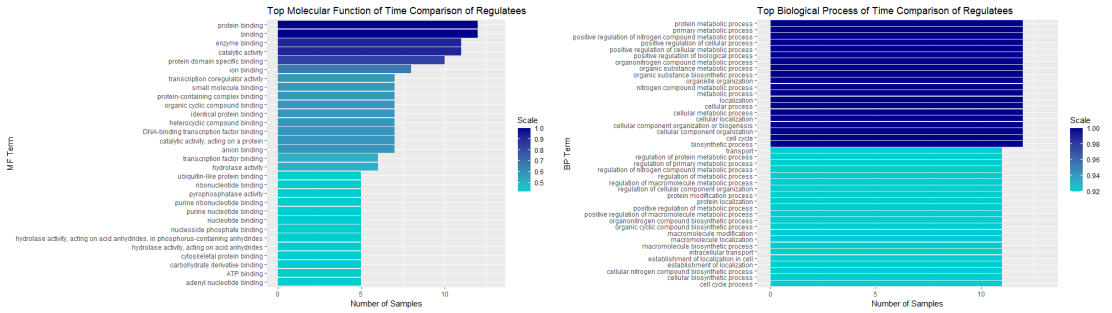



**Figure 3.20**: The bar plots depict the top pathways in the MF and BP categories based on the analysis of regulatees in time comparisons. The left plot represents the selected MF pathways, while the right plot displays the selected BP pathways. These pathways were chosen based on a p-adjusted value smaller than 0.02 and were observed in over 10% of the regulatees across all time point comparisons.

biological process (BP), and cellular component (CC) categories when analyzing regulatees in time point comparisons. Many of these pathways were shared across almost all tissues, indicating common regulatory processes at different developmental stages. In the MF category, pathways such as protein binding, ion binding, enzyme binding, and catalytic activity were consistently identified in most samples, which aligns with the findings from the analysis of expressed genes in time point comparisons. However, we also identified a notable pathway, 'transcription coregulator activity', which was present in both tissue comparisons and the regulatees analysis. This pathway may be specifically related to the unique feature of regulatees, as many of them are under the regulation of driver transcription factors (TFs).

## 3.6 cRAM Identification and Comparison Results

**Summary** By summarizing the results, it becomes evident that relying solely on pathway analysis in pairwise comparisons is insufficient for comprehensively understanding the functions of RAMs and boundaries in mouse development. However, we observed a high degree of similarity in RAM patterns and functions across all samples, particularly for expressed genes and regulatees. To further investigate this, we examined the consistency of RAM boundaries across

all time points for each tissue and all tissues for each time, defining them as consensus RAM boundaries. Surprisingly, we found that 83% of RAM boundaries were consistent and shared across all 72 samples. This highlights the robustness and stability of RAM boundaries throughout the developmental stages. Additionally, we identified a new pattern in cRAMs called the Merging and Splitting (MS) region.

### 3.6.1 Merging and Splitting Region Identification

We first identified cRAMs for each tissue and time point by considering those that were shared across over 50% of the samples. On average, we observed that 80% of boundaries belonged to cRAMs. Upon visual inspection using IGVs, we found that cRAMs exhibited similarities across different time points and tissues. We further categorized the cRAMs into three types: Identical Regions, Specific Regions, and Merging and Splitting (MS) Regions. Identical Regions demonstrated high consistency across all samples and remained relatively closed and less variable compared to other regions. Specific Regions, which appeared in less than half of the cRAMs, were not the main focus of our analysis. Our attention was primarily directed towards the MS Regions, which exhibited significant changes in their boundaries across time points or tissues.

To define MS regions, we required them to have over 15% shared RAM region and less than 85% RAM region coverage across all time points or tissues. Additionally, MS regions had to be equal to or greater than 3 bin sizes (750 kb) to focus on larger changing regions (Figure 3.21).

### 3.6.2 cRAM Analysis for Time and Tissue Comparisons

To further explore gene enrichment, we focused on genes that were consistently identified as expressed in the majority of samples. Similar to the findings in pairwise comparisons, we observed generic functional roles such as protein binding, catalytic activity, and generic biological processes in both time and tissue comparisons (Figure 3.22). However, we observed a significant

increase in the enrichment of cation and metal ion binding pathways in both tissue and time comparisons. This finding aligns with previous studies suggesting that cation binding proteins may play a role in the formation of RAM boundaries [12]. Prior research has indicated that multivalent cations contribute to reducing electrostatic repulsions between DNA chains, facilitating DNA condensation [36]. These cations can bind to specific DNA sequences and influence nucleosome positioning. The presence of cation binding pathways in the gene enrichment analysis suggests their potential involvement in transporting cations into the nucleus and binding to DNA, ultimately leading to chromatin condensation and affecting gene expression within the variable RAM regions.

Regarding the association with the nucleus and chromatin-related pathways, our analysis revealed that the cytoplasm was still the predominant cellular component (CC) term, while the nucleus showed less significance. However, it should be noted that many genes have functions in both the nucleus and cytoplasm. Therefore, we focused on genes involved in cation ion binding and located within the nucleus in both MS regions and variable regions in pairwise comparisons. We compared the overlap between the cation and metal ion binding gene database in the MS regions with genes associated with the nucleus, and we identified 39 genes in the time MS regions and 52 genes in the tissue MS regions. Notably, genes such as GATA3[37], HDAC2[38], and FOXP2[39] were among the identified genes, and they play important roles in histone modification and chromatin formation changes. For example, GATA3 has been shown to regulate both active and repressive histone modifications and is involved in chromatin reprogramming [37, 40].

We then identified the expressed genes that belong to cation binding and are located in the nucleus for pairwise comparisons. By combining all pairwise comparisons, we generated a list of 400 candidates for future downstream analysis. It is important to note that for the mouse data, the majority of genes in these lists lack experimental validation of their functions specifically within the nucleus. However, we once again found some important genes that appeared frequently, and some of them were also identified in the cRAM analysis, including GATA3, HDAC9[38], HDAC2, HNF4G[41] and FOXP2. These genes exhibit high expression levels in some or all

tissues and time points and are known to have a strong relationship with histone modification and chromatin structure.

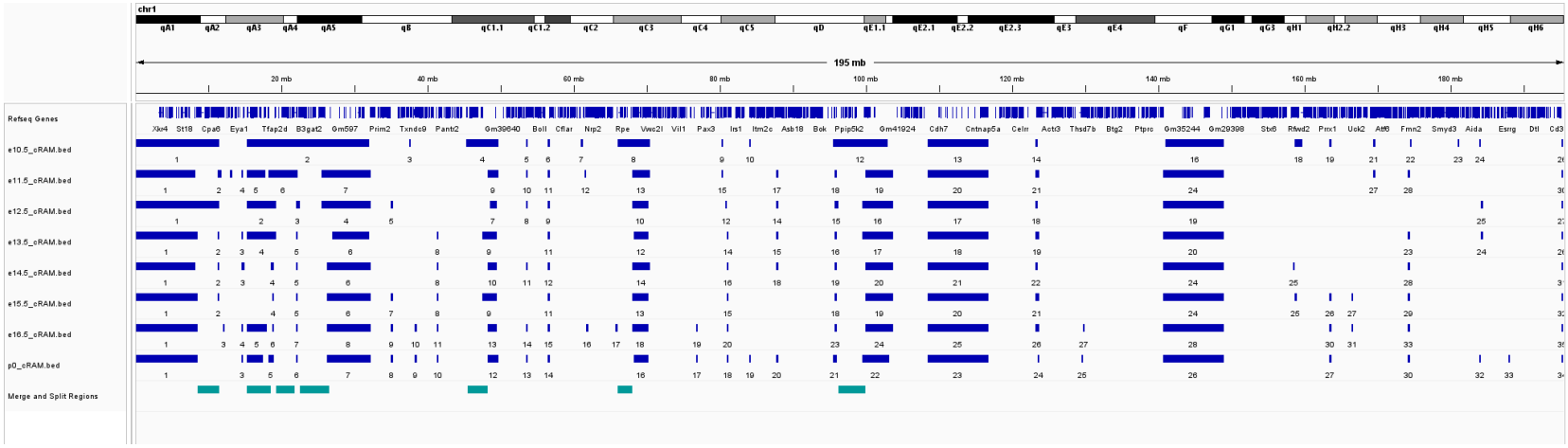



**Figure 3.21**: This IGV plot shows cRAMs for 8 different time points, with the last line indicating MS regions considered across all 8 time points.

