3D Genome Organization in the Developing Macaque Brain

By

AEDRIC LIM
THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

Integrative Genetics and Genomics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Megan Dennis, Chair

Alex Nord

Hong Ji

Committee in Charge

2022

Table of Contents

Abstract

A large proportion of genetic variation underlying differences in behavioral traits and neurodevelopmental disorders remains unknown despite considerable effort. Much work has focused on protein-coding regions, which comprise ~1.5% of primate genomes, compared to functional noncoding elements comprising ~40%. Transcriptional analyses of the developing brain in distantly related primates, such as humans and rhesus macaques, show that spatiotemporal expression patterns are largely conserved across lineages, with minor differences likely contributing to species' divergence. To delve into genomic mechanisms underlying gene regulation in the developing primate brain, we have generated transcriptional genomic datasets (3' Tag-Seq (n=3) and single-cell RNA-seq (n=1)) from diverse regions of fetal macaque brains (60 days gestation, late first/early second trimester) representing a time of early neurogenesis. The 3' Tag-Seq expression levels show correlation with previously published data of the same time point and comparable brain regions and highlights genes important in region-specific development. The single-cell RNA-seq data has revealed a number of different cell types, the majority of which are neuronal in all brain regions with the exception of the cerebellum.

We then queried the 3D genome organization of two different brain regions using a targeted protein mediated approach called Proximity Ligation-Assisted ChIP-Seq (PLAC-Seq) to identify DNA interactions within chromatin enriched at active promoters in macaque neural tissue. Our data generated from samples from the prefrontal cortex and cerebellum (n=3, 60 days gestation) shows low enrichment for the H3K4me3 antibody. Despite this low enrichment, we were able to identify 32,959 and 13,814 significant chromatin interactions for prefrontal cortex and cerebellum, respectively. The identified significant interactions in the prefrontal cortex were connected to over half of the genes implicated in ASD from two separate datasets, while

significant interactions in the cerebellum connected to about a third of those genes, providing putative cis regulatory elements driving regulation of these disease-associated genes in diverse brain regions.

Introduction

Recent advances in sequencing efforts of protein-coding regions have elucidated genetic variation contributing to brain development, neurodevelopmental disorders, and species differences. In spite of this collective effort, much of the noncoding regions of the genome remains understudied, leaving most functional genetic variation unknown. Previous studies of primate neural tissues have characterized the transcriptomic landscape. When coupled with chromatin marks of enhancers and promoters, putative regulatory elements and genes have been linked to brain development and neurodevelopmental disorders (An et al., 2018; Brandler et al., 2018; Turner et al., 2016; Zhou et al., 2019). However, these studies treat the genome as a linear model in contrast to the 3D structure that DNA exists as in the nucleus wrapped around protein as chromatin. Because of this, previous studies often fail to couple distal enhancers and promoters to the genes they may regulate.

Compared with nonhuman primates, the human brain has unique developmental patterns and anatomy (Sousa et al., 2017). These unique differences in humans can also lead to unique neurodevelopmental disorders. Using techniques like RNA sequencing (RNA-Seq) and microarrays, several studies to date have characterized gene transcription of neural tissue from primates and in human individuals with autism spectrum development (ASD), giving insight into impact of gene expression levels, but not into how those levels are regulated (An et al., 2018; Brawand et al., 2011; Gupta et al., 2014; Voineagu et al., 2011; Zhu et al., 2018). There have been very few studies to characterize genome organization using primary tissue in primates and even then, those have been limited. One study performed Hi-C of the frontoparietal cortex of three human fetuses at gestation weeks 17 and 18, dissected from the cortex plate and germinal zone (Won et al., 2016). They connected 3D genome maps with non-coding variants implicated

in schizophrenia, highlighting multiple candidate risk genes and pathways such as a distal GWAS loci that regulates *FOXG1* expression.

A subsequent study based on the experimental design of Won et al. (2016) took that previously published human and mouse data and compared it to newly generated ultra-deep Hi-C of three macaque fetuses at 84 days after gestation in order to identify human-specific chromatin structure changes that could play a role in human brain development and evolution (Luo et al., 2021). From this comparison, they identified 499 topologically associating domains (TADs) and 1,266 chromatin loops as human-specific chromatin structure changes. Of the human-specific loops identified, many were significantly enriched in enhancer-enhancer interactions, and subsequently those associated regulated genes were shown to be preferentially expressed in the subplate lamina. They further investigated the target gene of one of the human-specific loops, *EPHA7*, a gene previously implicated in regulating brain development, by disrupting the enhancer sequence using CRISPR-Cas9 and found that it impacts early neural circuit development. Additionally, many human sequence changes were found at novel TAD boundaries and loop anchors, which may lead to species-specific gene regulation, transcription factor binding sites, and chromatin structures.

Although the aforementioned study resulted in significant findings regarding primate brain development and neurodevelopmental disorders, it targeted only two cortical layers at a single developmental time point, which were subsequently merged into one larger dataset to increase resolution. This was due to the cost limitations of generating genome-wide Hi-C. To more affordably expand analyses of gene regulation beyond what was previously studied, targeted chromatin methods can be used to generate 3D genomic maps of actively expressed genes. One such method, Proximity Ligation-Assisted ChIP-Seq (PLAC-Seq) captures 3D

conformation using antibodies targeting specific proteins of interest (Fang et al., 2016). It is a similar protocol to chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) and HiChIP as these methods are used to identify long-range interactions at promoters and enhancers (Fullwood et al., 2009; Mumbach et al., 2016). The PLAC-Seq protocol is composed of *in situ* proximity ligation, chromatin immunoprecipitation (ChIP), and biotin pull-down followed by library construction and sequencing. While ChIA-PET requires hundreds of millions of cells for input, PLAC-Seq conducts proximity ligation in nuclei prior to chromatin shearing and immunoprecipitation and requires only a few million cells for input, drastically expanding the diversity of tissue and samples that can be used as input. The targeted nature of PLAC-seq also allows high resolution maps with relatively shallow sequencing when compared to whole genome Hi-C.

Since we are interested in brain development and disorders in humans, our experiment would ideally use fetal brain tissue from humans or our closest primate relatives, chimpanzees (Chimpanzee Sequencing and Analysis Consortium, 2005). However, because of the difficulties in obtaining fetal brain tissue from great apes, we used rhesus macaque fetal brain tissue. Macaque is the most widely studied nonhuman primate. Transcriptional analysis of diverse prenatal and postnatal brain regions of human and macaque show that distinct changes in expression across brain regions and developmental time are conserved across primate lineages (Bakken et al., 2015; Zhu et al., 2018). Macaque is a suitable model for studying the human brain, so gaining a detailed understanding of gene regulation in macaque through the study of noncoding regulatory elements can provide insight into human development and disease.

We focused our PLAC-seq efforts on the prefrontal cortex and cerebellum. Our main focus is on the prefrontal cortex because humans have a greatly expanded neocortex in

comparison to other primates, so the unique cognitive functions of humans are usually attributed to the neocortex (Rakic, 2009). In addition to unique cognitive functions, many neurodevelopmental disorders have been associated with the prefrontal cortex (Ouhaz et al., 2018). Conversely, the cerebellum is generally considered the oldest region of the brain in evolutionary terms, though it has been shown to have a strong pattern of correlated evolution with the neocortex (Barton & Venditti, 2017). Both regions have been linked to the etiology of ASD (Hoeft et al., 2011; Kelly et al., 2020).

Methods to characterize gene expression

We used 3' Tag-Seq to check for brain region gene expression consistency with previously published whole transcript RNA-Seq data from the PsychENCODE Consortium in order to verify consistent brain region isolations (PsychENCODE Consortium et al., 2015). Since 3' Tag-Seq focuses on sequencing only the 3' end of mRNA fragments and requires less RNA input, it theoretically costs less in terms of sequencing and tissue than whole transcript RNA-Seq (Lohman et al., 2016; Meyer et al., 2011). 3' Tag-Seq and similar RNA-Seq methods like Lexogen's QuantSeq have been shown to identify the same biological signatures as whole transcript RNA-Seq methods (Corley et al., 2019; Lohman et al., 2016). In Corley et al. (2019), QuantSeq and Illumina's TruSeq (the gold standard of whole transcript RNA-Seq) were compared using peripheral blood mononuclear cells. After using two different softwares, Salmon and Tophat2, for gene quantification and normalization, the different RNA-Seq methods were shown to have high correlation using Spearman's rank correlation, demonstrating that 3' Tag-Seq can be used to check for gene expression consistency with whole transcript RNA-Seq.

Single-cell RNA-Seq (scRNA-Seq) is a powerful emerging tool that enables cell type clustering and identification of heterogeneous samples (Kolodziejczyk et al., 2015). Differences in gene expression between individual cells, especially during development, can have a profound effect on function. With the importance of cell type specific expression for neurodevelopment in mind, we generated a macaque scRNA-Seq dataset to assess the cellular heterogeneity of the dissected brain regions.

Doublets are technical artifacts of scRNA-Seq caused when a droplet is filled with two or more cells and changes in accordance with input cell concentration which follows Poisson statistics (Bloom, 2018). Generally, fewer cells are sequenced in order to minimize the formation of doublets since they are known to affect scRNA-seq data analysis by overcounting gene expression (Ilicic et al., 2016; Stegle et al., 2015). To ensure that our measurement of cellular heterogeneity is as accurate as possible, we will remove doublets from our data before clustering and identifying cell types.

Autism spectrum development (ASD)

Leveraging conservation shown across primate lineages, we aim to understand the genome organization impacting regulation of important brain development genes in macaque in order to better understand it in humans. ASD is a neurodevelopmental disability with an early age of onset that describes individuals with specific combinations of social communication impairments, repetitive behaviors, and highly restricted interests. Previous studies so far have determined a myriad of genetic causes of risk, but with little practical benefits to date. Though worldwide prevalence of ASD is ~1%, an increase in ASD diagnoses has occurred over the last few decades, especially in high income countries (Lord et al., 2006).

Recently, there have been significant sequencing efforts in children with idiopathic ASD, a neurocognitive disorder where genetic mutations are believed to contribute to about 50% of cases (Colvert et al., 2015; Sandin et al., 2014). The Simons Foundation Autism Research Initiative (SFARI) has compiled much of this research into an evolving database called SFARI Gene in an effort to help researchers track genetic risk factors for ASD (Abrahams et al., 2013). The large volume of ASD databases and data complexity has made evaluation of genetic risk for ASD difficult. However, a more recent study from the Autism Sequencing Consortium (ASC) conducted the largest exome sequencing study of ASD to date (n = 35,584 total samples, 11,986 with ASD) and identified 102 high confidence risk genes at a false discovery rate of 0.1 or less (Satterstrom et al., 2020). They found that most of these high confidence genes are expressed and enriched early in excitatory and inhibitory neuronal lineages and mostly affect synapses or regulate other genes.

While many genes have been implicated in ASD, the majority of patients lack obvious protein-coding disease-causing variants despite having thousands of ASD genomes sequenced. This suggests that noncoding regulatory elements may have a significant impact on ASD. In a recent mutational assessment of ASD, only 11.2% of 2,620 cases carrying gene-impacting variants were determined to have a molecular basis of disease (C Yuen et al., 2017). Another whole-genome sequencing study of 53 ASD impacted families showed significant enrichment of *de novo* and rare mutations upstream of genes previously associated with ASD (Turner et al., 2016).

Comprehensively characterizing the 3D genome organization of neural tissue from macaque will have significant implications on our understanding of brain development and neurodevelopmental disorders in macaque as well as humans by identifying DNA physical

contacts between putative enhancers and gene promoters. It will provide an invaluable resource to the research community as it will help us characterize the unknown functions of previously and newly discovered noncoding regulatory elements. Studying the relationship between chromatin interactions and noncoding variants can help us better understand how noncoding variants affect genes. Understanding the role genome organization plays in brain development and neurodevelopmental disorders early in life may lead to further studies in adolescents and seniors.

Methods

Tissue collection and dissection

Tissue collection and dissections were carried out by Dr. Megan Dennis in collaboration with Dr. Alice Tarantal at the California National Primate Research Center (CNPRC). Rhesus macaque (Macaca mulatta) brain samples were collected postmortem from 5 individuals. Samples for 3 individuals were collected at 60 days gestation and samples for 2 individuals were collected at 150 days gestation. Brain samples were dissected into 5-6 distinct brain regions for the 60 days gestation individuals and 15 distinct brain regions for the 150 days gestation individuals. After dissection, 1 mL of tissue was used for papain which provides the best dissociation and viability of neurons. Tissue was incubated at 37°C for 30 min with an EBSS, papain, and DNase solution. Papain was deactivated using FBS/DNase solution and Hibernate E complete media followed by trituration of the mixture with a 10 mL serological pipette. Following neuron isolation, cells were mixed with Trypan blue and counted with Countess II. For PLAC-seq experiments, cells were then fixed in a 1% final concentration of formaldehyde, quenched with 125 mM glycine and washed with 1x PBS. At this point, cells were flash frozen

in liquid nitrogen and stored at -80°C until. For RNA-seq experiments, cells were resuspended in RNALater ready following manufacturer's protocol (Qiagen) and stored at 4ºC for one week followed by -80ºC for later RNA extraction. For scRNA-seq experiments, cells were fixed in methanol following 10X Genomics recommended protocol.

