A multi-omic approach to uncover enhancer-gene interactions in the human brain

Sophia Yan[1], Steve A. McCarroll[2,3,4,5], and Nicole B. Rockweiler[2,3]

[1]Newton South High School, Newton, MA, 02459, USA
[2]Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA
[3]Department of Genetics, Harvard Medical School, Boston, MA, 02115, USA
[4]Program in Neuroscience, Harvard Medical School, Boston, MA, 02215, USA
[5]Howard Hughes Medical Institute, Boston, MA 02115, USA

Abstract

Gene regulation is a complicated process, critical for maintaining cell type-specific functions by controlling RNA expression. Diseases such as cancer, neurological disorders, and autoimmunity stem from the mis-regulation of gene expression. Here, we computationally explored one of the primary regulators of gene expression, enhancers, at a single nucleus resolution, aiming to understand the cell type specific functions of enhancers. We've proposed a strategy to use single nucleus RNA-seq (snRNA-seq) to detect a canonical marker of active enhancers, enhancer RNA (eRNA), in the Brodmann area 46 (BA46) of 180 human brains across 1,217,965 nuclei. We assessed these putative enhancers by creating a scoring system to quantify their bidirectional transcription, a property of eRNA. We found a significant positive shift in the bidirectional score distributions between our putative enhancers and the nearby regions (Wilcoxon Signed Rank Test, *p*-value = 5e-143), providing confidence that our regions show enhancer behavior. We then utilized the unique power of our single nucleus sequencing data to explore the cell type-specificity of these putative enhancers and observed two-fold as many enhancers expressed in multiple cell types (n = 6,477) compared to cell type specific enhancers (n = 2,699). The mean expression level of cell type specific enhancers was also 29 times lower than ubiquitously expressed enhancers. In addition, we mapped these enhancers to their putative target genes by testing for correlation between putative eRNA expression and gene expression within the same topologically associating domain. Among the 4,147 potential enhancer-gene pairs we found across seven cell types using snRNA-seq, 116 (~3%) pairs are in chromatin loops and likely interact with each other in 3D space from a bulk Hi-C data analysis. Moreover, the enhancer-gene pairs identified in the multiome dataset are well-replicated in the snRNA-seq dataset across the same 20 donors, with the expression correlation values following the line of best fit $y = 0.400x$. Furthermore, increasing the sample size of snRNA-seq dataset to 160 donors yields similar correlation values when compared to the 20-donor snRNA-seq dataset (line of best fit of $y = 0.746x$), highlighting the robustness of snRNA-seq for studying enhancer-gene interactions. The putative enhancer-gene pairs provide insight into the complex regulatory networks in the brain, shedding light on how dysregulation of these regions may contribute to brain-related disorders. In addition, our framework for identifying putative active enhancers and enhancer-gene pairs is widely applicable to analyzing snRNA-seq data in other tissues and species.

Keywords: eRNA, gene regulation, snRNA-seq, transcriptional bidirectionality, Hi-C, human, brain

Table of Contents

Introduction

Gene regulation, the process of controlling gene expression in a cell, is one of the most critical and complicated mechanisms in the human body. It enables different cells to perform specialized functions despite having the same genetic material. Proper gene regulation is essential for maintaining homeostasis, and genetic disorders often arise from mutations in regulatory regions (i.e. promoters, enhancers, and silencers). For example, a mutation at the APP promoter significantly increases APP expression levels and the risk for Alzheimer[1].

One of the prominent methods used to explore gene expression is surveying the transcriptome — the complete set of RNA found in a cell. The prevailing understanding of the transcriptome categorizes RNA into two primary groups: coding and non-coding RNA. Coding RNA, which comprises less than 2% of the transcriptome, generally refers to messenger RNA (mRNA). mRNA is produced from the transcription of a target gene and serves as a blueprint for protein synthesis. Directly upstream of the target gene is the promoter, where RNA polymerase binds to when starting transcription of the gene. With enough proximity to the promoter, another regulatory region known as the enhancer is also able to modulate gene expression by providing binding sites for proteins like transcription factors (TFs) that promote transcription.

Interactions between enhancers and promoters can occur both in *cis*, when the promoter and enhancer are on the same chromosome, and *trans*, when they're on different chromosomes. *cis* interactions are well understood, with enhancers and promoters being brought close through chromatin loops formed by the ring-like cohesin protein. *trans* interactions, on the other hand, require complex spatial rearrangements to bring the promoter and enhancer into physical proximity.

However, enhancers themselves were not thought of as being capable of transcription until two studies

conducted in 2010 by Kim et al. and De Santa et al.[2,3] found evidence of enhancer RNA (eRNA), a type of non-coding RNA (ncRNA). While coding RNAs are only composed of mRNA, ncRNAs can be broken down into many different groups such as transfer, ribosomal, intronic, and long noncoding RNA. Their functions vary widely, from transmitting signals in and between cells to scaffolding structures within the cell.

It was quickly demonstrated that knockout of the target gene resulted in the wipeout of the corresponding eRNA molecule[2,4–9], and eRNA as a biomarker showed great promise in being able to link enhancer expression back to the target gene. Traditionally, it has been challenging to identify the enhancers that regulate a target gene because enhancers can regulate genes a megabase away.

However, eRNA is unstable and quickly degrading[10,11], making it difficult to detect. Assays like global run-on sequencing (GRO-seq) and cap analysis gene expression (CAGE) are commonly used to detect eRNA as they target nascent transcripts[10,12], allowing a full profile of transcriptional activity to be sequenced without the poly(A) tail bias snRNA-seq has[13]. The RNA-sequencing methods are undesirable for eRNA detection because they are built for targeting mRNA by utilizing the transcript's poly(A) tail; however, because eRNA transcripts have been found to have inconsistent properties with one another, two types of eRNA have emerged: 1-directional (1d) and 2-directional (2d) eRNA. 90% of eRNA molecules are 2d; they are unstable, small non-polyadenylated transcripts. The other 10% consists of 1d eRNA, which has higher stability, a longer length, and a poly(A) tail. These properties suggest that 1d-eRNA could be detected using snRNA-sequencing.

Furthermore, the ability to capture eRNA activity using single-cell sequencing is attractive both in terms of affordability and availability, not to mention that eRNA can be explored at an unprecedented cell type-level. Through a bidirectional transcription analysis of putative enhancers, we found evidence of eRNA in snRNA-seq data and annotated 4,147 enhancer-gene pairs, 116 (~3%) of which are confirmed by Hi-C data.