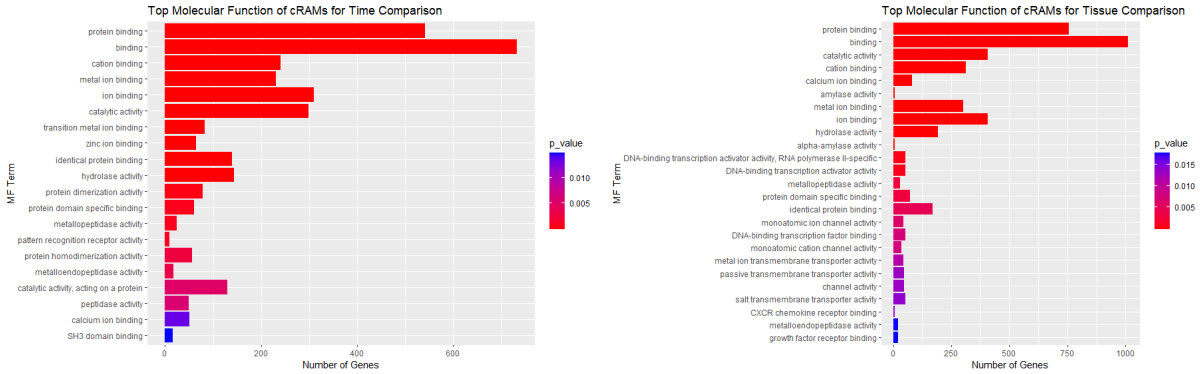



**Figure 3.22**: Bar plots show the top pathways of MF for expressed gene of cRAM time and tissue comparisons. The left plot represents the time comparison, while the right plot represents the tissue comparison. These pathways were selected based on a p-adjusted value smaller than 0.05.

# Chapter 4

# Discussions & Future Work

In this thesis, we focused on analyzing the peak density profiles of histone modification H3K27ac ChIP-seq in mouse development data. Utilizing the latest modal regulation-associated modules (RAMs), our aim was to understand how RAMs change as time points and tissues change and whether these changes in RAMs result in significant alterations in chromatin functions and gene expression. After thoroughly examining 72 mouse development samples, we identified several interesting features in RAMs during the developmental stages.

Firstly, we observed that the number of RAM boundaries in mouse development data was slightly smaller compared to those in humans when using the same bin size in 'findRAM'. This discrepancy could potentially be attributed to the shorter genome length in mice, which is approximately 14% shorter than that of humans.

Secondly, while over 60% of RAMs in human normal and cancer cells were identified as consensus RAMs (cRAMs), we found a higher proportion of cRAMs (80%) in mouse development samples, along with a higher consensus standard. This difference may be attributed to the high similarity between samples derived from the same tissue or time point in mouse development.

However, despite these differences, we also observed some similarities between RAM findings in human and mouse samples. Notably, RAM sizes for both humans and mice are much

larger than topologically associated domain (TAD) sizes and exhibit consistency across the two organisms. Additionally, the number of RAM boundaries showed proportionality to the whole genome length, with an average of 606 boundaries per cell for mice and approximately 720 boundaries for humans.

As a preliminary exploration of RAMs in mouse development data, our study yielded intriguing observations and proposed potential directions for future analysis and experimental validation of chromatin structure. In our pairwise comparisons, we did not find any avoidance or enrichment of expressed genes, differentially expressed genes (DEGs), or regulatees in variable RAM regions. Further gene enrichment analysis revealed that most expressed genes and regulatees in variable RAM regions were associated with generic molecular binding, biological processes, and cellular components involved in cell development, growth, and survival across tissue and time comparisons. For DEGs located within variable RAMs, despite their greater variability across samples, we still observed generic functions without identifying any specific roles for these genes.

The presence of significant pathway similarities across all pairwise comparisons implies that RAMs exhibit greater similarity in developmental stages across different tissues and times compared to normal and cancer cells. To validate this notion, we further identified consensus RAMs for each time and tissue stage. As expected, over 50% of the boundaries of cRAMs were shared within each stage, and more than 80% of boundaries were confirmed as identical cRAMs. These findings suggest that the similarities of RAMs in embryonic mouse cells are higher compared to normal and cancer cells (60%) in humans.

In our cRAM analysis, we observed a new pattern called merging and splitting (MS) regions. We further conducted enrichment analysis on these regions and found similarities across all pathway categories compared to the results obtained from pairwise comparisons. However, we observed a higher significance of both cation and metal ion binding in MS regions during cRAM analysis.

In the previous RAM project, there were suggestions that multivalent cations might

contribute to chromatin condensation, potentially leading to the formation of RAM boundaries[36, 12]. The increased significance of cation and metal ion binding in our analysis could serve as another indicator of how RAMs and boundaries are formed. In light of this, we examined all genes involved in cation binding within both cRAM and pairwise comparisons that are located in the nucleus. Overall, we identified over 400 distinct genes that fit this criteria. Notably, some of these genes, such as GATA3 and HDAC9, are well-known transcription factors that have been extensively linked to histone modifications and chromatin organization.

From our research on the RAM project using mouse development data, we have recognized several potential biases that could affect the results and conclusions. Firstly, we acknowledge that the identification software 'findRAM' is not yet perfect. While it can detect the most prominent boundaries based on extremely low peak densities, it still requires higher accuracy to differentiate RAMs and boundaries in regions with ambiguous peak density patterns that are near the cut-off value. Moreover, for a more comprehensive understanding of the functions of changing RAMs and their relationship to gene expression, it may be beneficial to consider the expression level of each peak in the H3K27ac density peak files. Instead of solely focusing on the number of peaks in each bin region, incorporating the expression levels can provide additional insights into the dynamics of RAMs and their impact on gene regulation.

Secondly, it is important to note that the availability of data for different time points in various tissues is still limited. For instance, the development of the heart starts at approximately e7.5 day, and the neural tube forms around e8.5 day. However, in our study, we only had data starting from e10.5 day. This limitation could potentially result in missing important RAM patterns that emerge in the early stages, thereby hindering our understanding of the mechanisms underlying RAM formation.

While RAM identification is not perfect, and data availability at different stages for various tissues is still not ideal, the findings from our analysis of mouse development data provide valuable insights. In this thesis, although we have not fully elucidated how RAMs form and the

purpose they serve, the observations have inspired us to continue studying RAMs and further explore the relationship between histone modification and chromatin structures.