3' Tag-Seq

Whole RNA was extracted from cells stored in RNALater using the Qiagen RNeasy extraction kit following the manufacturer's protocol. RNA integrity/quality was assayed using a Bioanalyzer and subsequent library was prepared and sequenced by the UC Davis Genome Center DNA Technologies Core. Barcoded 3' Tag-Seq libraries were prepared using the QuantSeq FWD kit (Lexogen, Vienna, Austria) for multiplexed sequencing according to the recommendations of the manufacturer. The fragment size distribution of the libraries was verified via micro-capillary gel electrophoresis on a Bioanalyzer 2100 (Agilent, Santa Clara, CA). The libraries were quantified by fluorometry on a Qubit instrument (LifeTechnologies, Carlsbad, CA), and pooled in equimolar ratios. Libraries were sequenced on one lane of a NextSeq500 sequencer (Illumina, San Diego, CA) with single-end 85 bp reads. The sequencing generated more than 4 million reads per library.

Raw reads data were downloaded from the PsychENCODE consortium to check for consistency with previously published data. For both the 3' Tag-Seq and PsychENCODE datasets, raw reads were trimmed and mapped to the rhesus macaque genome (Macaca mulatta, Feb. 2019 rheMac10) using two bioinformatics pipelines. For the first pipeline, we used Salmon to trim, map, and count reads (Patro et al., 2017). We then used tximport to normalize counts (Soneson et al., 2015). The second pipeline utilized Trim Galore to trim the raw reads, STAR to

map the reads, HTSeq to count reads, and edgeR to normalize counts (*Babraham Bioinformatics - Trim Galore!*, n.d.; Dobin et al., 2013; Putri et al., 2022; M. D. Robinson et al., 2010).

To measure consistency between our 3' Tag-Seq data and PsychENCODE whole RNA-Seq data, we calculated the Spearman's rank correlation coefficient of the expression levels for all genes with expression greater than zero. Biological replicates were combined by taking the average expression levels. For expression levels mapped from the Salmon pipeline, Spearman's rank correlation coefficients of the transcripts per million (TPMs) were calculated, and for expression levels mapped from the Trim Galore, STAR, HTSeq, edgeR pipeline, Spearman's rank correlation coefficients of the counts per million reads mapped (CPMs) were calculated. We then performed a differential expression (DE) analysis. We compared only the prefrontal cortex and cerebellum since those were the brain regions we focused on for PLAC-Seq.

scRNA-Seq

Cells from six dissected macaque brain region samples from one of the 60 days gestation individuals (189-0518) were fixed in methanol. Macaque methanol-fixed brain samples were submitted to the DNA technologies Core at the UC Davis Genome Center for library preparation using 10X Genomics Chromium with a targeted cell count of 10,000. Libraries were sequenced on a shared NovaSeq S4000 lane at UC Berkeley in paired-end mode at 150 bp read length.

Raw reads were counted with 10x Genomics Cell Ranger 5.0.1 using *cellranger count* and an expected cell count of 10,000 cells (Zheng et al., 2017). We used Seurat 4.1.0 to cluster and identify cell types following a standard workflow (Hao et al., 2021). Counts output from Cell Ranger was read in as a Seurat object and normalized with the global-scaling normalization

method “LogNormalize” and *scale.factor = 10000* followed by scaling the data. Doublets were removed using DoubletFinder which was integrated with the Seurat processing (McGinnis et al., 2019). Expected doublet rate was based on the estimated number of cells in each sample. Then, we performed linear dimensional reduction on the data and determined the dimensionality of the dataset. A graph-based clustering approach was used to cluster the cells, beginning with a K-nearest neighbor (KNN) graph and then using a non-linear dimensional reduction technique called Uniform Manifold Approximation and Projection (UMAP) to visualize the clusters. Cell type identification of the clusters was determined by comparing the top differentially expressed genes in each cluster to databases of previously determined gene markers. These databases included CellMarker, PanglaoDB, and CellKb (Franzén et al., 2019; Patil & Patil, n.d.; X. Zhang et al., 2019).

PLAC-Seq

PLAC-Seq protocol trials were initially carried out using the protocol from Fang et al. (2016) with lymphoblastoid cell lines (LCLs) from GM12878 crosslinked with 1% formaldehyde. Ultimately, we switched to the Arima-HiC+ kit (P/N A101020) to complete PLAC-Seq trials and generate PLAC-Seq data in accordance with the manufacturer’s protocols. Trials for the Arima-HiC+ kit were conducted using LCLs. Two technical replicates of LCLs were crosslinked with 2% formaldehyde at room temperature for 10 minutes in accordance with Arima’s protocol. After crosslinking, approximately 5 million cells were used as input for each replicate for PLAC-Seq.

The previously described dissected macaque brain samples were prepared via cell extraction for PLAC-Seq after brain region isolation. We note that macaque cells were

crosslinked using 1% formaldehyde at room temperature for 10 minutes following original guidance from Fang et al. (2016). PLAC-Seq trials were performed with two technical replicates of ~3 million cells from the parietal lobe (left hemisphere) of one of the 150 days gestation individuals (178-0503). PLAC-Seq data was generated for the right hemisphere prefrontal cortex (PFC) and cerebellum (CBC) of three biological replicates (Individuals 189-0247, 189-0249, and 189-0518). Each of the three prefrontal cortex biological replicates had a cell input of ~5 million cells. CBC1 had a cell input of ~5 million cells, CBC2 had a cell input of ~2.5 million cells and CBC3 had a cell input of ~3.3 million cells.

All LCL and macaque cells were stored at -80°C and thawed on ice until ready to begin PLAC-Seq. Crosslinked cells were then digested using Arima's restriction enzyme cocktail. Overhang ends were filled in with biotin-tagged nucleotide followed by proximity ligation of the ends to capture the chromatin structure with chimeric reads. The ligated DNA was then sheared by sonication to approximately 500 bp, ranging from 400-1000 bp. Sonication was carried out using a Covaris E220 with the following parameters: *setpoint temperature: 4°C; min/max temperature: 3-6°C; peak incident power: 75 W; duty factor 10%; cycles per burst: 200; treatment time: 600 sec*. Shearing was followed by binding of the antibody H3K4me3 (Millipore Cat #04-745), which has been previously validated by Arima and is highly enriched at active promoters. The proximally-ligated DNA was then immunoprecipitated with Dynabeads Protein A and purified with KAPA Pure Beads. The proximity ligated DNA was enriched for biotin-labeled fragments before library preparation with a custom Arima end repair and adapter ligation protocol that utilizes the Swift Biosciences Accel-NGS 2S Plus DNA Library Kit with indices from the Swift Biosciences Indexing Kit. Libraries were amplified for 12-16 PCR cycles using the KAPA Library Amplification Kit and purified with KAPA Pure Beads.

For the six macaque samples, barcoded libraries were pooled and sent to Novogene for partial lane sequencing with Illumina NovaSeq 6000. Libraries were sequenced in paired-end mode at 150 bp read length. A total of 386.7 Gb of raw data was sequenced for all six samples which produced over 200 million paired-end reads per sample. To analyze the PLAC-Seq data, two bioinformatics pipelines were considered to call loops, or significant interactions. Model-based Analysis of PLAC-seq and HiChIP (MAPS) uses Burrows-Wheeler Aligner, whichArima recommends for their restriction enzymes (Juric et al., 2019; H. Li & Durbin, 2009). HiC-Pro uses the Bowtie2 software for alignment and was not recommended for Arima restriction enzymes because mapping parameters were too strict (Langmead & Salzberg, 2012; Servant et al., 2015). Valid read pairs output from HiC-Pro is then used as input for FitHiChIP to call significant interactions (Bhattacharyya et al., 2019).

For the LCL trials, we utilized MAPS Arima Version 2.0 and HiC-Pro version 2.11.0 with FitHiChIP version 8.0 with reference ChIP-seq peaks from ensemble and aligned to human reference genome, hg19. For each individual macaque replicate, we used the same pipelines with inferred ChIP-seq peaks and aligned to the macaque reference genome, rheMac10. In order to use rheMac10 with MAPS, we generated the genomic features files using the the *Snakefile_multienzyme* file from MAPS and a bigwig mappability file that was generated from *gem-mappability* from the GEM library (*GEM Library - Browse /gem-Library at SourceForge.net*, n.d.). While Arima recommends defining protein localization peaks with ChIP-Seq from the same cells using the same antibody as the Arima-HiC+ experiments, we did not collect enough cells from macaque cerebellum for additional experiments. Therefore, we used inferred peaks generated from the macaque PLAC-Seq data using the Model-based Analysis for ChIP-Seq (MACS2) software and *PeakInferHiChIP.sh* script from FitHiChIP (Feng

et al., 2012; Y. Zhang et al., 2008). All samples analyzed with MAPS used the following parameters: *bin_size = 5000; binning_range=2000000; fdr = 2; generate_hic = 1; mapq = 30; length_cutoff = 1000; threads = 16*. All samples analyzed with the HiC-Pro used the following parameters: *FORMAT = phred33; MIN_MAPQ = 0; BOWTIE2_GLOBAL_OPTIONS = --very-sensitive -L 30 --score-min L,-0.6,-0.2 --end-to-end --reorder; BOWTIE2_LOCAL_OPTIONS = --very-sensitive -L 20 --score-min L,-0.6,-0.2 --end-to-end --reorder; BIN_SIZE = 20000 40000 150000 500000 1000000; MATRIX_FORMAT = upper; MAX_ITER = 100; FILTER_LOW_COUNT_PERC = 0.02; FILTER_HIGH_COUNT_PERC = 0; EPS = 0.1*. All samples analyzed with the FitHiChIP used the following parameters: *GCSize = 200; MappSize = 500; IntType = 3; BINSIZE = 5000; LowDistThr = 20000; UppDistThr = 2000000; QVALUE = 0.01; NBins = 200*. Hi-C matrix was produced from the HiC-Pro output using the *hicpro2juicebox.sh* utility file from HiC-Pro and visualized with Juicebox (J. T. Robinson et al., 2018).

We used HiCRep 1.11.0 to validate correlation of Hi-C interaction matrices of three biological replicates each for cerebellum and prefrontal cortex with stratum adjusted correlation coefficient (SCC) (Yang et al., 2017). SCC was measured at three different resolutions. For the 1000 kb resolution, SCC was measured with and without read depth correction. Data from biological replicates of each region were then merged to improve statistical power to call significant interactions with FitHiChIP.

As a quality control check, we used the *bedtools intersect* command from bedtools 2.29 to identify the top differentially expressed genes in prefrontal cortex and cerebellum that intersect with loop anchors from loops we called from the PLAC-Seq data (Quinlan & Hall, 2010). We then used it to identify high confidence genes implicated in ASD from the SFARI

Gene database and the Autism Sequencing Consortium (ASC) study that intersect with loop anchors from significant interactions we called from our PLAC-Seq data (Abrahams et al., 2013; Satterstrom et al., 2020). Finally, we used *bedtools intersect* to identify which de novo variants of the 255,106 identified from 1,902 quartet families from the Simons Simplex Collection intersect with significant interactions we called from our PLAC-Seq data (An et al., 2018; Fischbach & Lord, 2010). Rhesus prefrontal cortex and cerebellum significant interaction bed files were lifted over from rheMac10 to hg38 for analysis with high confidence genes implicated in ASD and de novo variants (Hinrichs et al., 2006). Loops were visualized using the WashU Epigenome Browser (D. Li et al., 2022).

Results

Assessing consistency of RNA-seq datasets

To determine consistency within our own generated expression data (3' Tag Seq) as well as with existing published results from the PsychENCODE Consortium (PEC) (PsychENCODE Consortium et al., 2015), we correlated RNA-seq datasets of biological replicates derived from various brain regions, developmental time points, and studies using two different analysis approaches. Figure 1 shows Spearman's rank correlations for all PEC and 3' Tag-Seq samples. 3' Tag-Seq samples include timepoints at 60 days gestation (T60) and 150 days gestation (T150). Correlations were calculated from TPMs counted and normalized from our 3' Tag-Seq samples and downloaded PEC raw reads by Salmon (Patro et al., 2017). All samples correlate well between replicates from the same study, brain region, and timepoints. Samples from the PEC dataset have a correlation >0.94 and T60 samples have a correlation >0.89. Samples for the T150

dataset have lower correlations with each other, ranging from 0.75-0.86. The two cerebellum regions at T150 correlate the most in that set, and seem to separate from the other samples.

We replicated our correlation analysis across the same datasets using an alternative RNA-seq analysis pipeline (utilizing Trim Galore, STAR, HTSeq, and edgeR) by calculating CPMs (*Babraham Bioinformatics - Trim Galore!*, n.d.; Dobin et al., 2013; Putri et al., 2022; M. D. Robinson et al., 2010). The overall trends were similar to correlations calculated from the Salmon pipeline. However, all correlations were higher with the CPMs from edgeR. This increase was also seen when we only compared the PEC and T60 datasets, as shown in Figure 3. When comparing PEC samples to T60 samples with TPMs from Salmon, correlations ranged from 0.62-0.69 whereas those same correlations with CPMs from edgeR ranged from 0.75-0.79. Since mapping rates were higher and Spearman's rank correlations were higher with normalized counts from HTSeq and edgeR, we decided to use the counts from HTSeq to carry out a differential expression analysis.

Overall, since we found that correlations between our dataset and PEC for the same brain regions and time points ranged from 0.75-0.79, we determined this correlation was insufficient to justify combining the two for a larger meta-analysis. Further, we observed more consistent results across biological replicates in cerebellum. When comparing PEC biological replicates and T60 biological replicates with CPMs from edgeR, correlations within PEC biological replicates ranged from 0.89-0.96, correlations within T60 biological replicates ranged from 0.85-0.90, whereas correlations across datasets were much lower ranging from 0.51-0.59. As such, moving forward we only used our 3'Tag seq data generated for T60 brain regions.