Results

1. Detecting eRNA through a bidirectional transcription scoring system

We analyzed the dorsolateral prefrontal cortex (Brodmann area 46, BA46) from two datasets: a single nucleus multiome ATAC + gene expression dataset comprising of 20 post-mortem human brain donors and 78,873 nuclei, and a 10x Chromium Single Nuclei 3′ dataset comprising of 180 donors and 1,217,965 nuclei. We combined enhancer annotations from FANTOM5 and the Enhancer Atlas databases to create a list of 29,091 potential brain-related enhancers. This list was then filtered to exclude enhancers that overlap GENCODE's protein-coding genes and lncRNA annotations, resulting in 28,509 putative enhancers.

We validated our putative eRNAs by assessing their bidirectional transcription activity, a distinguishing property of eRNA. We shifted our putative enhancers over by 1000bp (i.e. chr1:1,000-1,500 → chr1:2,500-3,000) as a control and measured the scores of both regions.

The bidirectional score consists of two parts — the divergence score and the balanced score (**Figure 1a**). We calculate a midpoint (P_c) based on read peaks instead of using gene coordinates as we are not confident in our putative enhancer boundaries (**Figure 1b**). The divergence score has an optimum score is 1, representing a perfect divergence of forward and reverse stranded reads; a score of 0 represents a complete convergence of the forward and reverse stranded reads. In the example shown in Figure 1b, the reverse-strand read peak is upstream of the forward-strand read peak, and the reads are generally diverging. However, our methodology does not force the reverse-strand read peak to be upstream of the forward-strand read peak, which is why some divergence scores show a convergence of reads.

The balanced score compares the number of forward and reverse stranded reads mapped to the enhancer. We expect balanced scores close to 0 for enhancers with bidirectional transcription, indicating there are similar amounts of forward and reverse stranded reads. -1 and +1 represent unidirectional transcription of reverse and forward reads, respectively.

We scored the putative enhancers and the shifted regions by multiplying the divergence score with an adjusted balanced score such that the resulting bidirectional score ranges from 0 to 1. This is done because the divergence score alone is not enough to classify if an enhancer displays bidirectional transcription, as enhancers showing strong diverging or converging reads are more likely to show unidirectional transcription (**Figure 1c**). The high counts at the extreme values of the divergence score shown in **Figure 1c** correspond to very negative balanced scores, indicating that many converging and diverging enhancers show unidirectional transcription instead of bidirectional transcription.

We note a positive shift in score distribution between the putative enhancers and their shifted regions, particularly as the bidirectional score increases, suggesting that it is more probable to see bidirectional transcription at our enhancer annotations compared to nearby intergenic regions (**Figure 1d**).

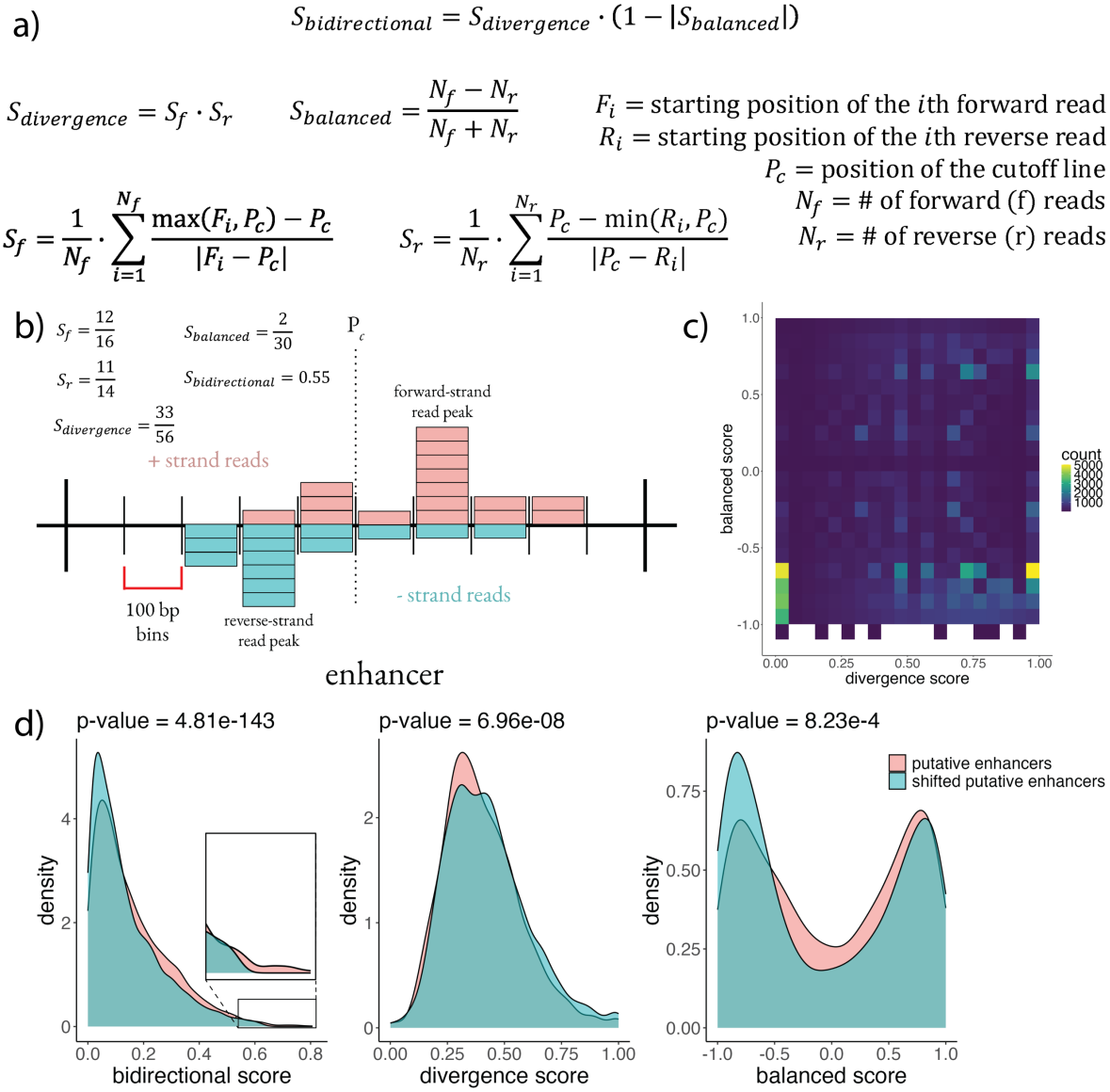

Figure 1. Bidirectional analysis of putative enhancers

a) Formulas for the scoring system specified in results and methods. $S_{divergence}$ is broken up into S_f and S_r (forward read score and reverse read score) for clarity, but S_f and S_r are not used individually. **b)** Schema for the bidirectional score calculation. **c)** 2D histogram of the divergence score versus the balanced score. White space indicates no counts were found for that region. **d)** Density plots of the score distributions for both putative and shifted putative enhancers. The enhancers are combined across cell type as each cell type demonstrates similar behavior. *p*-values were calculated using the Wilcoxon Signed Rank Test.

