MOLECULAR UNDERPINNINGS OF HUMAN BRAIN EVOLUTION AND COGNITION AT CELLULAR RESOLUTION

by

EMRE CAGLAYAN

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences
The University of Texas Southwestern Medical Center at Dallas
In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY
The University of Texas Southwestern Medical Center
Dallas, Texas

December, 2023

MOLECULAR UNDERPINNINGS OF HUMAN BRAIN EVOLUTION AND COGNITION AT CELLULAR RESOLUTION

Emre Caglayan, M.S.

The University of Texas Southwestern Medical Center at Dallas, 2023

Genevieve Konopka, Ph.D

ABSTRACT

Molecular and functional characterization of the human brain is challenging due to its experimental inaccessibility. Most of our understanding about human brain function relies on the assumption that biological processes uncovered in model organisms are conserved in humans. Comparisons of the human brain with non-human primate brains offer to both uncover the novelties in human brain evolution and better evaluate the insights obtained from model organisms about human brain function. To achieve this, high-throughput sequencing methods on post-mortem brain tissues provide a rewarding readout to understand human brain evolution at the molecular level. In addition to their use in comparative studies, these technologies were also utilized with a hope to understand molecular underpinnings of measurable human brain activity metrics. During my dissertation, I read relevant literature extensively **(Chapter 1)** and sought to understand human-specific epigenomic and transcriptomic changes at cellular resolution in the cortical brain **(Chapter 2)**. Additionally, after in-depth analysis of many human brain single-nuclei RNA-seq datasets, I found a pervasive ambient RNA contamination problem, and devised *in silico* solutions to tackle this problem. My efforts improved the analytical approach in the field as well as in my research **(Chapter 3)**. I have also been involved in efforts to identify transcriptomic correlates of brain activity in human subjects **(Chapters 4-5)**. After detailing these efforts, I discuss the implications of these findings, weigh their impact on our understanding of human brain function and offer ideas for further research **(Chapter 6)**.

Table of Contents

CHAPTER 1: Background

Note: The second part of this chapter – Molecular evolution of the human brain - is modified from a commissioned book chapter (accepted for publication) titled: "Differences in brain gene expression between humans and primates". It has been accepted for publication as a chapter in the book titled "*The evolutionary roots of human brain diseases*". This chapter is an edited version of my own original writing. It has been edited by my thesis supervisor Genevieve Konopka and has also been updated after feedback from the editors of the book.

> Caglayan, E. and Konopka, G. (2023) 'Differences in brain gene expression between humans and primates.', in *The evolutionary roots of human brain diseases*. Oxford University Press.

Phenotypic evolution of the human brain

Comparisons with other species, especially chimpanzees that shared a last common ancestor with humans ~6 million years ago[1], provide an opportunity to better categorize and understand the molecular and phenotypic evolutionary changes that are unique to humans. Phenotypic changes are especially important since prominent changes in the molecular landscape are often interpreted as causally linked to phenotypic changes. Despite the manifest uniqueness of human cognition, pinpointing the phenotypic changes unique to humans has been challenging. In this section, I categorize and provide an overview of these efforts.

Behavioral traits

Although cognitive differences between human and non-human primates are evident, cognitive skills have been surprisingly difficult to compare between human and non-human primates. A prominent point of discussion is whether the capacity to

understand the mental state of others – also known as theory of mind – is unique to humans[2,3]. This possibility has been examined in great apes. Chimpanzees are able to behave accordingly to newly acquired knowledge of another chimpanzee in multiple experimental paradigms. For example, when a subordinate chimpanzee observed food being hidden in sight of a dominant chimpanzee, the subordinate chimpanzee did not reach out to the food[3]. However, the subordinate chimpanzee reached out to the food if she observed the dominant chimpanzee did not witness where the food was hidden[3]. Interestingly, chimpanzees did not change their behavior when they observed another chimpanzee had a false belief (e.g., when the food was moved to a new location and the dominant chimpanzee did not witness this)[3]. These results indicate that chimpanzees may possess theory of mind, albeit to a limited degree compared to humans[3]. However, some critics have argued that these behaviors may arise from adaptive behavioral reflexes and not reflect an actual mental state of awareness[4]. More recently, great ape species were tested in a false belief paradigm by measuring the eye movements of the subjects that look in anticipation of a certain behavior (anticipatory looking)[5]. Contrary to previous research, the researchers found that great apes may possess false belief understanding[5], although others pointed out the high variability of results obtained with anticipatory looking test[6]. Other cognitive tests have also been performed on non-human primates. For example, to understand abstract thinking by measuring basic logical inference, the researcher hides food in one out of two cups and shows the subject which cup is empty. When presented with the choice, both apes and monkeys were able to select the cup with the food, indicating that non-human primates are able to do inferential reasoning[7,8].

Similar to theory of mind and other logical inferences, many cognitive capabilities that are effortlessly executed by humans can also be executed by other animals (both mammals and birds) including mental time travel, problem solving and toolmaking among others[9]. Therefore, the current understanding is that humans are mostly not unique in being able to perform cognitive tasks, but rather in how advanced they can perform these tasks compared to other species. For example, the first systematic behavioral observations of chimpanzees demonstrated that toolmaking - previously considered as a unique human trait – was shared with non-human primates[10]. Decades of subsequent research showed that many non-human species can make and/or use tools, but no species could demonstrate similar complexity of tool usage as humans[9]. Taken together, these results indicate that our closest genetic relatives may possess higher cognitive skills than previously appreciated but to a lesser extent than humans.

Perhaps the most salient cognitive ability that is considered to be uniquely human is language. Since non-human primates are not able to use vocalization for learning and complex communication, efforts to understand language capabilities in apes have focused on teaching sign language to newborn apes during their infancy and observing the ape's competency in language in comparison to human infants[9]. In two independent studies, researchers raised baby chimpanzees named Washoe and Nim who communicated only through American Sign Language. Washoe and Nim were reported to learn hundreds of signs and multisign utterances (mainly two-signs)[11,12]. It was noted that the breadth of sign utterances was too high to be explained by rote memorization[11]. Researchers also noted that Washoe was able to form a two-sign combination that was never taught before[13]. For example, when pointed to a swan on water, and asked *'what*

that', Washoe signaled *'water bird'*. As an alternative explanation, researchers raising Nim noted that Washoe may simply be signaling what she is observing (that is, *water* and *bird*) and not relating the two signs[11]. They also noted that Nim lacked many characteristics of language that typically developed in age-matched human infants. For example, mean length of sign utterances did not change in Nim's development. Nim's mean length of utterances were between 1 - 2, whereas they increased rapidly for human infants from 1 to 5 signs in 1–3-year-olds[11]. These behavioral metrics have been observed both in healthy controls and deaf infants who communicate with sign language[11]. When Nim uttered >2 signs, he often repeated one of the signs twice (e.g., *eat Nim eat*) and his longest utterance was a 16 sign with many repeats: *give orange me give eat orange me eat orange give me eat orange give me you*[11]. In contrast, human infants often add increasing semantic complexity and rarely repeat the words in a similar fashion[11]. Researchers also noted that Nim interrupted his teacher more often than human infants would and his utterances mostly followed his teacher's utterances[11]. Taken together, the authors claimed that apes lack the innate ability to develop language similar to humans which supports the hypothesis that language is a genetic endowment unique to humans[11,14].

Anatomical and stereological differences

In addition to the behavioral and cognitive complexity, human brain evolution is also characterized by enlargement in the last 2.5 million years that resulted in three times larger brains than our closest genetic relative chimpanzees[15]. Some studies also pointed out that brain weight to body weight ratio in the human brain is seven times larger than an average mammalian brain and three times larger than an average primate brain[16].

However, others note that large brain weight relative to body weight is not an indicator of higher cognitive ability, and it correlates with cognitive ability less than brain weight alone[17,18]. Some studies further note that brain mass and neuron number do not scale similarly in all lineages. For example, brain mass increases more compared to neuron number in the rodent lineage than in the primate or insectivore lineages[17]. In other words, primate brains scale to contain more neurons for the same weight, thus saving more space compared to rodent brains[19]. In contrast, non-neuronal cells scale similarly between lineages[17]. However, humans do not show extraordinary divergence in terms of their neuronal or non-neuronal cell numbers per brain mass compared to other primates[17]. The authors therefore argue that the human brain scales similarly to other primate brains in terms of cell counts per brain mass[17,19]. While the large size of the human brain may be linked to the evolution of higher cognitive capacity in humans, these results call for a reconsideration of the most appropriate measure to utilize allometric differences to understand the unique trends in human brain evolution.

In addition to high-level allometric comparisons, compositional differences (e.g., density and number of neuronal processes) also need to be examined to understand human brain evolution. Decades of research focused on varying brain regions and cortical layers while utilizing different technologies. As a result, most studies do not provide entirely overlapping conclusions with previous studies. For example, in one study, dendritic spine density in the neocortex was observed to be greater in human upper layer pyramidal neurons compared to macaques and marmosets[20]. Another study found similar dendritic spine density between human and chimpanzee cortex[21]. However, both studies found longer and more branched dendrites in humans compared to chimpanzees and

monkeys[20,21]. While this could reflect that dendritic spine density is similar between human and chimpanzees but different in monkeys, more comprehensive studies are needed to complement these results. Indeed, a more recent study provided a comprehensive comparison of synapse densities – but not spine densities – across multiple primate species which revealed higher synapse density in humans compared to all primates including chimpanzees[22]. Notably, synapse and dendritic spine density are higher in association areas than primary sensory cortex, which may be linked to their function in neural plasticity and higher order computations with potential consequences for human brain evolution[21,23].

In contrast to synapse and dendritic spine density, neuron density is lower in humans than other primates, although the difference is more subtle between humans and other apes[22]. This difference was also more prominent in visual cortex than inferior temporal cortex[22]. Lower neuron density in humans is also supported by recent spatial transcriptomics profiling that found both three fold less cell density and fewer neurons than glia in human cortex compared to mouse cortex[24]. Previous studies also showed higher glia to neuron ratio in humans compared to non-human primates using Nissl staining[25], however these results were not supported by quantitative analyses in single-cell approaches[26,27]. These results suggest a trend for less neuron density in humans and a need for more comprehensive study between humans and non-human primates to elucidate changes in neuron density, glia density and other cell type densities as well as their relative proportions. Importantly, while relative cell type proportions can be obtained by single-cell or single-nuclei sequencing methods, density measurements require

preservation of tissue space. Future comparative spatial transcriptomics studies across primate tissues will be informative to answer these questions.