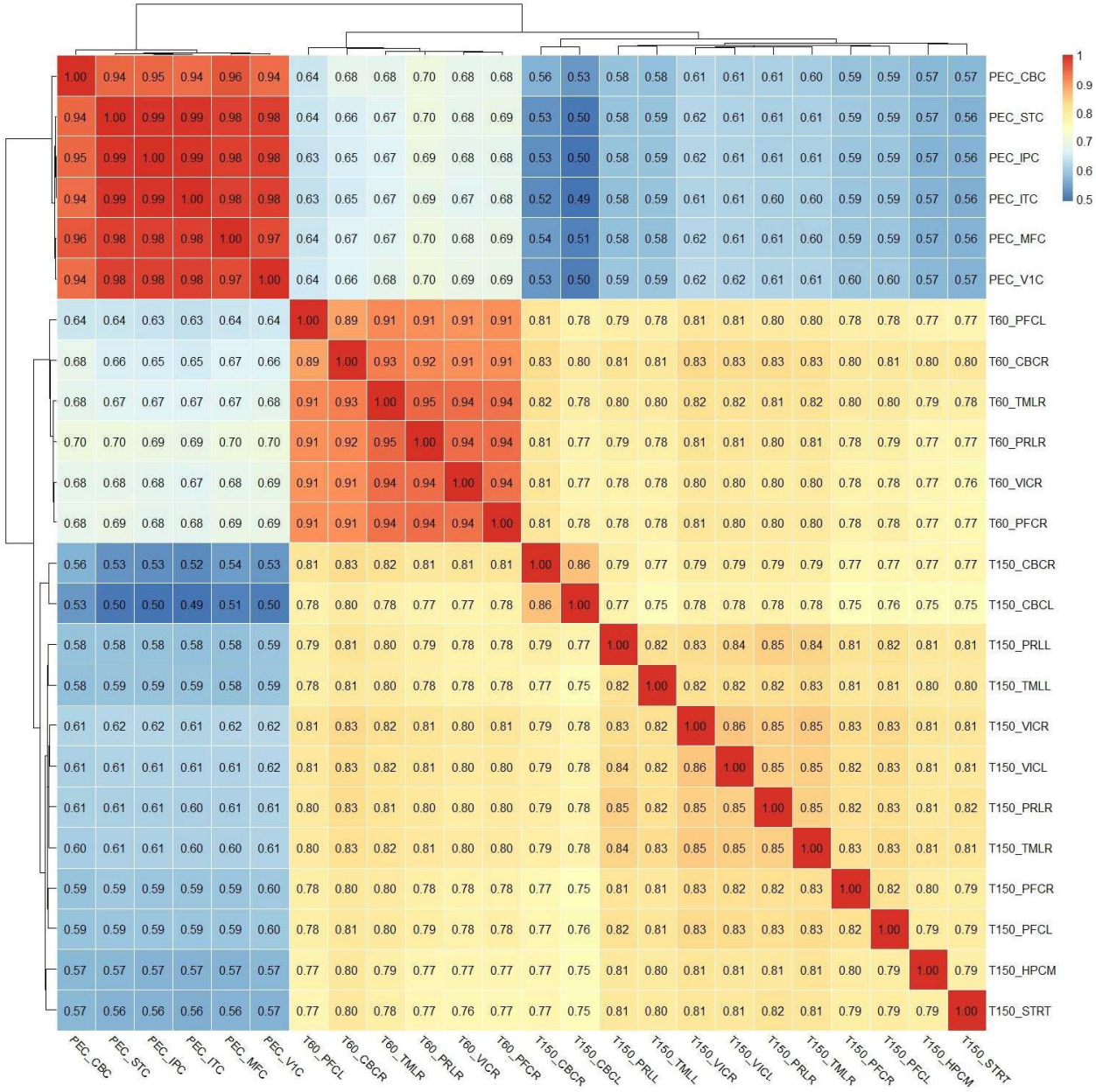

Figure 1. RNA-Seq Spearman's rank correlation for PsychENCODE (60 days gestation) and 3' Tag-Seq (60 and 150 days gestation) expression using TPM output from Salmon. PsychENCODE (PEC) brain regions: MFC – medial prefrontal cortex; IPC – inferior posterior parietal cortex; STC – superior temporal cortex; ITC – inferior temporal cortex; V1C – primary visual cortex; CBC – cerebellar cortex. 3' Tag-Seq 60 days gestation (T60) brain regions: prefrontal cortex left – PFCL; prefrontal cortex right – PFCR; cerebellum right – CBCR; temporal lobe right – TMLR; parietal lobe right – PRLR; visual cortex right – VICR. 3' Tag-Seq 150 days gestation (T150) brain regions: prefrontal cortex left – PFCL; prefrontal cortex right – PFCR; cerebellum left – CBCL; cerebellum right – CBCR; temporal lobe left – TMLL; temporal lobe right – TMLR; parietal lobe left – PRLL; parietal lobe right – PRLR; visual cortex left – VICL; visual cortex right – VICR; hippocampus – HPCM; striatum – STRT.

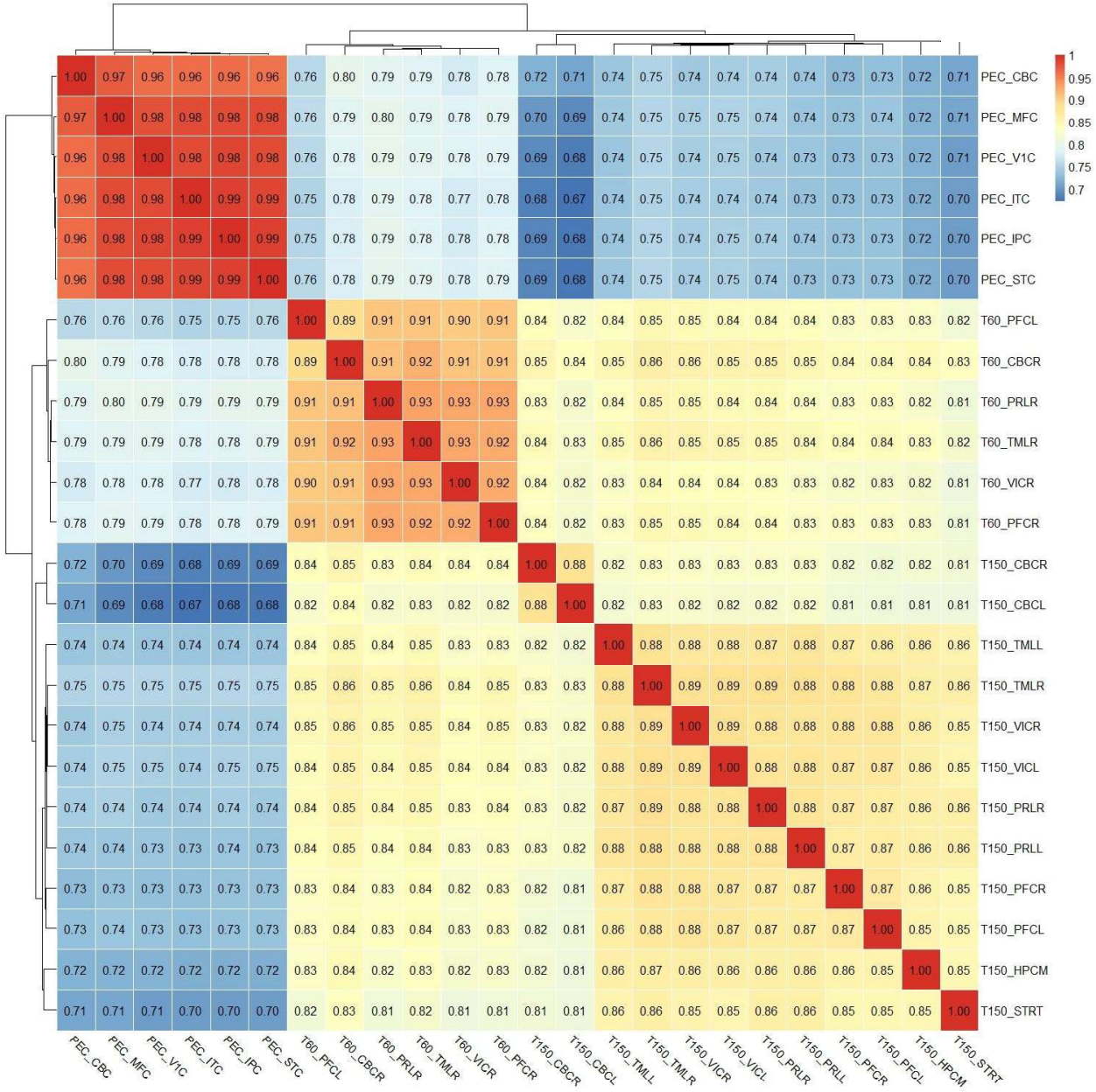

Figure 2. Spearman's rank correlation for PEC (60 days gestation) and 3' Tag-Seq (60 and 150 days gestation) expression of CPM output from edgeR. Reads were aligned with STAR and counted with HTSeq. Abbreviations for brain regions are defined in Figure 1.

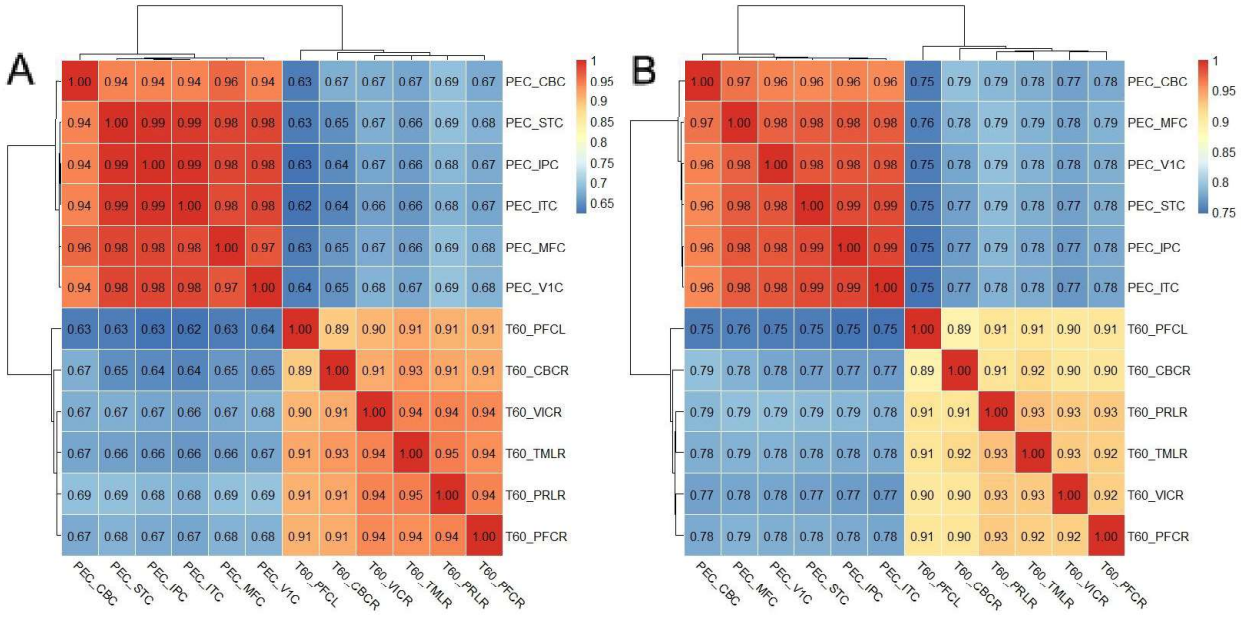

Figure 3. Spearman's rank correlation for PEC (60 days gestation) and 3' Tag-Seq (60 days gestation) expression. (A) Correlation of TPM output using Salmon. **(B)** Correlation of CPM output using edgeR. Reads were aligned with STAR and counted with HTSeq.

Differential expression (DE) analysis across prefrontal cortex and cerebellum

Based on the consistency of our dataset at the earlier developmental timepoint, counts from HTSeq were used to perform a DE analysis for the prefrontal cortex and cerebellum at 60 days gestation. We first looked at a multidimensional scaling (MDS) plot of all the biological replicates to determine the relationships between them by seeing how well they cluster together. As shown in Figure 4, the prefrontal cortex biological replicates cluster together along both dimensions, though PFC_60_3 separates slightly along dimension 2. The prefrontal cortex and cerebellum samples are separated well from each other along dimension 1. However, CBC_60_3 also has a large separation from the first two cerebellum samples along dimension 2. This result could be expected due to batch effects because the first two replicates were prepared together while the third replicates were prepared on a different day. However, since the separation

between prefrontal cortex and cerebellum samples was so large, we continued with grouping by brain region for the DE analysis.