2. Exploring enhancer expression across cell types

The assignment of cell types is commonly known to be determined by the expression of marker genes, and less often do we think of cell types as defined by the enhancer activity that regulates gene expression. Here we explore this concept of cell type-specific and ubiquitously expressed enhancers. We took inspiration from Zhang *et al.* who had utilized an eRNA expression heatmap to find highly active enhancers across various cancers[14].

We found similar numbers of cell type-specific (n = 2,699) and ubiquitous enhancers (n = 2,272) and close to two-fold as many somewhat specific enhancers (n = 4,205). Most of the somewhat specific enhancers are concurrently expressed in the neurons, suggesting that they regulate genes that control for neuron-specific functions and that cell-types with similar functions will share similar enhancer profiles (**Figure 2a**).

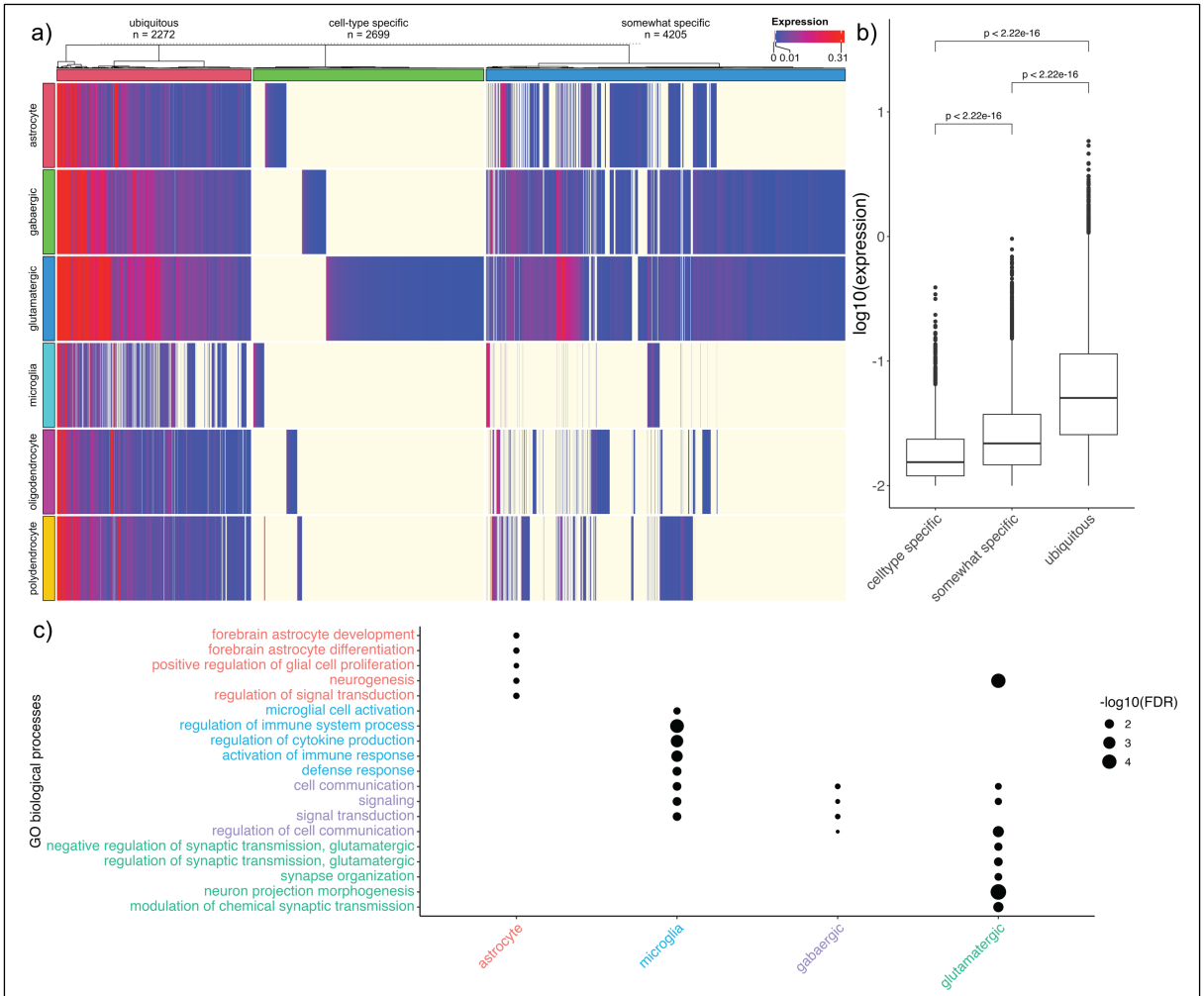

Figure 2. Classification of enhancers based on cell type specificity

a) Heatmap comparing the quantity and expression level of the different groups of enhancers across cell types. The three distinct groups of enhancers identified in the heatmap were cell type specific (expressed in 1 cell type), somewhat specific (expressed in 2-4 cell types), and ubiquitous (expressed in 5 or more cell types). **b)** Boxplot of the enhancers' eRNA RPM levels under each category defined in panel **a**. *p*-values were calculated using the Wilcoxon Rank-Sum Test. **c)** Gallery of a subset of significantly enriched pathways found in the set of genes closest to the cell type specific enhancers.

The ubiquitously expressed enhancers, which likely regulate housekeeping genes, have a mean expression level seven times higher than the somewhat specific enhancers and twenty-nine times higher than the cell-type specific enhancers. This implies a direct relationship between the conservation of enhancers and the magnitude of their expression, further suggesting that the genes associated with basic cellular functions are more intensely regulated compared to cell-type specific genes (**Figure 2b**).

We then focused on the cell-type specific enhancers, aiming to provide further evidence that the reads we're detecting show behavior of eRNAs. We ran gene ontology on the set of genes closest to each enhancer; as it is common for enhancers to regulate proximal genes, we expected to see an enrichment of cell-type specific pathways.

