Elucidating the Cell Type-Specific Gene Regulatory Landscape of the Developing Brain in the Investigation of Autism Spectrum Disorder-Related Genes

Abdool Al-Khaledi

Student I.D: 7880103

July 14, 2023

Bioinformatics & Biocomplexity MSc Thesis

Department of Translational Neuroscience, University Medical Center Utrecht Brain Center, UMCU

Supervisor: Onur Basak, PhD

Abstract

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental disorder distinguished by a spectrum of symptoms and severity levels, predominantly impacting social engagement and motor behavior. The study of ASD is significantly complicated by its exceptional heterogeneity, involving over 1000 genes and diverse developmental processes. However, the emergence of single-cell RNA sequencing (scRNA-seq) and the structure of Gene Regulatory Networks (GRNs) offer a promising pathway to navigate this heterogeneity. By combining various ASD-related transcripts under one regulatory mechanism, we can gain a more comprehensive understanding of the disorder. Capitalizing on these tools, this study aims to unite diverse ASD-related transcripts under a common regulatory framework, thereby offering a more holistic understanding of the disorder. Specifically, we sought to identify those cell type-specific GRNs in which ASD-related Transcription Factors (TFs) control a disproportionately greater number of unique target gene interactions compared to other cell types throughout neurodevelopment. We call this subset of unique ASD related TF-gene interactions the 'ASD regulon'. The research identified an enrichment in ASD regulon activity in certain populations of cells within the mouse brain, including Layer 4 and 6 neurons, interneurons, projection neurons, Cajal-Retzius cells, and several glial cell types. We further sought to delineate the effect of time on the activity of ASD regulons by modeling ASD regulon activity as a function of time. The progression of neurodevelopment was found to have a significant effect on the increase in ASD regulon activity in the developing forebrain of mice from prenatal to neonatal period, while a decrease in ASD regulon activity was observed throughout adolescent-adult whole mouse brain development. The study also identified Ctcf, a highly conserved, ubiquitously expressed protein, as a key driver of ASD regulon activity in the enriched cell types. Ctcf was found to orchestrate the expression of ~150-300 genes across the enriched cell types and was solely culpable for the designation of a cell type as ASD regulon enriched. However, these results were not reproducible in humans, highlighting the translational difficulties in investigating ASD in a murine system. The study acknowledges limitations such as the potential overlooking of crucial players in ASD etiology due to the focus on TF driven enrichment in ASD activity. In conclusion, the study developed a GRN reconstruction pipeline that serves as a tool for the investigation of cell type-specific changes in GRNs across the dynamic gene expression landscape of brain development and further identified cell types and time periods which warrant further investigation.

Plain language summary

Autism Spectrum Disorder (ASD) is a disorder of the developing brain that affects social interaction, communication, and behavior. It's a complex condition that involves a multitude of genes - over a thousand, in fact. This genetic complexity makes ASD challenging to study and understand. In our research, we used advanced scientific techniques to delve deeper into the genetic underpinnings of ASD. One of these techniques is single-cell RNA sequencing. This method allows us to examine the activity of individual cells in the brain. Imagine being able to listen in on the 'conversations' that each cell is having. This is essentially what single-cell RNA sequencing allows us to do. It gives us a detailed look at what each cell is doing and how it's behaving, which can provide valuable insights into complex conditions like ASD. We also utilized Gene Regulatory Networks (GRNs). If you think of a cell as a factory, then GRNs are the blueprints that show how different parts of the factory interact and influence each other. In the context of ASD, GRNs can help us understand how different genes interact and contribute to the disorder under one master framework. Our objective was to pinpoint which types of brain cells have a high amount of important ASD genes in their networks. We discovered that certain cells, including specific types of neurons (nerve cells) and glial cells (supportive cells in the nervous system), exhibited increased ASD-related activity. These cells are found in various regions of the brain, including the cortex, which is involved in many complex brain functions, including memory, attention, perceptual awareness, thought & language. We also found that the timing of brain development plays a significant role in ASD. During early brain development, ASD-related activity increased, but this activity decreased as the brain matured into adolescence and adulthood. Moreover, we identified a protein called Ctcf, which regulates gene activity, as a key player in ASD-related activity. Ctcf is like a conductor in an orchestra, guiding and coordinating the activity of various genes. In some individuals with ASD, Ctcf has a much more difficult time doing its job, leading to the symptoms of the disorder. However, when we attempted to apply these findings to humans, we encountered several challenges. The results from our mouse studies did not fully align with what we observed in humans. This highlights the difficulties in translating findings from animal models to human conditions and underscores the need for more human-based research, especially in the context of complicated disorders of the brain. Despite these hurdles, our research led to the development of a new tool for studying ASD - a GRN reconstruction pipeline. This tool allows us to examine how ASD related genes can influence different types of cells. It's like having a map that shows us how ASD changes the landscape of the brain at the cellular level. We believe this tool can be useful in advancing our understanding of ASD and could potentially be used to guide and analyze experiments in the lab.

Acknowledgments

I wish to express my deepest gratitude to the Department of Translational Neuroscience for fostering an environment of inclusivity and exploration. It has been a nurturing ground for a nascent scientist like me, providing invaluable opportunities to embark on this scientific journey.

Special thanks must be extended to my project supervisor, Onur Basak, for his unwavering support and generous guidance. His patience, expertise, and kindness have been instrumental in shaping the project. I am also profoundly grateful to my co-supervisor, Adrien Melquiond, whose mentorship has been invaluable in navigating the scientific and career-related complexities of a bioinformatician.

I would also like to extend my heartfelt appreciation to Elly Hol and the esteemed members of the Hol and Basak Labs. Their wisdom, continuous guidance, and unwavering support have enriched my research experience and academic growth.

The journey to complete this thesis has been an insightful voyage of learning, made possible only by the collective efforts of these individuals, in conjunction with the seminal work cited in this study. I am deeply thankful for their contributions.

Table of Contents

Introduction

Autism Spectrum Disorder

Autism Spectrum Disorder (ASD) represents a multifaceted neurodevelopmental condition that manifests through a range of symptoms affecting social interaction, communication, and behavior [1]. It was first described in detail by two pioneering figures in the mid-20th century. Austrian psychiatrist Leo Kanner was the first to outline the syndrome in 1943, identifying a group of children with profound social and communicative impairments that distinguished them from their peers. In his paper, "Autistic Disturbances of Affective Contact", Kanner detailed the uniqueness of the disorder, differentiating it from known psychiatric or neurological conditions of that time [2]. Autism was first added to the third edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-III) in 1980 as "Infantile Autism" and was classified under the category of "Pervasive Developmental Disorder" [3].

Concurrently to Kanner, across the Atlantic in Austria, psychiatrist Hans Asperger described a similar but distinct condition. His observations, which were characterized by less severe symptoms and the absence of language delays, were later recognized as Asperger's Syndrome and included in the DSM-IV in 1994 [4]. In the latest version; DSM-5, published in 2013, the separate diagnoses of "Autistic Disorder," "Asperger’s Disorder," and other conditions previously categorized under "Pervasive Developmental Disorders" were consolidated into the single diagnosis of "Autism Spectrum Disorder". This change reflected the scientific consensus that these conditions represent points along a continuum of neurodevelopmental disorders, varying primarily in severity and manifestation of symptoms [3-5]. I will further stress this point with a small example; Rett Syndrome, once possibly categorized as an atypical form of autism due to symptom overlap, has since been recognized as a distinct condition following the identification of its unique genetic basis in 1999 (MECP2 dysfunction) [6]. Today, Rett Syndrome and autism are acknowledged as separate disorders, despite sharing some similar symptoms [7]. It is tempting if not reasonable to speculate that over the coming years, certain ‘sub-types’ of autism will also be described, marked by the identification of a specific mechanistic underpinning (or the interplay thereof) which deterministically explains a certain aspect(s) of ASD phenotype(s).

The ASD diagnosis in DSM-5 is now made based on two main symptom domains: 1) persistent deficits in language acquisition and social interaction, and 2) restricted, repetitive patterns of behavior (ticks), interests, or activities [1]. ASD affects about 1% of the population worldwide, a rate which has increased in recent years, and one that is markedly higher in high-income countries [8-9]. Given the complex nature of ASD, it should come as no surprise that it often co-occurs with other neurological and cognitive disorders. ASD patients have a higher likelihood of presenting with Attention Deficit Hyperactivity Disorder (ADHD) (15%), epilepsy (13%), and intellectual disability (22%) [1]. Certain types of cancer have also been associated with ASD. Interestingly in the case of cancer, this increased risk is applicable to ASD Patients with co-occurring Intellectual Disability (ID) or birth defects [10]. Furthermore, ASD exhibits a male bias, with boys being diagnosed 4-5 times more often than girls [11]. Two major hypotheses have been proposed to explain this male predominance.

The first explanation suggests that ASD diagnostic criteria, originally benchmarked on predominantly male populations, might result in an underdiagnosis in girls. This is especially relevant since ASD can present differently in boys and girls [12]. The repercussion is an underdiagnosis of ASD in girls, leading to numerous false negatives. The second hypothesis posits a 'female protective effect.' This concept suggests, for example, that girls have higher baseline gene expression levels in key ASD genes compared to boys, indicating a higher threshold until the mosaic of genetic effects related to ASD can achieve penetrance [13]. Both of these hypotheses highlight the intricate nuances and challenges associated with diagnosing and understanding ASD.

Etiology of ASD

The etiology of ASD remains an active area of research with multiple theories proposed. Current consensus leans towards ASD as a multistage disorder of prenatal development involving several developmental processes with both environmental and genetic components playing significant roles [14]. In terms of environmental factors, prenatal and perinatal complications have been linked to an elevated risk of ASD. For example, maternal infections during pregnancy may increase the risk of the child developing ASD, possibly due to inflammatory responses affecting fetal brain development [15]. Moreover, exposure to toxins has also been implicated in ASD. Though the mechanism of action is not yet fully understood, it's thought that these toxins could interfere with normal neurodevelopmental processes [16]. Furthermore, studies suggest that older parents, particularly fathers, have a higher likelihood of having offspring with ASD, potentially due to increased risk of newly acquired mutations in sperm as men age [17].

The genetic landscape of ASD is characterized by a complex interplay of rare mutations with substantial effects and common variants with more subtle influences. The substantial heritability of ASD is widely acknowledged, as indicated by the high concordance rates observed in monozygotic twins and the increased occurrence of the disorder among siblings of affected individuals [18]. However, the genetic etiology of ASD remains largely unexplained, where only 1-2% of patients have an explainable genetic cause [19].

Significant progress has been made in identifying various genes that contribute to ASD pathogenesis. These genes affect a wide range of biological processes and developmental stages critical for normal neurodevelopment, such as neurogenesis, synaptogenesis and neural network formation [20]. Transcription factors (TFs), in particular, have a profound impact on the pathogenesis of ASD. A prime example of this is the Chd8 gene, which ranks among the most frequently mutated genes in ASD cases. This gene encodes for a protein that is part of the ATP-dependent chromatin-remodeling factors, a group integral to genetic regulation. In their research, Wang et al. demonstrated that a heterozygous knockout of Chd8 in cerebral organoids resulted in defective neural progenitor proliferation and differentiation. Complementing this, Durak et al. revealed in another study that Chd8 knockdown during cortical development led to unusual neuronal morphology and behaviors in adult mice [21-22]. Furthermore, several genes have been identified for the role in aberrant synaptogenic processes in ASD. PTEN for example, which is responsible for approximately 10% of ASD cases accompanied by macrocephaly, has been implicated in heightened microglial activation and synaptic pruning, possibly through its interaction with the mammalian target of rapamycin (mTOR) kinase [23-24].

Neurodevelopmental investigations into ASD have revealed significant alterations in brain morphology, particularly in cortical development. A notable longitudinal study by Zielinski et al. employed Magnetic Resonance Imaging (MRI) to compare cortical thickness in male ASD patients. The study proposed a dynamic model outlining the progression of cortical development in ASD patients, consisting of three main stages [25]:

1. Early childhood, where cortical thickness in ASD children initially mirrors that of typically developing children, but rapidly increases by ages 3-4, indicating unusually fast cortical expansion. Other studies have shown that functional MRI (fMRI) images taken from high-risk ASD patients at 6-months old correctly classified an ASD diagnosis at 2 years old in 57 out of 59 infants [26].
2. A transition from cortical expansion to region-specific cortical thinning, leading to 'pseudonormalization' of cortical thickness trajectories between 8 to 18 years of age.
3. Beginning in early adulthood and extending into middle age, this stage is characterized by reduced cortical thinning in individuals with ASD.

A growing tool-box

Single-cell RNA sequencing (scRNA-seq) and single-nucleus RNA-seq (snRNA-seq) have emerged as transformative tools that allow for the exploration of the molecular underpinnings of ASD at an unprecedented level of detail. By enabling the measurement of gene expression at the level of individual cells/nuclei, single cell/nucleus RNA-seq has revolutionized our understanding of cellular heterogeneity and function. This technology allows us to identify distinct cell types and states based on their unique gene expression profiles, providing a more nuanced view of gene ecxpression compared to bulk RNA sequencing. Bulk RNA sequencing averages gene expression across a multitude of different cell types, potentially masking the unique gene expression profiles of individual cells and obscuring the presence of rare cell types [27]. Recent studies have leveraged single cell omics to investigate ASD at the molecular level. For instance, a study by Velmeshev et al. used snRNA-seq to identify specific cell types in the brain that are associated with ASD, revealing that upper-layer excitatory neurons and cortico-cortical projection neurons in the prefrontal cortex show significant differential gene expression in postmortem ASD tissue compared to controls [28].

Research leveraging cell type-specific gene networks and scRNA-seq in the context of ASD is relatively sparse. However, a notable study by Pang et al. has made significant strides in this area. This study integrated neurodevelopmental disorder (NDD) genetics with scRNA-seq data to examine gene co-expression enrichment patterns of various NDD gene sets. The authors identified a critical convergence point in ASD and epilepsy during midfetal neural progenitor cell development in the cortex. Specifically, gestational week 10 was highlighted as a key period of convergence in ASD and epilepsy related co-expression enrichment patterns and implicated the both radial glia and intermediate progenitor cells [29].

Given the monumental number of genes and related studies implicated in ASD, maintaining a comprehensive record of all the genes involved can be a daunting task. In this regard, organizations like the Simons Foundation Autism Research Initiative (SFARI) are invaluable. SFARI keeps a comprehensive and evolving catalog of ASD-associated genes, curating more than 1,000 genes across three tiers of confidence [30]. Numerous studies have utilized this list in their ASD research, for instance, by overlaying the SFARI gene list over lists of upregulated/downregulated genes, which is exactly what was done in the Velmeshev et al. study described earlier [28].

However, identifying upregulated and downregulated ASD genes and co-expressed clusters, while informative, misses an essential piece of the puzzle: the regulatory relationships among these genes. This information is crucial for pinpointing the genes that are the causal drivers of ASD. For instance, if ten genes are found to be downregulated, it may be that only one of these is actually malfunctioning, and the downregulation of the remaining nine can be attributed to the regulatory interactions of this causal gene. Thus, a more comprehensive architecture is needed to unravel the molecular dynamics of ASD at the regulatory level.