Future work should involve modifying the computational pipeline 'findRAM' to enhance its complexity and accuracy. Additionally, additional experimental data is necessary to validate our hypothesis regarding RAM formation, such as investigating the exact function and effects of genes like GATA3, HDAC9, and FOXP2 located in variable RAMs and MS regions of cRAMs. Furthermore, exploring the role of genes like SVEP1, which are predicted to have calcium ion binding activity and chromatin binding activity, would contribute to a deeper understanding of RAMs and their significance in chromatin structure regulation. For future RAM studies, it would be valuable to analyze more RAM patterns in other eukaryotes such as Drosophila and plants, spanning from embryo stages to adult stages, and from normal cells to cancer cells. This analysis would help determine whether RAM patterns are conserved across species in different stages like TAD or if they exhibit species-specific patterns.

This thesis, in part, is a coauthored unpublished material with Wei Wang. The thesis author is the primary investigator and author of the material that appeared in this thesis.

# Appendix A

# Supplementary Tables

**Table A.1**: The table below shows the number of RAM boundaries presented in each sample. If '-' is indicated, it means that there is no available data on the ENCODE website for that particular sample.

| **Tissue/Time** | **e10.5** | **e11.5** | **e12.5** | **e13.5** | **e14.5** | **e15.5** | **e16.5** | **p0** |
|---|---|---|---|---|---|---|---|---|
| EFP | 556 | 556 | 572 | 600 | 621 | 605 | - | - |
| Forebrain | 565 | 612 | 624 | 621 | 625 | 667 | 648 | 654 |
| Midbrain | 580 | 603 | 610 | 631 | 621 | 640 | 638 | 616 |
| Hindbrain | 551 | 588 | 613 | 642 | 662 | 656 | 617 | 638 |
| NT | - | 590 | 516 | 649 | 591 | 625 | - | - |
| Heart | 564 | 621 | 644 | 664 | 654 | 646 | 651 | 654 |
| Limb | 562 | 552 | 603 | 590 | 588 | 595 | - | - |
| Liver | - | 537 | 558 | 562 | 545 | 582 | 604 | 638 |
| Lung | - | - | - | - | 627 | 655 | 660 | 617 |
| Kidney | - | - | - | - | 590 | 582 | 642 | 647 |
| Intestine | - | - | - | - | 619 | 597 | 652 | 590 |
| Stomach | - | - | - | - | 606 | 611 | 640 | 588 |

**Table A.2**: The table below lists all the ENCODE assays that were used in the analysis.

| Accession | assay | Biosample term name | Biosample age |
|---|---|---|---|
| ENCSR151APL | ChIP-seq | embryonic facial prominence | e10.5 |
| ENCSR401GRX | ChIP-seq | embryonic facial prominence | e11.5 |

**Table A.2**: The table below lists all the ENCODE assays that were used in the analysis.

| Accession | assay | Biosample term name | Biosample age |
|---|---|---|---|
| ENCSR813SCQ | ChIP-seq | embryonic facial prominence | e12.5 |
| ENCSR420MUV | ChIP-seq | embryonic facial prominence | e13.5 |
| ENCSR481SGM | ChIP-seq | embryonic facial prominence | e14.5 |
| ENCSR382DRK | ChIP-seq | embryonic facial prominence | e15.5 |
| ENCSR825ZJV | ChIP-seq | forebrain | e10.5 |
| ENCSR275KPI | ChIP-seq | forebrain | e11.5 |
| ENCSR966AIB | ChIP-seq | forebrain | e12.5 |
| ENCSR311YPF | ChIP-seq | forebrain | e13.5 |
| ENCSR320EEW | ChIP-seq | forebrain | e14.5 |
| ENCSR691NQH | ChIP-seq | forebrain | e15.5 |
| ENCSR428OEK | ChIP-seq | forebrain | e16.5 |
| ENCSR094TTT | ChIP-seq | forebrain | p0 |
| ENCSR582SPN | ChIP-seq | heart | e10.5 |
| ENCSR222IHX | ChIP-seq | heart | e11.5 |
| ENCSR123MLY | ChIP-seq | heart | e12.5 |
| ENCSR699XHY | ChIP-seq | heart | e13.5 |
| ENCSR360ANE | ChIP-seq | heart | e14.5 |
| ENCSR574VME | ChIP-seq | heart | e15.5 |
| ENCSR846PJO | ChIP-seq | heart | e16.5 |
| ENCSR675HDX | ChIP-seq | heart | p0 |
| ENCSR594JGI | ChIP-seq | hindbrain | e10.5 |
| ENCSR129LAP | ChIP-seq | hindbrain | e11.5 |
| ENCSR784TLR | ChIP-seq | hindbrain | e12.5 |

**Table A.2**: The table below lists all the ENCODE assays that were used in the analysis.

| Accession | assay | Biosample term name | Biosample age |
|---|---|---|---|
| ENCSR344HHI | ChIP-seq | hindbrain | e13.5 |
| ENCSR054JHZ | ChIP-seq | hindbrain | e14.5 |
| ENCSR066XFL | ChIP-seq | hindbrain | e15.5 |
| ENCSR797EYS | ChIP-seq | hindbrain | e16.5 |
| ENCSR332JYZ | ChIP-seq | hindbrain | p0 |
| ENCSR424END | ChIP-seq | intestine | e14.5 |
| ENCSR599GVS | ChIP-seq | intestine | e15.5 |
| ENCSR639DND | ChIP-seq | intestine | e16.5 |
| ENCSR642VYW | ChIP-seq | intestine | p0 |
| ENCSR057SHA | ChIP-seq | kidney | e14.5 |
| ENCSR711SVB | ChIP-seq | kidney | e15.5 |
| ENCSR357JII | ChIP-seq | kidney | e16.5 |
| ENCSR140YPL | ChIP-seq | kidney | p0 |
| ENCSR863VHE | ChIP-seq | limb | e10.5 |
| ENCSR897WBY | ChIP-seq | limb | e11.5 |
| ENCSR737QWV | ChIP-seq | limb | e12.5 |
| ENCSR905FFU | ChIP-seq | limb | e13.5 |
| ENCSR021ALF | ChIP-seq | limb | e14.5 |
| ENCSR988BRP | ChIP-seq | limb | e15.5 |
| ENCSR058DOA | ChIP-seq | liver | e11.5 |
| ENCSR136GMT | ChIP-seq | liver | e12.5 |
| ENCSR175KBJ | ChIP-seq | liver | e13.5 |
| ENCSR075SNV | ChIP-seq | liver | e14.5 |

**Table A.2**: The table below lists all the ENCODE assays that were used in the analysis.

| Accession | assay | Biosample term name | Biosample age |
|---|---|---|---|
| ENCSR479LFP | ChIP-seq | liver | e15.5 |
| ENCSR802RET | ChIP-seq | liver | e16.5 |
| ENCSR616TJM | ChIP-seq | liver | p0 |
| ENCSR452WYC | ChIP-seq | lung | e14.5 |
| ENCSR895BMP | ChIP-seq | lung | e15.5 |
| ENCSR140UEX | ChIP-seq | lung | e16.5 |
| ENCSR884MYD | ChIP-seq | lung | p0 |
| ENCSR989LUY | ChIP-seq | midbrain | e10.5 |
| ENCSR088UKA | ChIP-seq | midbrain | e11.5 |
| ENCSR252ONR | ChIP-seq | midbrain | e12.5 |
| ENCSR671NSS | ChIP-seq | midbrain | e13.5 |
| ENCSR254AHA | ChIP-seq | midbrain | e14.5 |
| ENCSR428GHF | ChIP-seq | midbrain | e15.5 |
| ENCSR553IWV | ChIP-seq | midbrain | e16.5 |
| ENCSR672ZXY | ChIP-seq | midbrain | p0 |
| ENCSR531RZS | ChIP-seq | neural tube | e11.5 |
| ENCSR891SAW | ChIP-seq | neural tube | e12.5 |
| ENCSR289SWJ | ChIP-seq | neural tube | e13.5 |
| ENCSR265NBM | ChIP-seq | neural tube | e14.5 |
| ENCSR241BSK | ChIP-seq | neural tube | e15.5 |
| ENCSR316CNR | ChIP-seq | stomach | e14.5 |
| ENCSR929SEW | ChIP-seq | stomach | e15.5 |
| ENCSR546ANT | ChIP-seq | stomach | e16.5 |