With FDR < 0.05 as cutoff, we found 33 significantly enriched pathways in astrocytes, 15 in GABAergic neurons, 98 in glutamatergic neurons, and 66 in microglia. Astrocytes, microglia, and glutamatergic neurons had particularly strong cell-type specific pathways with "forebrain astrocyte development", "microglial cell regulation", and

"regulation of synaptic transmission, glutamatergic" respectively (**Figure 2c**). GABAergic neurons had weaker cell-type specific signal and more general pathways such as "cell communication" and "signaling" that both microglia and glutamatergic neurons shared. In addition, oligodendrocytes, polydendrocytes, and endothelia cells did not have significantly enriched pathways. The gene ontology analysis is significantly biased to the cell-types with more cell-type specific enhancers, and this can be seen by the strong signal from glutamatergic neurons and the non-existent ones from oligodendrocytes and polydendrocytes. Though the analysis cannot prove the functional relevance of the putative enhancers, it does provide confidence to our enhancers as we see the cell types with enough signal display pathways that contribute to their cell-type specific function.

3. Mapping enhancers to genes and validating their correlations with Hi-C data

With reasonable confidence in our putative eRNAs, we next mapped these regions to potential genes that they could regulate. We focus on enhancer-gene interactions in *cis* by taking a brute force approach, comparing every gene and enhancer in the same topologically associating domain (TAD). Optimally, we would have used single-cell Hi-C data to create TAD boundaries for each of our cell types, but because of the assay's high cost and low accessibility, we were unable to do so. Instead, we used a publicly available bulk Hi-C dataset from the dorsolateral prefrontal cortex of an 87-year-old female brain[12], which contains the BA46 region our snRNA-seq data was acquired from, to create a set of general TAD boundaries that all cell types referenced. The ratio of significant enhancer-gene pairs versus the total number of pairs checked is not consistent across cell types and may be related to the different amounts of reads each cell type sequences (**Figure 3a**). This is best exemplified with the neurons, which often have more reads and demonstrate a higher ratio of correlated enhancer-gene pairs versus total pairs.

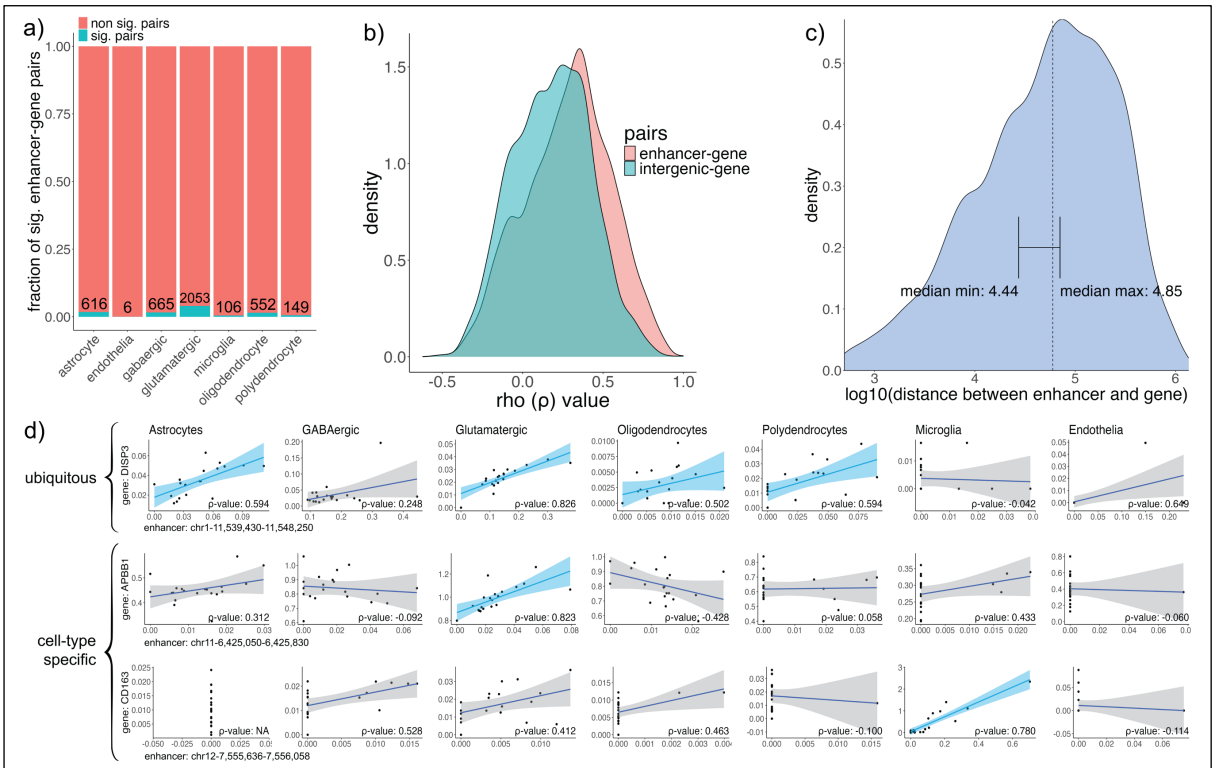

Figure 3. Spearman's correlation analysis on enhancer-gene pairs.

a) Relative bar-plot to show percentage of significant enhancer-gene pairs found versus total enhancer-gene pairs checked. **b)** Density plot to compare the distribution of Spearman's ρ scores between enhancer-gene and intergenic-gene pairs. These pairs are combined across cell types for easy visualization; results were consistent across cell types. **c)** Density plot showing the distribution of distance between the enhancer and gene in the significantly correlated pairs. Pairs are combined across cell type, and the error bar shows the range of median distances for each cell type. **d)** Representative enhancer-gene correlation pairs. A pair is colored blue if Spearman's $\rho \geq 0.5$ and at least 10 donors have positive gene and enhancer expression. Expression units on the *x* and *y* axis are averaged log-normalized UMI counts (specified further in section 4 of Methods).

To confirm that these enhancer-gene pairs are not a result of widespread transcription at stretches of open chromatin, we also calculate the correlation between two other regulatory region pairs: enhancer-intergenic regions and gene-intergenic regions. Intergenic regions are the shifted putative-enhancer regions as used in **Figure 1d**. We observe a non-trivial positive shift in Spearman's ρ between enhancer-gene pairs and intergenic-gene pairs (enhancer-gene versus intergenic-gene *p*-value = 0; Wilcoxon rank-sum tests) (**Figure 3b**). This result confirms that the enhancer gene pairs we identified are more likely to have real functional correlations.