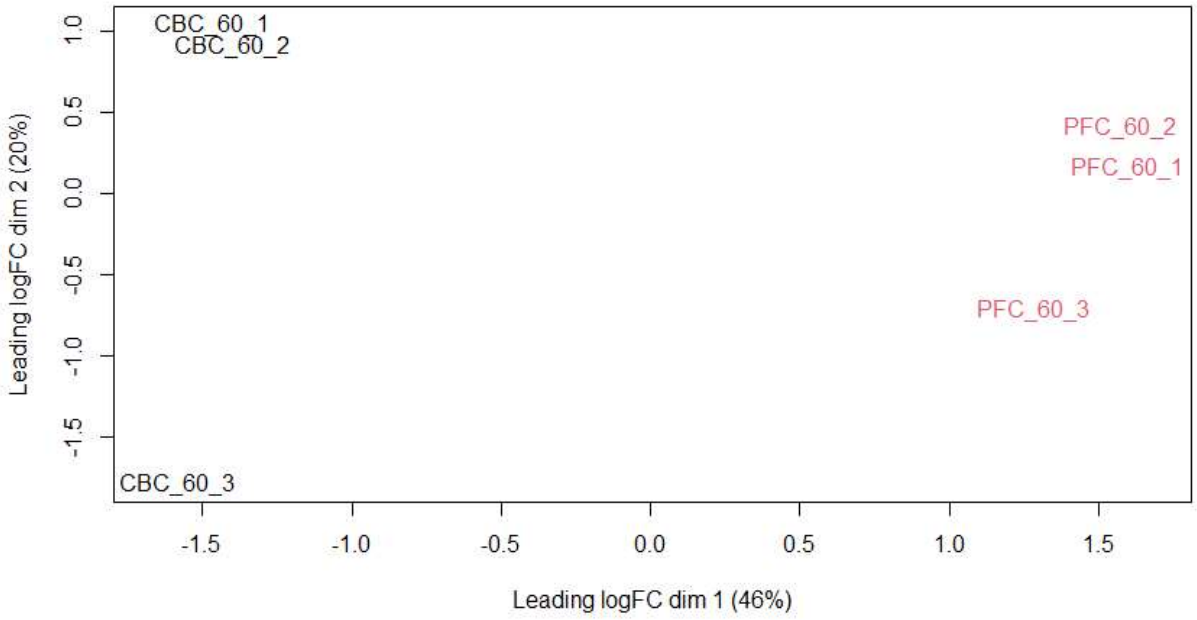

Figure 4. MDS plot of prefrontal cortex and cerebellum biological replicates. Cerebellum samples are separated from the prefrontal cortex samples along dimension 1. CBC_60_3 also separates from the other two cerebellum biological replicates along dimension 2.

We then performed a DE analysis to determine the most differentially expressed genes between prefrontal cortex and cerebellum. Figure 5 shows a heatmap of the z-scores for the top 40 differentially expressed genes with the color scheme scaled by row. As expected, genes highly expressed in the prefrontal cortex had low expression in cerebellum and vice versa. Additionally, z-scores for CBC_60_3 seemed to follow the same pattern as the other cerebellum samples despite not clustering well with them in the MDS plot. We then used these top differentially expressed genes for downstream analysis to see if there are any brain region specific differences in significant interactions identified by PLAC-Seq.

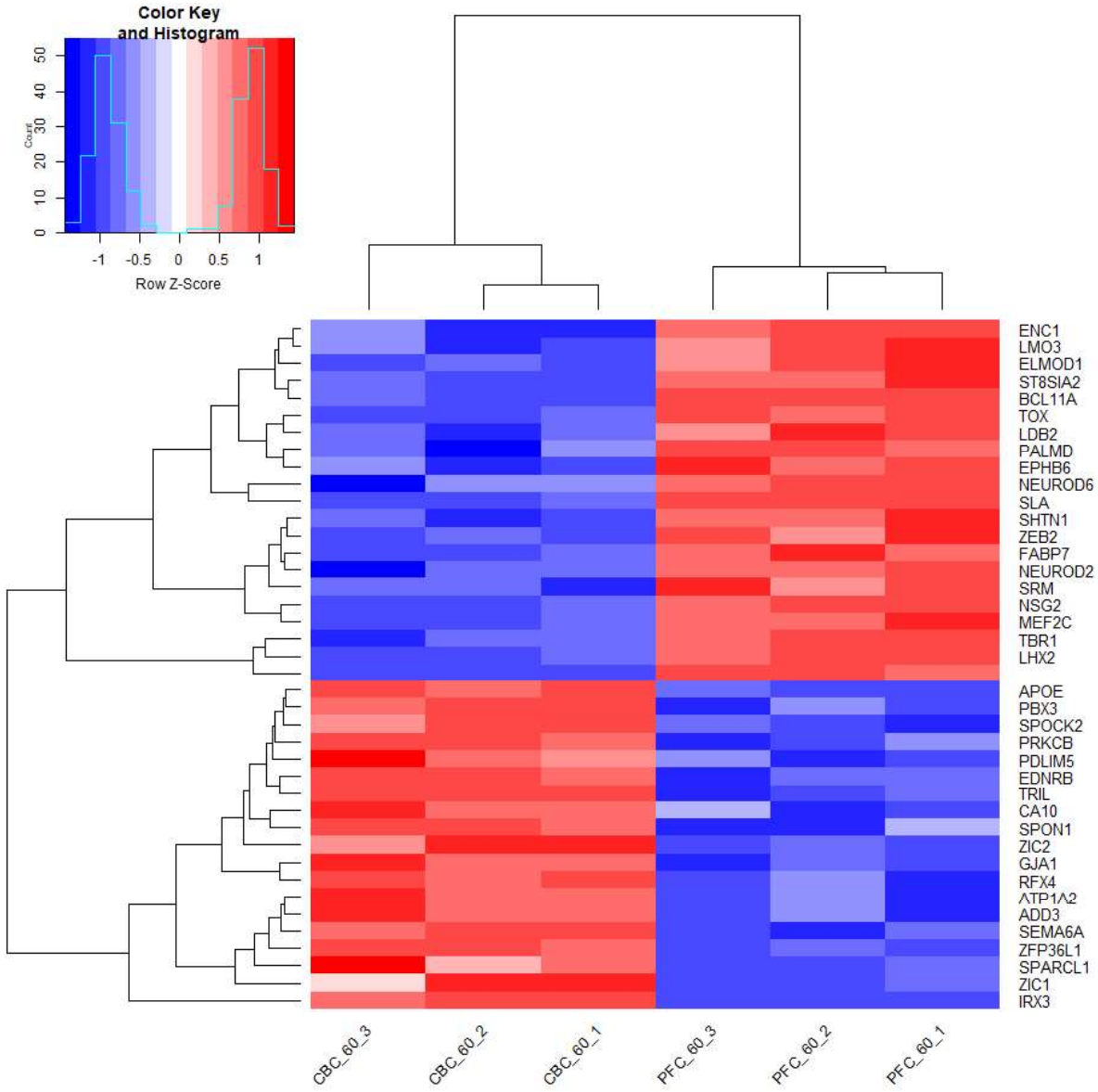

Figure 5. Heatmap of top 40 most differentially expressed genes when comparing prefrontal cortex and cerebellum. Z-scores and color scheme scaled by row.

Assessment of cell types within brain samples

With our increasing understanding of the importance of cell type specific expression for neurodevelopment, we set out to assess the cellular heterogeneity of the dissected brain regions by generating a macaque scRNA-Seq dataset. Before clustering and identifying cell types, we removed doublets using DoubletFinder in order to minimize its effect when measuring cellular

heterogeneity (McGinnis et al., 2019). The number of doublets removed was based on the number of cells counted by Cell Ranger. The number of expected doublets increases as the number of cells increases. Cerebellum had only 5,816 cells so we removed the least number of doublets from that brain region with only a 4.6% expected doublet rate. All the other samples had an expected doublet percent of 6.1% or 6.9%. In Figure 6, we checked the number of genes expressed for cells identified as doublets and singlets to confirm that we removed doublets instead of singlets. As expected, cells identified as doublets tended to have a higher number of genes expressed than cells identified as singlets.

Table 1. Doublets removed using DoubletFinder. Expected doublet percent is based on the number of cells in each sample.

Sample	Cell Count	Expected Doublet Percent	Number of Doublets	Cell Count With Doublets Removed
Parietal Lobe (R)	9,038	6.9%	624	8,414
Visual Cortex (R)	8,225	6.1%	502	7,723
Temporal Lobe (R)	7,952	6.1%	485	7,467
Cerebellum	5,816	4.6%	268	5,548
Prefrontal Cortex (L)	8,535	6.9%	589	7,946
Prefrontal Cortex (R)	7,820	6.1%	477	7,343

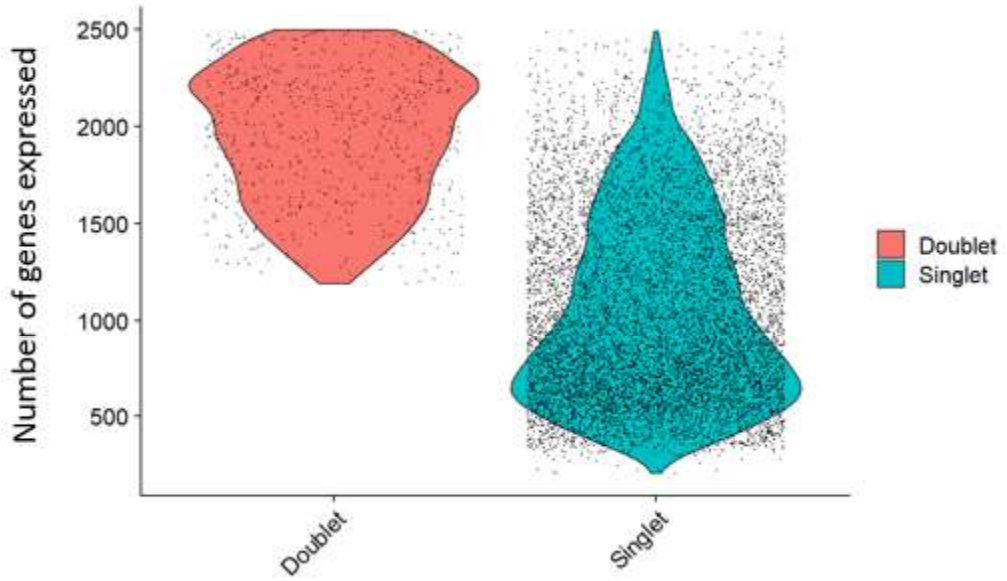

Figure 6. Number of genes expressed of doublets and singlets for parietal lobe (R). Cells identified as doublets show a higher number of genes expressed than singlets.

Figure 7 shows UMAP plots for six isolated macaque brain regions. All six brain regions show cellular heterogeneity. Despite having the least number of cells with only 5,548, the cerebellum has the most identified cell types with 16 clusters. The parietal lobe has 12 clusters while the other four brain regions have 11 clusters. All of the samples mostly comprised a variety of glia or neuronal cells. Clusters were assigned cell types by comparing previously determined gene markers from the CellMarker, PanglaoDB, and CellKb databases (Franzén et al., 2019; Patil & Patil, n.d.; X. Zhang et al., 2019). Of note, many of the top DE genes from our DE analysis represent gene markers that were used to identify cell types in prefrontal cortex and cerebellum (Supplementary Table S1). *NEUROD2* was found to be upregulated in the prefrontal cortex and was a gene marker for newborn excitatory neurons. Similarly, *SPARCL1*, upregulated in cerebellum, was a gene marker for astrocytes.

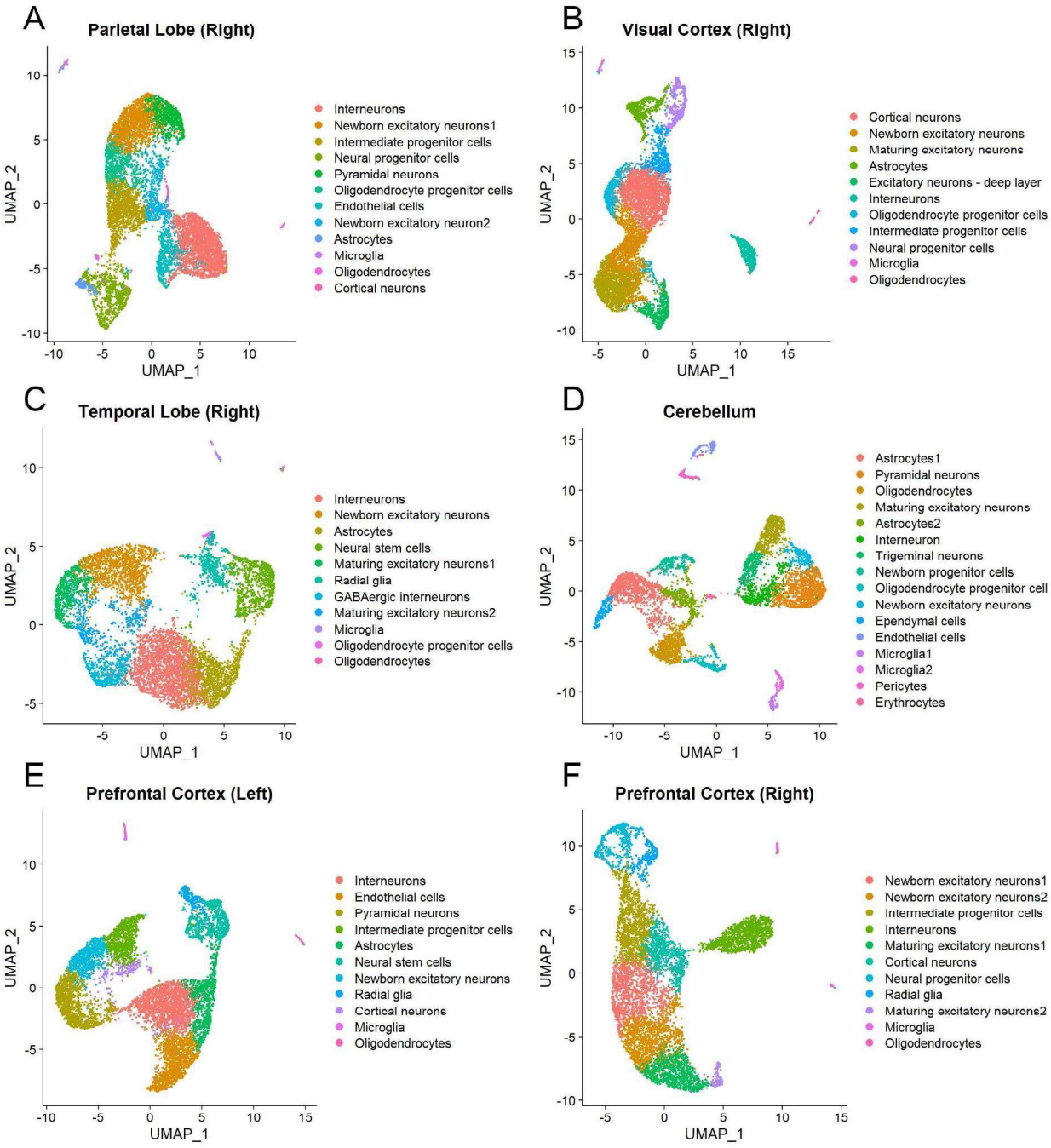

Figure 7. UMAP plot of gene expression relationships for macaque brain regions collected at 60 days gestation. (A) Parietal lobe of the right hemisphere (n = 1; 8,414 cells). Cells are colored by cluster identity into 12 clusters. **(B)** Visual cortex of the right hemisphere (n = 1; 7,723 cells). Cells are colored by cluster identity into 11 clusters. **(C)** Temporal lobe of the right hemisphere (n = 1; 7,467 cells). Cells are colored by cluster identity into 11 clusters. **(D)** Cerebellum (n = 1; 5,548 cells). Cells are colored by cluster identity into 16 clusters. **(E)** Prefrontal cortex of the left hemisphere (n = 1; 7,946 cells). Cells are colored by cluster identity into 11 clusters. **(F)** Prefrontal cortex of the right hemisphere (n = 1; 7,343 cells). Cells are colored by cluster identity into 11 clusters.