**Table A.2**: The table below lists all the ENCODE assays that were used in the analysis.

| Accession | assay | Biosample term name | Biosample age |
|---|---|---|---|
| ENCSR346FJG | ChIP-seq | stomach | p0 |
| ENCSR809VYL | RNA-seq | embryonic facial prominence | 10.5 days |
| ENCSR848HOX | RNA-seq | embryonic facial prominence | 11.5 days |
| ENCSR851HEC | RNA-seq | embryonic facial prominence | 12.5 days |
| ENCSR538WYL | RNA-seq | embryonic facial prominence | 13.5 days |
| ENCSR823VEE | RNA-seq | embryonic facial prominence | 14.5 days |
| ENCSR636CWO | RNA-seq | embryonic facial prominence | 15.5 days |
| ENCSR304RDL | RNA-seq | forebrain | 10.5 days |
| ENCSR160IIN | RNA-seq | forebrain | 11.5 days |
| ENCSR647QBV | RNA-seq | forebrain | 12.5 days |
| ENCSR970EWM | RNA-seq | forebrain | 13.5 days |
| ENCSR185LWM | RNA-seq | forebrain | 14.5 days |
| ENCSR752RGN | RNA-seq | forebrain | 15.5 days |
| ENCSR080EVZ | RNA-seq | forebrain | 16.5 days |
| ENCSR094TTT | RNA-seq | forebrain | p0 days |
| ENCSR049UJU | RNA-seq | heart | 10.5 days |
| ENCSR691OPQ | RNA-seq | heart | 11.5 days |
| ENCSR150CUE | RNA-seq | heart | 12.5 days |
| ENCSR284YKY | RNA-seq | heart | 13.5 days |
| ENCSR727FHP | RNA-seq | heart | 14.5 days |
| ENCSR597UZW | RNA-seq | heart | 15.5 days |
| ENCSR020DGG | RNA-seq | heart | 16.5 days |
| ENCSR675HDX | RNA-seq | heart | p0 days |

**Table A.2**: The table below lists all the ENCODE assays that were used in the analysis.

| Accession | assay | Biosample term name | Biosample age |
|---|---|---|---|
| ENCSR943LKA | RNA-seq | hindbrain | 10.5 days |
| ENCSR760TOE | RNA-seq | hindbrain | 11.5 days |
| ENCSR420QTO | RNA-seq | hindbrain | 12.5 days |
| ENCSR921PRX | RNA-seq | hindbrain | 13.5 days |
| ENCSR559TRB | RNA-seq | hindbrain | 14.5 days |
| ENCSR401BSG | RNA-seq | hindbrain | 15.5 days |
| ENCSR285WZV | RNA-seq | hindbrain | 16.5 days |
| ENCSR332JYZ | RNA-seq | hindbrain | p0 days |
| ENCSR932TRU | RNA-seq | intestine | 14.5 days |
| ENCSR370SFB | RNA-seq | intestine | 15.5 days |
| ENCSR848GST | RNA-seq | intestine | 16.5 days |
| ENCSR642VYW | RNA-seq | intestine | p0 days |
| ENCSR504GEG | RNA-seq | kidney | 14.5 days |
| ENCSR062VTB | RNA-seq | kidney | 15.5 days |
| ENCSR537GNQ | RNA-seq | kidney | 16.5 days |
| ENCSR140YPL | RNA-seq | kidney | p0 days |
| ENCSR968QHO | RNA-seq | limb | 10.5 days |
| ENCSR541XZK | RNA-seq | limb | 11.5 days |
| ENCSR216NEG | RNA-seq | limb | 14.5 days |
| ENCSR750YSX | RNA-seq | limb | 12.5 days |
| ENCSR347SQR | RNA-seq | limb | 13.5 days |
| ENCSR830IVQ | RNA-seq | limb | 15.5 days |
| ENCSR284AMY | RNA-seq | liver | 11.5 days |

**Table A.2**: The table below lists all the ENCODE assays that were used in the analysis.

| Accession | assay | Biosample term name | Biosample age |
|---|---|---|---|
| ENCSR648YEP | RNA-seq | liver | 12.5 days |
| ENCSR611PTP | RNA-seq | liver | 15.5 days |
| ENCSR448MXQ | RNA-seq | liver | 13.5 days |
| ENCSR867YNV | RNA-seq | liver | 14.5 days |
| ENCSR826HIQ | RNA-seq | liver | 16.5 days |
| ENCSR616TJM | RNA-seq | liver | p0 days |
| ENCSR039ADS | RNA-seq | lung | 14.5 days |
| ENCSR457RRW | RNA-seq | lung | 15.5 days |
| ENCSR992WBR | RNA-seq | lung | 16.5 days |
| ENCSR884MYD | RNA-seq | lung | p0 days |
| ENCSR764OPZ | RNA-seq | midbrain | 10.5 days |
| ENCSR908JWT | RNA-seq | midbrain | 12.5 days |
| ENCSR307BCA | RNA-seq | midbrain | 11.5 days |
| ENCSR792RJV | RNA-seq | midbrain | 13.5 days |
| ENCSR343YLB | RNA-seq | midbrain | 14.5 days |
| ENCSR557RMA | RNA-seq | midbrain | 15.5 days |
| ENCSR367ZPZ | RNA-seq | midbrain | 16.5 days |
| ENCSR672ZXY | RNA-seq | midbrain | p0 days |
| ENCSR337FYI | RNA-seq | neural tube | 11.5 days |
| ENCSR508GWZ | RNA-seq | neural tube | 12.5 days |
| ENCSR115TWD | RNA-seq | neural tube | 13.5 days |
| ENCSR928OXI | RNA-seq | neural tube | 14.5 days |
| ENCSR004XCU | RNA-seq | neural tube | 15.5 days |

**Table A.2**: The table below lists all the ENCODE assays that were used in the analysis.

| Accession | assay | Biosample term name | Biosample age |
|---|---|---|---|
| ENCSR290RRR | RNA-seq | stomach | 14.5 days |
| ENCSR906YQZ | RNA-seq | stomach | 15.5 days |
| ENCSR466KZY | RNA-seq | stomach | 16.5 days |
| ENCSR346FJG | RNA-seq | stomach | p0 days |

**Table A.3**: Number of Driver Transcript Factors Identification

| Stage/Tissue | Number of Driver TFs |
|---|---|
| e10.5 | 73 |
| e11.5 | 39 |
| e12.5 | 36 |
| e13.5 | 30 |
| e14.5 | 15 |
| e15.5 | 21 |
| e16.5 | 27 |
| p0 | 18 |
| EFP | 44 |
| Forebrain | 19 |
| Midbrain | 20 |
| Hindbrain | 29 |
| NT | 51 |
| Heart | 41 |
| Limb | 39 |
| Liver | 35 |
| Kidney | 86 |
| Lung | 75 |
| Intestine | 76 |
| Stomach | 80 |