To check how realistic our enhancer-gene pairs are, we looked at the average distance between the enhancers and genes in our significant correlations and compared it with the results in the literature. Most interactions occur within 1Mb, and the median distance is ~100Kb[15–17]. Our enhancer-gene pairs show interaction distances spanning between ~20Kb and ~70Kb, smaller than the reported median, but still a plausible distance (**Figure 3c**).

Using a single-nucleus resolution dataset enables us to find both enhancer-gene pairs that are specific to a single cell-type and found in multiple. For example, *DISP3*, which plays a role in neural proliferation and differentiation, shows strong correlation with an enhancer around 2Kb downstream of it in astrocytes, glutamatergic neurons, oligodendrocytes, and polydendrocytes. In contrast, *APBB1*, a gene active in glutamatergic synapses and involved in several processes, one of which being the generation of neurons, appears to be specifically regulated by an enhancer roughly 6Kb downstream of it. And *CD163*, a gene that encodes the CD163 protein normally found in monocytes and macrophages but can be found expressed in microglia in cases of HIV encephalitis and Alzheimer disease, is seen to be regulated in microglia by an enhancer much further downstream of it, close to 52Kb away (**Figure 3d**).

Table 1 provides a summary of all the enhancer-gene pairs found, as well findings on how many genes an enhancer typically regulates, as well as how many enhancers regulate a gene across cell types. Although enhancers can regulate more than one gene, we've found that most enhancers likely regulate one or two genes, and most genes are likely only regulated by one or two enhancers (**Table 1**).

Table 1. General statistics of the enhancer-gene pairs

Cell type	# of significant pairs	# of unique enhancers	# of unique genes	Average # of genes correlating with an enhancer	Average # of enhancers correlating with a gene	# of significant pairs with chromatin interaction	Probability of significant pairs with chromatin interaction
astrocytes	616	428	427	1.44	1.44	39	6.33%
GABAergic neurons	665	514	498	1.29	1.34	32	4.81%
glutamatergic neurons	2,053	1,364	1,231	1.51	1.67	56	2.73%
microglia	106	83	68	1.28	1.56	12	11.32%
oligodendrocytes	552	380	374	1.45	1.48	31	5.62%
polydendrocytes	149	115	113	1.30	1.32	15	10.07%

To further confirm the interaction between these potential significantly correlated enhancer-gene pairs, we check if they overlap a Hi-C contact-domain (**Figure 4**). Using Hi-C to provide confidence to these pairs also reduces the snRNA-seq bias that glutamatergic neurons demonstrate. Though they have over four-fold more significant enhancer-gene pairs compared to the other cell types, only 2.73% of those pairs show chromatin interaction with each other. Across cell types, the number of pairs that demonstrate chromatin interaction is remarkably similar. This is likely due to using the same Hi-C data for all cell-types; we are assuming every cell type has the same chromatin organization, which is unlikely. Nevertheless, the fact that there is a strong percentage of pairs (~6.81%) confirmed by Hi-C to be biologically relevant is encouraging for further exploration into these pairs.

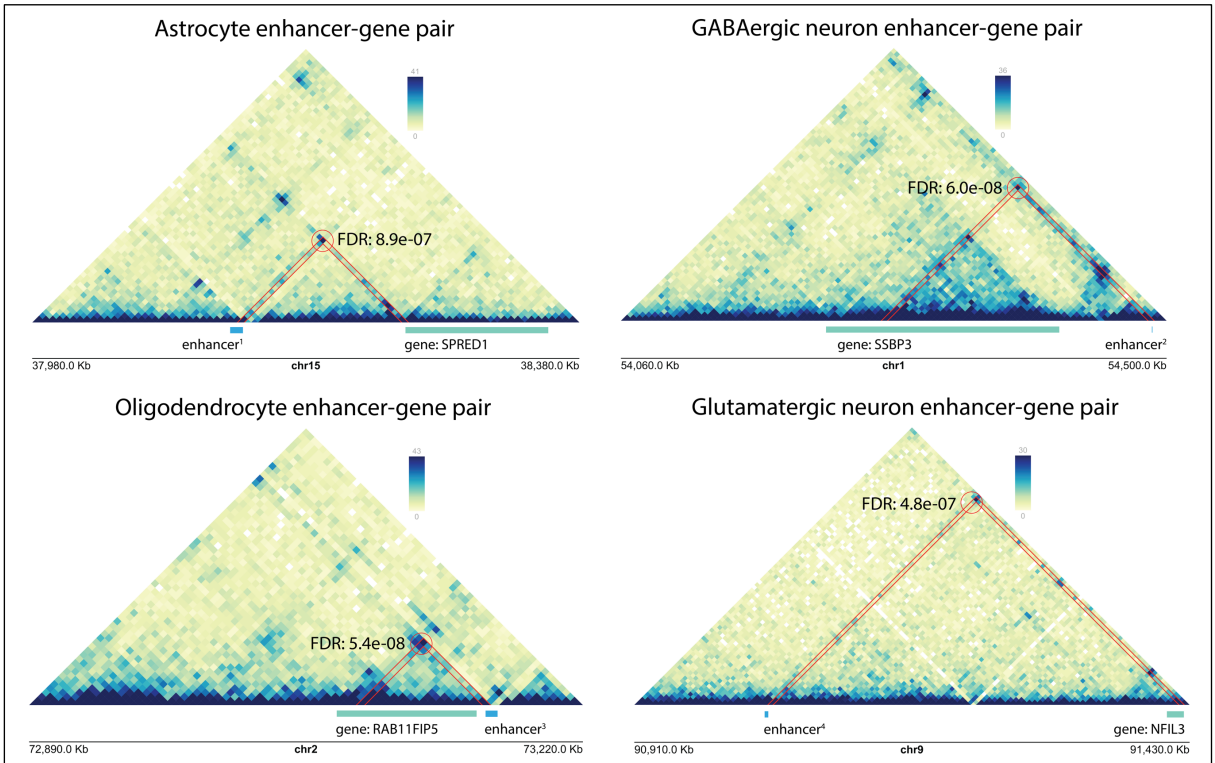

Figure 4. Hi-C plots of representative enhancer-gene pairs with strong chromatin contact levels. The list of contact domains (n = 4,394) was created using Arrowhead from the Juicer suite, and the contact-loops were acquired using HiCCUPS[15,19]. The contact-loops were filtered by fdrDonut score, which accounts for the neighboring background that forms a donut-shaped area around the loop. Our background was the TAD where the enhancer-gene pair was located. Key: enhancer[1] (chr15-38,124,660-38,134,070), enhancer[2] (chr1-54,488,096-54,488,652), enhancer[3] (chr2-73,162,020-73,169,160), enhancer[4] (chr9-91,032,030-91,035,450).