As I will detail in the next chapter, most comparative studies at the molecular level characterize the cells resident in the tissue, often at transcriptomic and epigenomic levels. However, this is a very limited scope for a brain tissue that is highly connected both within itself and with other brain regions. For example, subcortical neurons synapse onto cortical neurons within the neocortex and very little is known about whether and how these synaptic transmissions differ between species. Such biological processes are unlikely to be captured by comparative transcriptomics of tissues or nuclei. By immunostaining crucial receptor proteins across species, several studies have identified that humans and chimpanzees have more dense innervations of serotonergic, dopaminergic and cholinergic axons from subcortical regions onto prefrontal cortex compared to macaque monkeys[28-30]. Other studies have focused on the connectivity differences of language-relevant brain areas between humans and chimpanzees using brain imaging measurements (e.g magnetic resonance imaging or diffusion tensor imaging), although they offer limited resolution[31]. More studies are needed to elucidate differences in connectivity and molecular identity of connections compared to non-human primate brains.

Molecular evolution of the human brain

Phenotypic comparisons of the human brain with non-human primate brains revealed surprisingly few differences. It should be noted that comparing phenotypic readouts between species requires a priori knowledge about the importance of the phenotype for the overarching question (e.g., importance of dendritic spines for human

brain evolution). Some have noted that the human brain may possess many unique phenotypes that require unbiased detection methods[32,33]. Since the phenotypic changes that make us human are majorly contributed by – if not fully caused by – genetic changes, a closer examination of the functional consequences of human-specific genetic changes may permit discovery of previously unsuspected phenotypic changes[33]. Indeed, many studies have compared the sequence or regulation of human genome to non-human primate genomes with a focus on brain function. In the following section, I will focus on efforts to identify and characterize genetic and gene regulatory changes associated with human brain evolution.

Evolution and function of human-specific DNA sequences

Human-specific DNA sequence changes are, directly or indirectly, connected to human-specific gene regulatory changes, human-specific novel genes, and human-specific alterations of the protein structures. Genes encoding for proteins expressed in the brain are largely conserved in the human genome[34]. Notable exceptions have been an active area of research since altering amino acid composition of a single gene is more amenable to genetic modification in experimental systems than sequence changes in the non-coding genome[35]. However, in this chapter, we will focus on the gene regulatory effect of DNA sequence changes that comprise the millions of human-specific substitutions, insertions and deletions[36]. Since ~98% of the human genome is non-coding, most of these changes are in the non-coding genome, and understanding their functional role is an exciting and ongoing challenge.

Human-specific substitutions can be either functional genomic changes or neutral changes without a functional consequence. To determine functional genomic sequences

with increased divergence specifically in human evolution, one approach is to identify the genomic sequences that are highly conserved across many vertebrate species with low substitution rate indicating functionality, and then retain the regions that accumulated significantly more human-specific substitutions on this constrained background, indicating accelerated evolution of the conserved elements on the human lineage. The accumulation of studies utilizing this approach has yielded ~3000 human accelerated regions (HAR)[37]. While some studies have exclusively focused on non-coding sequences[38], other studies have carried out genome-wide analyses and found >90% of the accelerated regions to be non-coding[39], indicating that most HARs are likely to affect gene regulation rather than protein sequences.

Associating HARs to nearby genes for the identification of their potential functions revealed enrichments for genes involved in neuronal functioning, specifically in neurodevelopment[40]. In their analysis of 2649 non-coding HARs, Capra et al. found that 773 are predicted to be developmental enhancers and 251 of them are predicted to be active in brain[40]. Predicted enhancer activity in development motivated functional characterization of HARs with reporter assays through injection into mouse embryos[40,41]. These studies revealed activity of HARs in the developing brain, with the HARs driving different patterns of reporter activity compared to the ancestral state of the genomic region[40]. Even with ways to narrow down the regions to the most promising candidates, low-throughput methodology, such as reporter assays in mouse embryos, is a major roadblock for the functional characterization of HARs. More recently, studies have been able to parallelize the delivery of genomic constructs into cultured cells, allowing the high-throughput characterization of the effects of sequence changes on regulatory function

using MPRAs, or massively parallel reporter assays[42-44]. MPRAs function as reporter assays that produce RNA as a readout instead of fluorescence. Combined with RNA-seq, this technique allows parallelized screening of the activity of regulatory regions of interest[45]. Studies utilized MPRAs by delivering HARs and their corresponding chimpanzee sequences into neural stem cells or neural cells in culture. Two studies found that 50-60% of active HARs displayed significantly altered activity compared to chimpanzee sequences[42,43]. Another study found this to be 27.5%[44]. Importantly, the lack of differential activity could be due to the limitations of the culture systems or cell types being utilized, indicating that, even at ~50%, this is likely an underestimation of functional activity of human-specific sequence changes in HARs. Moreover, the functionality of the HARs to drive reporter expression was also similar between the same cell types from human / mouse[43], and human / chimpanzee[42], indicating that HAR functionality is primarily driven by the human-specific sequence changes and not by the *trans* effects of the cellular environment. In addition to uncovering the pattern of HAR activity, these studies also further characterized some HARs for their role in human neural stem cells. An example was a HAR-regulated gene, *PPP1R17,* that slows cell cycle progression in neural progenitor cells[43]. *PPP1R17* and other genes with human-specific functions (e.g *SRGAP2C* that promotes radial glia migration and increases spine density[46]), could be molecular factors responsible for neoteny in human brain development.

HARs comprise only a small portion of all human-specific genomic changes[44]. The sequencing of ancient human genomes has also allowed identification of human-specific substitutions that were ancestral in Neanderthals and Denisovans, pointing out regions that were likely changed more recently in modern human evolution[47]. Genes carrying

modern human-specific amino acid substitutions are enriched in neurodevelopmental function, similar to HARs[47]. Notably, three of these genes are associated with kinetochore of the mitotic spindle (*CASC5*, *KIF18A*, *SPAG5*)[47]. Both human-specific substitutions not linked to HARs and modern human-specific substitutions were also recently characterized by MPRAs[44,48]. Interestingly, human-gained enhancers (HGE) that have human-specific substitutions that are not necessarily characterized as HARs caused differential activity in 33.9% of the active HGEs[44]. Similarly, ~23% of active modern human-specific substitutions were differentially active compared to the ancestral sequences[48] and the genes associated with the loci of differentially active sequences are enriched for brain anatomy and function[48]. These results indicate that modern human-specific substitutions and human-specific substitutions in the non-HAR enhancers are also important sources of molecular evolution in human cells.

Gene regulatory changes – single gene comparisons

While genomic changes provide the ultimate resource for identifying the genomic underpinnings of human evolution and disease, the functional consequences of human-specific genomic changes are highly complex. Even if a genomic change is correlated with a detectable phenotype, uncovering the activity of this genomic change in different organs, tissues and cell types is an arduous task. Moreover, the majority of the genomic changes are non-coding and their interactions with gene promoters are mostly unknown. Genomic changes can also exert a functional impact indirectly by altering the expression of a gene that subsequently differentially alters the expression patterns or function of other genes. Methodologies have been insufficient to capture such a complex interaction since this would require high-throughput screening of molecular function throughout

development to adult stages and at high cellular and regulatory (from DNA to protein) resolution. However, gene expression and chromatin architecture can be investigated in whole tissues, and more recently, at single-cell resolution per tissue. Such breakthroughs have allowed investigators to understand human-specific gene regulatory novelties, especially in brain tissues. In this section, we will summarize the major findings from these studies that span more than a decade and discuss their relevance to the understanding of human brain disorders.

As mentioned above, human-specific changes are largely non-coding, but can also rarely be among coding sequences. Nonsynonymous changes in the coding sequences can potentially alter the function of the given protein, which can have indirect effects on the molecular landscape of the cell. Regulatory proteins, such as transcription factors and RNA-binding proteins, can be prioritized to test this hypothesis and uncover potential human-specific gene expression changes caused by these human-specific evolutionary novelties in regulatory proteins. Variants in the coding region of the transcription factor FOXP2 are associated with both language disorders[49] and in human evolution through positive selection[50]. Humanized FOXP2 mice that express the two human-specific amino acids show altered ultrasonic vocalizations (~30kHz-~100kHz), indicating that human-specific FOXP2 sequence functionality may have been altered in brain circuits that underlie motor-relevant behaviors in human evolution[51]. Another study tested whether human FOXP2 has differential transcriptional targets compared to chimpanzee FOXP2 ($FOXP2^{chimp}$) in cultured neuronal cells[52]. Interestingly, ~100 genes were differentially regulated by FOXP2 compared to $FOXP2^{chimp}$ indicating that many gene expression changes in humans can be driven by differential *trans* activity of a single regulatory

protein. These genes were also enriched in genes that are differentially expressed between human and chimpanzee brain tissues, underscoring the relevance of this finding for *in vivo* functions[52]. A more recent study characterized the modern human-specific coding sequence change in NOVA1, an RNA binding protein that regulates alternative splicing and is associated with neurological disorders[53]. Despite NOVA1's regulation of alternative splicing, which may not directly affect gene expression levels, the authors found 277 differentially expressed genes in brain organoids expressing the modern-humanized *NOVA1* compared to the ancestral *NOVA1*[53]. Another recent study using overexpression in mouse and ferret cortex as well as human brain organoids found that the modern-human version of *TKTL1* (*hTKTL1*), that codes for an enzyme in the glycolysis pathway, increases the production of basal radial glia and neurons compared to the ancestral variant[54]. While this study did not investigate whether there are gene regulatory changes associated with hTKTL1, the authors' phenotypic observations suggest that hTKTL1 evolution likely affects the wiring of the gene regulatory programs. Taken together, these studies have shown that human-specific changes in the function of a single regulatory protein can affect the expression patterns of many other genes, leading to phenotypic alterations, and indicating that the molecular networks of a given human brain cell can be very different than a comparable chimpanzee (or other non-human primates) cell.