# Appendix B

# Supplementary Figures

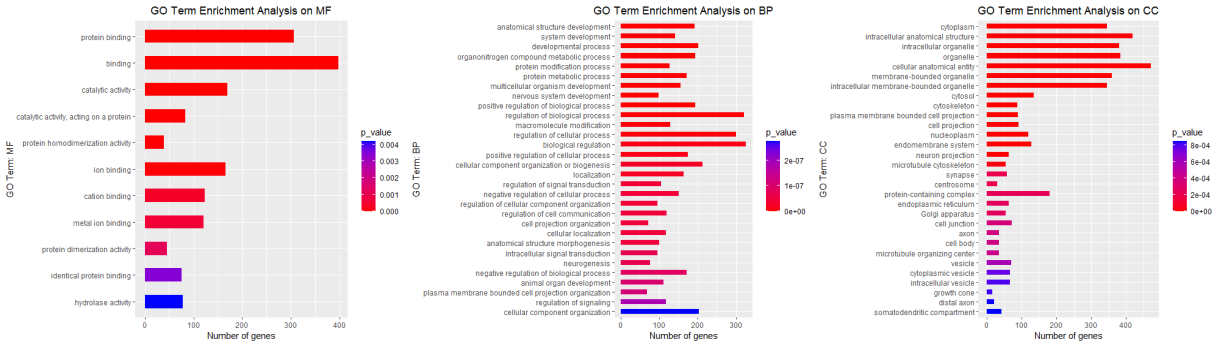



**Figure B.1**: This graph displays the top 30 detailed GO terms in three different categories for the comparison of neural tube and midbrain expressed genes at 14.5 days in variable RAM. The x-axis represents the number of genes associated with each pathway, while the y-axis represents the pathway names. The bars are color-coded based on the p-adjusted value.

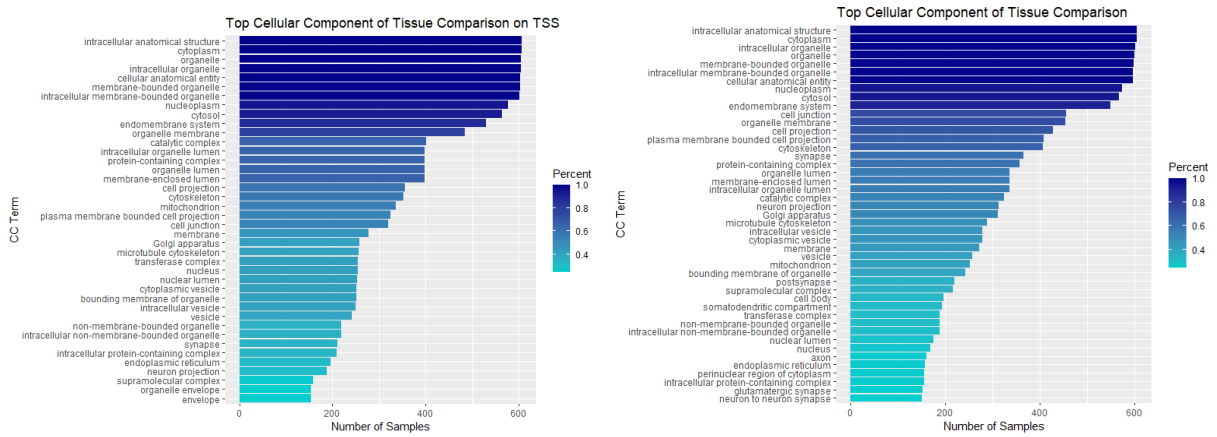



**Figure B.2**: Here are the selected pathways with a p-adjusted value smaller than 0.01 and presented in over 20% of all tissue comparisons. The left side shows the pathways using the TSS located in variable RAMs, while the right side shows the pathways using the expressed genes located in variable RAMs.

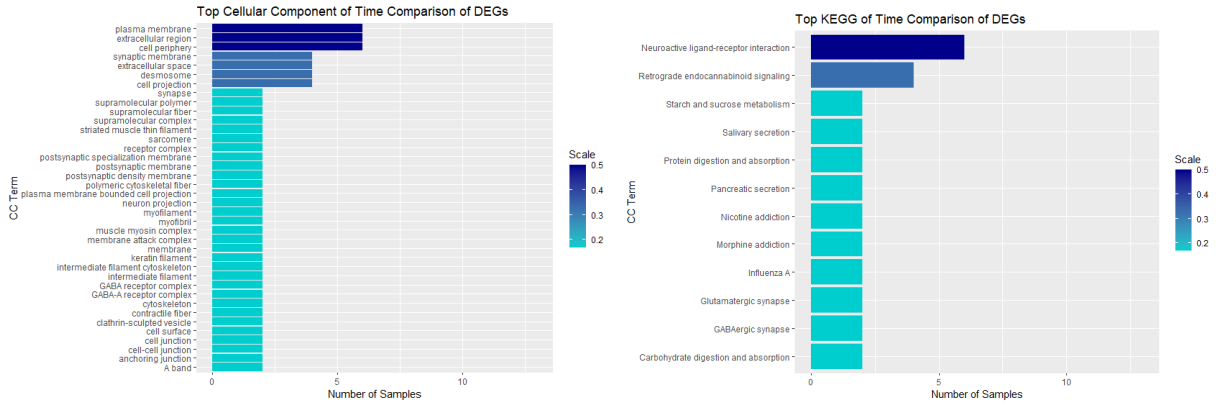



**Figure B.3**: Here are the pathways selected based on a p-adjusted value smaller than 0.02 and presented in over 10% for cellular component (CC) pathways and 10% for KEGG pathways of tissue comparisons. The left side shows the CC pathways in variable RAMs, while the right side shows the KEGG pathways in variable RAMs.