4. Assessing the reproducibility of enhancer-gene pairs

Though we can detect enhancer-gene pairs that are biologically interesting in our 20 donor multiome dataset, their strong correlation should hold across differing assays or sample sizes. As such, we used a larger 180-donor snRNA-seq dataset from the BA46 region that encompassed the 20 donors from the multiome dataset to re-run our analyses.

First, we compared the enhancer-gene pairs' Spearman ρ values between the same 20 donors across both the multiome dataset and the snRNA-seq dataset. An optimal slope of 1 for the line of best fit between the multiome and snRNA-seq Spearman ρ values would indicate perfect reproducibility of correlations between our enhancer-gene pairs. **Figure 5a** shows slope of 0.400, shallower than expected. This is likely due to sampling error and the small number of donors used in the analysis, but we do see a promising sign of reproducibility from the enhancer-gene pairs in the upper tail end of the Spearman ρ value density plots. They fit the line $y = x$ well, indicating that enhancer-gene pairs with high multiome ρ values are also likely to have high snRNA-seq ρ values. This trend is not present in **figure 5b**, our null model where we randomly pair an enhancer with a gene in the multiome dataset. In the null model, enhancer-gene pairs are equally as likely to have high snRNA-seq ρ value for all multiome ρ values.

For the sample size comparison, we use the 180-donor snRNA-seq dataset. We partitioned this dataset into 20-donor and 160-donor datasets to keep the enhancer-gene pairs' Spearman ρ values independent of each other. **Figure 5c** shows a steeper line of best fit, with a slope of 0.746, showcasing how reproducible our enhancer-gene pairs are with a large enough sample size. With the sample size comparison as well, we see enhancer-gene pairs that fit the line $y = x$, showing how enhancer-gene pairs with high 160-donor ρ values have similar 20-donor ρ values, which the sample size null model does not reflect (**Figure 5d**).

Figure 5e showcases some highly interesting glutamatergic enhancer-gene pairs in the upper quadrant of the ρ value density plots. We are particularly interested in these enhancer-gene pairs because of their high Spearman ρ values that manage to stay consistent in both the 20-donor multiome and 180-donor snRNA-seq dataset and the convincing chromatin interactions they show. *KCTD7* is a gene involved in the excitability of cortical neurons, *C5AR1* is a

biomarker of Alzheimer's disease, and *GBF1* plays a role in modulating vesicular trafficking in the Golgi apparatus. With these pairs corroborated across two different assays and 180 different donors, it is likely that they hold biological relevance and can be further explored to uncover their roles in aging and disease.

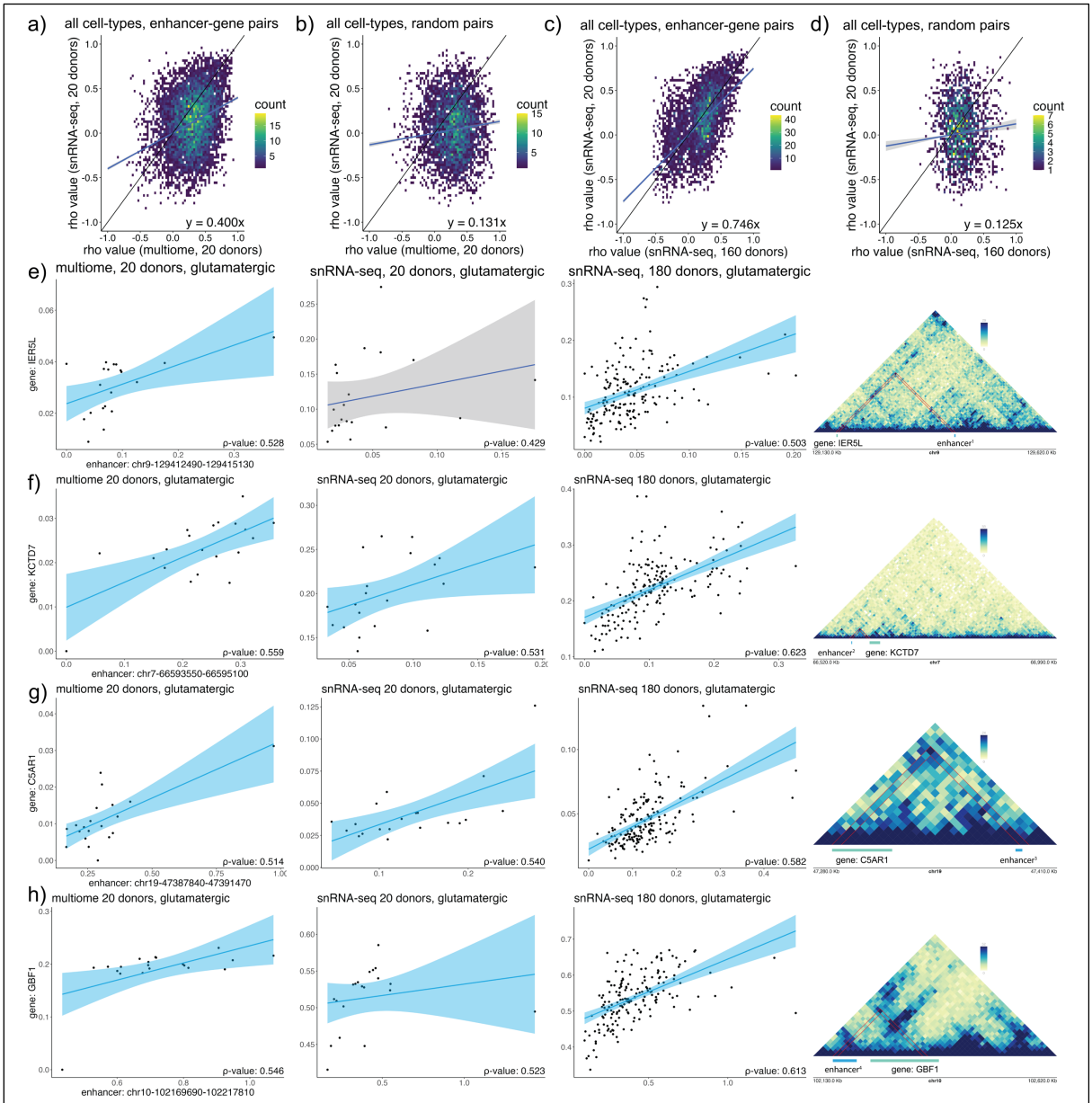

Figure 5. Strong reproducibility of enhancer-gene pairs across differing assays and sample sizes
a,b,c,d) Density plots of enhancer-gene pairs, with Spearman's ρ values on the axes between differing conditions of assays (multiome versus snRNA-seq) and sample size (20 donors versus 180 donors). If there are multiple enhancer-gene pairs with the same enhancer, the enhancer-gene pair with the highest absolute Spearman's ρ value is plotted. **e,f,g,h)** Four enhancer-gene pairs in glutamatergic neurons shown across three different conditions. Expression units on the *x* and *y* axis are averaged log-normalized UMI counts (specified further in section 4 of Methods).