Gene Regulatory Networks (GRNs)

The utility of gene lists like the SFARI gene list are further highlighted when coupled with an architecture which enables us to view interactions between genes. This architecture can take the form of a GRN, which is a directed network graph where the nodes represent transcription factors or genes, and the edges represent regulatory interactions (activatory/inhibitory) [31]. These interactions dictate the level of gene expression within a cell, thereby influencing biological processes and cellular functions. A 'regulon' is a term used in this analysis within the context of GRNs to denote a group of genes that are regulated by the same transcription factor.

The construction and investigation of GRNs has been the focus of numerous studies, and several methodologies have been developed for this purpose. For example, the DecoupleR method relies on the “finger-print” of a TF to infer its activity [32]. The authors behind this method reasoned that the expression of a TF is not informative on its own due to the often-low expression. Given how potent transcription factors are as catalysts, they are often present in relatively small numbers, leading to issues in power when analyzing these TFs and comparing their activity across cells. To tackle this, DecoupleR utilizes a Multivariate Linear Model (MLM) that works in conjunction with DoRothEA to infer TF activities [32]. DoRothEA provides a curated database of known interactions between transcription factors and their target genes, which serves as the foundation for the analysis [33]. Although the DecoupleR method does not explicitly generate a GRN, it can be leveraged in that capacity.

Other tools for GRN reconstruction include PySCENIC, ARACNe, CLR, and GENIE3. PySCENIC uses an ensemble of methods to identify regulons, relying on binding sites and a de novo approach to GRN reconstruction [34]. ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks) uses mutual information to infer regulatory relationships [35], while CLR (Context Likelihood of Relatedness) builds upon this by also considering the context of these relationships [36]. GENIE3 (GEne Network Inference with Ensemble of trees) uses a machine learning approach, specifically an ensemble of decision trees, to predict regulatory interactions [37]. The most recent contribution to this rapidly growing field comes from the developers of the DoRothEA network, who developed CollecTRI. The most notable advantage of this yet unpublished method is the expanded library size and possibility to split TFs into sub-units or complexes [38].

Objective

The objectives of the present study are twofold:

1- Our first aim is to reconstruct cell type-specific GRNs from publicly available and high-quality sc/snRNA-seq data of the developing brain. The reconstruction of these GRNs will provide a comprehensive map of TF-gene interactions within specific cell types.

2- Upon the successful construction of these networks, we intend to leverage them against the SFARI gene list to answer the following key questions:

   a. Are there specific cell types whose GRNs contain a significantly higher number of genes that are uniquely targeted by ASD TFs?

      Reasoning: In the context of gene regulation, a mutation in a TF can have a profound impact on the genes it regulates. If these ASD-related TFs are mutated, it could lead to the improper regulation of their target genes, potentially disrupting normal cellular functions and contributing to the development of ASD. This inference is strengthened in the case where the target gene is exclusively targeted by ASD TFs. We refer to these cell types as 'ASD regulon enriched'. Where an ASD regulon denotes the set of an ASD TF and target genes exclusive to ASD TFs.

   b. Can we model ASD regulon activity as a function of the total network activity & time?

      Reasoning: Understanding the temporal dynamics of ASD-regulon activity could provide valuable insights into the developmental trajectory of ASD, potentially identifying key periods of vulnerability during brain development.

Methods

Data

Several datasets were leveraged in the current analysis; Table 1 provides an overview. These include high resolution developing and adolescent mouse brain atlases [39-40]. As well as data on the developing neo-cortex of the mouse [41] and human cortical development from 2nd trimester until adulthood [42].

Table 1. Mouse and Human Brain Development Data Sets Analyzed in the Study.

Data	Species	Region	Cells	Cell types	Time period
La Manno et al. 2021	Mouse	Whole brain- with possibility to subset.	~350,000	748	E7-E18
Di Bella et al. 2021	Mouse	Neo-cortex	~80,000	24	E10-P4
Zeisel et al. 2018	Mouse	Whole brain	~160,000	265	P16-P60
Velmeshev et al. 2022	Human	Cortex	~350,000	28	2nsd trimester - Adult

Table 1. Summary of the 4 datasets analyzed in this study encompassing both mouse and human brain development. The table presents information on the species, brain regions analyzed, number of cells captured, the variety of cell types identified, and the corresponding time periods or developmental stages covered by each respective dataset.

DoRothEA network & DecoupleR's MLM

DoRothEA is a comprehensive resource that provides a curated database of TF - target gene interactions. Each row in the DoRothEA network represents an interaction, with the 'source' column indicating the TF and the 'target' column indicating the target gene. Each interaction is associated with a confidence level, ranging from A (highest confidence) to E (lowest confidence), and a 'weight' that indicates whether the interaction is activatory (positive weight) or inhibitory (negative weight). It is worth noting that the magnitude of the weight has no clear relation to binding affinity, and is instead used to numerically represent the confidence (A = 1, B = 0.5, C = 0.33) of the interaction. The interactions in DoRothEA are curated from a variety of sources, including literature mining, ChIP-seq peak overlap, and motif enrichment analysis, providing a robust and comprehensive overview of TF-target gene interactions. In total, the DoRothEA network catalogues 1399 TFs and 27,979 target genes across 5 confidence levels.

DecoupleR is a programming library containing an ensemble of computational tools, one such tool is the MLM, which leverages the DoRothEA network to infer TF activities. For each cell in a given dataset, DecoupleR sets up a MLM where the response variable is the observed gene expression level of the target gene, and the predictor variable is the associated TF weight, as provided by DoRothEA. The MLM is then fitted to the gene expression data for each cell, with the goal of finding the coefficient that minimizes the sum of the squared residuals across the system of linear equations describing the expression of the target genes. We consider this coefficient the activity of that TF for that specific cell.

Let’s consider a single cell. For this cell, we have expression data for 100 target genes of the transcription factor P53. Let's denote these gene expression levels as G1, G2, ..., G100. Let’s assume that the binding weight of P53 for all of these genes is the same, and is given as +0.5.

In the MLM used by DecoupleR, the expression level of each target gene is modeled as a linear function of the activity of the transcription factor P53. This can be represented as follows:

G1 = 0.5 * Activity_P53 + e1

G2 = 0.5 * Activity_P53 + e2

...

G100 = 0.5 * Activity_P53 + e100

Here, Activity_P53 represents the inferred activity of P53, and e1, e2, ..., e100 are the error terms for each gene. These error terms represent the difference between the observed gene expression levels and the levels predicted by the model. The goal of the MLM is to find the value of Activity_P53 that minimizes the sum of the squared error terms (i.e. $(e1)^2$ $+(e2)^2$ …. + $(e100)^2$). This operation is performed for each TF (one TF at a time) and for each cell (one cell at a time). DecoupleR outputs these TF activities in a standard anndata object format (CellsxTFs) where the rows represent cells, the columns represent TFs with an identified foot-print in the original CellxGene matrix and the entries contain the model coefficients, which represent the TF activities.

Cell type specific GRN Pipeline overview

Our bioinformatics pipeline aims to construct cell type-specific networks using a sc/snRNA-seq CellxGene matrix as input. The pipeline proceeds through a series of well-defined steps.

Initially, we load the data and filter out cells with percent mitochondrial gene counts greater than 20%. Mitochondria, ribosomal, and blood genes are subsequently removed due to their potential to introduce noise and interfere with downstream analysis. We further filter out cells labeled 'Undefined' or 'Low quality'.

At this early juncture, we can introduce a 'for loop' for individual analysis across age categories or bypass the loop to analyze cells/cell-types across all time periods together. Irrespective of the approach, the pipeline filters cell types with less than 5 cells, cells that express insufficient genes (min_genes = 200) and genes detected in only a few cells (min_cells = 5). Following this, we normalize the data with the 'normalize_total' function from the Scanpy package and log-transform it using the 'log1p' function.

To initialize the MLM, we use the 'get_dorothea ' function from the DecoupleR package to load the DoRothEA network, specifying either 'Mouse' or 'Human' as the organism depending on the dataset. We limit the retrieval to the top 3 confidence interaction levels ('A', 'B', 'C') to reduce false negatives, although we acknowledge the potential for missed interactions. We then calculate the TF activity for each cell using the 'run_mlm' function.

Subsequently, the pipeline extracts inferred TF activities from the MLM output to form a CellxTFs anndata object. Quality control on the activity matrix removes TFs absent in the original data, TFs present in fewer than 20 cells, and cells expressing no TFs. Using the 'summarize_acts' function, the average TF activities are calculated across cell types. This function provides a parameter, min_std, which enables the specification of a minimum standard deviation of TF activity across cells. We set min_std to 0 to retain all active TFs and perform our own TF filtering.

The next step involves performing cutoffs for TF activity, TF expression, and gene expression. We calculate the 25th percentile of non-zero values in the relevant matrices to achieve this. We designate a TF as 'active' in a cell-type if it surpasses both expression and activity cutoffs. Similarly, we introduce a gene expression cutoff, where genes that do not meet the 25th percentile of expression in their cell-type are excluded from the target gene pool for the network reconstruction.

Using the list of active TFs and genes for each cell type, we customize the DoRothEA network per cell type. TF/gene combinations that meet the cutoff criteria become part of the cell type-specific network. We only include activatory interactions in our network, this is due to the fact that inhibitory relationships are less straight-forward to infer, as the ‘magnitude’ of TF activity will ultimately depend on how high the expression of the gene(s) is/are. We note that positive interactions represent the vast majority of interactions in the DoRothEA network.

Finally, the pipeline applies additional filtering to the constructed networks, excluding TFs with fewer than 5 interactions in a specific cell type. This ensures that the retained interactions mirror the output of the MLM, which only considers TFs targeting at least 5 genes. The pipeline outputs these cell type-specific networks as CSV files, ready for further downstream analysis and visualization.

Note: Due to the large temporal gap in sampling human cortical development (2nd trimester-Adult), the large number of cells in this dataset, and relatively low cell-type resolution across this long developmental period. The human cortical dataset was only analyzed one time point at a time.

Network metrics

In order to assess ASD related regulatory activity; a set of metrics pertaining to the activity of ASD genes/TFs in each cell type specific network was formulated to aid downstream analysis. These include 3 main metrics:

1- GRN activity:

This metric reflects the overall activity of the GRN within each cell type. It is determined by quantifying the number of edges in the network (or rows present in the network data frame). The purpose of this metric is to capture the comprehensive activity of the GRN.

2- ASD regulon activity

This metric is calculated by summing the edges where: A TF present in the SFARI gene list (i.e., ASD TF) is targeting a gene, which is not targeted by a non-ASD TF. This metric was devised to determine if ASD TFs are more active in certain cell-types. The implication with maintaining edges only targeted by ASD TFs is to more confidently infer an effect on these target genes in the case of a dysfunction in an ASD TF, since these genes are not targeted by non-ASD TFs.

3- ASD activity:

This metric is calculated by summing the interactions where either the 'source' or 'target' is present is in the SFARI gene list and represents a comprehensive view of the ASD related TF-gene interactions.

As well as some supplementary metrics which will be described as well:

ASD genes: This metric is calculated by summing the interactions where the 'target' is an ASD gene. This metric comprehensively represents the involvement of ASD genes in each cell type specific GRN.

Free-floating ASD genes: This metric is calculated by summing the interactions where an ASD gene is connected to a non-ASD TF. It was developed as a measure to help determine if certain cell types expressed a higher number of ASD genes bound to non-ASD TFs. This is an example of a measure which can be used to investigate non-ASD TFs in the context of ASD.

Statistical testing

Mann-Whitney U test

All statistical testing for differences between two groups were performed using the Mann-Whitney U test. The Mann-Whitney U test, also known as the Wilcoxon rank-sum test, is a non-parametric statistical test that is used to compare two independent samples to determine whether there is a significant difference between their distributions. In our case, we also utilize this test for its ability to handle parametric as well as non-parametric data, such that we perform the same test on our multiple datasets.

The test first ranks all the observations from both groups together, from smallest to largest. Then, for each group, the ranks of the observations are summed. The test statistic, U, is calculated based on these rank sums. The U statistic represents the number of times a value from the first group is less than a value from the second group. Moreover, the null hypothesis states that the distributions of both groups are equal, meaning that there is a 50% chance that an observation from one group is less than an observation from the other group. If the U statistic is significantly different from what would be expected under the null hypothesis, then the null hypothesis is rejected, indicating a significant difference between the two groups ($p<0.05$).

Median Absolute Deviation (MAD) test

In order to identify and excise outlying cell types in GRN activity, we employed the non-parametric MAD test. The MAD test is a robust measure of statistical dispersion. It operates by calculating the median of absolute differences from the data's median. A threshold of 3 MAD units was used in this study, such that any cell type with a MAD score of more than 3 was considered an outlier and subsequently removed from the analysis. This procedure was important to ensure that we accounted for cell types with over/under sampled GRN activity which, if not addressed, could potentially skew our downstream analyses' results. For instances where each time point was evaluated individually, we did not employ outlier detection methods. This was primarily due to the inherent variability in network sizes across different developmental stages, making outlier detection less straightforward in these cases.

Shapiro-wilk test

Normality of the data was assessed in all cases utilizing the Shapiro-wilk test for normality. The null hypothesis of the test is that the data presented is drawn from a normal distribution. If p-value < 0.05 the null hypothesis is rejected, suggesting that the data does not follow a normal distribution.

Gene set enrichment analysis (GSEA)

EnrichR is a widely used tool for GSEA and is accessible through the GSEAPY python library [43-45]. GSEA identifies whether predefined sets of genes show statistically significant differences between two biological states. In the context of EnrichR, these predefined sets of genes are contained within various libraries, the one used for the purposes of this study is the Elsevier Pathway Collection. In the case of this collection, the gene sets represent various biological protein pathways curated from the scientific literature. When a list of genes (for example, differentially expressed genes from an experiment) is input into EnrichR, an enrichment test is performed against the various gene sets in the chosen library. This is done by calculating a p-value for each gene set, which represents whether the given overlap between the input genes and the gene set can be explained by chance alone. The resulting p-values are then adjusted for multiple testing using the Benjamini-Hochberg (BH) correction to control for the False Discovery Rate (FDR) given the large number of tests performed. In all cases GSEA was conducted on a regulon-by-regulon basis where a gene set is defined as an ASD TF and that TFs exclusive target genes, while the BH correction was applied to the p-values of each tested regulon separately.

Ordinary least squares regression (OLS)

In this study, we employed Ordinary Least Squares (OLS) regression to estimate and control for the effects of GRN activity and time on ASD regulon activity. An advantage of the regression analysis is that it further isolates the component of our predictor variable which cannot be explained by GRN activity or time.

The implementation of OLS in this study was done using the Python library statsmodels. The dependent variable was ASD regulon activity, and the independent variables were the GRN activity and time. The data was structured in long format, with each row representing a unique combination of cell type and time point. It should be noted that for each dataset, the 'time' variable was recoded for use in the regression analysis, for example 'E10': '1', 'E11': '2', 'E12': '3' … etc.

The model was further specified with the formula 'ASD regulon activity ~ GRN activity + time'. Lastly, the model was then fitted to each dataset using robust covariance estimation (cov type='HC3'). This estimator was used to provide robust standard errors in the presence of heteroscedasticity, which refers to the situation where the variability of the error term in a regression model is not constant across all levels of the independent variables. This is included to aid in the robustness of the regression analysis across different data sets.