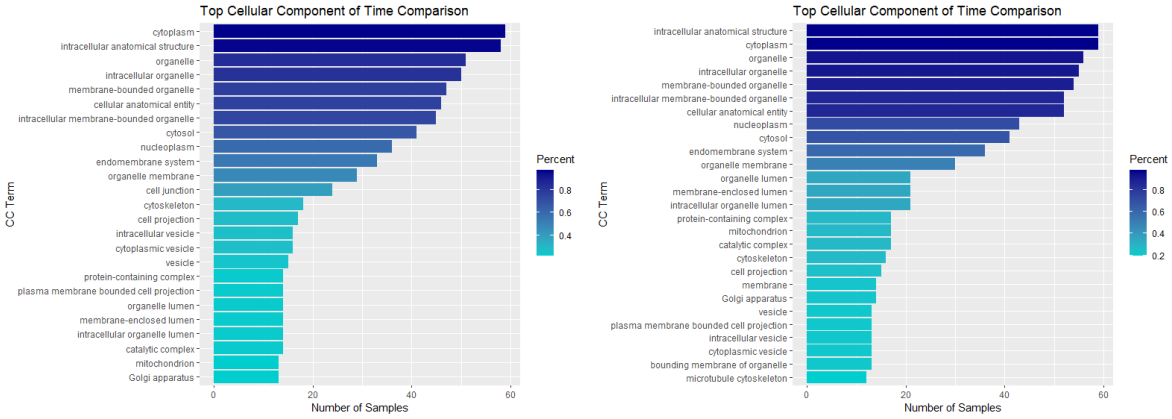
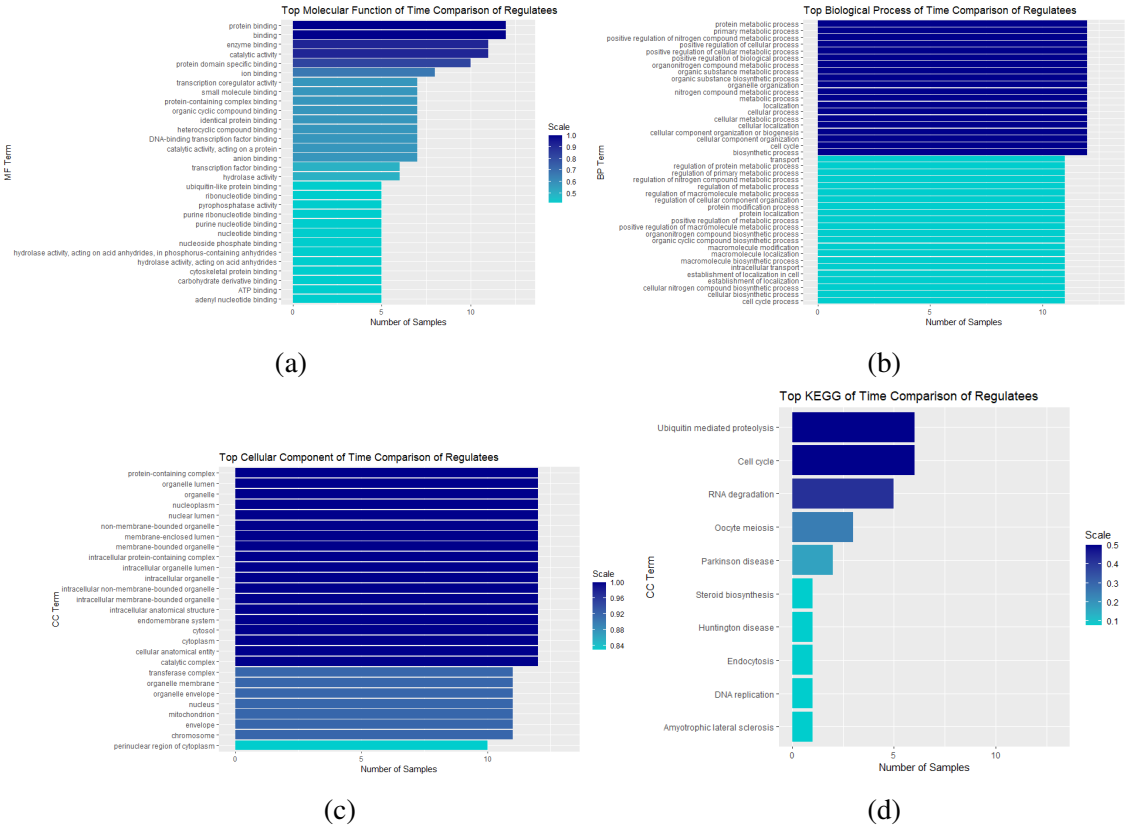



**Figure B.4**: Here are the CC pathways selected based on a p-adjusted value smaller than 0.01 and presented in over 15% of all tissue comparisons. The left side displays the pathways using the TSS located in variable RAMs, while the right side shows the pathways using the expressed genes located in variable RAMs for time comparison.




(a) (b)

(c) (d)

**Figure B.5**: Here are the pathways selected based on a p-adjusted value smaller than 0.02 and presented in over 40% for MF, 90% for BP, 83% for CC, and 10% for KEGG of 12 tissues with merged time comparisons:(a) MF pathways in variable RAMs;(b) BP pathways in variable RAMs;(c) CC pathways in variable RAMs;(d) KEGG pathways in variable RAMs

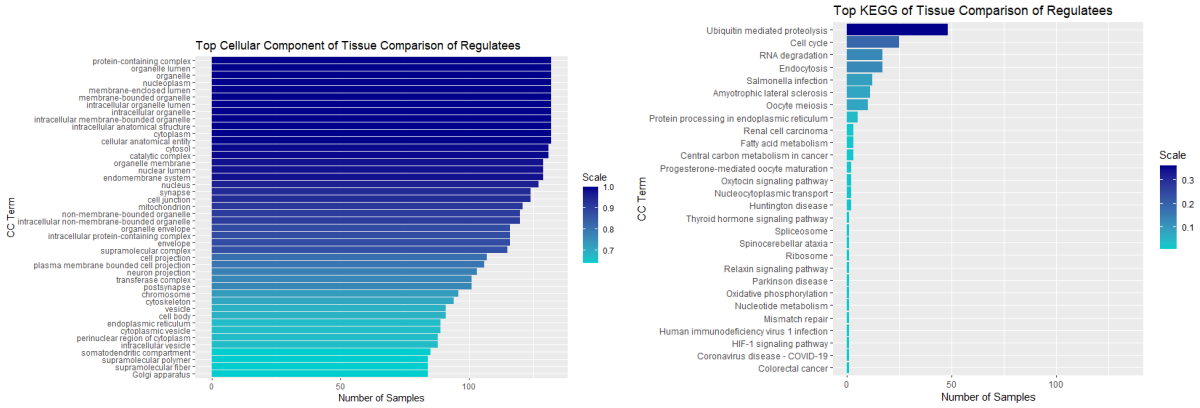
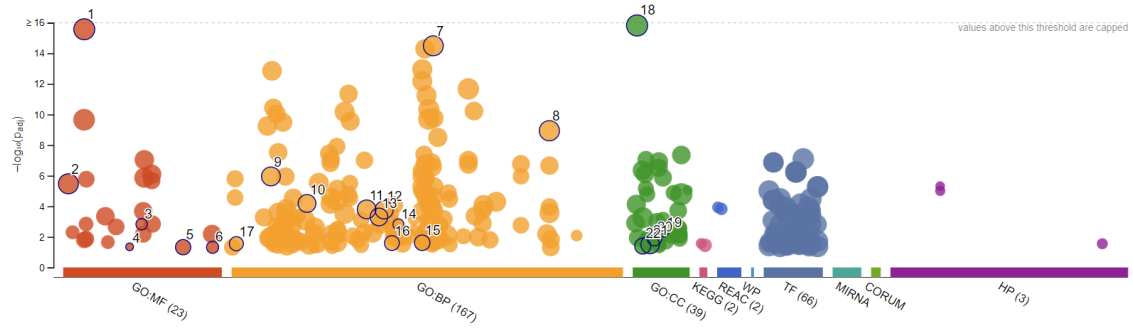
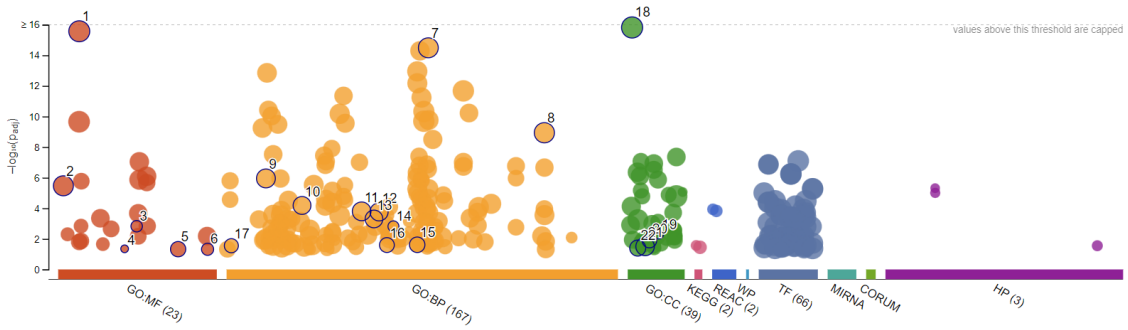



**Figure B.6**: Here are the pathways selected based on the p-adjusted value smaller than 0.02 and presented in over 20% of CC and all KEGG tissue comparisons. The left side displays CC pathways in variable RAMs, while the right side shows KEGG pathways in variable RAMs.