Gene regulatory changes - transcriptomic comparisons in the developing brain

The effects of human-specific coding sequence changes can be considered as trans-effects as they alter gene expression through a diffusible molecule whereas a HAR regulating a nearby gene's expression is considered a cis-effect. Strikingly, the studies of

protein coding evolution show that functional changes, even in a single protein, can affect the expression patterns of many other genes. It is not feasible, if not currently impossible, to predict the complex regulatory landscape of human brain cells compared to non-human primate brain cells by DNA sequence data alone. Human-specific phenotypes at different developmental and cellular levels are also challenging to delineate using DNA sequence alone without comprehension of the regulatory landscape of each gene in humans and non-human primates in a given biological context. To overcome these challenges, many studies have adopted a more direct approach to understanding the human-specific molecular functionality by comparing the transcriptomes of humans and non-human primates. Importantly, these studies use brain tissue from developing and adult humans and non-human primates. While initial studies utilized microarray technology, more recent studies have adopted RNA-sequencing, as it is not prone to biases in the pre-determined sequences central to hybridization-based microarray technology. Here, we outline these comparative transcriptomics studies and discuss key findings.

The human brain is larger than a non-human primate brain and has unique cognitive capabilities. These phenotypes are proposed to be due to the heterochronous development (i.e. altered developmental rate or timing) of human features, including brain cell types. To understand whether there is molecular support for these changes, and to characterize human-specific molecular alterations related to heterochrony, studies have compared the transcriptomes of the brains of humans and non-human primates in early development.

An early study focused on the early postnatal development of the dorsolateral prefrontal cortex (DLPFC) using tissue from humans, chimpanzees and rhesus

macaques, and found an excess number of genes that show delayed expression in humans relative to chimpanzee and rhesus-macaque[55]. Later studies also found heterochronous transcriptomic changes in human brain tissues. One study showed that the peak expression of synaptic genes in the prefrontal cortex was delayed until 5 years in humans, whereas this was achieved at ~1 year in chimpanzees and rhesus macaques[56]. However, a morphological study showed prolonged synaptic maturation in chimpanzees until 5 years old, arguing that the original study may have suffered from low sample sizes in chimpanzees[57]. A more recent study compared the transcriptomes of human and rhesus macaque brains across prenatal and postnatal development by matching the chronological ages of humans and non-human primates according to their transcriptomic profile[58]. Comparing the heterochrony in the transcriptomic signatures of five major biological processes (neurogenesis, neuronal differentiation, astrogliogenesis, synaptogenesis, myelination), the authors found that synaptogenesis related genes were not delayed but accelerated in human neocortex[58]. These studies may have yielded different results due to variabilities in the readout (gene expression versus neuronal morphology) and analytical approaches to match the chronological age between species.

In contrast to tissue-based comparisons, *in vitro* studies using induced pluripotent stem cell (iPSC) derived neurons offer a more controlled setting to study heterochrony in human and non-human primate neuronal development. Studies differentiating human and chimpanzee neurons from iPSCs consistently reported slower maturation in human neurons both in terms of neuronal morphology (e.g. dendritic length) and also in terms of neuronal function (e.g. synaptic firing)[59-61]. Transcriptomic comparisons also revealed that genes related to neuronal maturation were differentially expressed in human compared

to non-human primate neurons[60]. While this lends support for the initial observation of delayed transcriptomic upregulation of neuronal development in humans[56], these studies did not explicitly test whether the heterochronous genes in *in vivo* development matched the heterochronous genes in *in vitro* development. Importantly, transcriptomic profiles of monolayer and organoid culture systems have been shown to largely correspond to prenatal development unless cultured for ~1 year[62], while comparative transcriptomic studies targeting neurodevelopment were often from postnatal tissues older than 1 year. Nevertheless, in vitro studies have shown that human neuronal maturation is slower than chimpanzee neuronal maturation outside of their tissue environment, indicating that this property is an intrinsic feature of human neurons.

Another heterochronic biological process associated with human development is myelination. Myelination in the central nervous system is mediated through oligodendrocytes that mature postnatally. Comparisons of myelination levels in human and chimpanzee cortical gray matter throughout postnatal development showed that myelination is prolonged in the human brain, extending beyond late adolescence, whereas it peaked before sexual maturation in chimpanzees[63]. A carbon dating study of human oligodendrogenesis also showed that oligodendrocyte generation is prolonged in the gray matter of the cortex until ~40 years old, whereas oligodendrocyte generation peaked at ~5 years old in white matter, indicating that the observation of prolonged myelination in humans might be specific to the gray matter of the cortex[64]. Transcriptomic comparisons have also shown delayed increased expression of myelination related genes in humans compared to rhesus macaque[58]. A recent study found that the cortical gray matter of the adult human brain has a higher ratio of OPCs (oligodendrocyte progenitor

cells) and lower ratio of mature oligodendrocytes compared to non-human primates[65], indicating that humans may retain a larger oligodendrocyte progenitor pool, and therefore a likely higher progenitor capacity, compared to non-human primates even in late adulthood.

The developing brain has many different cell types progressing through various stages of maturation at a given time point. Most investigations that focused on development have been at the tissue level, thus human-specific regulatory patterns throughout development are currently being investigated at the cell type level. One study compared fetal human and rhesus macaque brain at single-cell resolution in dorsolateral prefrontal cortex but noted the high variability of expression patterns at single-cell resolution and utilized bulk transcriptomes to identify differentially expressed genes between the species and investigated their expression across cell types[58]. The authors found 14 differentially expressed genes in humans including *TRIM54* – a gene encoding a protein important in axonal growth – expressed in excitatory neurons at lower levels in humans than rhesus macaque[58]. Given that even a single regulatory gene can be responsible for over a hundred gene expression changes[52,53], this number is likely an underestimation and future studies are needed to elucidate the human-specific regulatory changes more accurately.

Gene regulatory changes - transcriptomic comparisons in the adult brain

The differential maturation of the human brain indicates that the molecular architecture of the mature human brain is likely vastly different from that of a non-human primate brain. Post-mortem tissues from the adult brain are more readily accessible to researchers than those from the developing brain, especially for endangered non-human

primates such as chimpanzees. Therefore, comparative transcriptomic studies on the adult brain have been more frequent and thus more insightful than the limited studies from the developing brain. Additionally, these studies have uncovered the immense cellular heterogeneity in the adult human brain, which is largely reflected in non-human primate brains[27,65]. This section aims to outline these efforts and provide an overview of human-specific molecular features in the adult human brain.

Tissue-level comparisons of human-specific gene expression changes were identified by comparing anatomically matched tissues typically from human, chimpanzee and rhesus macaque brains, although some studies included more species including bonobo[26] and gorilla[66] as well as a new world monkey marmoset as an outgroup to rhesus macaque (an old world monkey)[27,67]. Since gene expression differences between humans and chimpanzees can be either due to a change in humans or chimpanzees, studies have used another non-human primate (often rhesus macaque) as an outgroup species and identified human-specific gene expression changes as the genes that are consistently up / down regulated in both human-chimpanzee and human-rhesus macaque comparisons[68,69]. Focusing on multiple brain regions from cortical and subcortical regions, these studies have identified hundreds of human-specific gene expression differences. While human-specific gene expression alterations are often reproducible across studies[68], quantitative comparisons have not always yielded the same conclusion. For example, focusing on three brain regions, frontal pole from neocortex, caudate nucleus and hippocampus, one study found more human-specific gene expression differences in the frontal pole compared to the caudate nucleus or hippocampus[68]. However, another study identified a greater number of human-specific gene expression changes in the

striatum and thalamus compared to cortical regions[69]. A more recent study conducted a comparative survey of 33 anatomical brain regions and found a greater number of human-specific gene expression changes in the cerebral cortex, hypothalamus and cerebellar gray and white matter regions compared to striatal regions[26]. Variability across such studies could be explained by differences in the exact anatomical regions used (e.g., striatum encompasses caudate nucleus, putamen, and nucleus accumbens), as well as differences in the analytical and experimental approaches. A meta-analysis that starts from the raw data across these and other studies could be instructive with respect to any potential differences in analytical methods.

Several studies have sought to prioritize specific differentially expressed genes in a number of ways. One approach is to carry out assessment of co-expression networks (co-expressed genes that are likely co-regulated). Multiple studies using this approach have found that these gene modules are often not conserved between humans and non-human primates[68,69]. Notably, one study identified a gene module that contained *FOXP2* as a hub gene (a gene with the highest level of correlation with the other genes in the module), providing further support that *FOXP2* may have important human-specific functions[68]. Another hub gene in a human-specific gene module was *CLOCK*, which is implicated in psychiatric diseases in addition to its role in circadian rhythms[68]. Human-specific molecular changes in the striatum also revealed that genes involved in dopamine biosynthesis (tyrosine hydroxylase: *TH* and DOPA decarboxylase: *DDC*) were human-specifically upregulated[69]. Interestingly, *TH* was downregulated in several great apes but not in humans in the neocortex, indicating that it was likely downregulated in the great ape divergence but upregulated again in human lineage[69]. A recent single-cell

comparative study also found that cell types orthologous to *TH* expressing human SST (somatostatin) neurons are present in chimpanzees but lack *TH* expression. This is in contrast to human, rhesus macaque and marmoset SST neurons that express *TH*[27]. While great apes other than chimpanzees were not examined in this study, both previous and more recent findings suggest that great apes lack *TH*+ neurons. Therefore, *TH* upregulation (and SST+TH+ neurons) may have been convergently evolved in humans with currently unknown functional consequences. Thus, these studies have uncovered human-specific molecular features at the tissue level.