Results

Distributional analysis.

Developing mouse forebrain atlas

Analysis of the forebrain section of the La Manno developing mouse brain atlas began with examining the distributions of GRN activities (Fig. 1(a)), ASD activities (Fig. 1(b)), and ASD regulon activities (Fig. 1(c)) across the 469 identified cell types in the dataset.

Fig. 1

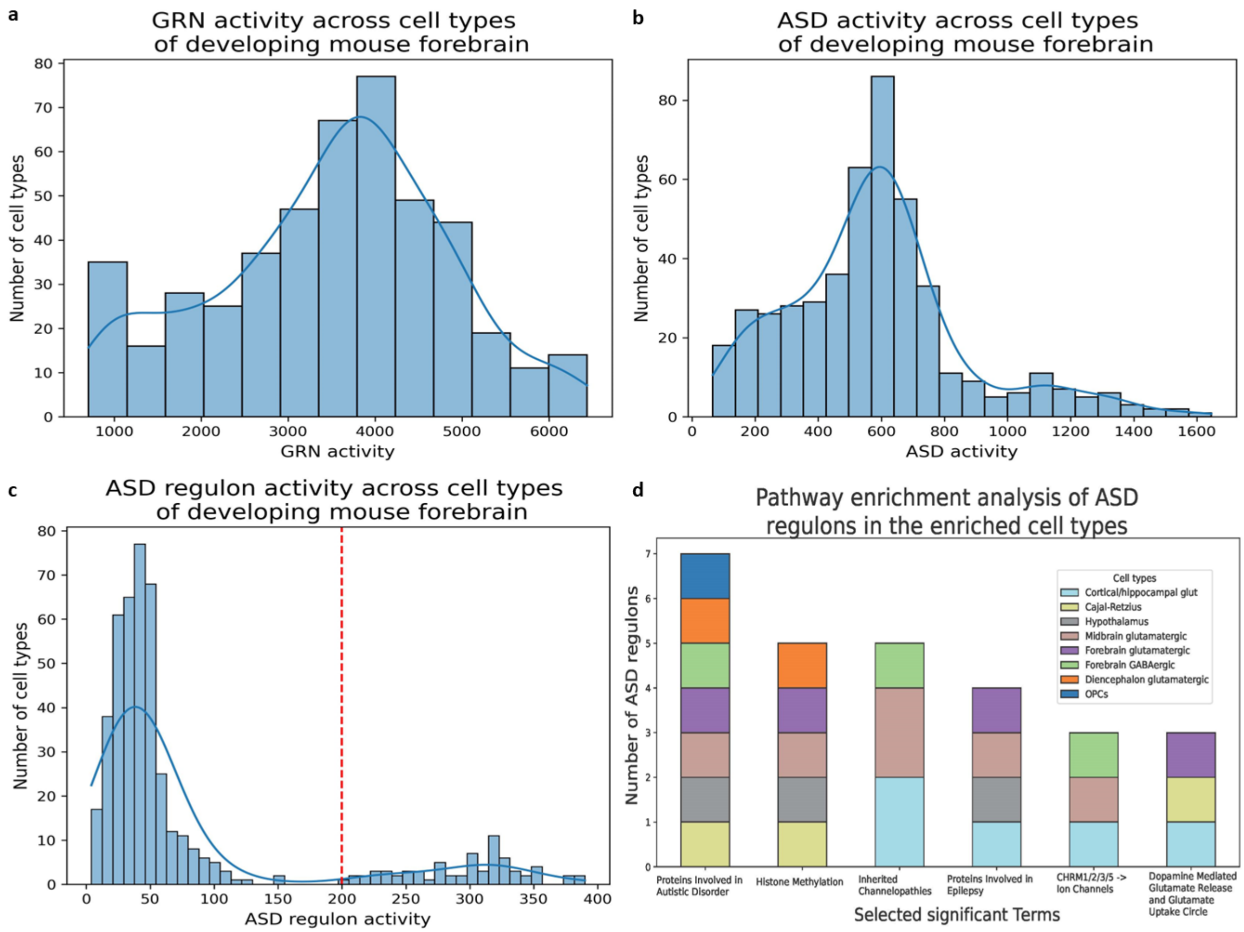

Figure 1. Distributional analysis of the forebrain section of the developing mouse brain (E9-E18) and subsequent GSEA of selected cell-types. a) A histogram depicting the distribution of GRN activity across 469 identified cell types. b) Histogram of ASD activity across cell types. c) Histogram depicting the distribution of ASD regulon activity. A cutoff point of 200 (red line) separates group 1 (high ASD regulon activity) and group 2 (low ASD regulon activity). d) GSEA analysis of ASD regulons on a representative sample of group 1 cell-types. The y-axis represents the number of regulons which returned a significant enrichment for the biological process described on the x-axis.

These distributions were found to be non-normal. A qualitative analysis identifies a relatively homogenous distribution in GRN activity, which is notably perturbed by the formation of a longer tail when examining ASD activity. The distribution of ASD regulon activity results in the formation of a second-peak, with 69/469 cell types exhibiting a markedly higher ASD regulon activity after a period of distributional quiescence.

The cell types contributing to the formation of the second peak were categorized into group 1 and compared with the rest, which we classified as group 2. Mann-Whitney U tests were performed to analyze differences in ASD regulon and GRN activity between these groups. The results demonstrated a significant difference (p-value = 3.2e-40) in ASD regulon activity with group 1 having larger activity levels than group 2. However, no significant difference was detected in terms of GRN activity between the two groups (p-value = 0.06865). Additionally, there was no significant difference in the distribution of free-floating ASD genes between these two groups. Further analysis revealed that group 1 had larger quantities of unique ASD TFs, ASD genes, and ASD activity compared to group 2. Enriched cell types were mainly glutamatergic and GABAergic in nature. We also identified 3 Cajal-Retzius clusters and a single instance of Oligodendrocyte Precursor Cells (OPCs). GSEA was conducted on each ASD regulon on a representative sample of the enriched cell types ('Neur521’: Cortical/hippocampal glutamatergic, ‘Neur677’: Cajal-Retzius, ‘Neur698': Hypothalamus, 'Neur706': Midbrain glutamatergic, ‘Neur718': Forebrain glutamatergic, 'Neur748': Forebrain GABAergic, 'Neur794': Diencephalon glutamatergic, and ‘OPC5’: OPCs). The results were extensive, with hits for several biological processes involving cell signaling, proliferation, cancer and various neurological conditions. The 5 most frequently returned hits across all queried regulons include: Local Estrogen Production in Endometriosis, Toll-like Receptors in Sterile Inflammation, TGFBR -> ATF/GADD/MAX/TP53 Signaling, Proteins Involved in Autistic Disorder and EGFR/ERBB3 -> MEF/MYOD/NFATC/MYOG Signaling. A set of selected biological processes relevant to the current study and their prevalence across the enriched cell types are presented in Figure 1(c). Ctcf was identified as targeting ~200 genes when it was present in the network, whereas most ASD TFs had less than 10 interactions with target genes in the cell type specific networks (see supplementary materials 1 for details).

Developing mouse neocortex

Subsequent examination shifted focus to the developing mouse neo-cortex (Fig. 2). Shapiro-wilk test on the distribution of GRN activity followed a normal distribution (p-value=0.0536) (Fig. 2(a)). However, the distribution of ASD activity and ASD regulon activity failed the normality test (Fig. 2(b-c)). As opposed to a longer tail, plotting ASD activity in this dataset separates the data into shallow peaks. Furthermore, we observed that 12 out of 22 cell types contribute to the formation of a clearly distinct second peak in the distribution of ASD regulon activity. These cell types, listed in the GSEA (Fig. 2(d)) span both early and late stages of prenatal cortical development, including Neural Progenitors (NP), Cajal-Retzius cells, deep layer cortical projection neurons, and a single instance of oligodendrocytes. Using the same group designation as in the forebrain section, we performed Mann-Whitney U tests on these two groups. Once again, group 1 (cell types beyond the red line forming the second peak) exhibited a significantly larger ASD regulon activity compared to group 2, as evidenced by a p-value of 8.57e-05. Similar to the forebrain section, no significant difference was found in GRN activity (p-value = 0.575) between the two groups.

Fig. 2

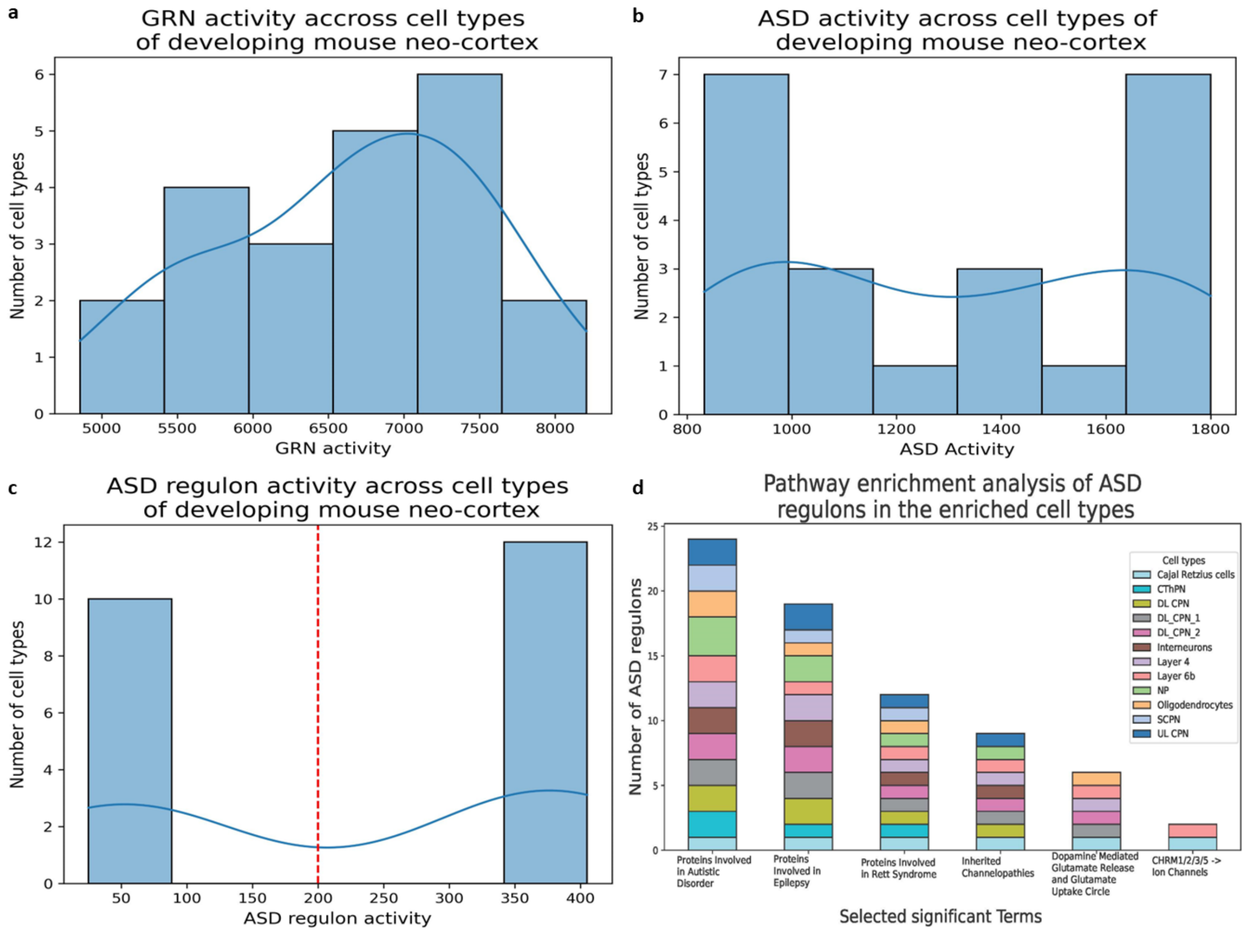

Figure 2. Distributional analysis of the developing mouse neo-cortex (E10-P4) and subsequent GSEA of enriched cell-types. a) A histogram depicting the distribution of GRN activity across 22 identified cell types. b) Histogram of ASD activity across cell types. c) Histogram depicting the distribution of ASD regulon activity. A cutoff point of 200 (red line) separates group 1 (high ASD regulon activity) and group 2 (low ASD regulon activity). d) GSEA analysis of ASD regulons on a representative sample of group 1 cell-types. The y-axis represents the number of regulons which returned a significant enrichment for the biological process described on the x-axis.

Additionally, the distribution of free-floating ASD genes and the number of ASD genes did not show significant variation between the groups. Further exploration revealed that group 1 had significantly higher unique ASD TFs and ASD activity than group 2. GSEA was conducted on ASD regulons of all 12 enriched cell types (Cajal-Retzius cells, Interneurons, Layer (L) 4 and 6b neurons, Neural Progenitors (NP), oligodendrocytes and several projection neurons clusters including: Cortico-Thalamic Pojection Neurons (CThPN), Deep-Layer Cortical Projection Neurons (DL CPN), Striato-Cortical Projection Neurons and lastly Upper-Layer Projection Neurons (UL CPN). The most frequently returned biological processes include: Proteins Involved in Autistic Disorder, Proteins Involved in Epilepsy, Myostatin-IGF1 Crosstalk in Skeletal Muscles, WNT Canonical Signaling Activation in Cancer, and WNT Canonical Signaling. Ctcf was again identified as a driver of enrichment in ASD regulon activity (supplementary materials 2).

Developing mouse whole brain atlas

A holistic examination of the La Manno developing mouse brain which combines data for the developing forebrain, midbrain and hindbrain regions demonstrated non-normal distributions of GRN activity, ASD activity, and ASD regulon activity across 661 cell types, shown in Figure 3.