Discussion

We find that snRNA-seq has potential for detecting eRNA by creating our own scoring system to check the bidirectionality of an enhancer. The bidirectionality test is often done with GRO-seq, so we used papers that had already created their own scoring system for inspiration[10,12]. The idea of separating the enhancer into 100bp bins came from Kristjánsdóttir *et al*[12]. In their application, they forced the reverse-strand read peak to be upstream of the forward-strand

read peak, which eliminates the possibility of converging reads. Although this makes sense from a biological standpoint as there is no reason for the reads to converge, we included the possibility to compare the frequency of converging reads versus diverging reads in enhancers. We expected that more enhancers would show diverging read behavior, however the two groups seemed to be of equal amount. We believe a bin size of 100bp is too large for some enhancers as both the forward and reverse strand read peaks might map to the same bin. We expect the peak of enhancers that demonstrate converging read behavior to drop significantly with a suitable bin size, which we will explore further.

We demonstrate the power of eRNA as an indicator of enhancer activity and find that there are more somewhat specific enhancers than both cell type specific and ubiquitously expressed enhancers; furthermore, the ubiquitously expressed enhancers show a much higher expression level, implying that there is stronger regulation on the housekeeping genes as compared to cell type specific ones. Further confidence is provided to our putative cell type specific enhancers by running gene ontology on the surrounding genes; we observe enriched pathways linking back to cell type specific functions. We believe the small number of cell type specific enhancers detected for oligodendrocytes, polydendrocytes, and endothelia cells resulted in the lack of significantly enriched pathways. This theory is further supported with the highly significant pathways found for glutamatergic neurons as they had the largest number of cell type specific enhancers by far.

Then, to map these enhancers back to their target genes, we used Spearman's correlation analysis to find significantly linked enhancer-gene pairs; these pairs provided more insight into the enhancer-gene landscape of different cell types, revealing that an enhancer does not necessarily regulate the same gene across cell types. From the 4,147 significantly correlated enhancer-gene pairs we found, only 116 demonstrated significant chromatin contact; however, as different cell types have different chromatin arrangements, we believe that this number would be higher if we had Hi-C data at a cell type specific resolution.

Our enhancer-gene pairs also show good consensus across different assays (snRNA-seq and multiome) as well as sample sizes (20 donors to 180 donors), particularly in the upper and lower tails of the density plots, which contain the enhancer-gene pairs we are most interested in exploring. We have demonstrated the reproducibility of our enhancer-gene pairs and their potential for understanding how diseased states or aging could lead to the mis-regulation of certain genes.

We present here the cell type specific enhancer-gene pairs in brain and propose a strategy of mapping enhancers to genes at a cell type specific level using snRNA-seq. We plan to further look at the role of these enhancer-gene pairs in schizophrenia, bipolar disorder, and Alzheimer's. We will also assess the sensitivity and specificity of our framework by running it on other published RNA-seq datasets that have validated enhancer regions[20,21]. We hope that our project opens doors for finding enhancer-gene pairs at a cell type level using the widely available snRNA-seq data.

Methods

1. Biospecimen selection

Frozen blocks containing Brodmann area 46 (BA46) from postmortem human brain tissue were obtained from the Harvard Brain Tissue Resource Center (HBTRC), a biorepository of the NIH NeuroBioBank. Tissue donations were acquired, stored, and distributed in compliance to applicable state and federal guidelines and regulations. Donors with evidence of gross and/or macroscopic brain changes, or with clinical history indicative of cerebrovascular accident or neurological disorders other than schizophrenia were excluded. Twenty donors with both snRNA-seq and snATAC-seq data were selected for the multiome dataset (**Table M1**). For more information on the 180-donor snRNA-seq dataset, please see Ling, E. *et al*[24].

Table M1: Donor Demographics

donor ID	sex	age	post-mortem interval	schizophrenia (sz) status
1	female	82	15.7	control
2	female	79	8.03	sz
3	female	62	10.75	sz
4	male	69	15.95	sz
5	male	65	18.57	control

6	male	64	20.75	control
7	male	60	33.58	sz
8	male	56	18.51	control
9	female	50	20.25	control
10	male	50	33	sz
11	female	88	13.33	sz
12	male	65	20.92	control
13	male	66	20.85	control
14	male	67	23.12	control
15	female	84	22.65	control
16	male	73	20.88	control
17	male	77	25.33	sz
18	female	70	Not Available	sz
19	male	58	26.28	sz
20	female	76	27.66	sz

2. Single-nucleus library preparation and data processing

Nuclei isolation was performed using protocol.io[22]. To minimize technical variation, nuclei from all donors were pooled prior to library construction. From the pooled nuclear suspension, eight single nuclei multiome libraries were created using the Chromium Next GEM Single Cell Multiome ATAC + Gene Expression protocol (version CG000338 Ref F).

snRNA-seq data was processed using the standard Drop-seq (v.2.4.1) workflow[13]. Dropulation[23] (v.2.4.1) was used to demultiplex the pooled samples into individual donors using transcribed SNPs as described in Ling *et al.*[24] Cell types were assigned using scPred[25] (v1.9.2) models trained on data from Ling *et al.* Nuclei that were confidently assigned a single donor and cell type were retained for further analysis (N = 78,809).

snATAC-seq data was processed by Cell Ranger ARC (v2.0.0; reference data version GRCh38-2020-A) to generate alignment files.

3. eRNA and mRNA identification strategy

Usually, H3K4me1 and H3K27ac histone modifications combined with DNase I hypersensitive sites — sites of open chromatin where TFs and cofactors can gain easy access to — are markers for enhancers. Many databases such as ENCODE, FANTOM5, and the Enhancer Atlas have already compiled the CHIP and ATAC-seq data and identified high confidence regulatory regions. We used Zhang et al.[14] annotations of the FANTOM5 enhancers to find enhancers active in brain tissue samples. Another database we used to find brain-related enhancers was the Enhancer Atlas[26]; the Atlas had already categorized their enhancers into 277 different tissues and cell types, so we selected six categories relevant to the human adult brain: astrocyte, BE2C, cerebellum, CNCC, hNCC, NH-A. Enhancers that overlapped with ENCODE's hg38 blacklist region[27] were removed using *bedtools*[28]. In total, we found 29,091 enhancer candidates that had relation to the brain. mRNA transcription regions were defined from GENCODE 46's protein-coding gene annotations[29]. Reads that did not overlap an enhancer, gene, or ncRNA regions were defined as intergenic.