Brain tissues are notable for containing multiple cell types. The major cell types include neurons, astrocytes, microglia, oligodendrocytes and OPCs (oligodendrocyte progenitor cells), with many subtypes within each major cell type (especially neurons). Tissue level comparisons combine transcripts from all cell types in a tissue, potentially masking any cell type-specific effects. To circumvent this problem, some studies have isolated nuclei from post-mortem tissue and used flow cytometry to sort several major cell types per tissue sample across species[70,71]. Comparing differential gene expression in oligodendrocytes, one study found that previous tissue level comparisons failed to capture the human-specific gene expression changes in oligodendrocytes[70]. The authors also showed evidence for an increased number of human-specific gene expression changes in oligodendrocytes compared to neurons[70]. Another study distinguished excitatory and inhibitory neurons from each other and found that many genes with human-specific expression in one cell type was not altered in the other[71]. Broad cell type categories (excitatory neurons, inhibitory neurons, astrocytes, oligodendrocytes, OPCs, microglia) have been further examined across species using single-nuclei RNA-

sequencing (snRNA-seq), finding that glial gene expression changes may be more human-specific compared to neuronal gene expression changes despite having similar evolutionary rates of change when all branches are considered[26]. These results could indicate that glial functions may have been altered in a more human-specific manner. Given the recent discoveries on neuron-glia interactions (e.g., both OPCs and microglia can engulf neuronal processes and affect the specificity and turnover of neuronal connections[72,73]), glial function could also indirectly alter the neuronal and neural function of the human brain.

snRNA-seq of post-mortem brains has facilitated cellular characterization of brain regions. High quality datasets have shown that each of the broad cell type categories of the human cortex (explained above) contain further transcriptomically distinct subtypes[74]. Neurons are especially heterogeneous with more than a dozen distinct subtypes for both excitatory and inhibitory neurons[74]. Comparisons between human and non-human primate brains have revealed that the complex and heterogeneous diversification of neuronal subtypes are largely conserved across species[27,65,75]. However, the abundances of subtypes are not always uniform across species. For example, upper layer excitatory neurons are more abundant in human and chimpanzee compared to other non-human primates[27,75]. Several less abundant subtypes of excitatory neurons, inhibitory neurons, astrocytes and microglia are also absent in certain species, including a human-specific microglia subtype[27]. Transcriptomic comparisons of conserved cell types have shown that cell type identity is more conserved among inhibitory neurons compared to excitatory neurons across species[75]. Another study showed that most human-specific gene expression changes are only observed in one or a few subtypes (both in excitatory

and inhibitory neurons), and failure to disentangle neuronal subtypes masked these changes[65]. One example is human-specific upregulation of *FOXP2* in only two (out of 14 detected) excitatory subtypes in the posterior cingulate cortex (PCC)[65]. Strikingly, this upregulation was not observed in a similar comparative transcriptomic study from DLPFC, and comparisons of PCC with other cortical regions showed higher *FOXP2* levels in PCC, indicating that subtype-specific *FOXP2* upregulation in humans is also region-specific[65]. In addition to its neuronal subtype-specific and region-specific upregulation, *FOXP2* is also human-specifically upregulated in microglia[27]. These results revealed novel human-specific changes in the levels of critical regulatory genes, motivating future studies to characterize the functional consequences of these novelties in human evolution.

Gene regulatory changes - epigenomic comparisons in the adult brain

Non-coding genomic elements function to regulate gene expression. These genomic elements are also referred to as gene regulatory elements (GREs), and they can be further classified based upon their precise functioning (e.g., enhancers, promoters, silencers). GRE function can be detected by the presence of histone markers, chromatin state (open or closed) or the level of DNA methylation. It is possible to profile these markers through various high throughput assays and compare the level of epigenomic readout across species. While epigenomic profiling may not be as informative as transcriptomic profiling – since little is known about the functions of non-coding regions – the results of such profiling can help to pinpoint the GREs that function human-specifically and provide further mechanistic insight into the regulatory evolution of the human brain.

Histone proteins are physically associated with DNA and can be modified to facilitate or obstruct the function of a given DNA sequence. Functionally active DNA

sequences (such as enhancers and promoters) are free of histone proteins, and the histone proteins flanking these regions are typically modified with H3K27ac or H3K4me1[76]. Several comparative studies have captured the active GREs with an antibody against H3K27ac and these DNA sequences were profiled in post-mortem human and non-human primates brain tissues using chromatin immunoprecipitation sequencing (ChIP-seq)[67,71,77]. Similar to genes, GREs are also largely conserved across species at the sequence level; however, their activity can vary between species, indicating that the conserved GREs were modulated during evolution to create novel molecular networks[77]. One study showed that regulatory gains in adult hominins (common in human and chimpanzee) are enriched in elements that regulate oligodendrocyte gene expression, further implicating human-specific changes in the oligodendrocyte lineage[67]. Another study compared the epigenomes of prenatal human and rhesus macaque brains by profiling enhancers (H3K27ac) and promoters (H3K27ac and H3K4me2) and found enrichments for elements that are linked to genes involved in neuronal proliferation and migration among the human-gained enhancers[78].

Another, more stable epigenetic modification that can change gene expression is DNA methylation. The majority of DNA methylation occurs on CpG sites, however non-CG methylation (CH methylation) can also modulate gene expression. Multiple studies have shown that CH methylation is enriched in brain tissue, in particular in neurons, and accumulates during the development of neural circuitry[79,80]. Interestingly, human neurons contain more CH methylation compared to chimpanzee neurons whereas CG methylation levels are similar between the two species[79]. These studies reveal another layer of human-specific gene expression regulation in the brain.

Multiple modalities of the epigenome can be profiled with high-throughput assays at the tissue level, however achieving cellular resolution in epigenomics has been a technical challenge. Recently, a highly efficient assay for profiling open-chromatin genomic regions (ATAC-seq) has been optimized for single-cell sequencing. A recent study compared the transcriptome and epigenome of human, chimpanzee and rhesus macaque brains at cellular resolution[65]. Focusing on the GREs with human-specific accessibility changes, the authors found that elements that had gains in accessibility specifically in human upper layer excitatory neurons are enriched for FOS / JUN transcription factor motifs. Since FOS / JUN are immediately transcribed upon neuronal depolarization and target hundreds of genes[81], altered accessibility of putative FOS / JUN targets indicate that gene regulation upon neuronal depolarization has likely undergone human-specific modifications, specifically in upper layer excitatory neurons that are important for higher order cognition[82]. The authors also found that human-specific chromatin accessibility gains in deep layer excitatory neurons are enriched for FOX transcription factor motifs, including FOXP2, that are factors consistently implicated in neurodevelopment and cognitive functions[83]. Utilizing the comparative genomic sequence datasets, the authors further showed that human accelerated regions (HARs) and modern human-specific variants are enriched within human-specific chromatin changes. Interestingly, while HAR enrichment was observed in all cell types, modern variant enrichment was specific to an upper layer excitatory subtype, indicating potentially more cell type specificity in recent human evolution[65].

Taken together, these studies show past and present efforts to understand human brain evolution at the cellular and molecular level by using comparative transcriptomics and epigenomics **(Figure 1)**.

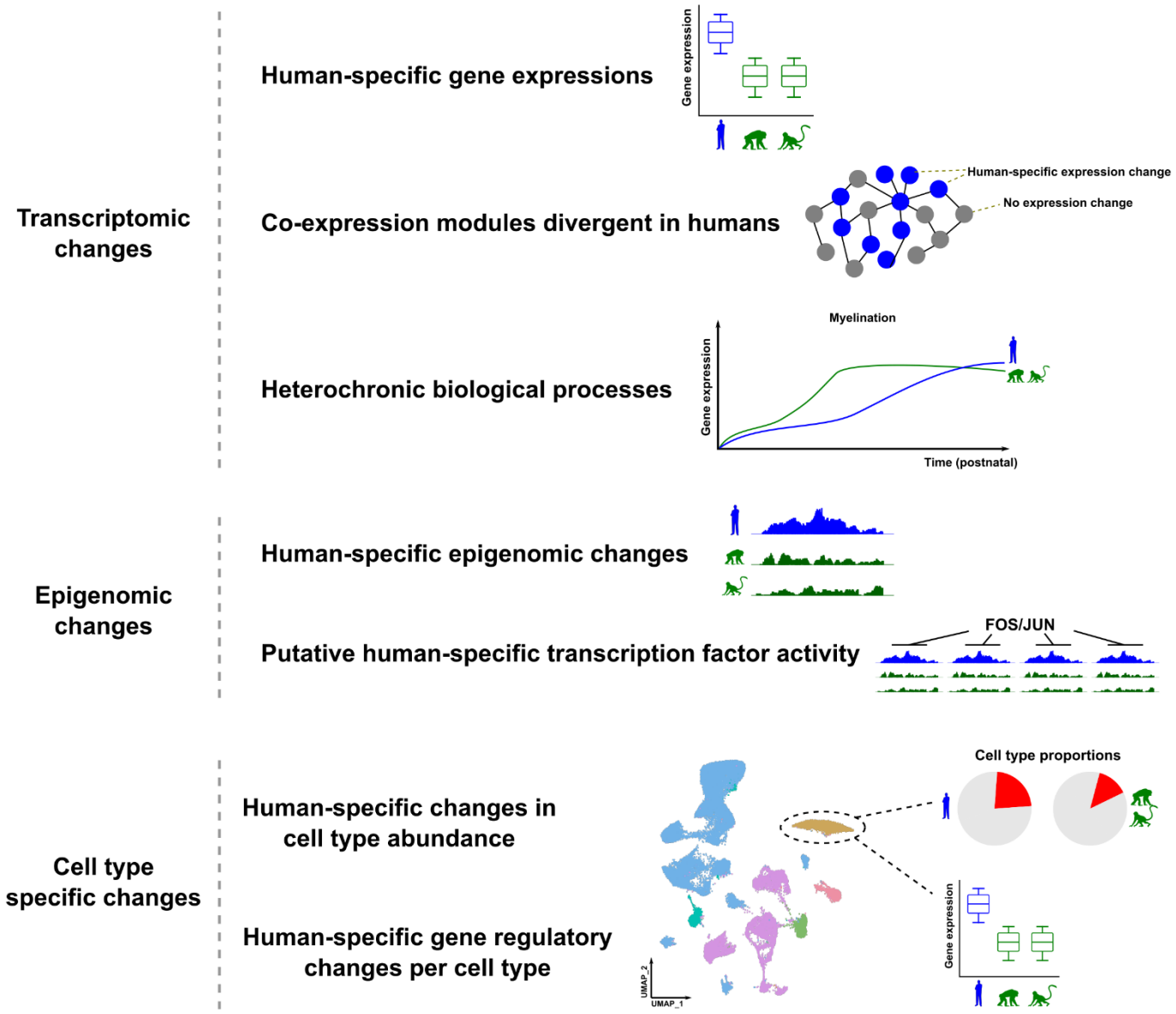


Figure 1: Summary of approaches to identify human-specific cellular and gene regulatory changes.