Fig. 3

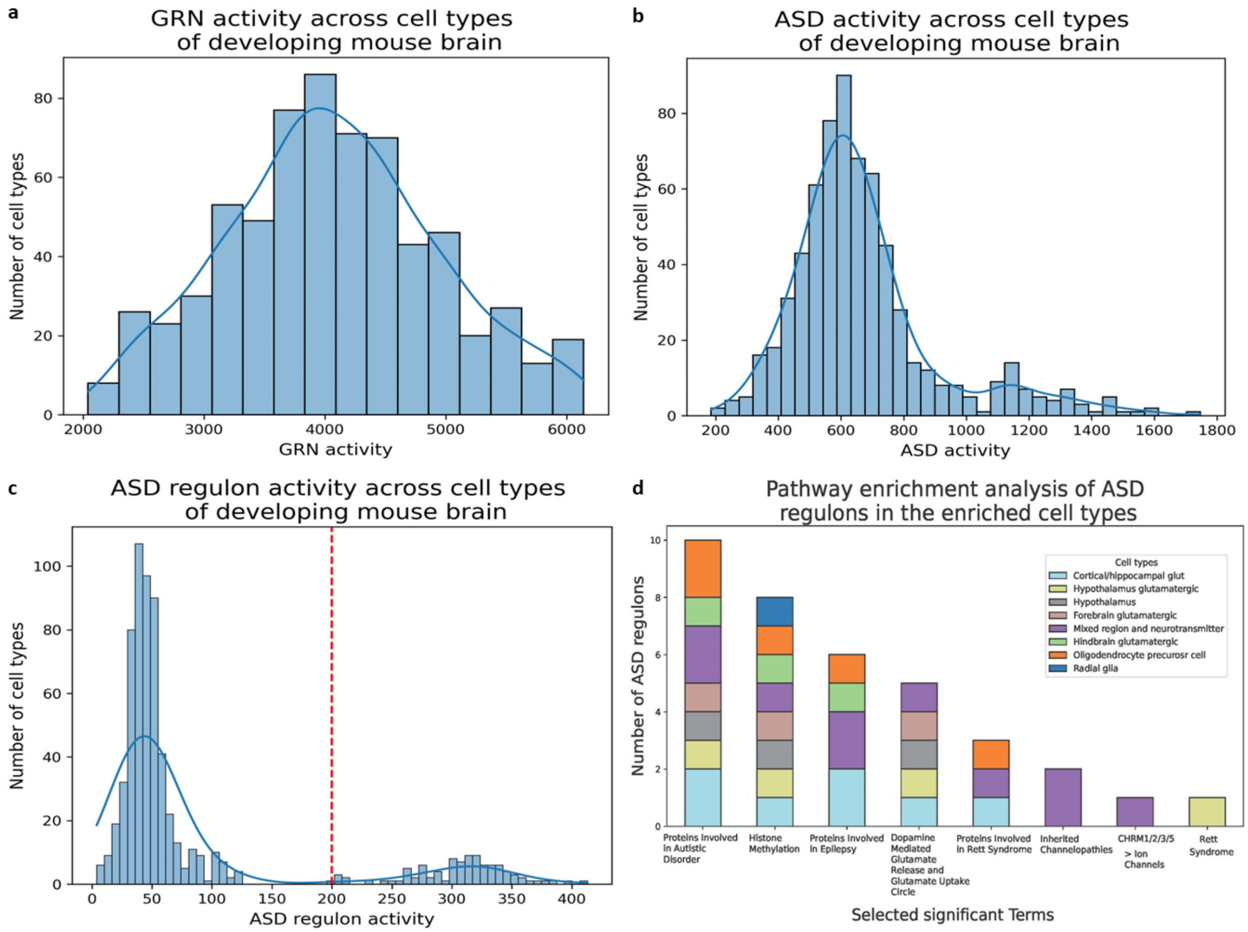

Figure 3. Distributional analysis of the developing mouse whole brain (E7-E18) and subsequent GSEA of selected cell-types. a) A histogram depicting the distribution of GRN activity across 661 identified cell types. b) Histogram of ASD activity across cell types. c) Histogram depicting the distribution of ASD regulon activity. A cutoff point of 200 (red line) separates group 1 (high ASD regulon activity) and group 2 (low ASD regulon activity). d) GSEA analysis of ASD regulons on a representative sample of group 1 cell-types. The y-axis represents the number of regulons which returned a significant enrichment for the biological process described on the x-axis.

It was observed that 95 out of 661 cell types were responsible for the formation of a second peak in the distribution of ASD regulon activity which qualitatively resembles the distribution of the forebrain section described earlier. Most notably by the presence of a relatively narrow distribution containing hundreds of cell types, a period of distributional quiescence and finally a shallower and wider second peak. The analysis of the whole brain revealed an enrichment of non-forebrain-specific cell types, including those from both midbrain and hindbrain regions. Similar to previous analyses, cell types forming a second peak (ASD regulon activity >200) were classified into group 1 and compared against the remaining cell types, denoted as group 2. Mann-Whitney U tests highlighted a significant difference in ASD Regulon activity between the two groups (p-value = 5.98e-55), with group 1 exhibiting greater activity than group 2. However, as seen in the forebrain and neo-cortex analyses, no significant difference was observed in GRN activity between the two groups (p-value = 0.867). In terms of the distribution of free-floating ASD genes, no significant difference was found between the groups (p-value = 0.098). Additional analyses showed that group 1 had significantly larger quantities of unique ASD TFs, ASD genes, and ASD activity compared to Group 2.

Due to the large number of identified cell types, similar to the forebrain section of the developing mouse brain atlas; a representative sample of cell-types was chosen for GSEA, these include: 'Neur523': Cortical/hippocampal glutamatergic, 'Neur649': Hypothalamus glutamatergic, 'Neur677': Hypothalamus, ‘Neur759’: Hindbrain glutamatergic, ‘Neur721’: Forebrain glutamatergic, ‘OPC5’: OPCs, ‘RglF2’: Radial glia and ‘Neur730”; Mixed region and neurotransmitter. The most frequently returned significantly enriched biological processes of the ASD regulons in group 1 include ERK5/MAPK7 Signaling, TGFBR -> ATF/GADD/MAX/TP53 Signaling, EGFR/ERBB3 -> MEF/MYOD/NFATC/MYOG Signaling, IGF1 Role in Muscle Hypertrophy, IGF1R -> MEF/MYOD/MYOG Signaling. Lastly, the presence of Ctcf alone, similar to the previous results, was sufficient to drive ASD regulon enrichment in this data (supplementary materials 3).

Adolescent mouse brain atlas

Investigation into the Zeisel adolescent mouse brain mimicked the non-normal distributions of GRN activity, ASD activity, and ASD regulon activity observed in previous analyses. Here, a pronounced second peak in the distribution of ASD regulon activity was identified, represented by 160 out of the 257 cell types. Similar to the whole brain analysis, the adolescent mouse brain also demonstrated an enrichment of non-forebrain-specific cell types, encompassing both midbrain and hindbrain regions, the results of which are shown on the following page (Fig. 4).

The cell types forming the second peak (group 1) were examined and compared with the remaining cell types (group 2). As expected from the previous analyses, Mann-Whitney U tests revealed a significant difference in ASD regulon activity (p-value = 3.86e-41), with group 1 again demonstrating larger activity levels. No significant difference in GRN activity (p-value = 0.269) was detected between the groups. Similarly, the distribution of free-floating ASD genes did not vary significantly (p-value = 0.509). Further comparative analysis revealed that group 1 had larger quantities of unique ASD TFs, ASD genes, and ASD activity than group 2. Consistent with all previous analyses, Ctcf was identified as the main driver of this enrichment (supplementary materials 4).

Fig. 4

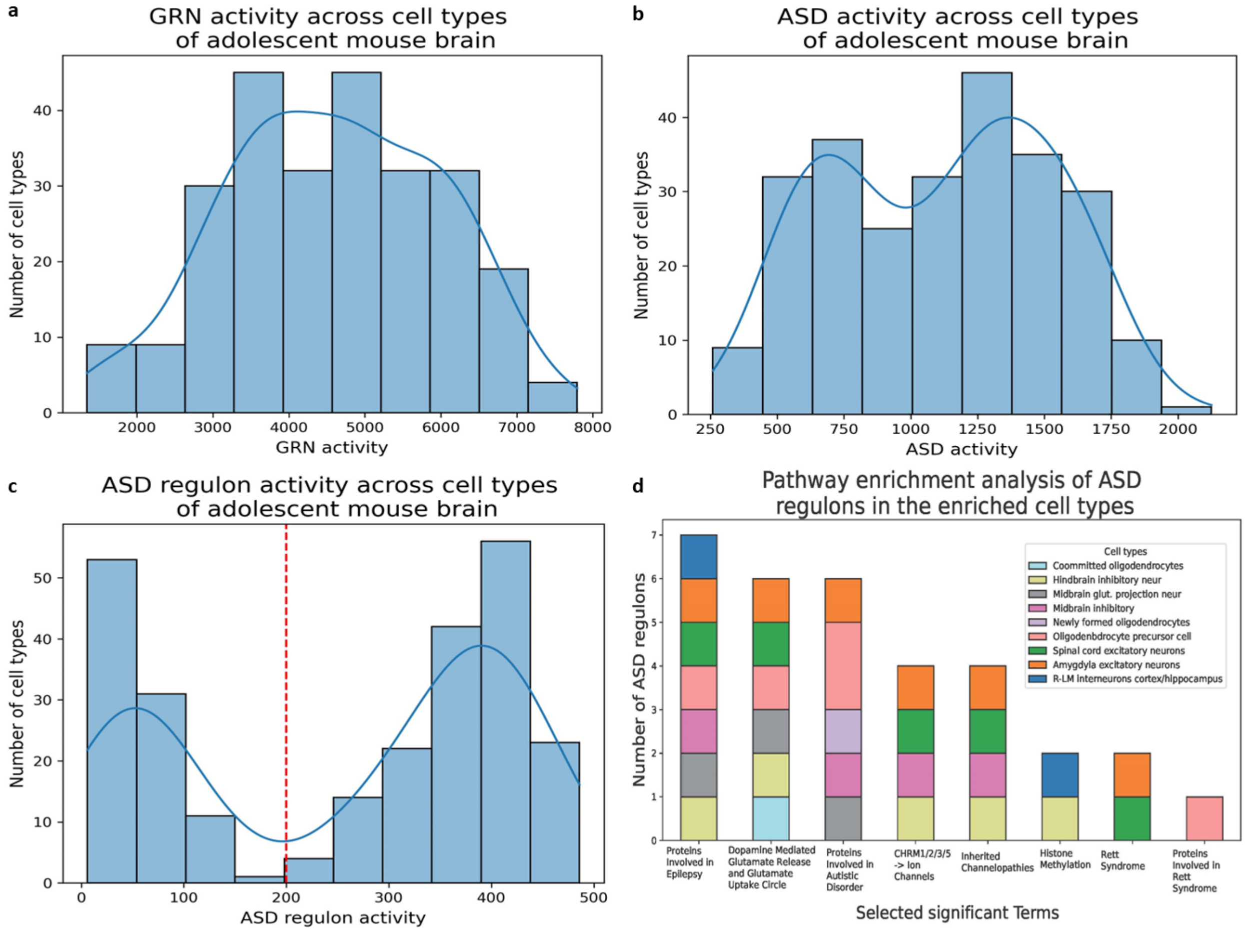

Figure 4. Distributional analysis of the developing adolescent mouse whole brain (P19-P60) and subsequent GSEA of selected cell-types. a) A histogram depicting the distribution of GRN activity across 257 identified cell types. b) Histogram of ASD activity across cell types. c) Histogram depicting the distribution of ASD regulon activity. A cutoff point of 200 (red line) indicates the cutoff point between group 1 (high ASD regulon activity) and group 2 (low ASD regulon activity). d) GSEA analysis of ASD regulons on a representative sample of group 1 cell-types. The y-axis represents the number of regulons which returned a significant enrichment for the biological process described on the x-axis.

GSEA was performed on a representative sample of group 1 cell types including committed, precursor and newly formed oligodendrocytes ('COP2', 'OPC', 'NFOL2'), hindbrain and midbrain inhibitory neurons ('HBINH9', 'MEINH5'), excitatory neurons of the spinal cord and amygdala ('SCGLU2', 'TEGLU22'), midbrain projection neurons ('MEGLU14') and cortex/hippocampus interneurons ('TEINH10'). Finally, the five most frequently enriched biological processes elucidated through the GSEA include Androgen Receptor/Akt Signaling, Alzheimer's Disease, beta-Catenin/Androgen Receptor Signaling in Prostate cancer, Clear Cell Ovarian Carcinoma, WNT Canonical Signaling Activation in Cancer.

Human cortical development

The Velmeshev et al. dataset, composed of human cortical development samples, required a unique analytical approach due to the considerable temporal gap in sampling the human cortex and the relative lack of cell type resolution (28 cell types) given the number of cells (n = 349,312). Accordingly, this dataset was analyzed one time point at a time, revealing a relatively consistent distribution of GRN activity across all time periods, with the exception of the 1-2 years stage, which demonstrated a leftward skew (Fig. 5).

Fig. 5

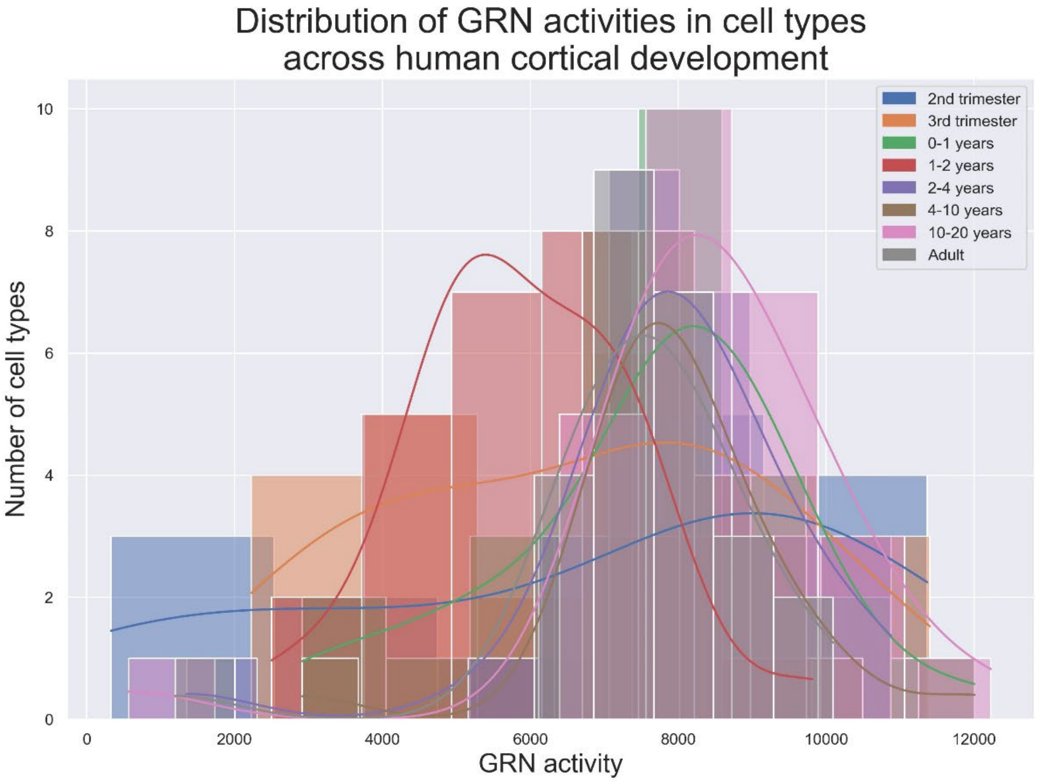

Figure 5. Grouped histogram of cell type specific GRN activity of the developing human cortex (2nd trimester – Adult) across 28 cell types. Each color represents the cell type specific distribution of GRN activity in a specific developmental time point. GRN activity is calculated by counting the total number of edges in each cell type specific network.