To better quantify the cellular heterogeneity of our samples, we looked at cell class abundance. Cell types were classified into five broad classes: blood cells, endothelial cells, glia, neural progenitor cells (NPCs), or neurons. The cells for scRNA-seq were processed with papain dissociation to specifically enrich neuronal cells, so most of the cells identified were neuronal. Only the cerebellum and prefrontal cortex (left) had a neuron abundance less than 0.50. If we consider NPCs and neurons together, the cerebellum is the only brain region with an abundance less than 0.50 for those two cell classes.

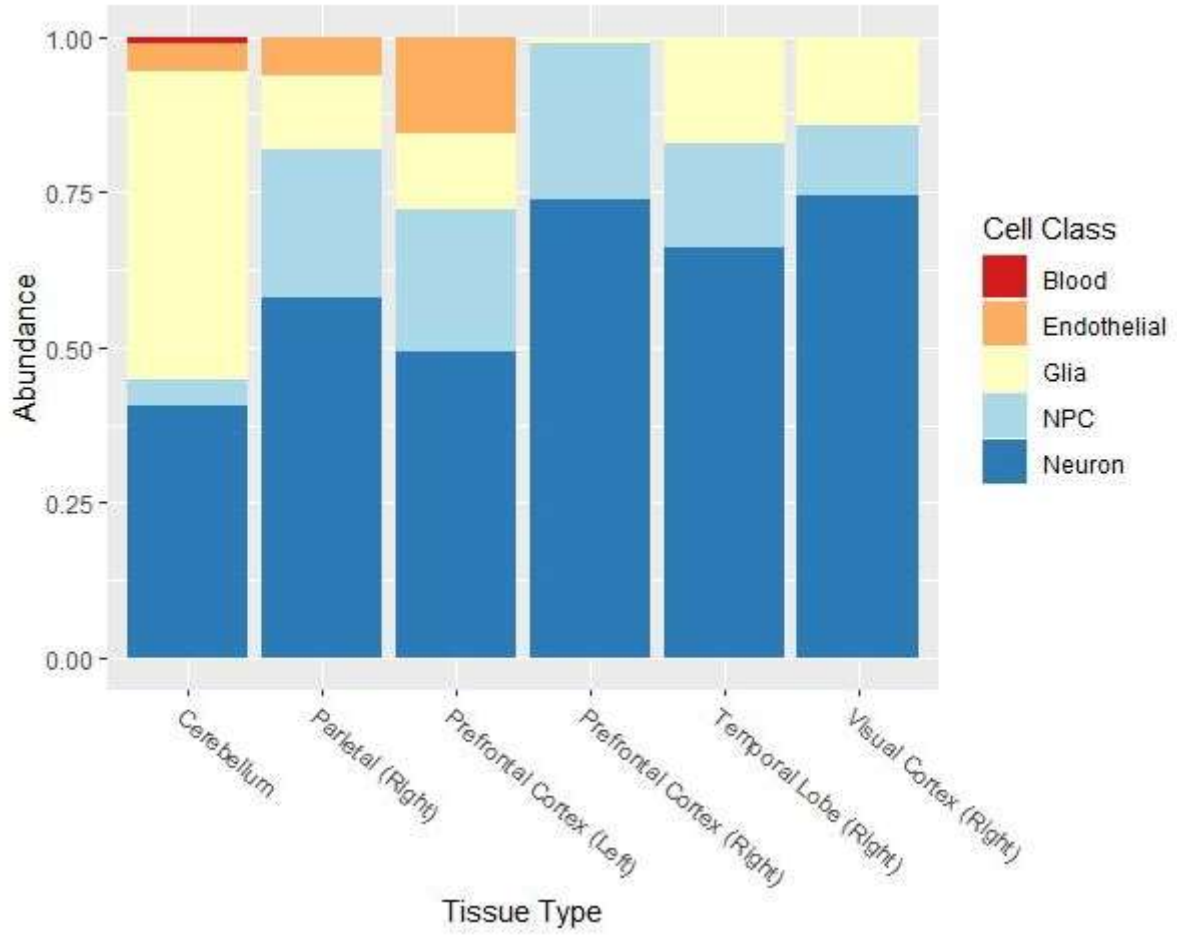

Figure 8. Cell class abundance for 6 brain regions (n=1) at 60 days gestation. Cell types identified by Seurat were categorized into 5 larger cell classes.

Assessment of genome organization in prefrontal cortex and cerebellum

In order to understand putative cis-regulatory elements contributing to gene regulation in prefrontal cortex and cerebellum, we used PLAC-Seq to generate targeted 3D genomic maps. To begin, we performed PLAC-Seq on an LCL GM12878 to verify our ability to perform the method. As a quality control check, we wanted to assess whether there were any issues with alignment of valid read pairs before calling loops, or significant interactions, by looking at the Hi-C interaction matrix. Figure 9 takes all the valid pairs for PFC1 from HiC-Pro and visualizes a Hi-C matrix with Juicebox (J. T. Robinson et al., 2018). The Hi-C matrices confirmed that there were no concerns with the alignment step and that the Hi-C portion of PLAC-Seq worked.

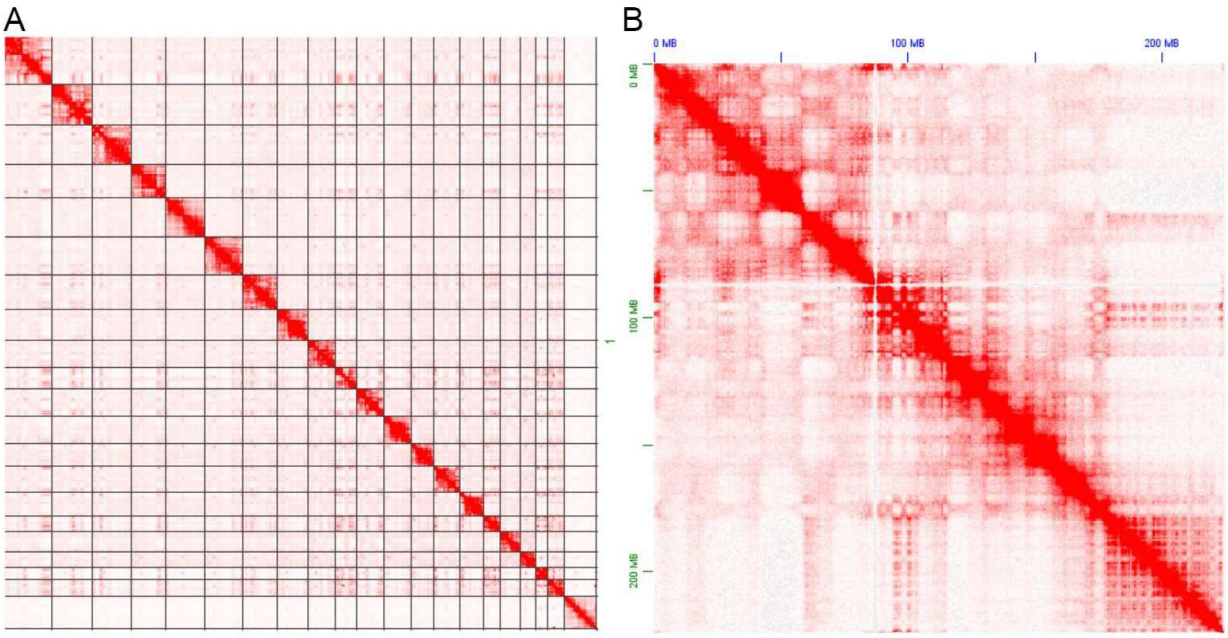

Figure 9. Hi-C matrix of macaque brain sample PFC1. (A) Hi-C matrix of all chromosomes. **(B)** Hi-C matrix of chromosome 1.

As a quality control check for the ChIP-Seq, we looked at the enrichment at inferred ChIP peaks for H3K4me3. In Figure 10, the LCLs (GM12878) has the highest enrichment at ChIP peaks with peak coverage over 4x greater than average genome-wide coverage. Of the

macaque samples, PFC3 and CBC3 had the best ChIP enrichment with greater than 3x peak coverage compared to the average genome coverage. The rest of the macaque samples had peak coverage of 2x or less, indicating very low enrichment of H3K4me3.

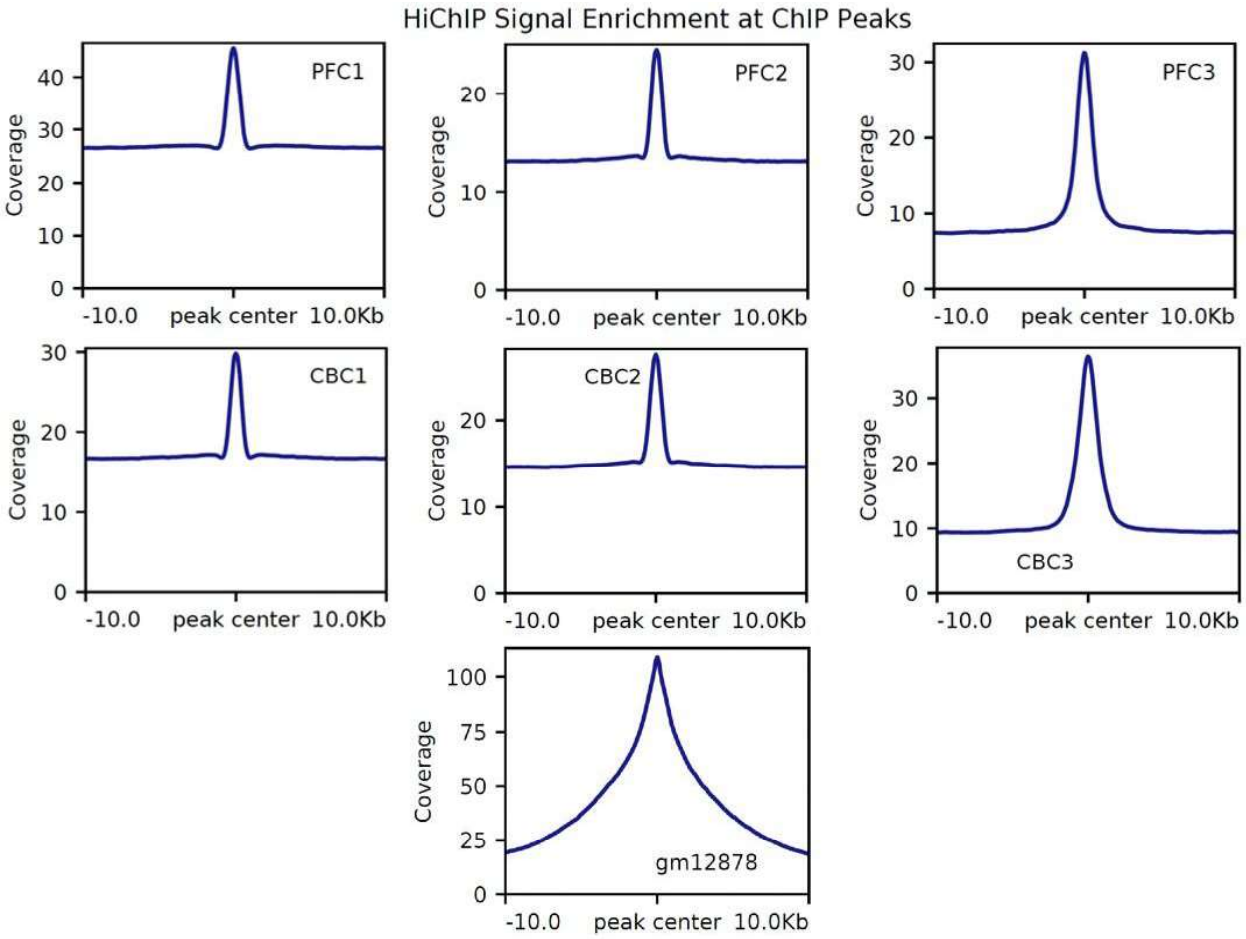

Figure 10. Enrichment at ChIP peaks from MAPS. PFC3 and CBC3 show the highest enrichment at ChIP peaks of the macaque samples.