| ID | Source | Term ID | Term Name | $p_{adj}$ (query_1) |
|---|---|---|---|---|
| 1 | GO:MF | GO:0005515 | protein binding | $2.712\times10^{-16}$ |
| 2 | GO:MF | GO:0003824 | catalytic activity | $3.407\times10^{-6}$ |
| 3 | GO:MF | GO:0038187 | pattern recognition receptor activity | $1.495\times10^{-3}$ |
| 4 | GO:MF | GO:0033782 | 24-hydroxycholesterol 7alpha-hydroxylase activity | $4.485\times10^{-2}$ |
| 5 | GO:MF | GO:0061630 | ubiquitin protein ligase activity | $4.724\times10^{-2}$ |
| 6 | GO:MF | GO:0140303 | intramembrane lipid transporter activity | $4.805\times10^{-2}$ |
| 7 | GO:BP | GO:0051179 | localization | $3.255\times10^{-15}$ |
| 8 | GO:BP | GO:1901564 | organonitrogen compound metabolic process | $1.164\times10^{-9}$ |
| 9 | GO:BP | GO:0007166 | cell surface receptor signaling pathway | $1.126\times10^{-6}$ |
| 10 | GO:BP | GO:0016477 | cell migration | $6.566\times10^{-5}$ |
| 11 | GO:BP | GO:0035556 | intracellular signal transduction | $1.593\times10^{-4}$ |
| 12 | GO:BP | GO:0042592 | homeostatic process | $1.725\times10^{-4}$ |
| 13 | GO:BP | GO:0040011 | locomotion | $4.993\times10^{-4}$ |
| 14 | GO:BP | GO:0044790 | suppression of viral release by host | $1.533\times10^{-3}$ |
| 15 | GO:BP | GO:0048511 | rhythmic process | $2.432\times10^{-2}$ |
| 16 | GO:BP | GO:0043583 | ear development | $2.543\times10^{-2}$ |
| 17 | GO:BP | GO:0001508 | action potential | $2.809\times10^{-2}$ |
| 18 | GO:CC | GO:0005737 | cytoplasm | $1.530\times10^{-16}$ |
| 19 | GO:CC | GO:0046581 | intercellular canaliculus | $4.875\times10^{-3}$ |
| 20 | GO:CC | GO:0034703 | cation channel complex | $1.300\times10^{-2}$ |
| 21 | GO:CC | GO:0031982 | vesicle | $3.153\times10^{-2}$ |
| 22 | GO:CC | GO:0015629 | actin cytoskeleton | $4.010\times10^{-2}$ |

**Figure B.7**: This is the enrichment analysis for driver GO:term for expressed genes in cRAM MS region for tissue comparisons.




| ID | Source | Term ID | Term Name | $p_{adj}$ (query_1) |
|---|---|---|---|---|
| 1 | GO:MF | GO:0005515 | protein binding | $2.712\times10^{-16}$ |
| 2 | GO:MF | GO:0003824 | catalytic activity | $3.407\times10^{-6}$ |
| 3 | GO:MF | GO:0038187 | pattern recognition receptor activity | $1.495\times10^{-3}$ |
| 4 | GO:MF | GO:0033782 | 24-hydroxycholesterol 7alpha-hydroxylase activity | $4.485\times10^{-2}$ |
| 5 | GO:MF | GO:0061630 | ubiquitin protein ligase activity | $4.724\times10^{-2}$ |
| 6 | GO:MF | GO:0140303 | intramembrane lipid transporter activity | $4.805\times10^{-2}$ |
| 7 | GO:BP | GO:0051179 | localization | $3.255\times10^{-15}$ |
| 8 | GO:BP | GO:1901564 | organonitrogen compound metabolic process | $1.164\times10^{-9}$ |
| 9 | GO:BP | GO:0007166 | cell surface receptor signaling pathway | $1.126\times10^{-6}$ |
| 10 | GO:BP | GO:0016477 | cell migration | $6.566\times10^{-5}$ |
| 11 | GO:BP | GO:0035556 | intracellular signal transduction | $1.593\times10^{-4}$ |
| 12 | GO:BP | GO:0042592 | homeostatic process | $1.725\times10^{-4}$ |
| 13 | GO:BP | GO:0040011 | locomotion | $4.993\times10^{-4}$ |
| 14 | GO:BP | GO:0044790 | suppression of viral release by host | $1.533\times10^{-3}$ |
| 15 | GO:BP | GO:0048511 | rhythmic process | $2.432\times10^{-2}$ |
| 16 | GO:BP | GO:0043583 | ear development | $2.543\times10^{-2}$ |
| 17 | GO:BP | GO:0001508 | action potential | $2.809\times10^{-2}$ |
| 18 | GO:CC | GO:0005737 | cytoplasm | $1.530\times10^{-16}$ |
| 19 | GO:CC | GO:0046581 | intercellular canaliculus | $4.875\times10^{-3}$ |
| 20 | GO:CC | GO:0034703 | cation channel complex | $1.300\times10^{-2}$ |
| 21 | GO:CC | GO:0031982 | vesicle | $3.153\times10^{-2}$ |
| 22 | GO:CC | GO:0015629 | actin cytoskeleton | $4.010\times10^{-2}$ |

**Figure B.8**: This is the enrichment analysis for driver GO:term for expressed genes in cRAM MS region for time comparisons.

# Bibliography


[1] D. U. Gorkin, D. Leung, and B. Ren, "The 3d genome in transcriptional regulation and pluripotency.," *Cell stem cell*, vol. 14 6, pp. 762–75, 2014.

[2] D. G. Lupiáñez, M. Spielmann, and S. Mundlos, "Breaking tads: How alterations of chromatin domains result in disease.," *Trends in genetics : TIG*, vol. 32 4, pp. 225–237, 2016.

[3] J.-P. Fortin and K. D. Hansen, "Reconstructing a/b compartments as revealed by hi-c using long-range correlations in epigenetic data," *Genome Biology*, vol. 16, 2015.

[4] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, pp. 376 – 380, 2012.

[5] A. Pombo and N. Dillon, "Three-dimensional genome architecture: players and mechanisms," *Nature Reviews Molecular Cell Biology*, vol. 16, pp. 245–257, 2015.

[6] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. S. Sandstrom, B. E. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. A. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *Science*, vol. 326, pp. 289 – 293, 2009.

[7] Y. Zhu, Z. Chen, K. Zhang, M. Wang, D. Medovoy, J. W. Whitaker, B. Ding, N. Li, L. Zheng, and W. Wang, "Constructing 3d interaction maps from 1d epigenomes," *Nature Communications*, vol. 7, 2016.

[8] J. Huang, E. Marco, L. Pinello, and G. cheng Yuan, "Predicting chromatin organization using histone marks," *Genome Biology*, vol. 16, 2015.

[9] H. Ashoor, X. Chen, W. Rosikiewicz, J. Wang, A. W. Cheng, P. Wang, Y. Ruan, and S. Li, "Graph embedding and unsupervised learning predict genomic sub-compartments from hic chromatin interaction data," *Nature Communications*, vol. 11, 2020.

[10] Z. J. Liu, W. R. Legant, B.-C. Chen, L. Li, J. B. Grimm, L. D. Lavis, E. Betzig, and R. Tjian, "3d imaging of sox2 enhancer clusters in embryonic stem cells," *eLife*, vol. 3, 2014.

[11] A. Pancholi, T. Klingberg, W. Zhang, R. Prizak, I. Mamontova, A. Noa, M. Sobucki, A. Y. Kobitski, G. U. Nienhaus, V. Zaburdaev, and L. Hilbert, "Rna polymerase ii clusters form in line with surface condensation on regulatory chromatin," *Molecular Systems Biology*, vol. 17, 2021.

[12] L. Zheng and W. Wang, "Regulation associated modules reflect 3d genome modularity associated with chromatin activity," *Nature Communications*, vol. 13, 2022.