Normality tests were conducted for GRN activity across cell types at each time point. These revealed a normal distribution during the 2nd trimester (p-value = 0.125), 3rd trimester (p-value = 0.222), 0-1 years (p-value = 0.322), 1-2 years (p-value = 0.268), 2-4 years (p-value = 0.371), 10-20 years (p-value = 0.251). and the adult stage (p-value = 0.904). However, the distribution was not normal during the 4-10 years period (p-value = 0.032). The divergence from normality during this period suggests the presence of outliers which should be further assessed. In contrast, normality tests for ASD regulon at each time point revealed non-normal distributions in all cases except the 2nd trimester (p-value = 0.298) and 0-1 years (p-value = 0.103) periods. Plotting ASD activity and ASD regulon activity resulted in a rightward skew in the distribution with no discernable cutoff, seen in Fig. 6(a). In contrast, our analyses of mouse data further demonstrate a bi modal distribution when considered one time-point at a time (Fig. 6. b-d).

Fig. 6

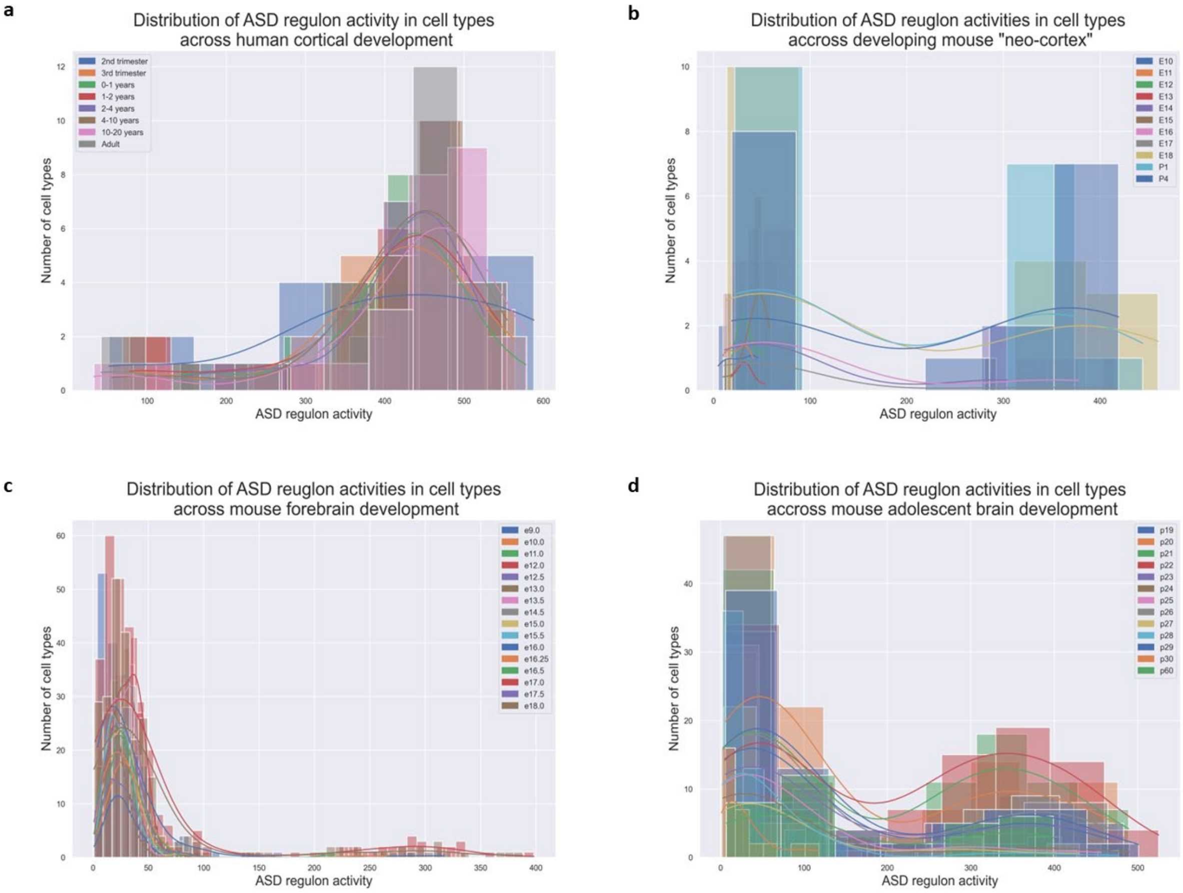

Figure 6. Grouped histogram of cell type-specific ASD regulon activity across: a) Human cortical development (2nd trimester – Adult), b) Mouse neo-cortex development (E10-P4), c) Mouse cortical development (E9-E18), d) Mouse adolescent mouse brain development (P19-P60). Each color represents the cell type specific distribution of ASD regulon activity in a specific developmental time point.

Regression analysis.

Developing mouse forebrain atlas

An OLS regression analysis was conducted on the cell type-specific GRN data from the forebrain section of La Manno et al. developing mouse brain atlas. The ASD regulon activity was modeled as a function of GRN activity and time, the results of which are illustrated in Figure 7.

Fig. 7

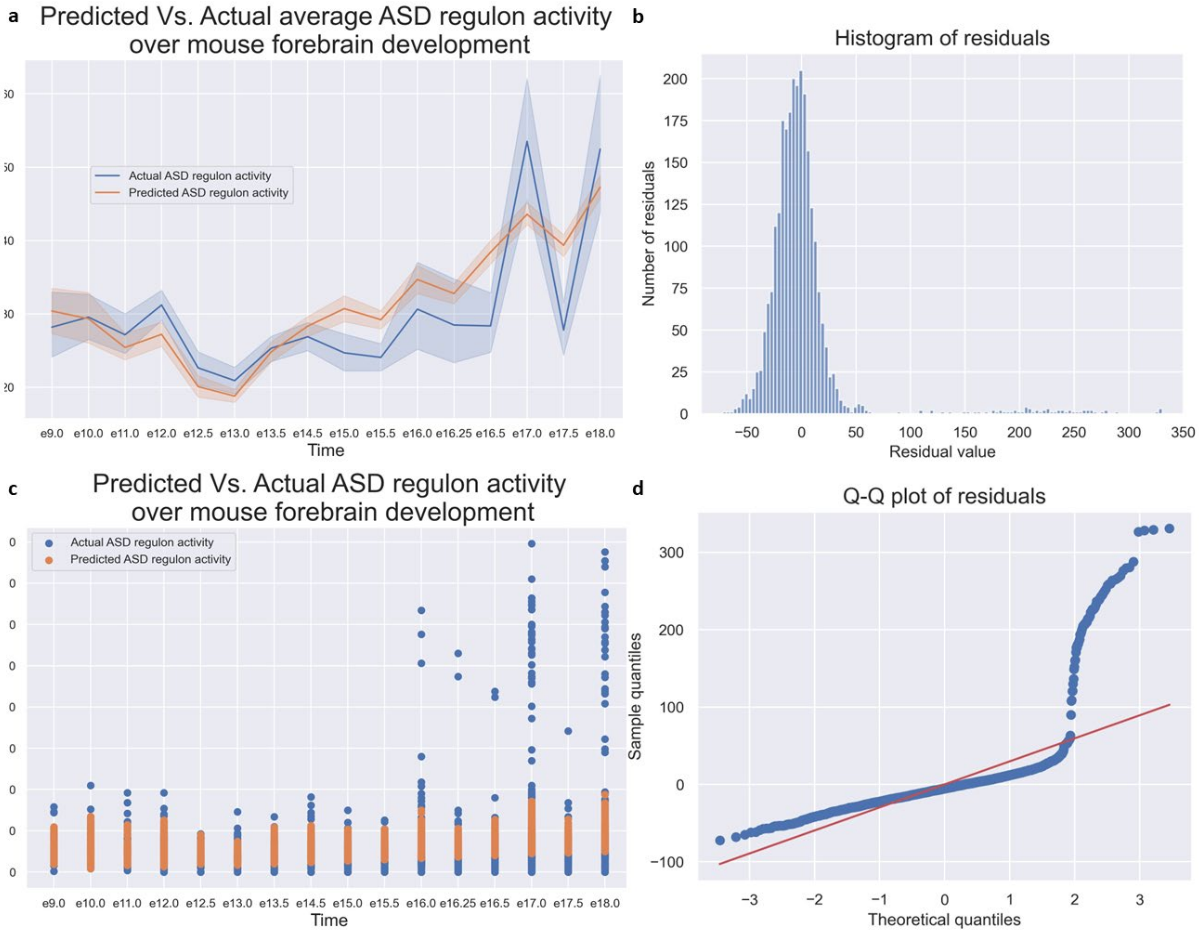

Figure 7. Regression analysis of ASD regulon activity as a function of GRN activity and time in the developing mouse forebrain (E9-E18). (a) Average ASD regulon activity (blue) and model-predicted activity (orange) over mouse brain development stages. (b) Histogram showing distribution of residuals from the OLS regression. (c) Marker plot contrasting actual and model-predicted ASD regulon activity for each cell type at various time points. (d) QQ plot of residuals, with red line denoting a theoretical normal distribution. The model was implemented using statsmodels with HC3 covariance.

The model had an Adjusted R-squared value of 0.105, indicating that about 10.5% of the variance in ASD regulon activity could be explained by the GRN activity and time. The F-statistic for the model was 53.64, with a probability value (Prob F-statistic) of 1.51e-23, significantly less than 0.05, indicating the model's overall significance, i.e., it performed better than a model with random variables. A total of 2557 observations were analyzed in this model. The regression coefficient for GRN activity was 0.0107, with a corresponding p-value of 0.00, indicating a significant effect of GRN activity on ASD regulon activity. Similarly, the coefficient for time was 1.6507, with a p-value of 0.00, signifying that the effect of time on ASD regulon activity was positive and significant. Regarding the residuals, they were found not to be normally distributed and exhibited a bias towards very positive values. This implies that the ASD Regulon activity was generally under-predicted in these cell-type/time-point observations.

Developing mouse neocortex

The OLS regression analysis on the developing neo cortex returned an adjusted R-squared value of 0.216, indicating that approximately 21.6% of the variability in ASD Regulon activity was accounted for by the GRN activity and time. The fit and residuals are presented below in Figure 8. The F-statistic was 23.13, and the Prob F-statistic was significant at 2.56e-9 (<0.05). The analysis incorporated a total of 133 observations. The regression coefficient for GRN activity was 0.02, with an associated p-value of 0.837 (>0.05), making the effect of GRN activity on ASD Regulon activity inconclusive based on this model. In contrast, the time coefficient was substantial at 21.3679, with a p-value of 0.00, demonstrating a statistically significant effect of time on ASD regulon activity. Lastly, the residuals were found to be non-normally distributed.

Fig. 8

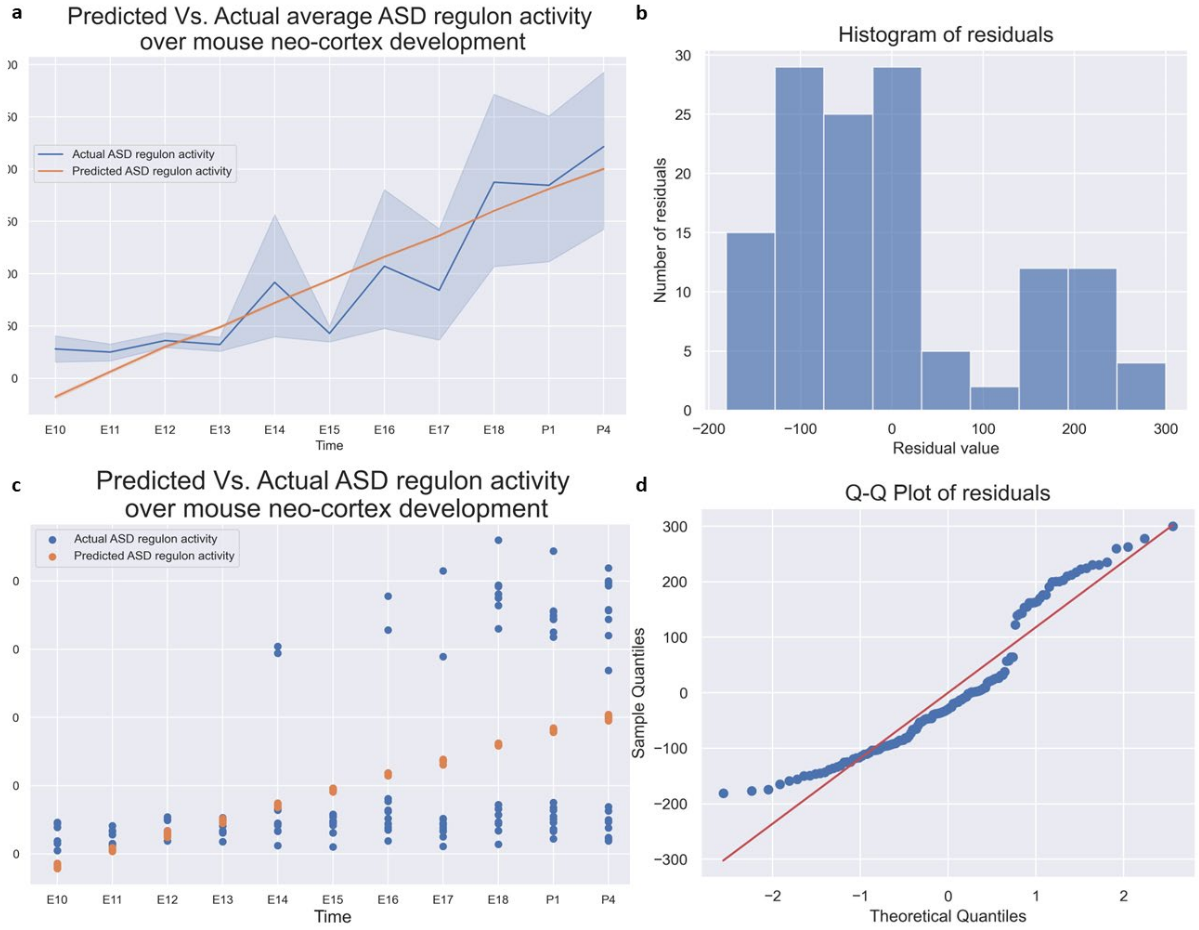

Figure 8. Regression analysis of ASD regulon activity as a function of GRN activity and time in the developing mouse neo cortex (E10-P4). (a) Average ASD regulon activity (blue) and model-predicted activity (orange) over mouse brain development stages. (b) Histogram showing distribution of residuals from the OLS regression. (c) Marker plot contrasting actual and model-predicted ASD regulon activity for each cell type at various time points. (d) QQ plot of residuals, with red line denoting a theoretical normal distribution. The model was implemented using statsmodels with HC3 covariance.

Adolescent mouse brain atlas

The OLS regression analysis was extended to the Zeisel et al. adolescent mouse brain. For this dataset, the model returned an Adjusted R-squared value of 0.170, suggesting that about 17% of the variability in ASD Regulon activity is explained by the GRN activity and time. The F-statistic was calculated as 102.8, with an associated Prob F-statistic of 6.61e-41 (<0.05), confirming its significance.