We began processing our PLAC-Seq data with the MAPS pipeline because it uses the Burrows-Wheeler Aligner (BWA) software which is the preferred aligner for Arima's restriction enzymes. As shown in Table 2, reads from the LCL trial had a high mapping rate and MAPS called 44,132 loops from inferred peaks. However, MAPS was unable to call loops for any of the macaque samples. Although the mapping rate was lower for macaque reads compared to the

LCL reads, it was not significantly lower where we would expect it to cause an issue with the downstream analysis. This could suggest there was an issue later in the pipeline, such as with user generated genomic feature files for the rheMac10 genome. Several MAPS trials with loosened parameters such as a higher FDR and broad inferred peaks still resulted in no loops called.

We then utilized the HiC-Pro and FitHiChIP pipeline to call loops, also known as significant interactions as called by FitHiChIP. HiC-Pro uses Bowtie2 for alignment, which Arima found was more strict than BWA. However, we found that the mapping rate using HiC-Pro was higher for all of our samples. Interestingly, MAPS had identified almost 10x more loops than FitHiChIP despite having a slightly lower mapping rate. We were able to call significant interactions for all macaque samples with FitHiChIP, though still significantly less than the number of loops called by MAPS for the LCL trial despite having much greater read depth coverage.

Since we were unable to call any significant interactions with MAPS, we decided to move forward with the HiC-Pro and FitHiChIP pipeline. While we were able to call some significant interactions with this pipeline despite low enrichment, we predicted merging the biological replicates would allow us to improve the statistical power to call significant interactions as merging the data would not only increase overall read depth coverage, but enrichment at peaks as well. In order to confidently merge the biological replicates, we confirmed the consistency of the Hi-C interaction matrices across the biological replicates. Typical measurements for correlation such as the Pearson correlation coefficient or Spearman's rank correlation coefficient disregard spatial features found in Hi-C data because they only consider point interactions. To avoid these pitfalls, we used HiCRep to assess reproducibility and

consistency of the three biological replicates for cerebellum and prefrontal cortex (Yang et al., 2017). HiCRep introduced stratum adjusted correlation coefficient (SCC) which quantifies the similarity between Hi-C interaction matrices. While based on the Pearson correlation coefficient, SCC takes into account factors such as domain structure and distance dependence.

We first measured SCC at 1000 kb with no read depth correction as shown in Figure 11. SCC comparing biological replicates ranged from 0.98-0.99 for prefrontal cortex and 0.96-0.99 for cerebellum while SCC comparing brain regions showed a dip in the range from 0.94-0.98. CBC3 stood out as an outlier as it had the lowest SCC values when compared to the other cerebellum biological replicates and the prefrontal cortex. We then tested SCC with read depth correction and at higher resolutions. As expected, the SCC values dropped as we increased resolution. Overall, the trends of the SCC values across the different resolutions did not change, but the ranges of SCC widened as we increased the resolution which made it easier to identify outliers. CBC3 continued to have the lowest SCC values, demonstrating that it has the least correlation with all other biological replicates according to the Hi-C interaction matrices. This is in line with results from the MDS plot for 3' Tag-Seq expression from Figure 4 which shows that CBC3 is separated from the prefrontal cortex as well as the other two cerebellum samples.

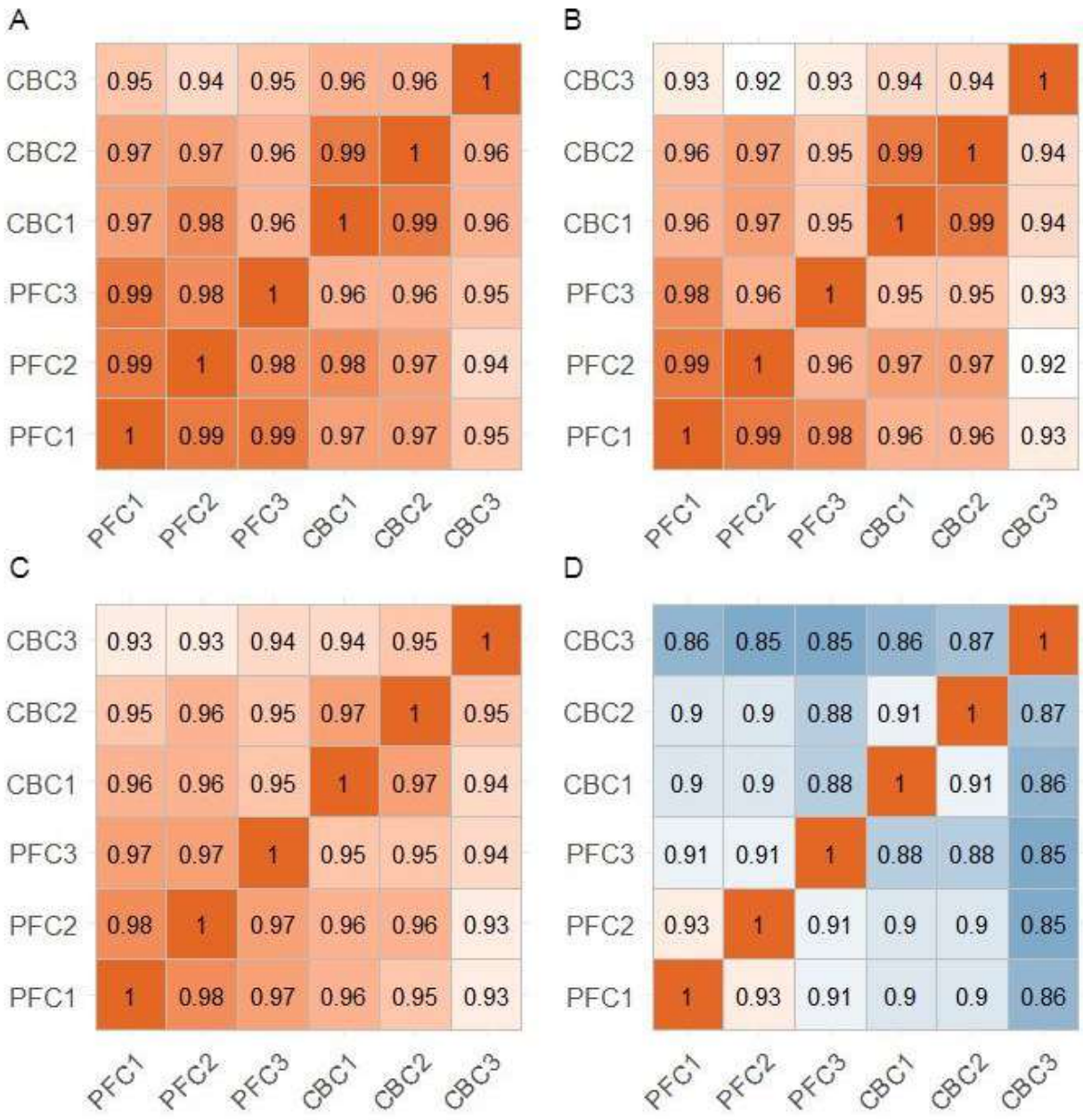

Figure 11. Stratum adjusted correlation coefficient (SCC) of cerebellum and prefrontal cortex biological replicates at different resolutions. All SCC plots include read depth correction unless noted otherwise. **(A)** 1000 kb without read depth correction. **(B)** 1000 kb. **(C)** 100 kb. **(D)** 10 kb.

As expected, after merging the HiCPro valid pairs output from the three biological replicates for both prefrontal cortex and cerebellum, we were able to call significantly more

loops for both brain regions. As shown in Table 2, for the prefrontal cortex, FitHiChIP called 32,959 significant interactions while 13,814 significant interactions were called for the cerebellum.

Table 2. Summary stats of LCL trial and macaque PLAC-Seq data using MAPS and HiC-Pro FitHiChIP pipelines. LCL reads were mapped to the hg19 reference genome. Macaque reads were mapped to the rheMac10 reference genome.

Sample	Number of paired-end reads	Mapping rate	Average genome read depth	Significant interactions
MAPS pipeline				
LCL	64,376,806	88.62%	2.73	44,132
PFC1	316,833,246	80.72%	25.82	0
PFC2	162,044,442	77.82%	12.73	0
PFC3	205,216,917	75.95%	15.74	0
CBC1	215,622,758	77.86%	16.95	0
CBC2	187,402,006	79.19%	14.98	0
CBC3	201,880,439	81.09%	16.53	0
HiC-Pro FitHiChIP Pipeline				
LCL	64,376,806	93.50%	2.88	4,665
PFC1	316,833,246	93.00%	29.75	12,807
PFC2	162,044,442	90.85%	14.86	161
PFC3	205,216,917	90.10%	18.67	5,116
CBC1	215,622,758	91.50%	19.92	973
CBC2	187,402,006	91.30%	17.27	1,930
CBC3	201,880,439	86.25%	17.58	3,087
PFC merged	1,368,189,210	92.62%	63.28	32,959
CBC merged	1,209,810,406	89.69%	54.77	13,814

Before looking into how the called significant interactions are connected to ASD, we identified significant interactions connected to the top DE genes from our 3' Tag-Seq data. While nearly none of the promoters of our DE genes had significant interactions connected to them, we did identify XX DE genes from cerebellum with significant interacting regions. In Figure 12, *PBX3* which is a top DE gene upregulated in the cerebellum, has six connected loops in the cerebellum but no connected loops in the prefrontal cortex. This pattern was also seen in *SEMA6A* and *TOX*. However, unlike *PBX3* and *SEMA6A* which are upregulated in cerebellum, *TOX* was upregulated in the prefrontal cortex.

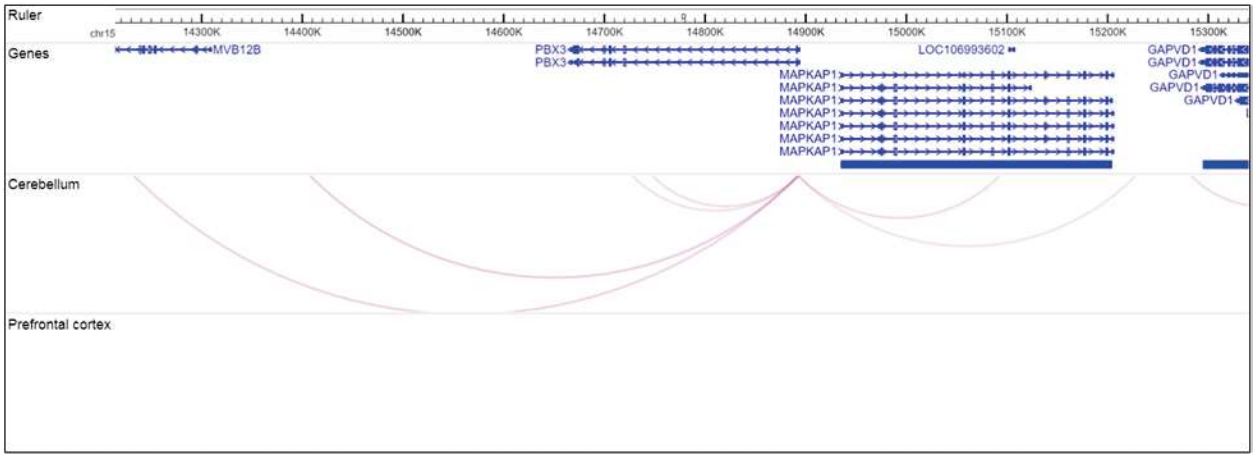

Figure 12. *PBX3* connected to loops identified in the cerebellum. No loops are connected in the prefrontal cortex.

ASD implicated genes and variants connected to significant interactions

In order to assess the impact of chromatin organization on ASD, we first looked at the significant interactions identified from our PLAC-Seq data connected to genes that have been previously implicated in ASD with high confidence. These genes were pulled from two data sources: a 102 gene dataset from the Autism Sequencing Consortium (ASC-generated WES sequencing data) (Satterstrom et al., 2020) and a 1,044 gene dataset from the SFARI Gene database (Abrahams et al., 2013). Table 3 shows that over half of the promoters of these genes

from both datasets had a connection to significant interactions in the prefrontal cortex, with 69 genes from ASC and 544 genes from SFARI. Many genes were connected to significant interactions in both the prefrontal cortex and cerebellum such as *FOXP1*, *SYNGAP1*, and *KDM5B* (Supplementary Tables S2-S5). The significant interactions identified in the cerebellum connected to fewer genes, with only 34 genes from ASC and 274 genes from SFARI. Overall, significantly more genes, such as *CHD8* and *GRIN2B*, and significant interactions were connected in the prefrontal cortex than cerebellum which was expected since nearly three times more significant interactions were identified in the prefrontal cortex.

Table 3. Number of loop anchors that intersect with genes implicated in ASD from the ASC-generated WES sequencing data and SFARI Gene database. ASC-generated WES sequencing data has 102 genes. SFARI Gene database has 1044 genes.