4. Normalizing UMI read counts

Our gene and enhancer DGEs were normalized using Seurat v4.4.2:

```
NormalizeData(dge, normalization.method = "LogNormalize", scale.factor = 10000)
```

To create a metacell for donor *i* and cell type *j*, we summed the normalized read counts for all cells from donor *i* of cell type *j* and divided the sum by the number of cells. Donor metacells were used in our enhancer-gene correlation to reduce noise from individual cells while maintaining overall trends.

5. Shifted region justification and bidirectional scoring system

We exclude shifted regions that overlap known genes and non-coding RNA (ncRNA) annotations to ensure that the shifted regions won't pull more reads than our putative enhancer regions.

To calculate a bidirectional score, the midpoint of the enhancer is often assumed to be the transcriptional start site (TSS). However, as the candidate enhancer boundaries are not confidently defined, we cannot assume the midpoint to be the TSS of the enhancer. Instead, we cut the enhancer into 100 bp bins[12]; our midpoint (P_c) is calculated from the two bins with the respective highest number of forward and reverse stranded reads.

The divergence score is calculated by multiplying the ratio of forward reads mapping to the right of P_c versus the total number of forward reads with the ratio of reverse strands mapping to the left of P_c versus the total number of reverse reads.

Code availability

Code is available upon request.

Acknowledgements

I'd like to express my gratitude to Emi Ling, a postdoctoral fellow in the McCarroll lab; she generated and performed the initial data processing of the BA46 datasets and was very patient in helping me navigate through them.

Heartfelt thanks to Yong Hoon Kim, another postdoctoral fellow in the McCarroll lab, for his lessons on eRNA, introducing the concept of topologically associating domains, and presenting me with the idea of using them in my own project.

Many thanks to Steve McCarroll who proposed the bidirectional transcription analysis and reinforced the need to validate our putative enhancer annotations. He provided much needed guidance for the direction of this project.

Thanks to my mentor, Nicole Rockweiler, who was essential for the conception of this project, everything from background research and helping me find public consortiums, to introducing me to the McCarroll lab snRNA-seq processing pipeline and the Broad Server.

Finally, thanks to Professor Slava Gerovitch and the MIT Primes Computational Biology program for making this project possible.

References

1. Theuns, J. *et al.* Promoter Mutations That Increase Amyloid Precursor-Protein Expression Are Associated with Alzheimer Disease. *Am. J. Hum. Genet.* **78**, 936 (2006).
2. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
3. De Santa, F. *et al.* A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers. *PLoS Biol.* **8**, e1000384 (2010).
4. Ren, C. *et al.* Functional annotation of structural ncRNAs within enhancer RNAs in the human genome: implications for human disease. *Sci. Rep.* **7**, 1–15 (2017).
5. Cichewicz, M. A. *et al.* MUNC, an Enhancer RNA Upstream from the MYOD Gene, Induces a Subgroup of Myogenic Transcripts in trans Independently of MyoD. *Mol. Cell. Biol.* (2018) doi:10.1128/MCB.00655-17.
6. Lam, M. T. Y. *et al.* Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. *Nature* **498**, 511–515 (2013).
7. Li, W. *et al.* Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**, 516–520 (2013).
8. Kaikkonen, M. U. *et al.* Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Mol. Cell* **51**, 310 (2013).

9. eRNAs Are Required for p53-Dependent Enhancer Activity and Gene Transcription. *Mol. Cell* **49**, 524–535 (2013).
10. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
11. Rahnamoun, H., Orozco, P. & Lauberth, S. M. The role of enhancer RNAs in epigenetic regulation of gene expression. *Transcription* **11**, 19 (2020).
12. Kristjánsdóttir, K., Dziubek, A., Kang, H. M. & Kwak, H. Population-scale study of eRNA transcription reveals bipartite functional enhancer architecture. *Nat. Commun.* **11**, 1–12 (2020).
13. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202 (2015).
14. Zhang, Z. *et al.* Transcriptional landscape and clinical utility of enhancer RNAs for eRNA-targeted therapy in cancer. *Nat. Commun.* **10**, 1–12 (2019).
15. Aiden, E. ENCSR165UJN. The ENCODE Data Coordination Center (2022).
16. Lettice, L. A. *et al.* Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proceedings of the National Academy of Sciences* **99**, 7548–7553 (2002).
17. Furlong, E. E. M. & Levine, M. Developmental enhancers and chromosome topology. *Science* (2018) doi:10.1126/science.aau0320.
18. Mora, A., Sandve, G. K., Gabrielsen, O. S. & Eskeland, R. In the loop: promoter–enhancer interactions and bioinformatics. *Brief. Bioinform.* **17**, 980 (2016).
19. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *cels* **3**, 95–98 (2016).
20. Lyu, R. *et al.* Quantitative analysis of cis-regulatory elements in transcription with KAS-ATAC-seq. *Nat Commun.* **15**, 6852 (2024).
21. Yao, L. *et al.* A comparison of experimental assays and analytical methods for genome-wide identification of active enhancers. *Nat Biotechnol* **40**, 1056–1065 (2022).
22. McCarroll, S., Ling, E. & Goldman, M. Village nuclei isolation with myelin removal v1. (2023) doi:10.17504/protocols.io.4r3l22e3xl1y/v1.
23. Wells, M. F. *et al.* Natural variation in gene expression and viral susceptibility revealed by neural progenitor cell villages. *Cell Stem Cell* **30**, (2023).
24. Ling, E. *et al.* A concerted neuron–astrocyte program declines in ageing and schizophrenia. *Nature* **627**, 604–611 (2024).
25. Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q. & Powell, J. E. scPred: accurate supervised method for cell type classification from single-cell RNA-seq data. *Genome Biol.* **20**, 1–17 (2019).
26. Gao, T. & Qian, J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.* **48**, D58–D64 (2019).
27. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Rep.* **9**, 1–5 (2019).
28. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
29. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res.* **49**, (2021).
30. Ge, S. X., Jung, D. & Yao, R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36**, 2628 (2020).