Fig. 9

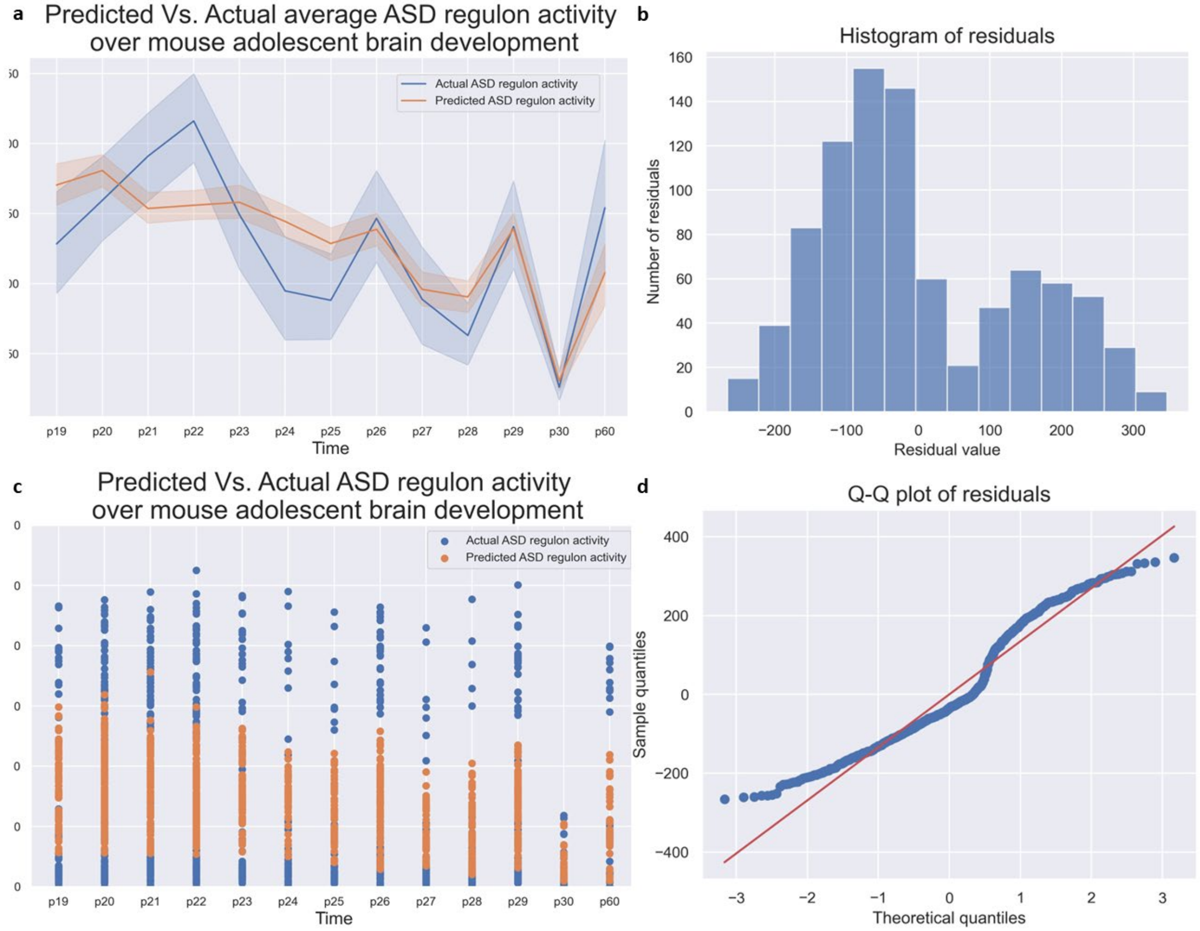

Figure 9. Regression analysis of ASD regulon activity as a function of GRN activity and time across adolescent mouse brain development (P19-P60). (a) Average ASD regulon activity (blue) and model-predicted activity (orange) over mouse brain development stages. (b) Histogram showing distribution of residuals from the OLS regression. (c) Marker plot contrasting actual and model-predicted ASD regulon activity for each cell type at various time points. (d) QQ plot of residuals, with red line denoting a theoretical normal distribution. The model was implemented using statsmodels with HC3 covariance.

With a total of 900 observations, the GRN activity coefficient was 0.0400, with a p-value of 0.0, thus indicating a significant effect of GRN activity on ASD Regulon activity. The time coefficient was -4.4074, again with a p-value of 0.0, implying a significant negative relationship between time and ASD Regulon activity. The resulting residuals from this analysis were not normally distributed.

Human cortical development.

Lastly, the analysis was applied to the Velmeshev et al. human cortical development dataset. The Adjusted R-squared of the model for this dataset was calculated as 0.077, indicating that about 7.7% of the variance in ASD Regulon activity is explained by GRN activity and time.

Fig. 10

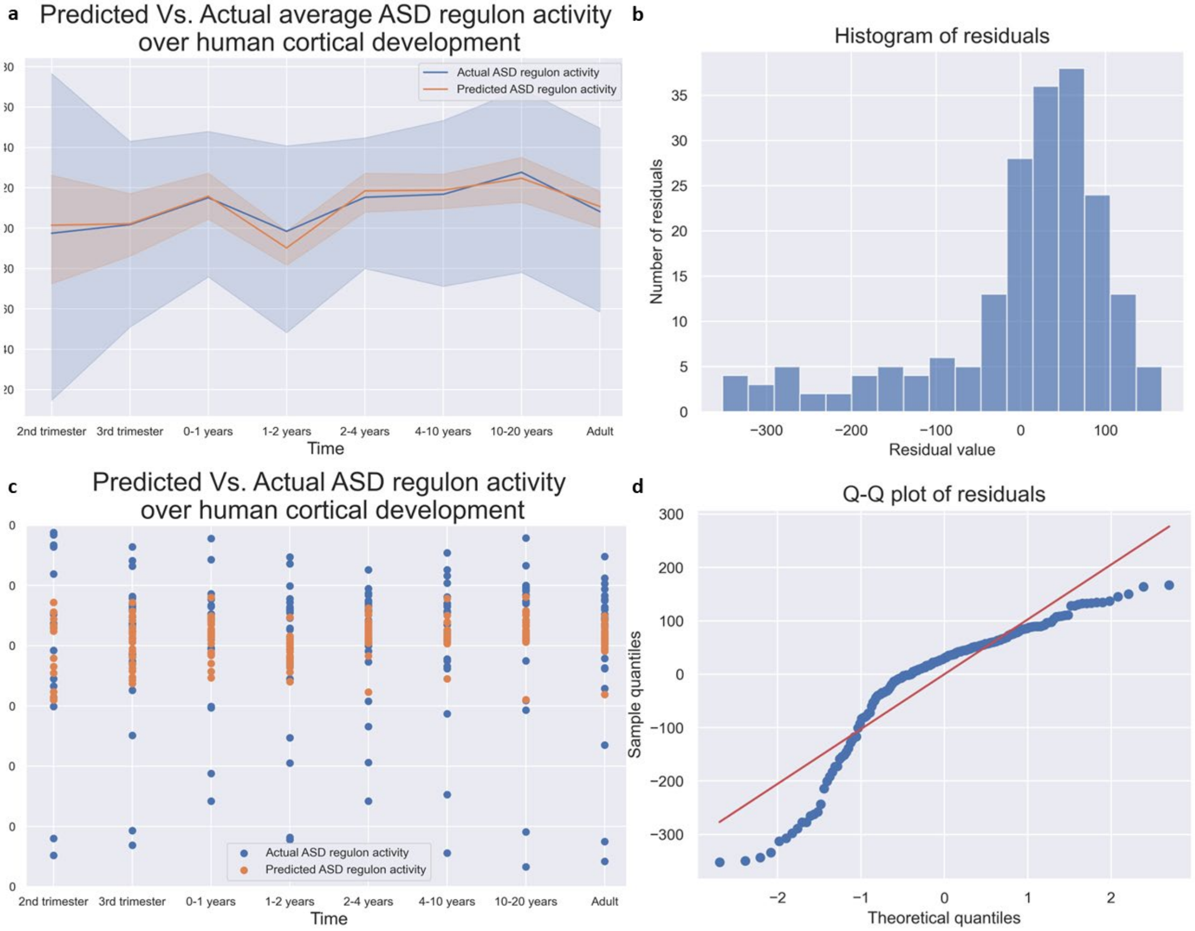

Figure 10. Regression analysis of ASD regulon activity as a function of GRN activity and time across the developing human cortex (2nd trimester – Adult). (a) Average ASD regulon activity (blue) and model-predicted activity (orange) over mouse brain development stages. (b) Histogram showing distribution of residuals from the OLS regression. (c) Marker plot contrasting actual and model-predicted ASD regulon activity for each cell type at various time points. (d) QQ plot of residuals, with red line denoting a theoretical normal distribution. The model was implemented using statsmodels with HC3 covariance.

The F-statistic was calculated as 3.483, with the Prob F-statistic of 0.0327 (<0.05), signifying the statistical significance of the model in this context. A total of 197 observations were included in the analysis. The GRN activity coefficient was computed as 0.0147, with a p-value of 0.013, indicating a significant influence of GRN activity on ASD regulon activity. However, the time coefficient was -0.5336, with a p-value of 0.901, which is inconclusive in its effect on the dependent variable. In line with the previous models, the residuals of this analysis were not normally distributed, suggesting potential deviations from the model's assumptions.

Discussion

Distributional analysis

The identification of ASD regulon-enriched cell types throughout the course of mouse brain development is consistent with existing literature that postulates ASD as a neurodevelopmental disorder that progresses through multiple stages. However, the discovery of such enriched cell types in both the adolescent and adult mouse brains challenges the notion that these effects are confined solely to the developmental stages. This finding corresponds well with studies that have reported ongoing brain abnormalities in ASD patients extending into adulthood. Additionally, the observed enrichment in certain cell types, both globally across the brain and specific to certain regions, support the concept of ASD as a multi-process disorder involving the whole brain, with region specific effects as well [20]. This enrichment in ASD regulon activity was found independent of differences in GRN activity, indicating that this increased ASD regulon activity is not due to differences in the total size of the network.

Ctcf, identified as a key driver of ASD regulon activity in this study, is a highly conserved, ubiquitously expressed protein, which uses its 11 zinc fingers to interact with DNA across ~20,000 binding sites in the human genome, for activation and/or inhibition [46]. In our analysis, Ctcf was found to target a substantial range of 180-300 genes across the enriched cell types for activation. As a benchmark, the threshold for ASD regulon activity (red line; indicated by the emergence of a second peak) was 200 across the mouse brain data, implying that the interactions involving Ctcf alone are capable of escalating the identified cell types to this second peak. Several studies have reported genetic variants of Ctcf in individuals diagnosed with ASD, identifying de novo loss-of-function variants and numerous de novo missense variants in the gene [47]. The implications of Ctcf as a driver of ASD regulon activity will be further expounded upon in the subsequent sections of this discussion.

The gene set enrichment analysis reveals several significantly enriched protein pathways involved across different developmental stages and areas of the mouse brain. Key findings include enrichment for "Proteins Involved in Autistic Disorder", emphasizing the role of altered protein networks in ASD and "Proteins Involved in Epilepsy" suggesting shared molecular underpinnings between these disorders. Furthermore, the presence of pathways such as "WNT Canonical Signaling" hints at the crucial role of neurodevelopmental pathways in the etiology of ASD, as dysregulation of these pathways could potentially affect neuronal proliferation and migration. Altogether, these enrichments indicate a diverse set of potentially dysregulated pathways and protein networks in ASD, encompassing mechanisms of neurodevelopment, cellular signaling, and disease-related proteins.

Our findings for the human cortical development data, on the other hand, were largely inconclusive. Despite the variation in cutoff stringency by observing every time-point independently, we did not observe a secondary peak in the distribution of ASD regulon activity. This is in stark contrast to the mouse brain results, which demonstrated a brain-wide enrichment spanning both the developing and adolescent mouse brain. Across 3 different mouse brain datasets of variable developmental time frames, the cutoff point of ASD regulon activity = 200 delineated the cell types with a significantly higher ASD regulon activity. This highlights a lack of the translational abilities of the model to tackle human data, which will be expounded upon when discussing limitations. Lastly, in regards to the main culprit of ASD regulon enrichment in our results; Ctcf was present in almost every identified cell type specific network. It was suspected that the over-arching effect of Ctcf presented noise in the results. However, removing Ctcf from the human data did not affect the distributional analysis. On the other hand, removing Ctcf from the mouse brain analysis also removes the observed second peaks.

Regression analysis

Our regression analysis revealed that time plays a significant role in the activity of ASD regulons in the developing mouse forebrain. For each unit of time progression, we observed an increase in ASD regulon activity, meaning that a greater number of genes are being targeted for activation. This increase is further substantiated by similar findings in the developing neocortex data, where a comparable brain region was sampled across a similar time frame. Such a time-dependent increase in ASD regulon activity was notably absent in the adolescent mouse brain atlas. Instead, we observed a significant negative effect with respect to time on ASD regulon activity during adolescence. It's important to note that in all datasets (except neo-cortex), an increase in GRN activity resulted in an increase in ASD regulon activity. This effect is potentially attributable to the sampling of more ASD genes as the GRN increases in size, rather than indicating a direct relationship between GRN activity and ASD regulon activity. In the case of the expansive human cortical development data spanning second trimester - adult, we noted a negative coefficient with respect to time. However, it was not statistically significant, making any conclusions based on this observation uncertain. Indeed, a visual inspection of ASD regulon activity over time demonstrates the homogeneity of the activity scores across development (Fig. 10 (a)).

In terms of understanding the neurodevelopmental processes, an increased ASD regulon activity over time suggests that the progression of neurodevelopment could be coupled with heightened ASD-related gene regulation. This could imply a potential vulnerability window during which dysregulated gene expression might contribute to the emergence of ASD-related phenotypes. It also provides a clue about the dynamic nature of ASD at the molecular level, with its impact potentially varying across different developmental stages.

The OLS results suggest that a more complex model may be required to accurately capture the underlying behavior in ASD regulon activity. This assertion is primarily based on two specific observations made during the analysis. First, the residuals of our models exhibited deviations from normality across all datasets. This deviation undermines a core assumption of linear regression models. Particularly, the residuals in the developing mouse forebrain were highly skewed, suggesting that the model consistently underestimated the ASD regulon activities of these cell types.

Second, the visual inspection of the model fit highlights the model's inability to accurately capture the dynamic and complex patterns of ASD regulon activity. More specifically, the predicted activity tended to be more linear and did not mirror the recurrent peaks and troughs observed in the actual ASD regulon activity. This discrepancy is most noticeable in the developing neocortex data. Here, the model's predicted ASD activity closely resembled a straight line, which starkly contrasted with the several peaks present in the actual ASD regulon activity.

Such consistent inconsistencies indicate that the current linear model may not be sufficient to encapsulate the nuanced behavior of ASD regulon activity over developmental time. These observations underscore the need for a more sophisticated modeling approach that can better represent the dynamic changes and complex patterns inherent in ASD regulon activity during brain development.