Brain region	**# of genes**	**# of loops connected**
ASC-generated WES sequencing data		
Prefrontal cortex	69	337
Cerebellum	34	98
SFARI Gene		
Prefrontal cortex	544	2348
Cerebellum	274	855

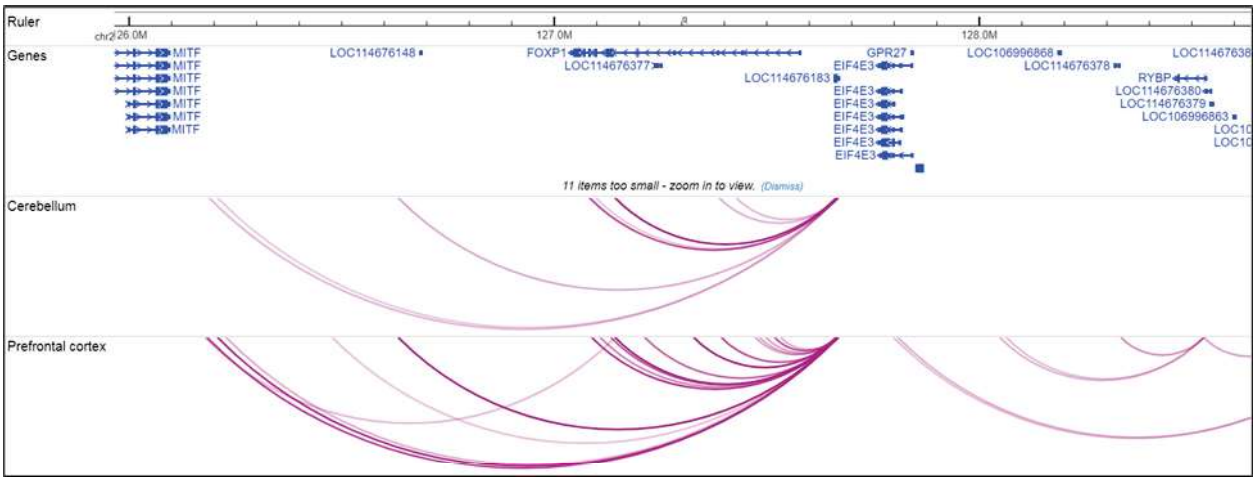

Figure 13. Loops connected to *FOXP1*. Five loops connected in cerebellum. 12 loops connected in prefrontal cortex.

We next overlaid our interacting regions from prefrontal cortex and cerebellum with identified de novo variants from affected probands and unaffected siblings sequenced as part of the Simons Simplex Collection (SSC) (An et al., 2018). From this, we identified six total unique de novo variants falling within promoters or putative cis regulatory elements of top ASD candidate genes (as defined by our PLAC-seq interactions). Two de novo variants were identified in cerebellum and four in prefrontal cortex, with one de novo variant identified in both regions. In Figure 15, comparing frequencies between probands and unaffected siblings in prefrontal cortex and cerebellum shows no association of loop connection to a proband de novo variant (p-value = 1).

Table 4. Number of loops that intersect with de novo variants identified in ASD quartet families from the SCC. 255,106 de novo variants identified in 1902 families were overlaid with 13,814 and 32,959 loops from cerebellum and prefrontal cortex, respectively.

Chr	Position	Ref	Alt	Type	Phenotype	Nearest gene	# of loops connected
Cerebellum							
chr5	151446389	G	A	SNV	proband	*SLC36A1*	5
chr8	69666932	CT	C	Indel	sibling	*SULF1*	1
Prefrontal Cortex							
chr5	151446389	G	A	SNV	proband	*SLC36A1*	7
chr8	30669624	T	C	SNV	proband	*GTF2E2*	1
chr8	69666932	CT	C	Indel	sibling	*SULF1*	1
chr12	1689584	T	C	SNV	sibling	*ADIPOR2*	2
chr16	48703313	C	CT	Indel	sibling	*AC007611.1*	1

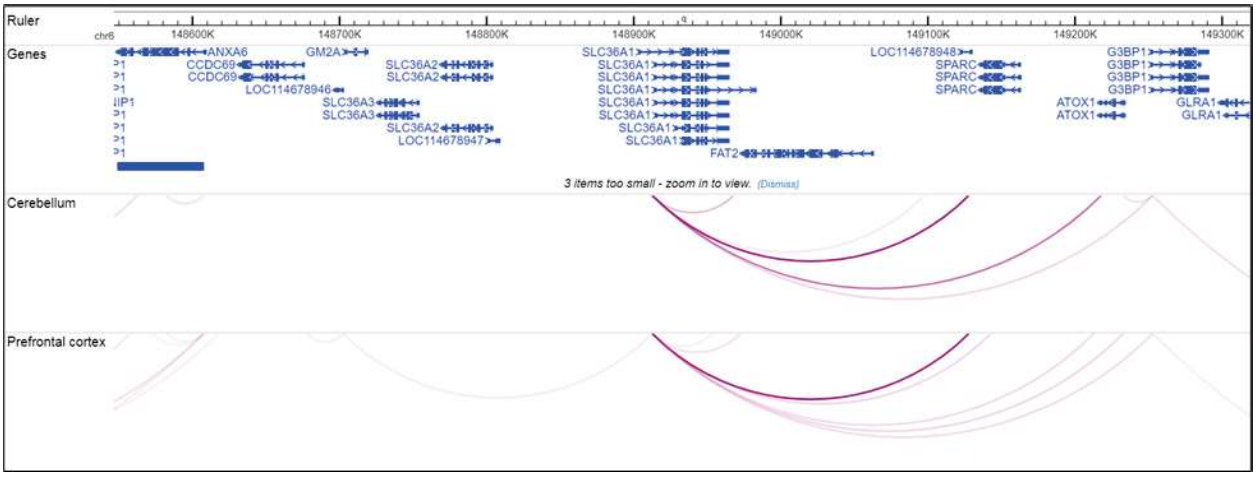

Figure 14. Loops connected to a de novo variant nearest to *SLC36A1*. Five loops connected in cerebellum. Seven loops connected in prefrontal cortex.

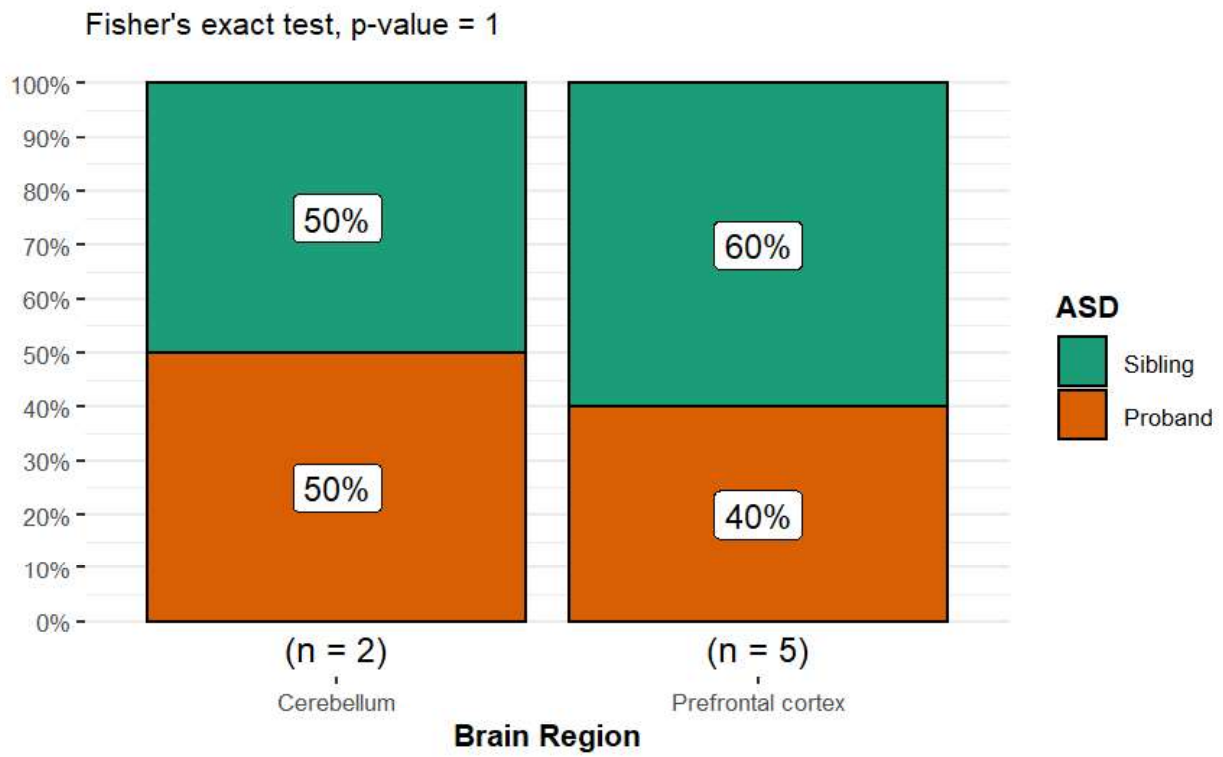

Figure 15. Fisher's exact test. No association of de novo variants in specific brain regions based on proband variant frequencies.

Discussion

Our study aimed to investigate the underlying chromatin organization differences that affect neurodevelopment and associated disorders. Although we only generated targeted 3D genomic maps for one early developmental time point, we were able to identify tens of thousands of significant chromatin interactions in the prefrontal cortex and cerebellum at 60 days gestation, many of which are connected to genes that have been previously implicated in ASD. Despite the limited number of significant interactions called, when overlaid with de novo variants previously identified in ASD quartet families we were able to connect six de novo variants.

We were able to call significant interactions using the HiC-Pro FitHiChIP pipeline, despite having low ChIP enrichment for our macaque samples. We would have been able to call

more significant interactions if we had greater ChIP enrichment as we saw in the LCL trial. One possible explanation for this lower ChIP enrichment is that the macaque brain samples were crosslinked using 1% formaldehyde in accordance with Fang et al. 2016 compared to 2% formaldehyde fixation used in the Arima-HiC+ kit for the LCL trial. MAPS was able to call 44,132 significant interactions for the LCL trial compared to only 32,959 and 13,814 for the prefrontal cortex and cerebellum respectively even though the LCL trial only had 64 million reads compared to over 1.2 and 1.3 billion reads for the macaque samples. However, we acknowledge this comparison is not exact as the significant interactions for LCLs and macaque were not called using the same pipeline and as mentioned before, there may be an issue with the genome files for macaque when using MAPS that affected the downstream analysis. Regardless, future PLAC-Seq runs for macaque brain samples should be prepped using 2% formaldehyde for crosslinking.

One way to continue to use our PLAC-seq data and improve the called interactions is to enhance it using computational methods based on deep neural networks. Several methods including HiCPlus, HiCNN, HiCNN2, DeepHiC and Variationally Encoded Hi-C Loss Enhancer (VEHiCLE), have been developed to enhance the sequencing depth of Hi-C data and have also been shown to work with targeted protein mediated chromatin interaction methods like PLAC-Seq data (Huang et al., 2022). They found that most models produced enhanced datasets that lead to improved detection of long-range chromatin interactions. While models can be trained using Hi-C data or PLAC-Seq data, model performance was higher when trained with PLAC-Seq data. Using one of these Hi-C data enhancement methods could uncover new insights from our existing dataset.

We also aimed to characterize gene expression of our isolated macaque brain regions using 3' Tag-Seq. We checked the consistency of our 3' Tag-Seq data with previously published data from the PEC using Spearman's rank correlation with the goal of incorporating the PEC data into our analysis. However, our correlations were not high enough between datasets to confidently merge them, despite having downloaded the raw data and treating each dataset with the same analysis. There are several potential factors that contributed to this lower correlation. The PEC used whole RNA-Seq to generate their dataset while we used 3' Tag-Seq. While this could be a contributing factor, previous studies have shown that expression levels of these two methods are comparable (Danielsson et al., 2015). Differences and variability between dissections was likely a bigger factor in the differences between the two datasets since brain regions compared were dissected slightly differently. T150 samples had the lowest correlations within a dataset, but this is expected as there were more dissected and diverse brain regions. We also did not have PLAC-Seq data for any of the T150 samples, so 3' Tag-Seq data of T150 samples was excluded from downstream analysis.

With our scRNA-Seq dataset, we set out to determine the cellular heterogeneity of our macaque brain regions. As expected, most of the cell types identified were neuronal cells because of the papain dissociation preparation. Overall, the identified cell types corresponded to the correct brain regions showing accuracy of the brain region isolations. In the future, one possible use for this scRNA-Seq dataset is in deconvolution methods. As single cell technology has improved recently, the importance of cell type composition has become increasingly apparent when trying to understand function, development, and disease. Bulk RNA-Seq datasets are typically confounded by differences in cell type proportions, so researchers have developed many computational methods to infer cell type proportions from bulk transcriptomics data.

While older deconvolution methods like CIBERSORT and dtangle infer cell type proportions from bulk RNA-Seq data, some newer methods like MuSiC and SCDC use scRNA-Seq data as reference (Avila Cobos et al., 2020). The newer methods generally had comparable performance to the best performing bulk methods, with some significantly down run time. Our scRNA-Seq dataset can be used in those newer deconvolution methods where a scRNA-Seq dataset is required.

References

Abrahams, B. S., Arking, D. E., Campbell, D. B., Mefford, H. C., Morrow, E. M., Weiss, L. A., Menashe, I., Wadkins, T., Banerjee-Basu, S., & Packer, A. (2013). SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). In *Molecular Autism* (Vol. 4, Issue 1). https://doi.org/10.1186/2040-2392-4-36

An, J.-Y., Lin, K., Zhu, L., Werling, D. M., Dong, S., Brand, H., Wang, H. Z., Zhao, X., Schwartz, G. B., Collins, R. L., Currall, B. B., Dastmalchi, C., Dea, J., Duhn, C., Gilson, M. C., Klei, L., Liang, L., Markenscoff-Papadimitriou, E., Pochareddy, S., … Sanders, S. J. (2018). Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science*, *362*(6420). https://doi.org/10.1126/science.aat6576

Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P., & De Preter, K. (2020). Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature Communications*, *11*(1), 5650.