CHAPTER 2: Molecular features driving cellular complexity in human brain evolution

Published as:

Caglayan, E.*, Ayhan, F.*, Liu, Y. *et al.* Molecular features driving cellular complexity of human brain evolution. *Nature* **620**, 145–153 (2023). https://doi.org/10.1038/s41586-023-06338-4

*co-first authors

Molecular features driving cellular complexity of human brain evolution

Emre Caglayan[1,2*], Fatma Ayhan[1,2*], Yuxiang Liu[1,2,] Rachael Vollmer[1,2], Emily Oh[1,2], Chet C. Sherwood[3], Todd M. Preuss[4,5], Soojin V. Yi[6] and Genevieve Konopka[1,2]

[1]Department of Neuroscience, UT Southwestern Medical Center, Dallas, TX 75390, USA.

[2]Peter O'Donnell Jr. Brain Institute, UT Southwestern Medical Center, Dallas, TX 75390, USA.

[3]Department of Anthropology and Center for the Advanced Study of Human Paleobiology, The George Washington University, Washington, DC, USA.

[4]Division of Neuropharmacology and Neurologic Diseases, Yerkes National Primate Research Center, Emory University, Atlanta, GA, USA.

[5]Department of Pathology, Emory University School of Medicine, Atlanta, GA, USA.

[6] Department of Ecology, Evolution, and Marine Biology, Department of Molecular, Cell and Developmental Biology, and Neuroscience Research Institute, University of California, Santa Barbara, CA 93106, USA.

*equal contribution

Corresponding authors: Genevieve Konopka (genevieve.konopka@utsouthwestern.edu), Soojin V. Yi (soojinyi@ucsb.edu),

Abstract

Human-specific genomic changes contribute to the unique functionalities of the human brain[1-5]. The cellular heterogeneity of the human brain[6,7] and the complex regulation of gene expression highlight the need to characterize human-specific molecular features at cellular resolution. Here, we analyzed single-nuclei RNA-sequencing and single-nuclei open chromatin sequencing (ATAC-seq) datasets in human, chimpanzee, and rhesus macaque brain tissue from posterior cingulate cortex. We show a human-specific increase of oligodendrocyte progenitor cells (OPCs) and a decrease of mature oligodendrocytes across cortical tissues. Human-specific regulatory changes were accelerated in OPCs and we highlight key biological pathways that may be associated with the proportional changes. We also identify human-specific regulatory changes in neuronal subtypes, which reveal human-specific upregulation of *FOXP2* in only two of the neuronal subtypes. We additionally identify hundreds of new human accelerated genomic regions associated with human-specific chromatin accessibility changes. Our data also reveal that FOS / JUN and FOX motifs are enriched in the human-specifically accessible chromatin regions of excitatory neuronal subtypes. Together, we reveal multiple novel mechanisms underlying the evolutionary innovation of human brain at cell type resolution.

Main text: Phenotypic differences between humans and our closest extant relatives, including chimpanzees and other great apes, are driven by a combination of regulatory and coding sequence changes[1]. These genomic underpinnings of human brain evolution can be elucidated by genome-wide comparisons with non-human primate species. Previous studies have profiled the transcriptome of brain tissues in bulk to identify human-specific gene expression changes[2-4]. These findings highlighted human-specific changes including in synaptogenesis[3,5] and myelination[5,8]. However, brain tissue has tremendous cellular heterogeneity[6,9]. Therefore, single-cell genomics approaches are required to identify the full scope of human-specific gene regulatory changes. While previous studies have explored comparisons of epigenome or transcriptome between humans and other species[7,10-15], a systematic identification of human-specific epigenomic and transcriptomic changes at cellular resolution is lacking. To address this gap of knowledge and assign changes to the human lineage, here we profiled the transcriptomes and epigenomes of adult tissue from posterior cingulate cortex from humans and chimpanzees (*Pan troglodytes*) and included rhesus macaques (*Macaca mulatta*) as an outgroup. Our single nuclei RNA-sequencing (snRNA-seq) and single nuclei open chromatin profiling-sequencing (snATAC-seq) revealed significant changes in proportions of cells in the oligodendrocyte lineage, uncovering thousands of human-specific regulatory changes. We further assessed the association of these regulatory changes with the underlying human-specific substitutions, providing critical links between changes in DNA sequence and function in human brain evolution at cellular resolution. We also uncovered specific enrichment of the immediate early gene transcription factors FOS and JUN motifs in human-specific chromatin accessibility gains, indicating human specificity in activity dependent gene regulation. These results shed light on previously unknown cellular dimensions of human brain evolution.

Results

To identify molecular and cellular changes accompanying human brain evolution, we examined the evolution of Brodmann area 23 (BA23) by applying snRNA-seq and snATAC-seq approaches to the same samples. Notably, BA23 is part of the posterior cingulate cortex, a hub region in the default mode network[16] that is involved in higher-order cognitive process such as theory of mind and has been implicated in schizophrenia[17]. Despite such importance, there is currently no detailed study of BA23 at single-cell resolution. We detected 148,540 nuclei

using snRNA-seq (Human: 41,397, Chimpanzee: 53,539, Macaque: 53,604) and 73,486 nuclei using snATAC-seq (Human: 28,630, Chimpanzee: 20,703, Macaque: 24,153) after quality control **(Extended Figures 1, Supplementary Table 1, Methods)**. We annotated major cell types **(Extended Figures 1-2)** and neuronal subtypes (14 excitatory subtypes and 8 inhibitory subtypes) **(Extended Figures 3)** across species in both snRNA-seq and snATAC-seq (see Methods).

Proportional changes in oligodendrocytes

Evolutionary changes can arise from both proportional[13] and/or gene regulatory[2,5,10-12,18] changes of cell types. Compared to non-human primates, the human brain has prolonged myelination and altered gene regulation in the oligodendrocyte lineage[5,8,18], indicating possible changes in human-specific cell type abundances. Assessing the proportional changes of the oligodendrocyte lineage in single-cell genomics can be particularly challenging since glia express fewer transcripts than neurons as evidenced by fewer unique molecular identifiers (UMIs) **(Extended Figure 1E)** and are thus more prone to filtering during empty droplet removal with a UMI cutoff. To overcome this bias, we used a low UMI cutoff after empty droplet removal (see Methods) and calculated the percentage of mature oligodendrocytes (MOLs) and oligodendrocyte progenitor cells (OPCs) compared to all glia. We found a human-specific increase in OPC abundance and a human-specific decrease in MOL abundance while astrocytes and microglia were not significantly altered **(Figure 1A, Supplementary Table 2)**. To confirm this finding using an independent method that preserves tissue anatomy, we performed single molecule fluorescence in situ hybridization (smFISH), which validated a significant increase of OPC and a significant decrease of MOL populations in humans compared to chimpanzees **(Figure 1B-D)**. We then examined data from other cortical regions that were previously profiled. Re-analysis of a snRNA-seq dataset from anterior cingulate cortex[11] yielded a concordant result with our finding **(Figure 1E, Extended Figure 4A-B)**. We further validated this with smFISH using anterior cingulate cortical tissue from humans and chimpanzees **(Figure 1F-H).** Since we quantified the signal from all layers of the cortex, we also divided the columnar images into sections which revealed similar trends in both cortical regions **(Extended Figure 4C-F),** indicating that the result is not driven by uneven sampling of cortical layers. In addition, we examined a bulk RNA-sequencing data set of the entire oligodendrocyte lineage in the dorsolateral prefrontal cortex[18]. Using deconvolution, we found a higher OPC/MOL ratio in humans, regardless of which species was used as reference **(Figure 1I, Extended Figure**

4G). Re-analysis of a comparative dataset[7] in primary motor cortex tissue yielded similarly increased proportions of OPCs and decreased proportions of MOLs in humans compared to marmosets **(Figure 1J)** and a similar, trend compared to rhesus macaques **(Figure 1K)**. Notably, we did not observe similar abundance changes in the caudate nucleus[11] or dentate gyrus[19] **(Extended Figure 4H-I)**. Together, these results show that adult human brain cortical regions have proportionally more OPCs and fewer MOLs compared to non-human primates.

Gene regulatory changes in OPCs

To understand the gene regulatory novelties in the human lineage, we identified human-specific gene expression alterations (HS-Genes: HS-Up-Genes and HS-Down-Genes) and human-specific chromatin accessibility level alterations in cis-regulatory elements (HS-CREs: HS-Open-CREs and HS-Closed-CREs) per cell type **(Extended Figures 4J, 5A-B, Supplementary Tables 3-4, see Methods)**. Focusing on the oligodendrocyte lineage, we found an increased relative abundance of HS changes compared to CS changes (chimpanzee specific) in OPCs than MOLs in both snRNA-seq and snATAC-seq **(Figure 2A).** Applying a similar approach to the anterior cingulate cortex[11] yielded similarly accelerated evolution in OPCs **(Figure 2B)** as well as a significant overlap with our results **(Extended Figure 4K-L).**

Among the human-specific regulatory changes in OPCs, HS-Down-Genes are enriched in cytoskeletal activity **(Figure 2C; Supplementary Table 5)**, which is crucial for OPC migration and oligodendrocyte differentiation[20]. We posited that marker genes in committed oligodendrocyte progenitors (COPs) may also have been altered in human OPCs. We identified 15 COP marker genes that are common across all species in our dataset and in two additional human datasets[21-23]. Two COP markers, *SH3RF3* and *KIF21A* were HS-Down-Genes in OPCs **(Fig 2D-F)**. In line with the enrichment of cytoskeletal genes, *KIF21A* is a kinesin motor protein that is involved in microtubule function, whereas *SH3RF3* encodes for a SH3 domain containing protein with ubiquitin ligase activity, a process also implicated in oligodendrocyte maturation[24]. We also identified a HS-Closed-CRE in OPCs near the transcriptional start site (TSS) of *SH3RF3,* potentially linked to the human-specific downregulation of this gene **(Figure 2G)**. Interestingly, snRNA-seq from the frontal cortex of adult mice showed that most primate COP markers exhibit upregulation in COPs or NFOLs (newly formed oligodendrocytes) compared to OPCs,

except for *Sh3rf3*, indicating potential primate-specificity **(Fig 2H-J)**. Together, these results highlight key regulatory changes in human OPCs which may underlie human-specific proportional changes in the oligodendrocyte lineage.