In light of the literature

Our results have shown an interesting overlap with findings from previous studies implicating cell types in ASD. To expand on a study mentioned in the introduction, which utilized single nucleus RNA sequencing on post-mortem cortical samples (prefrontal cortex/cingulate cortex) from ASD patients and control. The authors reported dysregulated gene expression in L2/3 neurons, L4 neurons, interneurons, and microglia [28]. This aligns well with our analysis of the mouse neocortex, where we observed a significant enrichment in ASD regulon activity in L4 neurons and interneurons. However, some discrepancies exist. While the authors of the aforementioned study concluded an effect on upper layer cortical neurons, our analysis identified layer 6 cortical cells as an enriched cell type as well, which is a lower-level layer. This disparity suggests a broader scope of potential neuronal involvement in ASD and emphasizes the need to understand the specific roles of different neuronal layers in ASD pathology. The Gene Ontology (GO) analysis of differentially expressed genes (DEGs) in the study also identified enrichment for cell surface receptor signaling, cellular growth/motility, and GABA signaling. These findings somewhat align with our gene set enrichment analysis (GSEA) results, which highlighted the involvement of several pathways, such as local estrogen production, Toll-like receptor signaling, and TGF-beta signaling. Both analyses point towards altered cellular signaling and growth dynamics as potential key players in ASD etiology.

In a study more directly comparable to our mouse brain analysis, which involved a perturb-seq analysis targeting 35 ASD-related genes in developing mouse embryos, the authors noted that the perturbation of nine ASD genes (Adnp, Ank2, Ash1l, Chd8, Gatad2b, Pogz, Scn2a1, Stard9, and Upf3b) significantly affected layer 4 and 5 projection neurons [48]. Interestingly, none of these genes were identified as ASD TFs in our analysis, meaning they did not contribute to the classification of ASD enrichment in our data. Despite this, we found several clusters of interneurons, projection neurons, and layer 4 neurons within our enriched cohort. The fact that independent analyses led to the identification of the same cell types further supports a role for a dysfunction of these cell types in ASD.

The authors also emphasized the role of glial dysfunction in ASD, implicating oligodendrocyte progenitors instead of microglia as in the previous study. This expands our understanding of the role of non-neuronal cells in ASD and aligns with our previous findings indicating the contribution of non-neuronal cells, such as OPCs, microglia and oligodendrocytes. Collectively, these comparisons demonstrate the complex and multifaceted nature of ASD, encompassing a diverse range of cell types and molecular mechanisms.

Even though Ctcf has been acknowledged as a high-confidence ASD gene, this alone does not explain or substantiate its dominant role in driving ASD regulon activity. Its lack of brain-specific expression further muddies conclusions regarding the etiology of ASD arising from Ctcf mutations. Yet, the TF has been causally linked to a rare neurodevelopmental disorder, namely ' Ctcf -related neurodevelopmental disorder,' with less than 1 in 1,000,000 prevalence and currently 107 reported cases in the literature [49-50]. These patients often present with a spectrum of symptoms ranging from mild developmental delay to severe intellectual disability, and some have been diagnosed with ASD. Notably, analysis of the blood transcriptome of these patients revealed 2161 downregulated genes [49].

In terms of genetic susceptibility, a GWAS meta-analysis has pointed towards Ctcf as a gene enriched for heritable SNPs [51]. While a 2017 study investigating hotspot missense mutations in neurodevelopmental disorders noted a clustering of these mutations between the 4th and 7th zinc finger of the protein [52]. Lastly, Ctcf has been shown to interact with the ASD related Chd8 chromatin modifier [53], although this interaction was not captured in our networks.

These collective findings hint at the potential for even minor dysfunction in Ctcf to profoundly affect a large number of genes and contribute to the development of ASD-related phenotypes such as intellectual disability and motor behavior dysfunction. However, it's important to underscore that although Ctcf dominated the activity of the enriched networks in this study, it did not serve as a hub to unite the various ASD-related TFs under a common regulatory framework.

The present study builds upon previous research conducted in the Basak lab, where Bram Schouten carried out a comprehensive investigation aimed at determining if autism-related genes were integrated within GRNs of cortical lineage cells. An aspect of his work investigated the 'ASD-TF regulon,' which filters for interactions involving ASD-related TFs, drawing close, but not exact parallels to our 'ASD regulon' concept.

A notable point of convergence between our studies is the recognition of Tcf4 as a ‘hub’ of connectivity within Cajal-Retzius cells. In Shouten's study, Tcf4 was identified as a hub for the high number of interactions (8) in the network. In contrast, our work expands upon this, demonstrating that Tcf4 targets 18 genes and is, in turn, targeted by four ASD TFs (Zeb2, Tcf12, Atf2, Pou2f2) in E12.0 mouse forebrain Cajal-Retzius cells. Shouten's analysis also distinguished cell types enriched for both activatory and inhibitory interactions, noting particular enrichment of inhibitory interactions in ectodermal, blood, and radial glia cells. Intriguingly, our analysis did not identify a similar enrichment in these cell types. This discrepancy suggests that our results might shift under different analytical conditions, such as when considering inhibitory interactions, underlining the nuanced and context-dependent nature of GRNs in the study of ASD. Adding to these observations, a notable divergence in our analysis revolves around the transcription factor Ctcf. Despite Ctcf not returning a significant result in Shouten's ASD-TF regulon analysis, he identified it as a TF enriched for ASD gene targets using the Fisher's exact test. This finding highlights Ctcf as a potential convergence point for diverse ASD-related genetic influences, underscoring its potential significance in the multifaceted genetic landscape of ASD.

While ASD is primarily categorized as a neurodevelopmental disorder, our understanding of its pathogenesis spans across neurodevelopment and continues well into adulthood. This perspective is reinforced by regional brain changes that extend beyond the early developmental period, as was described in the introduction. An example of the influence of time on ASD activity was reported in the authors of the study who produced the human cortical data under investigation. This study, which gauged ASD activity by assessing the differential expression of ASD-related genes, observed two prominent peaks in ASD activity, namely in 2nd trimester and adult.

In line with these findings, our analysis of the adolescent mouse brain data identified several cell types that were significantly enriched in ASD regulon activity in the adolescent-adult mouse brain. The detection of ASD activity at such late stages implies that the mechanisms leading to ASD do not simply halt post-development, but may continue to evolve. This leads to the idea that context of these cell types within their developmental timeline could hold significant importance for understanding their role in ASD. The question arises, does the presence of ASD-enriched cell types in later developmental stages suggest a prolonged period of susceptibility to ASD-related dysfunctions, or do they reflect a late onset of mechanistic processes contributing to the disorder? These observations challenge the traditional notion of ASD as a purely developmental disorder and suggest a broader temporal window for both the progression and possible treatment of ASD.

Limitations

The limitations of this study are multifaceted and stem from a variety of sources including the nature of the data and resources employed, the assumptions baked into the analytical pipeline, and the narrow scope of our investigation. Beginning with the data, the variation in brain regions, time-frames, and gender of mice across datasets restricts the robustness of our results. To bolster the validity of our findings, comparable data that samples the same brain regions over the same time-frames and accounts for gender variability in ASD needs to be utilized. Additionally, the discrepancies in cluster resolution and functional annotation between different datasets impede the drawing of solid conclusions when comparing them. A high cluster resolution can provide a high degree of separation between cells, but this results in many enriched cell types which then need to be unified under one umbrella. While a low cluster resolution can fail to delineate the heterogenous population of cell types and states. This issue is particularly relevant in the case of the human cortical development dataset. In this dataset, cell type annotations were not present. Instead, the authors sub sampled the original data, and clustered each sub sample. In this analysis, we combine the annotation from the different sub-samples into the whole data. This is relevant since the clustering results are likely to change if conducted on the whole data from the start, which is likely affecting our results.

Regarding resources, our reliance on the DoRothEA network presents several limitations. Primarily developed with cancer research in mind, the DoRothEA network may not adequately reflect the regulatory landscape of brain development. This especially affects our interpretation of Ctcf as a driver of enrichment. Since the strong signal we observe in Ctcf might be an artifact of the TFs high involvement in cancer. Furthermore, while the curations for the mouse DoRothEA network are arguably more reliable due to the preponderance of knockout studies in mice, this network likely oversimplifies the regulatory relationships between TFs and genes, especially in humans. Our exclusive focus on activatory interactions further limits the representativeness of our networks.

Turning to the pipeline, our analysis makes several simplifying assumptions that might limit the accuracy of our results. One key assumption is that gene expression levels directly correlate with the magnitude of regulatory interactions, an assumption inherent in the MLM model employed. In reality, genes can exhibit critical cellular functions even at relatively low expression levels. Moreover, our pipeline does not utilize a weight to measure the interaction between TFs and ASD genes, possibly hindering the identification of major drivers of regulatory activity beyond the binary (edge/no edge) designation. The significant enrichment of hundreds of biological processes returned by the GSEA, many related to cell growth/proliferation and cancer, also presents challenges in discerning which processes are truly relevant to ASD and which are artifacts of the developmental nature of the data and the cancer related DoRothEA network.

Lastly, the narrow perspective of our study warrants acknowledgment. We have focused primarily on TF driven enrichment in ASD activity and have not delved extensively into investigating ASD genes, the ASD-related motifs that can interconnect them, or interactions with non-coding regions. As a result, our analysis could overlook crucial players in ASD etiology. Future research should broaden its scope to incorporate different avenues of investigation, fully recognizing the complexity of ASD as a multifactorial disorder with likely involvement from several different mechanism of gene regulation.

Future direction

Looking forward, there are several avenues to explore that could enhance our understanding of ASD in the context of GRNs. One promising prospect involves the utilization of the CollecTRI database in place of DoRothEA. Although currently unpublished, CollecTRI offers a more extensive catalogue of interactions and provides the capacity to partition TFs into complexes. Comparing the results of our analysis with those obtained via CollecTRI or other GRN reconstruction methods could further validate our findings.

Adjustments to our analytical approach could also yield improvements. For instance, instead of employing a cell type-specific gene expression/activity cutoff, we could consider all identified interactions and discard those common to all cell types. This approach could provide a more nuanced understanding of 'cell type specific' networks and potentially illuminate less obvious interactions. Moreover, incorporating weights and inhibitory interactions could offer a more comprehensive view of the GRNs. To better model ASD regulon activity, we recommend employing splines or non-linear regression which may be more adept at capturing the complex relationships that likely exist within ASD regulon activity. This approach could provide a more nuanced understanding of how ASD develops and progresses over time.

Finally, the potential utility of the GRN pipeline extends beyond this study. The GRN architecture could serve as a valuable resource for both experimentalists and data researchers. Experimentalists can browse through specific cell-type networks to investigate their gene of interest in a topological framework, while informatically inclined researchers can integrate these networks with network/molecular dynamics simulations to probe deeper into the complex dynamics of ASD. Based on the observation that perturbation studies were employed to benchmark both DoRothEA and its successor, CollecTRI; the pipeline likely possesses significant potential for application in cell type-specific perturbation analyses. This may prove particularly beneficial for the host institute, where several organoid studies are currently being conducted.

Code Availability

The code to generate cell type specific GRNs, as well as perform the subsequent distributional and regression analyses is publicly available under The GNU General Public License. Interested parties can access and download the code from the following GitHub repository: https://github.com/AbdoolAK/GRN_reconstruction

Supplementary Materials

In the attached 'Supplementary Materials' folder accompanying this document, details regarding the ASD regulon activity, ASD TFs, and the count of genes targeted by Ctcf are provided for each cell type. These details have been compiled in table format for each of the four datasets studied in the ‘whole data’ analysis:

1. Developing mouse forebrain.
2. Developing mouse neo-cortex.
3. Developing mouse whole brain.
4. Adolescent mouse brain.

References

1. Hirota T, King BH. Autism Spectrum Disorder: A Review. JAMA. 2023 Jan 10;329(2):157-168. doi: 10.1001/jama.2022.23661. PMID: 36625807.
2. Leo Kanner, "Autistic Disturbances of Affective Contact," The Nervous Child 2 (1943):217-50.
3. Masi A, DeMayo MM, Glozier N, Guastella AJ. An Overview of Autism Spectrum Disorder, Heterogeneity and Treatment Options. Neurosci Bull. 2017 Apr;33(2):183-193. doi: 10.1007/s12264-017-0100-y. Epub 2017 Feb 17. PMID: 28213805; PMCID: PMC5360849.
4. Szatmari P. The classification of autism, Asperger's syndrome, and pervasive developmental disorder. Can J Psychiatry. 2000 Oct;45(8):731-8. doi: 10.1177/070674370004500806. PMID: 11086556.
5. Owen MJ, O'Donovan MC. Schizophrenia and the neurodevelopmental continuum:evidence from genomics. World Psychiatry. 2017 Oct;16(3):227-235. doi: 10.1002/wps.20440. PMID: 28941101; PMCID: PMC5608820.
6. Amir RE, Van den Veyver IB, Wan M, Tran CQ, Francke U, Zoghbi HY. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat Genet. 1999 Oct;23(2):185-8. doi: 10.1038/13810. PMID: 10508514.
7. Einspieler C, Sigafoos J, Bölte S, Bratl-Pokorny KD, Landa R, Marschik PB. Highlighting the first 5 months of life: General movements in infants later diagnosed with autism spectrum disorder or Rett Syndrome. Res Autism Spectr Disord. 2014 Mar;8(3):286-291. doi: 10.1016/j.rasd.2013.12.013. Epub 2014 Jan 9. PMID: 29770159; PMCID: PMC5951269.
8. Talantseva OI, Romanova RS, Shurdova EM, Dolgorukova TA, Sologub PS, Titova OS, Kleeva DF, Grigorenko EL. The global prevalence of autism spectrum disorder: A three-level meta-analysis. Front Psychiatry. 2023 Feb 9;14:1071181. doi: 10.3389/fpsyt.2023.1071181. PMID: 36846240; PMCID: PMC9947250.
9. Hirota T, King BH. Autism Spectrum Disorder: A Review. JAMA. 2023 Jan 10;329(2):157-168. doi: 10.1001/jama.2022.23661. PMID: 36625807.
10. Liu Q, Yin W, Meijsen JJ, Reichenberg A, Gådin JR, Schork AJ, Adami HO, Kolevzon A, Sandin S, Fang F. Cancer risk in individuals with autism spectrum disorder. Ann Oncol. 2022 Jul;33(7):713-719. doi: 10.1016/j.annonc.2022.04.006. Epub 2022 Apr 14. PMID: 35430370.
11. Werling DM, Geschwind DH. Sex differences in autism spectrum disorders. Curr Opin Neurol. 2013 Apr;26(2):146-53. doi: 10.1097/WCO.0b013e32835ee548. PMID: 23406909; PMCID: PMC4164392.
12. Hattier MA, Matson JL, Tureck K, Horovitz M. The effects of gender and age on repetitive and/or restricted behaviors and interests in adults with autism spectrum disorders and intellectual disability. Res Dev Disabil. 2011 Nov-Dec;32(6):2346-51. doi: 10.1016/j.ridd.2011.07.028. Epub 2011 Aug 6. PMID: 21824745.
13. Robinson EB, Lichtenstein P, Anckarsäter H, Happé F, Ronald A. Examining and interpreting the female protective effect against autistic behavior. Proc Natl Acad Sci U S A. 2013 Mar 26;110(13):5258-62. doi: 10.1073/pnas.1211070110. Epub 2013 Feb 19. PMID: 23431162; PMCID: PMC3612665.
14. Masini E, Loi E, Vega-Benedetti AF, Carta M, Doneddu G, Fadda R, Zavattari P. An Overview of the Main Genetic, Epigenetic and Environmental Factors Involved in Autism Spectrum Disorder Focusing on Synaptic Activity. Int J Mol Sci. 2020 Nov 5;21(21):8290. doi: 10.3390/ijms21218290. PMID: 33167418; PMCID: PMC7663950.
15. Zerbo O, Qian Y, Yoshida C, Grether JK, Van de Water J, Croen LA. Maternal Infection During Pregnancy and Autism Spectrum Disorders. J Autism Dev Disord. 2015 Dec;45(12):4015-25. doi: 10.1007/s10803-013-2016-3. PMID: 24366406; PMCID: PMC4108569.