Babraham Bioinformatics - Trim Galore! (n.d.). Retrieved August 13, 2022, from https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Bakken, T. E., Miller, J. A., Luo, R., Bernard, A., Bennett, J. L., Lee, C.-K., Bertagnolli, D., Parikshak, N. N., Smith, K. A., Sunkin, S. M., Amaral, D. G., Geschwind, D. H., & Lein, E. S. (2015). Spatiotemporal dynamics of the postnatal developing primate brain transcriptome. *Human Molecular Genetics*, *24*(15), 4327–4339.

Barton, R. A., & Venditti, C. (2017). Rapid Evolution of the Cerebellum in Humans and Other Great Apes. *Current Biology: CB*, *27*(8), 1249–1250.

Bhattacharyya, S., Chandra, V., Vijayanand, P., & Ay, F. (2019). Identification of significant chromatin contacts from HiChIP data by FitHiChIP. *Nature Communications*, *10*(1), 4221.

Bloom, J. D. (2018). Estimating the frequency of multiplets in single-cell RNA sequencing from cell-mixing experiments. *PeerJ*, *6*, e5578.

Brandler, W. M., Antaki, D., Gujral, M., Kleiber, M. L., Whitney, J., Maile, M. S., Hong, O., Chapman, T. R., Tan, S., Tandon, P., Pang, T., Tang, S. C., Vaux, K. K., Yang, Y., Harrington, E., Juul, S., Turner, D. J., Thiruvahindrapuram, B., Kaur, G., … Sebat, J. (2018). Paternally inherited cis-regulatory structural variants are associated with autism. *Science*, *360*(6386), 327–331.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S., & Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, *478*(7369), 343–348.

Chimpanzee Sequencing and Analysis Consortium. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, *437*(7055), 69–87.

Colvert, E., Tick, B., McEwen, F., Stewart, C., Curran, S. R., Woodhouse, E., Gillan, N., Hallett, V., Lietz, S., Garnett, T., Ronald, A., Plomin, R., Rijsdijk, F., Happé, F., & Bolton, P. (2015). Heritability of Autism Spectrum Disorder in a UK Population-Based Twin Sample. *JAMA Psychiatry* , *72*(5), 415–423.

Corley, S. M., Troy, N. M., Bosco, A., & Wilkins, M. R. (2019). QuantSeq. 3' Sequencing combined with Salmon provides a fast, reliable approach for high throughput RNA expression analysis. *Scientific Reports*, *9*(1), 18895.

C Yuen, R. K., Merico, D., Bookman, M., L Howe, J., Thiruvahindrapuram, B., Patel, R. V., Whitney, J., Deflaux, N., Bingham, J., Wang, Z., Pellecchia, G., Buchanan, J. A., Walker,

S., Marshall, C. R., Uddin, M., Zarrei, M., Deneault, E., D'Abate, L., Chan, A. J. S., … Scherer, S. W. (2017). Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nature Neuroscience*, *20*(4), 602–611.

Danielsson, F., James, T., Gomez-Cabrero, D., & Huss, M. (2015). Assessing the consistency of public human tissue RNA-seq data sets. *Briefings in Bioinformatics*, *16*(6), 941–949.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* , *29*(1), 15–21.

Fang, R., Yu, M., Li, G., Chee, S., Liu, T., Schmitt, A. D., & Ren, B. (2016). Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Research*, *26*(12), 1345–1348.

Feng, J., Liu, T., Qin, B., Zhang, Y., & Liu, X. S. (2012). Identifying ChIP-seq enrichment using MACS. *Nature Protocols*, *7*(9), 1728–1740.

Fischbach, G. D., & Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron*, *68*(2), 192–195.

Franzén, O., Gan, L.-M., & Björkegren, J. L. M. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database: The Journal of Biological Databases and Curation*, *2019*. https://doi.org/10.1093/database/baz046

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., Chew, E. G. Y., Huang, P. Y. H., Welboren, W.-J., Han, Y., Ooi, H. S., Ariyaratne, P. N., Vega, V. B., Luo, Y., Tan, P. Y., … Ruan, Y. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, *462*(7269), 58–64.

GEM library - Browse /gem-library at SourceForge.net. (n.d.). Retrieved August 13, 2022, from https://sourceforge.net/

Gupta, S., Ellis, S. E., Ashar, F. N., Moes, A., Bader, J. S., Zhan, J., West, A. B., & Arking, D. E. (2014). Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nature Communications*, *5*, 5748.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., 3rd, Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., … Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, *184*(13), 3573–3587.e29.

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., Hillman-Jackson, J., Kuhn, R. M., Pedersen, J. S., Pohl, A., Raney, B. J., Rosenbloom, K. R., Siepel, A., Smith, K. E., Sugnet, C. W., … Kent, W. J. (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, *34*(Database issue), D590–D598.

Hoeft, F., Walter, E., Lightbody, A. A., Hazlett, H. C., Chang, C., Piven, J., & Reiss, A. L. (2011). Neuroanatomical differences in toddler boys with fragile x syndrome and idiopathic autism. *Archives of General Psychiatry*, *68*(3), 295–305.

Huang, L., Yang, Y., Li, G., Jiang, M., Wen, J., Abnousi, A., Rosen, J. D., Hu, M., & Li, Y. (2022). A systematic evaluation of Hi-C data enhancement methods for enhancing PLAC-seq and HiChIP data. *Briefings in Bioinformatics*, *23*(3). https://doi.org/10.1093/bib/bbac145

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., & Teichmann, S. A. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*, *17*, 29.

Juric, I., Yu, M., Abnousi, A., Raviram, R., Fang, R., Zhao, Y., Zhang, Y., Qiu, Y., Yang, Y., Li, Y., Ren, B., & Hu, M. (2019). MAPS: Model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. *PLoS Computational Biology*, *15*(4), e1006982.

Kelly, E., Meng, F., Fujita, H., Morgado, F., Kazemi, Y., Rice, L. C., Ren, C., Escamilla, C. O., Gibson, J. M., Sajadi, S., Pendry, R. J., Tan, T., Ellegood, J., Albert Basson, M., Blakely, R. D., Dindot, S. V., Golzio, C., Hahn, M. K., Katsanis, N., … Tsai, P. T. (2020). Regulation of autism-relevant behaviors by cerebellar–prefrontal cortical circuits. In *Nature Neuroscience* (Vol. 23, Issue 9, pp. 1102–1110). https://doi.org/10.1038/s41593-020-0665-z

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., & Teichmann, S. A. (2015). The Technology and Biology of Single-Cell RNA Sequencing. In *Molecular Cell* (Vol. 58, Issue 4, pp. 610–620). https://doi.org/10.1016/j.molcel.2015.04.005

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.

Li, D., Purushotham, D., Harrison, J. K., Hsu, S., Zhuo, X., Fan, C., Liu, S., Xu, V., Chen, S., Xu, J., Ouyang, S., Wu, A. S., & Wang, T. (2022). WashU Epigenome Browser update 2022. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkac238

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. In *Bioinformatics* (Vol. 25, Issue 14, pp. 1754–1760). https://doi.org/10.1093/bioinformatics/btp324

Lohman, B. K., Weber, J. N., & Bolnick, D. I. (2016). Evaluation of TagSeq, a reliable low-cost alternative for RNAseq. *Molecular Ecology Resources*, *16*(6), 1315–1321.

Lord, C., Risi, S., DiLavore, P. S., Shulman, C., Thurm, A., & Pickles, A. (2006). Autism from 2 to 9 years of age. *Archives of General Psychiatry*, *63*(6), 694–701.

Luo, X., Liu, Y., Dang, D., Hu, T., Hou, Y., Meng, X., Zhang, F., Li, T., Wang, C., Li, M., Wu, H., Shen, Q., Hu, Y., Zeng, X., He, X., Yan, L., Zhang, S., Li, C., & Su, B. (2021). 3D Genome of macaque fetal brain reveals evolutionary innovations during primate corticogenesis. *Cell*, *184*(3), 723–740.e21.

McGinnis, C. S., Murrow, L. M., & Gartner, Z. J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems*, *8*(4), 329–337.e4.

Meyer, E., Aglyamova, G. V., & Matz, M. V. (2011). Profiling gene expression responses of coral larvae (Acropora millepora) to elevated temperature and settlement inducers using a novel RNA-Seq procedure. *Molecular Ecology*, *20*(17), 3599–3616.

Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J., & Chang, H. Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods*, *13*(11), 919–922.

Ouhaz, Z., Fleming, H., & Mitchell, A. S. (2018). Cognitive Functions and Neurodevelopmental Disorders Involving the Prefrontal Cortex and Mediodorsal Thalamus. *Frontiers in Neuroscience*, *12*, 33.

Patil, A., & Patil, A. (n.d.). *CellKb Immune: a manually curated database of mammalian hematopoietic marker gene sets for rapid cell type identification*. https://doi.org/10.1101/2020.12.01.389890

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419.

PsychENCODE Consortium, Akbarian, S., Liu, C., Knowles, J. A., Vaccarino, F. M., Farnham,

P. J., Crawford, G. E., Jaffe, A. E., Pinto, D., Dracheva, S., Geschwind, D. H., Mill, J., Nairn, A. C., Abyzov, A., Pochareddy, S., Prabhakar, S., Weissman, S., Sullivan, P. F., State, M. W., … Sestan, N. (2015). The PsychENCODE project. *Nature Neuroscience*, *18*(12), 1707–1712.

Putri, G. H., Anders, S., Pyl, P. T., Pimanda, J. E., & Zanini, F. (2022). Analysing high-throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics* , *38*(10), 2943–2945.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. In *Bioinformatics* (Vol. 26, Issue 6, pp. 841–842). https://doi.org/10.1093/bioinformatics/btq033

Rakic, P. (2009). Evolution of the neocortex: a perspective from developmental biology. *Nature Reviews. Neuroscience*, *10*(10), 724–735.

Robinson, J. T., Turner, D., Durand, N. C., Thorvaldsdóttir, H., Mesirov, J. P., & Aiden, E. L. (2018). Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Systems*, *6*(2), 256–258.e1.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* , *26*(1), 139–140.

Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Larsson, H., Hultman, C. M., & Reichenberg, A. (2014). The Familial Risk of Autism. In *JAMA* (Vol. 311, Issue 17, p. 1770). https://doi.org/10.1001/jama.2014.4144

Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., Stevens, C., Reichert, J., Mulhern, M. S., Artomov, M., Gerges, S., Sheppard, B., Xu, X., Bhaduri, A., Norman, U., … Buxbaum, J. D. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell*, *180*(3), 568–584.e23.

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J., & Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology*, *16*, 259.

Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, *4*, 1521.

Sousa, A. M. M., Meyer, K. A., Santpere, G., Gulden, F. O., & Sestan, N. (2017). Evolution of the Human Nervous System Function, Structure, and Development. *Cell*, *170*(2), 226–247.

Stegle, O., Teichmann, S. A., & Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews. Genetics*, *16*(3), 133–145.

Turner, T. N., Hormozdiari, F., Duyzend, M. H., McClymont, S. A., Hook, P. W., Iossifov, I., Raja, A., Baker, C., Hoekzema, K., Stessman, H. A., Zody, M. C., Nelson, B. J., Huddleston, J., Sandstrom, R., Smith, J. D., Hanna, D., Swanson, J. M., Faustman, E. M., Bamshad, M. J., … Eichler, E. E. (2016). Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *American Journal of Human Genetics*, *98*(1), 58–74.

Voineagu, I., Wang, X., Johnston, P., Lowe, J. K., Tian, Y., Horvath, S., Mill, J., Cantor, R. M., Blencowe, B. J., & Geschwind, D. H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. In *Nature* (Vol. 474, Issue 7351, pp. 380–384). https://doi.org/10.1038/nature10110

Won, H., de la Torre-Ubieta, L., Stein, J. L., Parikshak, N. N., Huang, J., Opland, C. K., Gandal,

M. J., Sutton, G. J., Hormozdiari, F., Lu, D., Lee, C., Eskin, E., Voineagu, I., Ernst, J., & Geschwind, D. H. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature*, *538*(7626), 523–527.

Yang, T., Zhang, F., Yardımcı, G. G., Song, F., Hardison, R. C., Noble, W. S., Yue, F., & Li, Q. (2017). HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Research*, *27*(11), 1939–1949.

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., Luo, T., Xu, L., Liao, G., Yan, M., Ping, Y., Li, F., Shi, A., Bai, J., Zhao, T., Li, X., & Xiao, Y. (2019). CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Research*, *47*(D1), D721–D728.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, *9*(9), R137.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*, 14049.

Zhou, J., Park, C. Y., Theesfeld, C. L., Wong, A. K., Yuan, Y., Scheckel, C., Fak, J. J., Funk, J., Yao, K., Tajima, Y., Packer, A., Darnell, R. B., & Troyanskaya, O. G. (2019). Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nature Genetics*, *51*(6), 973–980.

Zhu, Y., Sousa, A. M. M., Gao, T., Skarica, M., Li, M., Santpere, G., Esteller-Cucala, P., Juan, D., Ferrández-Peral, L., Gulden, F. O., Yang, M., Miller, D. J., Marques-Bonet, T., Imamura Kawasawa, Y., Zhao, H., & Sestan, N. (2018). Spatiotemporal transcriptomic divergence across human and macaque brain development. *Science*, *362*(6420). https://doi.org/10.1126/science.aat8077