Neuronal subtype specificity of evolution

We identified 14 subtypes of excitatory neurons and 8 subtypes of inhibitory neurons across species in both snRNA-seq and snATAC-seq **(Extended Figures 3)**. Unlike the oligodendrocyte lineage, we found that the neuronal subtype abundances were largely conserved across species **(Extended Figures 3B,H, Supplementary Table 2)**. The rates of gene regulatory changes were similar between human and chimpanzee lineages across most subtypes **(Extended Figure 5C-D)**. However, a few neuronal subtypes displayed signatures of human-specific acceleration in the epigenome (e.g. L2-3_1) or the transcriptome (e.g. L5-6 FEZF2_1) **(Extended Figure 5C-D)**.

We observed a high heterogeneity of HS changes among neuronal subtypes **(Extended Figure 5E)**. Since most previous comparative studies lacked cellular resolution at the subtype level, we assessed reproducibility between the previous bulk comparisons[12,18] and the subtype-resolved comparisons. While we found an overall enrichment between the species-specific genes across different studies **(Extended Figure 6A-C)**, bulk studies consistently showed low overlap with the more subtype-specific HS changes **(Extended Figure 6D-E)**. Notably, when we pooled the excitatory subtypes, our power to detect subtype-specific HS changes were also substantially reduced **(Extended Figure 6F-G)**. Therefore, most neuronal HS changes are not shared by more than a few subtypes and are masked in bulk approaches.

Subtype-specific evolution of *FOXP2*

We examined human-specific expression of transcription factors that are altered in only a few subtypes and found that *FOXP2*, a key transcription factor known for its roles in the development of cortical-striatal circuits related to speech and language and human brain evolution[25,26], showed human-specific upregulation in two

excitatory subtypes **(Figure 3A)**. This contrasted with the previous comparative studies of adult cortex that did not find a significant difference in the *FOXP2* expression between human and chimpanzee neurons[11,12,14,18]. Among these two subtypes, the L5-6_THEMIS_1 subtype (the most abundant THEMIS+ subtype, also marked by *C1QL3,* **Extended Figure 3E**) displayed low levels of *FOXP2* in non-human primates **(Figure 3A)**. We used smFISH to independently validate this finding in intact tissues and confirmed both more *FOXP2* and *THEMIS* co-positive cells in human compared to chimpanzee **(Figure 3B-C)**, and more *FOXP2*+ puncta in human *THEMIS*+ cells but not in *THEMIS*- cells **(Figure 3B,D)**. Interestingly, a recent study found similar *FOXP2* levels across species in all neuronal subtypes of the dorsolateral prefrontal cortex[14]. Corroborating this result, we also found significantly lower levels of *FOXP2* in the *THEMIS*+ *C1QL3*+ neurons of prefrontal cortex and anterior cingulate cortex in an independent dataset[23] **(Figure 3E-F)**. These results suggest that subtype-specific upregulation of *FOXP2* is also brain region-specific. Notably, some of the experimentally validated *FOXP2* downstream targets (*VLDLR*, *SRPX2*, *CNTNAP2*, *MET*, *DISC1*)[25] are not human-specifically altered in these two subtypes, indicating potentially distinct FOXP2 gene regulation among neuronal subtypes in the cortex **(Figure 3A; Supplementary Table 3)**. Two previously identified FOXP2 targets, *CNTNAP2* and *MET* are human-specifically upregulated in layer 4 subtypes **(Figure 3A)**. These results indicate a previously underappreciated neuronal subtype heterogeneity of key functional regulators in human brain evolution.

Co-evolution of chromatin and RNA

We then investigated the overall association between chromatin accessibility changes and gene expression changes. We found that the association between human specific gene expression and chromatin accessibility changes was the strongest at promoters and declined with the distance from the TSS **(Extended Figure 7A)**. This trend was observed only among the gains and losses that are concordant between the genome and the transcriptome (HS-Up-Gene and HS-Open-CRE/HS-Down-Gene and HS-Closed-CRE) but not in the discordant overlaps (e.g. HS-Up-Gene and HS-Closed-CRE) **(Extended Figure 7A)**. Overlaps for concordant, but not discordant, gains or losses were significant for nearly all subtypes **(Extended Figure 7B-D)**. Together these results show that HS-CREs are significantly associated with HS-Genes and this association is stronger if the former is near TSS and both are altered in the same direction.

To further refine associations between CREs and HS-Genes, we scanned the 500kb vicinity of each HS-Gene for HS-CREs that are altered in the same direction in the same cell type. This analysis assigned at least one HS-CRE to 26% of HS-Genes across cell types **(Supplementary Table 6)**. Focusing on the *FOXP2* gene and surrounding genomic regions, we identified four HS-Open-CREs in the L5-6_THEMIS_1 subtype. Two of these CREs are also close to another HS-Up-Gene (*MDFIC*) **(Figure 3G, Supplementary Table 6)**. Among the other two, one resides within a *FOXP2* intron, whereas the other one is ~244kb away from the nearest *FOXP2* TSS. To identify putative targets of human-specific *FOXP2* upregulation, we then retained HS-CREs that have a FOXP2 motif and are associated with an HS-Gene in the same subtype. This analysis yielded 47 genes for the L5-6_THEMIS_1 subtype and 14 genes for the L4-6_RORB_2 subtype **(Extended Figure 7E-F, Supplementary Table 6)**. We note that the *FOXP2* upregulation in L5-6_THEMIS_1 is greater than in L4-6_RORB_2 (logFC 0.8 and 0.4, respectively), and our analysis identified 3.35-fold more putative FOXP2 targets in L5-6_THEMIS_1 than in L4-6_RORB_2 (human-specific changes are only 1.8-fold more in L5-6_THEMIS_1 than L4-6_RORB_2). We further highlighted 7 genes that are not altered in the other 12 subtypes, similar to *FOXP2* itself **(Extended Figure 7E)**. Together, these results provide a list of potential epigenomic alterations associated with transcriptomic alterations in human brain evolution.

Novel human accelerated regions

A goal of comparative genomic studies is to connect the changes at genomic sequences to functional changes. We therefore focused on human accelerated regions (HARs)[27], which are genomic regions that have significantly accelerated sequence evolution in the human lineage[28]. We found that 30% of published HARs overlapped the CREs in our dataset (~2.5-fold excess compared to randomized background, p-value < 0.05. **Extended Figure 8A**), reaffirming the significance of these regions in human brain evolution[10,29,30]. Published HARs within CREs also showed modest but significant enrichment in HS-CREs in several cell types **(Figure 4A)**. However, these published HARs utilize sequence evolution without consideration of a specific tissue. Leveraging the snATAC-seq dataset, we hypothesized that we could find many accelerated genomic regions by performing HAR analysis restricted to the CREs we identified (see Methods). The odds ratio of published HAR and HS-CRE association

is ~1.4, which was achieved in our analysis with an unadjusted p-value cutoff of 0.001 (**Figure 4B, Supplementary Table 7**). We note that, in contrast to previous genome-wide approaches, this focused approach to define HARs allows us to relax statistical criteria (unadjusted $p < 0.001$) without reducing the effect sizes observed in published HARs, while simultaneously enhancing validity by linking substitution changes to functional changes (i.e. HS-CREs). We named these segments "cortical HARs" since the cellular composition of cortical brain regions are similar and we found that CREs from other cortical regions show a high degree of overlap with our dataset **(Extended Figure 8B)**. Many published HARs are also cortical HARs **(Extended Figure 8C)** and we identified >3 fold more HS-CREs overlapping a cortical HAR than overlapping a published HAR **(Extended Figure 8D)**. Cortical HARs were also significantly enriched in HS-CREs from most cell types **(Figure 4C),** and we highlight some notable examples of HS-CRE associated HARs that are important for synaptic (*CELF4*)[31] or oligodendrocyte (*NRG3*)[32] function **(Extended Figure 8E-F)**. Together, these results demonstrate a significant association between sequence divergence and chromatin accessibility in human evolution and provide hundreds of novel HARs accompanying chromatin accessibility change at cell type resolution in the human brain.

Chromatin evolution in modern humans

Comparison of anatomically modern human genomes to those of archaic humans permits the identification of 'modern human-specific' variants with unknown functional consequences[33]. We thus investigated the associations between modern human-specific variants and chromatin changes in the brain. In total, we identified 12,161 modern human-specific variants associated with HS-CREs **(Supplementary Table 8)**, which was a significant enrichment ($p= 0.007$, see Methods). Among the cell types, we found significant a enrichment only in upper layer excitatory neurons **(Figure 4D)**.

To compare the enrichments of modern human-specific variants to those that are specific to the entire human lineage (termed 'human-specific' henceforth), we first identified ~1.5 million human-specific substitutions within the CREs **(Supplementary Table 9)**. Similar to the HARs, human-specific substitutions were significantly enriched in HS-CREs, and we noted the example of *GRIK4*, which encodes a glutamate receptor subunit

implicated in brain disease[34] **(Extended Figure 8G-H)**. As expected, human-specific substitutions also encompassed ~88% of previously identified modern human-specific variants **(Extended Figure 8I)**. To reduce the confounding effects of sample sizes, we randomly down-sampled human-specific substitutions to match the number of modern human-specific variants and calculated their association with HS-CREs per cell type. This analysis revealed greater associations between modern human-specific variants and upper layer HS-CREs compared to the substitutions along the entire human lineage **(Figure 4E, Extended Figure 8J)**. Gene ontology enrichment analysis of HS-CREs with modern human-specific variants revealed the ephrin receptor signaling pathway as the only ontological enrichment **(Figure 4F-G)**. These results indicate that modern human-specific variants are associated with human-specific CRE changes.