16. Skogheim TS, Weyde KVF, Engel SM, Aase H, Surén P, Øie MG, Biele G, Reichborn-Kjennerud T, Caspersen IH, Hornig M, Haug LS, Villanger GD. Metal and essential element concentrations during pregnancy and associations with autism spectrum disorder and attention-deficit/hyperactivity disorder in children. Environ Int. 2021 Jul;152:106468. doi: 10.1016/j.envint.2021.106468. Epub 2021 Mar 22. PMID: 33765546.
17. Atsem S, Reichenbach J, Potabattula R, Dittrich M, Nava C, Depienne C, Böhm L, Rost S, Hahn T, Schorsch M, Haaf T, El Hajj N. Paternal age effects on sperm FOXK1 and KCNA7 methylation and transmission into the next generation. Hum Mol Genet. 2016 Nov 15;25(22):4996-5005. doi: 10.1093/hmg/ddw328. PMID: 28171595; PMCID: PMC5418740.
18. Ronald A, Hoekstra RA. Autism spectrum disorders and autistic traits: a decade of new twin studies. Am J Med Genet B Neuropsychiatr Genet. 2011 Apr;156B(3):255-74. doi: 10.1002/ajmg.b.31159. Epub 2011 Jan 13. PMID: 21438136.
19. Abrahams BS, Geschwind DH. Advances in autism genetics: on the threshold of a new neurobiology. Nat Rev Genet. 2008 May;9(5):341-55. doi: 10.1038/nrg2346. Erratum in: Nat Rev Genet. 2008 Jun;9(6):493. PMID: 18414403; PMCID: PMC2756414.
20. Courchesne E, Pramparo T, Gazestani VH, Lombardo MV, Pierce K, Lewis NE. The ASD Living Biology: from cell proliferation to clinical phenotype. Mol Psychiatry. 2019 Jan;24(1):88-107. doi: 10.1038/s41380-018-0056-y. Epub 2018 Jun 22. PMID: 29934544; PMCID: PMC6309606.
21. Wang P, Mokhtari R, Pedrosa E, Kirschenbaum M, Bayrak C, Zheng D, Lachman HM. CRISPR/Cas9-mediated heterozygous knockout of the autism gene CHD8 and characterization of its transcriptional networks in cerebral organoids derived from iPS cells. Mol Autism. 2017 Mar 20;8:11. doi: 10.1186/s13229-017-0124-1. PMID: 28321286; PMCID: PMC5357816.
22. Durak O, Gao F, Kaeser-Woo YJ, Rueda R, Martorell AJ, Nott A, Liu CY, Watson LA, Tsai LH. CHD8 mediates cortical neurogenesis via transcriptional regulation of cell cycle and Wnt signaling. Nat Neurosci. 2016 Nov;19(11):1477-1488. doi: 10.1038/nn.4400. Epub 2016 Oct 3. PMID: 27694995; PMCID: PMC5386887.
23. Skelton PD, Stan RV, Luikart BW. The Role of PTEN in Neurodevelopment. Mol Neuropsychiatry. 2020 Apr;5(Suppl 1):60-71. doi: 10.1159/000504782. Epub 2020 Jan 21. PMID: 32399470; PMCID: PMC7206585.
24. Bourgeron T. A synaptic trek to autism. Curr Opin Neurobiol. 2009 Apr;19(2):231-4. doi: 10.1016/j.conb.2009.06.003. Epub 2009 Jun 21. PMID: 19545994.
25. Zielinski BA, Prigge MB, Nielsen JA, Froehlich AL, Abildskov TJ, Anderson JS, Fletcher PT, Zygmunt KM, Travers BG, Lange N, Alexander AL, Bigler ED, Lainhart JE. Longitudinal changes in cortical thickness in autism and typical development. Brain. 2014 Jun;137(Pt 6):1799-812. doi: 10.1093/brain/awu083. Epub 2014 Apr 22. PMID: 24755274; PMCID: PMC4032101.
26. Emerson RW, Adams C, Nishino T, Hazlett HC, Wolff JJ, Zwaigenbaum L, Constantino JN, Shen MD, Swanson MR, Elison JT, Kandala S, Estes AM, Botteron KN, Collins L, Dager SR, Evans AC, Gerig G, Gu H, McKinstry RC, Paterson S, Schultz RT, Styner M; IBIS Network; Schlaggar BL, Pruett JR Jr, Piven J. Functional neuroimaging of high-risk 6-month-old infants predicts a diagnosis of autism at 24 months of age. Sci Transl Med. 2017 Jun 7;9(393):eaag2882. doi: 10.1126/scitranslmed.aag2882. PMID: 28592562; PMCID: PMC5819345.
27. Macaulay IC, Voet T. Single cell genomics: advances and future perspectives. PLoS Genet. 2014 Jan 30;10(1):e1004126. doi: 10.1371/journal.pgen.1004126. PMID: 24497842; PMCID: PMC3907301.
28. Velmeshev D, Schirmer L, Jung D, Haeussler M, Perez Y, Mayer S, Bhaduri A, Goyal N, Rowitch DH, Kriegstein AR. Single-cell genomics identifies cell type-specific molecular changes in autism. Science. 2019 May 17;364(6441):685-689. doi: 10.1126/science.aav8130. PMID: 31097668; PMCID: PMC7678724.

29. Pang K, Wang L, Wang W, Zhou J, Cheng C, Han K, Zoghbi HY, Liu Z. Coexpression enrichment analysis at the single-cell level reveals convergent defects in neural progenitor cells and their cell-type transitions in neurodevelopmental disorders. Genome Res. 2020 Jun;30(6):835-848. doi: 10.1101/gr.254987.119. Epub 2020 Jun 18. PMID: 32554779; PMCID: PMC7370880.
30. Banerjee-Basu S, Packer A. SFARI Gene: an evolving database for the autism research community. Dis Model Mech. 2010 Mar-Apr;3(3-4):133-5. doi: 10.1242/dmm.005439. PMID: 20212079.
31. Fiers MWEJ, Minnoye L, Aibar S, Bravo González-Blas C, Kalender Atak Z, Aerts S. Mapping gene regulatory networks from single-cell omics data. Brief Funct Genomics. 2018 Jul 1;17(4):246-254. doi: 10.1093/bfgp/elx046. PMID: 29342231; PMCID: PMC6063279.
32. Badia-I-Mompel P, Vélez Santiago J, Braunger J, Geiss C, Dimitrov D, Müller-Dott S, Taus P, Dugourd A, Holland CH, Ramirez Flores RO, Saez-Rodriguez J. decoupleR: ensemble of computational methods to infer biological activities from omics data. Bioinform Adv. 2022 Mar 8;2(1):vbac016. doi: 10.1093/bioadv/vbac016. PMID: 36699385; PMCID: PMC9710656.
33. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 2019 Aug;29(8):1363-1375. doi: 10.1101/gr.240663.118. Epub 2019 Jul 24. Erratum in: Genome Res. 2021 Apr;31(4):745. PMID: 31340985; PMCID: PMC6673718.
34. Kumar N, Mishra B, Athar M, Mukhtar S. Inference of Gene Regulatory Network from Single-Cell Transcriptomic Data Using pySCENIC. Methods Mol Biol. 2021;2328:171-182. doi: 10.1007/978-1-0716-1534-8_10. Erratum in: Methods Mol Biol. 2021;2328:C1. PMID: 34251625.
35. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006 Mar 20;7 Suppl 1(Suppl 1):S7. doi: 10.1186/1471-2105-7-S1-S7. PMID: 16723010; PMCID: PMC1810318.
36. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007 Jan;5(1):e8. doi: 10.1371/journal.pbio.0050008. PMID: 17214507; PMCID: PMC1764438.
37. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PLoS One. 2010 Sep 28;5(9):e12776. doi: 10.1371/journal.pone.0012776. PMID: 20927193; PMCID: PMC2946910.
38. Müller-Dott S, Tsirvouli E, Vázquez M, Ramirez Flores R.O, Badia-i-Mompel P, Fallegger R, Lægreid A, Saez-Rodriguez J. Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. bioRxiv. 2023. doi: 10.1101/2023.03.30.534849
39. La Manno G, Siletti K, Furlan A, Gyllborg D, Vinsland E, Mossi Albiach A, Mattsson Langseth C, Khven I, Lederer AR, Dratva LM, Johnsson A, Nilsson M, Lönnerberg P, Linnarsson S. Molecular architecture of the developing mouse brain. Nature. 2021 Aug;596(7870):92-96. doi: 10.1038/s41586-021-03775-x. Epub 2021 Jul 28. PMID: 34321664.
40. Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, van der Zwan J, Häring M, Braun E, Borm LE, La Manno G, Codeluppi S, Furlan A, Lee K, Skene N, Harris KD, Hjerling-Leffler J, Arenas E, Ernfors P, Marklund U, Linnarsson S. Molecular Architecture of the Mouse Nervous System. Cell. 2018 Aug 9;174(4):999-1014.e22. doi: 10.1016/j.cell.2018.06.021. PMID: 30096314; PMCID: PMC6086934.
41. Di Bella DJ, Habibi E, Stickels RR, Scalia G, Brown J, Yadollahpour P, Yang SM, Abbate C, Biancalani T, Macosko EZ, Chen F, Regev A, Arlotta P. Molecular logic of cellular diversification in the mouse cerebral cortex. Nature. 2021 Jul;595(7868):554-559. doi: 10.1038/s41586-021-03670-5. Epub 2021 Jun 23. Erratum in: Nature. 2021 Aug;596(7873):E11. PMID: 34163074; PMCID: PMC9006333.

42. Velmeshev D, Perez Y, Yan Z, Valencia JE, Castaneda-Castellanos DR, Schirmer L, Mayer S, Wick B, Wang S, Nowakowski TJ, Paredes M, Huang E, Kriegstein A. Single-cell analysis of prenatal and postnatal human cortical development. bioRxiv. 2022. doi: 10.1101/2022.10.24.513555.
43. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;14:128.
44. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44(W1):W90-7. doi: 10.1093/nar/gkw377.
45. Xie Z, Bailey A, Kuleshov MV, Clarke DJB, Evangelista JE, Jenkins SL, Lachmann A, Wojciechowicz ML, Kropiwnicki E, Jagodnik KM, Jeon M, Ma’ayan A. Gene set knowledge discovery with Enrichr. Curr Protoc. 2021;1:e90. doi: 10.1002/cpz1.90.
46. Bao L, Zhou M, Cui Y. CTCFBSDB: a CTCF-binding site database for characterization of vertebrate genomic insulators. Nucleic Acids Res. 2008 Jan;36(Database issue):D83-7. doi: 10.1093/nar/gkm875. Epub 2007 Nov 2. PMID: 17981843; PMCID: PMC2238977.
47. Krumm N, Turner TN, Baker C, Vives L, Mohajeri K, Witherspoon K, Raja A, Coe BP, Stessman HA, He ZX, Leal SM, Bernier R, Eichler EE. Excess of rare, inherited truncating mutations in autism. Nat Genet. 2015 Jun;47(6):582-8. doi: 10.1038/ng.3303. Epub 2015 May 11. PMID: 25961944; PMCID: PMC4449286.
48. Jin X, Simmons SK, Guo A, Shetty AS, Ko M, Nguyen L, Jokhi V, Robinson E, Oyler P, Curry N, Deangeli G, Lodato S, Levin JZ, Regev A, Zhang F, Arlotta P. In vivo Perturb-Seq reveals neuronal and glial abnormalities associated with autism risk genes. Science. 2020 Nov 27;370(6520):eaaz6063. doi: 10.1126/science.aaz6063. PMID: 33243861; PMCID: PMC7985844.
49. Konrad EDH, Nardini N, Caliebe A, Nagel I, Young D, Horvath G, Santoro SL, Shuss C, Ziegler A, Bonneau D, Kempers M, Pfundt R, Legius E, Bouman A, Stuurman KE, Õunap K, Pajusalu S, Wojcik MH, Vasileiou G, Le Guyader G, Schnelle HM, Berland S, Zonneveld-Huijssoon E, Kersten S, Gupta A, Blackburn PR, Ellingson MS, Ferber MJ, Dhamija R, Klee EW, McEntagart M, Lichtenbelt KD, Kenney A, Vergano SA, Abou Jamra R, Platzer K, Ella Pierpont M, Khattar D, Hopkin RJ, Martin RJ, Jongmans MCJ, Chang VY, Martinez-Agosto JA, Kuismin O, Kurki MI, Pietiläinen O, Palotie A, Maarup TJ, Johnson DS, Venborg Pedersen K, Laulund LW, Lynch SA, Blyth M, Prescott K, Canham N, Ibitoye R, Brilstra EH, Shinawi M, Fassi E; DDD Study; Sticht H, Gregor A, Van Esch H, Zweier C. CTCF variants in 39 individuals with a variable neurodevelopmental disorder broaden the mutational and clinical spectrum. Genet Med. 2019 Dec;21(12):2723-2733. doi: 10.1038/s41436-019-0585-z. Epub 2019 Jun 26. PMID: 31239556; PMCID: PMC6892744.
50. Valverde de Morales HG, Wang HV, Garber K, Cheng X, Corces VG, Li H. Expansion of the genotypic and phenotypic spectrum of CTCF-related disorder guides clinical management: 43 new subjects and a comprehensive literature review. Am J Med Genet A. 2023 Mar;191(3):718-729. doi: 10.1002/ajmg.a.63065. Epub 2022 Dec 1. PMID: 36454652; PMCID: PMC9928606.
51. Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. Mol Autism. 2017 May 22;8:21. doi: 10.1186/s13229-017-0137-9. PMID: 28540026; PMCID: PMC5441062.

52. Geisheker MR, Heymann G, Wang T, Coe BP, Turner TN, Stessman HAF, Hoekzema K, Kvarnung M, Shaw M, Friend K, Liebelt J, Barnett C, Thompson EM, Haan E, Guo H, Anderlid BM, Nordgren A, Lindstrand A, Vandeweyer G, Alberti A, Avola E, Vinci M, Giusto S, Pramparo T, Pierce K, Nalabolu S, Michaelson JJ, Sedlacek Z, Santen GWE, Peeters H, Hakonarson H, Courchesne E, Romano C, Kooy RF, Bernier RA, Nordenskjöld M, Gecz J, Xia K, Zweifel LS, Eichler EE. Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains. Nat Neurosci. 2017 Aug;20(8):1043-1051. doi: 10.1038/nn.4589. Epub 2017 Jun 19. PMID: 28628100; PMCID: PMC5539915.
53. Ishihara K, Oshimura M, Nakao M. CTCF-dependent chromatin insulator is linked to epigenetic remodeling. Mol Cell. 2006 Sep 1;23(5):733-42. doi: 10.1016/j.molcel.2006.08.008. PMID: 16949368.