[13] A. Barral and J. Déjardin, "The chromatin signatures of enhancers and their dynamic regulation," *Nucleus*, vol. 14, 2023.

[14] P. J. Park, "Chip–seq: advantages and challenges of a maturing technology," *Nature Reviews Genetics*, vol. 10, pp. 669–680, 2009.

[15] T. Kouzarides, "Chromatin modifications and their function," *Cell*, vol. 128, pp. 693–705, 2007.

[16] A. Srivastava, R. Timsina, S.-J. Heo, S. W. Dewage, S. Kirmizialtin, and X. Qiu, "Structure-guided dna–dna attraction mediated by divalent cations," *Nucleic Acids Research*, vol. 48, pp. 7018 – 7026, 2020.

[17] C. A. Davey and T. J. Richmond, "Dna-dependent divalent cation binding in the nucleosome core particle," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 11169 – 11174, 2002.

[18] U. Raudvere, L. Kolberg, I. V. Kuzmin, T. Arak, P. Adler, H. Peterson, and J. Vilo, "g:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)," *Nucleic Acids Research*, vol. 47, pp. W191 – W198, 2019.

[19] R. Phengchat, H. Takata, K. Morii, N. Inada, H. Murakoshi, S. Uchiyama, and K. Fukui, "Calcium ions function as a booster of chromosome condensation," *Scientific Reports*, vol. 6, 2016.

[20] I. Dunham, A. Kundaje, S. Aldred, P. Collins, C. Davis, F. Doyle, C. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. Lajoie, S. Landt, B.-K. Lee, F. Pauli Behn, K. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, and E. Birney, "The encode project consortium: An integrated encyclopedia of dna elements in the human genome. 2012. nature 489: 57–74," *Nature*, vol. 489, pp. 57–74, 09 2012.

[21] A. Frankish, M. Diekhans, I. Jungreis, J. Lagarde, J. Loveland, J. M. Mudge, C. Sisu, J. C. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, C. Boix, S. Carbonell Sala, F. Cunningham, T. Di Domenico, S. Donaldson, I. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, K. L. Howe, T. Hunt, O. G. Izuogu, R. Johnson, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, F. C. Riera, M. Ruffier, B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, M. Y. Wolf, J. Xu, Y. Yang, A. Yates, D. Zerbino, Y. Zhang, J. Choudhary, M. Gerstein,

R. Guigó, T. J. P. Hubbard, M. Kellis, B. Paten, M. L. Tress, and P. Flicek, "GENCODE 2021," *Nucleic Acids Research*, pp. D916–D923, 2020.

[22] W. S. Cleveland and E. Grosse, "Computational methods for local regression," *Statistics and Computing*, vol. 1, pp. 47–62, 1991.

[23] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov, "Integrative genomics viewer," *Nature biotechnology*, vol. 29, pp. 24 – 26, 2011.

[24] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "Star: ultrafast universal rna-seq aligner," *Bioinformatics*, vol. 29 1, pp. 15–21, 2013.

[25] B. Li and C. N. Dewey, "Rsem: accurate transcript quantification from rna-seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, pp. 323 – 323, 2011.

[26] P. A. Ewels, A. Peltzer, S. Fillinger, H. Patel, J. Alneberg, A. Wilm, M. U. Garcia, P. D. Tommaso, and S. Nahnsen, "The nf-core framework for community-curated bioinformatics pipelines," *Nature Biotechnology*, vol. 38, pp. 276–278, 2020.

[27] T. Juven-Gershon, J.-Y. hsu, J. W. M. Theisen, and J. T. Kadonaga, "The rna polymerase ii core promoter - the gateway to transcription.," *Current opinion in cell biology*, vol. 20 3, pp. 253–9, 2008.

[28] J. Wang, C. Liu, Y. xuan Chen, and W. Wang, "Taiji-reprogram: a framework to uncover cell-type specific regulators and predict cellular reprogramming cocktails," *NAR Genomics and Bioinformatics*, vol. 3, 2021.

[29] P. D. Thomas, D. Ebert, A. Muruganujan, T. Mushayahama, L.-P. Albou, and H. Mi, "Panther: Making genome-scale phylogenetics accessible to all," *Protein Science*, vol. 31, pp. 22 – 8, 2021.

[30] M. Kanehisa, M. Furumichi, Y. Sato, M. Kawashima, and M. Ishiguro-Watanabe, "Kegg for taxonomy-based analysis of pathways and genomes," *Nucleic Acids Research*, vol. 51, pp. D587 – D592, 2022.

[31] R. Fairless, S. K. Williams, and R. Diem, "Dysfunction of neuronal calcium signalling in neuroinflammation and neurodegeneration," *Cell and Tissue Research*, vol. 357, pp. 455–462, 2014.

[32] L. Zwierzchowski, M. Czlonkowska, E. Siadkowska, and A. Guszkiewicz, "The role of calcium in dna synthesis and development of mouse preimplantation embryos in vitro.," *Folia biologica*, vol. 40 3-4, pp. 103–8, 1992.

[33] R. Fairless, S. K. Williams, and R. Diem, "Calcium-binding proteins as determinants of central nervous system neuronal vulnerability to disease," *International Journal of Molecular Sciences*, vol. 20, 2019.

[34] M. Kaufman and V. Navaratnam, "Early differentiation of the heart in mouse embryos.," *Journal of anatomy*, vol. 133 Pt 2, pp. 235–46, 1981.

[35] S. E. Pryor, V. Massa, D. Savery, N. D. E. Greene, and A. J. Copp, "Convergent extension analysis in mouse whole embryo culture.," *Methods in molecular biology*, vol. 839, pp. 133–46, 2012.

[36] I. Koltover, K. Wagner, and C. R. Safinya, "Dna condensation in two dimensions.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97 26, pp. 14046–51, 2000.

[37] G. Wei, B. J. Abraham, R. Yagi, R. Jothi, K. Cui, S. Sharma, L. Narlikar, D. L. Northrup, Q. Tang, W. E. Paul, J. Zhu, and K. Zhao, "Genome-wide analyses of transcription factor gata3-mediated gene regulation in distinct t cell types.," *Immunity*, vol. 35 2, pp. 299–311, 2011.

[38] F. Carrier, "Chromatin modulation by histone deacetylase inhibitors: Impact on cellular sensitivity to ionizing radiation.," *Molecular and cellular pharmacology*, vol. 5 1, pp. 51–59, 2013.

[39] L. T. Gray, Z. Yao, T. N. Nguyen, T. K. Kim, H. Zeng, and B. Tasic, "Layer-specific chromatin accessibility landscapes reveal regulatory networks in adult mouse visual cortex," *eLife*, vol. 6, 2017.

[40] M. Takaku, S. A. Grimm, T. Shimbo, L. E. Perera, R. Menafra, H. G. Stunnenberg, T. K. Archer, S. Machida, H. Kurumizaka, and P. A. Wade, "Gata3-dependent cellular reprogramming requires activation-domain dependent recruitment of a chromatin remodeler," *Genome Biology*, vol. 17, 2016.

[41] I. Sandovici, N. H. Smith, M. D. Nitert, M. A. Ackers-Johnson, S. Uribe-Lewis, Y. Ito, R. H. Jones, V. E. Marquez, W. J. Cairns, M. Tadayyon, L. O'Neill, A. Murrell, C. Ling, M. Constância, and S. E. Ozanne, "Maternal diet and aging alter the epigenetic control of a promoter–enhancer interaction at the hnf4a gene in rat pancreatic islets," *Proceedings of the National Academy of Sciences*, vol. 108, pp. 5449 – 5454, 2011.