Activity-response elements in human CREs

Transcription factors (TFs) are key components in evolution and disease. We found enrichments of diverse TF binding motifs in HS-Open-CREs across neuronal subtypes **(Extended Figure 9A-B, Supplementary Table 10)**. Notably, we observed significant enrichments for FOS / JUN motifs in the upper layer excitatory neurons and for FOX motifs in the lower layer excitatory neurons **(Figure 5A-B)**. We further identified TFs that may be functional at these HS-CRE target sites by examining the accessibility of each enriched TF within each family **(Extended Figure 9C-D**, **Figure 5A-B)**.

FOS / JUN TFs are immediately transcribed upon neuronal depolarization and target hundreds of CREs[35,36]. Since FOS / JUN TFs respond to environmental stimuli, we decided to test whether FOS / JUN TF enrichment in HS-Open-CREs is driven by environmental factors. We first asked whether greater post-mortem-interval (PMI) in human tissues compared to chimpanzee and rhesus macaque tissues, a limitation to many similar studies[12,14,18], is driving this enrichment. To test this, we substituted our human snATAC-seq dataset (named PMI_24) with a surgical human dataset from middle temporal gyrus that has no PMI (PMI_0)[37]. We similarly found all excitatory subtypes in this dataset and identified HS-CREs which displayed highly significant overlaps with the PMI_24 HS-CREs **(Extended Figure 9E-G)** as well as enrichments of similar motifs **(Extended Figure 9H)**. Similar to the PMI_24 dataset, HS-Open-CREs in the upper layer excitatory neurons were highly enriched

in FOS / JUN motifs **(Extended Figure 9I)**. These results show that FOS / JUN enrichments in upper layer excitatory HS-Open-CRE are not driven by PMI differences.

To provide an orthogonal test for a possible environmental effect on FOS / JUN motif enrichments, we asked if HS-Open-CREs with FOS / JUN motifs also contain signatures of accelerated evolution. If FOS / JUN motif enrichments in HS-Open-CREs are driven by environmental factors, human-specific substitutions within the HS-Open-CREs with FOS / JUN motif occurrences should be depleted compared to other HS-Open-CREs. Contrary to this expectation, we found a significant excess of HS-substitutions and accelerated evolution when HS-Open-CREs with FOS / JUN motifs were compared to the non-specific (NS) CREs **(Figure 5C-D)**. FOX targets were also more divergent in humans compared NS-CREs with or without FOX motifs **(Figure 5E-F)**. An example of a human-specific gain of a FOS / JUN motif within a human-accelerated region is shown near *MTHFD2L* **(Figure 5G)**, which encodes a key enzyme in the one-carbon metabolism associated with neurotransmitter synthesis[38]. Taken together, these results do not support a possible environmental cause.

In summary, we have uncovered proportional and gene regulatory changes in human brain evolution using single cell genomics and have linked human-specific DNA sequence divergence, chromatin accessibility and gene expression at cellular resolution.

Discussion

In this study, we delineated epigenomic and transcriptomic features of human brain evolution at cell type resolution. We found that the adult human cortex had an increased proportion of oligodendrocyte progenitor cells and a decreased proportion of mature oligodendrocytes compared to non-human primates. Focusing on neurons, we showed that many human-specific changes were found in only a few neuronal subtypes, and demonstrated human-specific up-regulation of *FOXP2* in two neuronal subtypes. We also associated genomic sequence changes with HS-CREs at cellular resolution and identified hundreds of novel HARs that were associated with open chromatin in the adult brain. Furthermore, we identified increased FOS / JUN transcription

factor targets among the HS-Open-CREs in the upper layer excitatory neurons, emphasizing a previously underappreciated temporal dimension of human-specific molecular traits.

Previous studies showed prolonged myelination in human brain development compared to chimpanzees and rhesus macaques[5,8]. Correspondingly, the production of myelinating oligodendrocytes reaches a plateau in individuals older than ~40 years old in gray matter[39]. Interestingly, we observed proportionally higher OPCs in humans compared to chimpanzees and rhesus macaques even though individuals in our dataset are all in their mid- to late- adulthood (humanized age, **Supplementary Table 1)**. We also found that COPs, cells that denote active oligodendrocyte generation[40], are extremely rare (only 74 nuclei in all species), indicating low levels of oligodendrocyte generation in all species in our samples. We hypothesize that the higher proportion of OPCs and lower proportion of MOLs can contribute to neural plasticity in the human brain by altering myelination patterns. Non-canonical functions of OPCs such as pruning axonal branches and contributing to synaptic function have been recently described[41], indicating that increased numbers of OPCs in the human brain may serve functions other than providing a reservoir for mature oligodendrocytes. We note that a recent study found more divergence between species for MOLs compared to OPCs by comparing the gene expression correlations[42]. The discrepancy with our results could be due to differences in the brain regions analyzed, sorting strategy (NeuN- sorted versus not sorted), as well as in the analytical pipeline (e.g., correlations vs. differential gene expression). We also note that our approach separates human-specific changes to chimpanzee-specific changes as a measure of human-specificity, making it better tailored to highlight the changes in human lineage.

Single-cell sequencing facilitates characterization of regulatory changes in all cell types. However, we recently discovered that neuronal ambient RNAs contaminate glial cell types and require rigorous removal before identification of differentially expressed genes[21], as such contamination can skew the differential gene expression results[43]. We found that a recent study[14] shows evidence of human-specific differences in the level of ambient RNA contamination in glial cell types, indicating the importance of ambient contamination removal **(Extended Figure 10A-B)**. Among the studies with human-chimpanzee comparisons, our snRNA-seq dataset is thus far the only one to remove ambient RNA contamination[11,14]. Upon removal of ambient RNAs, we

uncovered that cytoskeletal activity and ubiquitin ligase activity through *SH3RF3* are specifically decreased in human OPCs **(Figure 2C-G)**. Both biological processes are linked to oligodendrocyte maturation, indicating that such functions might be linked to the human-specific OPC increase[20,24]. These results also suggest that an evolutionary modification in human brain may have been achieved through a loss of function in OPCs[44].

We found a subtype- and human-specific upregulation of *FOXP2* which may be unique to the posterior cingulate cortex. We also note that most FOX TF motifs are enriched in the HS-Open-CREs in *THEMIS*+ *C1QL3*+ neurons **(Figure 5B)**, and while the FOXP2 motif enrichment itself was not significant, this could be ascribed to possible variations of FOXP2 binding sites in different tissues. Indeed, we previously showed that FOXP2 can act both as a repressor or activator via heterodimerization with other TFs at distinct DNA motifs[45]. In addition, a recent study identified human-specific *FOXP2* upregulation in microglia[14], with a similar trend in our dataset **(Supplementary Table 3)**, suggesting a previously undescribed potential role of FOXP2. These results provide further insights into the role of *FOXP2* in human brain evolution.

Interestingly, association between human-specific gene expression changes and chromatin accessibility changes was significant only between the concordant changes but not between discordant changes (**Extended Figure 7**). While CREs can act as repressors and cause downregulation of their target genes (which would lead to discordant overlaps), the repressor activity of the CRE recruits histone deacetylases and closes the chromatin accessibility on its target regions[46], which could lead to overall closed accessibility, thus manifesting as a concordant change.

We discovered a novel enrichment of FOS and JUN family motifs in specifically cortical upper layer excitatory HS-Open-CREs. Late activity-regulated genes are known to be evolutionarily divergent[35], [47-49] and display high cell type specificity[50]. Along with the previous studies, our results underscore the need for more direct experiments to understand how adult human cortical cells respond to neuronal activity, and the underlying evolutionary trajectories. We also note that some of our analyses are limited to BA23 and future comparative

studies from other brain regions are needed. Overall, our results provide a comprehensive roadmap for delineating functional regulatory mechanisms of human brain evolution at cellular resolution.

References

1 King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107-116, doi:10.1126/science.1090005 (1975).

2 Konopka, G. *et al.* Human-specific transcriptional networks in the brain. *Neuron* **75**, 601-617, doi:10.1016/j.neuron.2012.05.034 (2012).

3 Liu, X. *et al.* Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques. *Genome Res* **22**, 611-622, doi:10.1101/gr.127324.111 (2012).

4 Sousa, A. M. M. *et al.* Molecular and cellular reorganization of neural circuits in the human lineage. *Science* **358**, 1027-1032, doi:10.1126/science.aan3456 (2017).

5 Zhu, Y. *et al.* Spatiotemporal transcriptomic divergence across human and macaque brain development. *Science* **362**, doi:10.1126/science.aat8077 (2018).

6 Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61-68, doi:10.1038/s41586-019-1506-7 (2019).

7 Bakken, T. E. *et al.* Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111-119, doi:10.1038/s41586-021-03465-8 (2021).

8 Miller, D. J. *et al.* Prolonged myelination in human neocortical evolution. *Proc Natl Acad Sci U S A* **109**, 16480-16485, doi:10.1073/pnas.1117943109 (2012).

9 Jakel, S. *et al.* Altered human oligodendrocyte heterogeneity in multiple sclerosis. *Nature* **566**, 543-547, doi:10.1038/s41586-019-0903-2 (2019).

10 Jeong, H. *et al.* Evolution of DNA methylation in the human brain. *Nat Commun* **12**, 2021, doi:10.1038/s41467-021-21917-7 (2021).

11 Khrameeva, E. *et al.* Single-cell-resolution transcriptome map of human, chimpanzee, bonobo, and macaque brains. *Genome Res* **30**, 776-789, doi:10.1101/gr.256958.119 (2020).

12 Kozlenkov, A. *et al.* Evolution of regulatory signatures in primate cortical neurons at cell-type resolution. *Proc Natl Acad Sci U S A* **117**, 28422-28432, doi:10.1073/pnas.2011884117 (2020).

13 Krienen, F. M. *et al.* Innovations present in the primate interneuron repertoire. *Nature* **586**, 262-269, doi:10.1038/s41586-020-2781-z (2020).

14 Ma, S. *et al.* Molecular and cellular evolution of the primate dorsolateral prefrontal cortex. *Science*, eabo7257, doi:10.1126/science.abo7257 (2022).

15 Mendizabal, I. *et al.* Comparative Methylome Analyses Identify Epigenetic Regulatory Loci of Human Brain Evolution. *Mol Biol Evol* **33**, 2947-2959, doi:10.1093/molbev/msw176 (2016).

16 Li, W., Mai, X. & Liu, C. The default mode network and social understanding of others: what do brain connectivity studies tell us. *Front Hum Neurosci* **8**, 74, doi:10.3389/fnhum.2014.00074 (2014).

17 Wang, D. *et al.* Altered functional connectivity of the cingulate subregions in schizophrenia. *Transl Psychiatry* **5**, e575, doi:10.1038/tp.2015.69 (2015).

18 Berto, S. *et al.* Accelerated evolution of oligodendrocytes in the human brain. *Proc Natl Acad Sci U S A* **116**, 24334-24342, doi:10.1073/pnas.1907982116 (2019).

19 Franjic, D. *et al.* Transcriptomic taxonomy and neurogenic trajectories of adult human, macaque, and pig hippocampal and entorhinal cells. *Neuron* **110**, 452-469 e414, doi:10.1016/j.neuron.2021.10.036 (2022).

20 Brown, T. L. & Verden, D. R. Cytoskeletal Regulation of Oligodendrocyte Differentiation and Myelination. *J Neurosci* **37**, 7797-7799, doi:10.1523/JNEUROSCI.1398-17.2017 (2017).

21 Caglayan, E., Liu, Y. & Konopka, G. Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. *Neuron*, doi:10.1016/j.neuron.2022.09.010 (2022).

22 Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* **36**, 70-80, doi:10.1038/nbt.4038 (2018).

23 Velmeshev, D. *et al.* Single-cell genomics identifies cell type-specific molecular changes in autism. *Science* **364**, 685-689, doi:10.1126/science.aav8130 (2019).

24 Fumagalli, M. *et al.* The ubiquitin ligase Mdm2 controls oligodendrocyte maturation by intertwining mTOR with G protein-coupled receptor kinase 2 in the regulation of GPR17 receptor desensitization. *Glia* **63**, 2327-2339, doi:10.1002/glia.22896 (2015).

25 den Hoed, J., Devaraju, K. & Fisher, S. E. Molecular networks of the FOXP2 transcription factor in the brain. *EMBO Rep* **22**, e52803, doi:10.15252/embr.202152803 (2021).

26 Konopka, G. *et al.* Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature* **462**, 213-217, doi:10.1038/nature08549 (2009).

27 Doan, R. N. *et al.* Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. *Cell* **167**, 341-354 e312, doi:10.1016/j.cell.2016.08.071 (2016).

28 Franchini, L. F. & Pollard, K. S. Human evolution: the non-coding revolution. *BMC Biol* **15**, 89, doi:10.1186/s12915-017-0428-9 (2017).

29 Capra, J. A., Erwin, G. D., McKinsey, G., Rubenstein, J. L. & Pollard, K. S. Many human accelerated regions are developmental enhancers. *Philos Trans R Soc Lond B Biol Sci* **368**, 20130025, doi:10.1098/rstb.2013.0025 (2013).

30 Girskis, K. M. *et al.* Rewiring of human neurodevelopmental gene regulatory programs by human accelerated regions. *Neuron*, doi:10.1016/j.neuron.2021.08.005 (2021).

31 Wagnon, J. L. *et al.* CELF4 regulates translation and local abundance of a vast set of mRNAs, including genes associated with regulation of synaptic function. *PLoS Genet* **8**, e1003067, doi:10.1371/journal.pgen.1003067 (2012).

32 Lundgaard, I. *et al.* Neuregulin and BDNF induce a switch to NMDA receptor-dependent myelination by oligodendrocytes. *PLoS Biol* **11**, e1001743, doi:10.1371/journal.pbio.1001743 (2013).

33 Prufer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43-49, doi:10.1038/nature12886 (2014).

34 Arora, V. *et al.* Increased Grik4 Gene Dosage Causes Imbalanced Circuit Output and Human Disease-Related Behaviors. *Cell Rep* **23**, 3827-3838, doi:10.1016/j.celrep.2018.05.086 (2018).

35 Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182-187, doi:10.1038/nature09033 (2010).

36 Yap, E. L. & Greenberg, M. E. Activity-Regulated Transcription: Bridging the Gap between Neural Activity and Behavior. *Neuron* **100**, 330-348, doi:10.1016/j.neuron.2018.10.013 (2018).

37 Berto, S. *et al.* Gene-expression correlates of the oscillatory signatures supporting human episodic memory encoding. *Nat Neurosci* **24**, 554-564, doi:10.1038/s41593-021-00803-x (2021).

38 Ducker, G. S. & Rabinowitz, J. D. One-Carbon Metabolism in Health and Disease. *Cell Metab* **25**, 27-42, doi:10.1016/j.cmet.2016.08.009 (2017).

39 Yeung, M. S. *et al.* Dynamics of oligodendrocyte generation and myelination in the human brain. *Cell* **159**, 766-774, doi:10.1016/j.cell.2014.10.011 (2014).

40 Marques, S. *et al.* Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science* **352**, 1326-1329, doi:10.1126/science.aaf6463 (2016).

41 Buchanan, J. *et al.* Oligodendrocyte precursor cells ingest axons in the mouse neocortex. *Proc Natl Acad Sci U S A* **119**, e2202580119, doi:10.1073/pnas.2202580119 (2022).

42 Jorstad, N. L. *et al.* Comparative transcriptomics reveals human-specific cortical features. *bioRxiv*, 2022.2009.2019.508480, doi:10.1101/2022.09.19.508480 (2022).

43 Berg, M. *et al.* FastCAR: Fast Correction for Ambient RNA to facilitate differential gene expression analysis in single-cell RNA-sequencing datasets. *bioRxiv*, 2022.2007.2019.500594, doi:10.1101/2022.07.19.500594 (2022).

44 McLean, C. Y. *et al.* Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**, 216-219, doi:10.1038/nature09774 (2011).

45 Hickey, S. L., Berto, S. & Konopka, G. Chromatin Decondensation by FOXP2 Promotes Human Neuron Maturation and Expression of Neurodevelopmental Disease Genes. *Cell Rep* **27**, 1699-1711 e1699, doi:10.1016/j.celrep.2019.04.044 (2019).

46 Yang, C. C. *et al.* Discovering chromatin motifs using FAIRE sequencing and the human diploid genome. *BMC Genomics* **14**, 310, doi:10.1186/1471-2164-14-310 (2013).

47 Ataman, B. *et al.* Evolution of Osteocrin as an activity-regulated factor in the primate brain. *Nature* **539**, 242-247, doi:10.1038/nature20111 (2016).

48 Pruunsild, P., Bengtson, C. P. & Bading, H. Networks of Cultured iPSC-Derived Neurons Reveal the Human Synaptic Activity-Regulated Adaptive Gene Program. *Cell Rep* **18**, 122-135, doi:10.1016/j.celrep.2016.12.018 (2017).
49 Qiu, J. *et al.* Evidence for evolutionary divergence of activity-dependent gene expression in developing neurons. *Elife* **5**, doi:10.7554/eLife.20337 (2016).
50 Hrvatin, S. *et al.* Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat Neurosci* **21**, 120-129, doi:10.1038/s41593-017-0029-5 (2018).
51 Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049, doi:10.1038/ncomms14049 (2017).
52 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
53 Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006-1007, doi:10.1093/bioinformatics/btt730 (2014).
54 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2014).
55 Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491-499, doi:10.1101/gr.209601.116 (2017).
56 Fleming, S. J., Marioni, J. C. & Babadi, M. CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets. 791699, doi:10.1101/791699 %J bioRxiv (2019).
57 Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res* **49**, D884-D891, doi:10.1093/nar/gkaa942 (2021).
58 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821, doi:10.1016/j.cell.2019.05.031 (2019).
59 *Picard Toolkit*, <http://broadinstitute.github.io/picard/> (2019).
60 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).
61 Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief Bioinform* **14**, 144-161, doi:10.1093/bib/bbs038 (2013).
62 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
63 Lareau, C. A., Ma, S., Duarte, F. M. & Buenrostro, J. D. Inference and effects of barcode multiplets in droplet-based single-cell assays. *Nat Commun* **11**, 866, doi:10.1038/s41467-020-14667-5 (2020).
64 Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat Methods* **18**, 1333-1341, doi:10.1038/s41592-021-01282-5 (2021).
65 Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289-1296, doi:10.1038/s41592-019-0619-0 (2019).
66 Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol Cell* **71**, 858-871 e858, doi:10.1016/j.molcel.2018.06.044 (2018).
67 Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67**, 1 - 48, doi:10.18637/jss.v067.i01 (2015).
68 Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16**, 278, doi:10.1186/s13059-015-0844-5 (2015).
69 Chen, Y., Lun, A. T. & Smyth, G. K. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Res* **5**, 1438, doi:10.12688/f1000research.8987.2 (2016).
70 McCarthy, D. J., Campbell, K. R., Lun, A. T. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179-1186, doi:10.1093/bioinformatics/btw777 (2017).
71 Gontarz, P. *et al.* Comparison of differential accessibility analysis strategies for ATAC-seq data. *Sci Rep* **10**, 10150, doi:10.1038/s41598-020-66998-4 (2020).
72 Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun* **10**, 380, doi:10.1038/s41467-018-08023-x (2019).
73 Mendizabal, I. *et al.* Cell type-specific epigenetic links to schizophrenia risk in the brain. *Genome Biol* **20**, 135, doi:10.1186/s13059-019-1747-7 (2019).

74 van Arensbergen, J., van Steensel, B. & Bussemaker, H. J. In search of the determinants of enhancer-promoter interaction specificity. *Trends Cell Biol* **24**, 695-702, doi:10.1016/j.tcb.2014.07.004 (2014).
75 Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287, doi:10.1089/omi.2011.0118 (2012).
76 Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics* **33**, 2381-2383, doi:10.1093/bioinformatics/btx183 (2017).
77 Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res* **46**, D260-D266, doi:10.1093/nar/gkx1126 (2018).
78 A, S. motifmatchr: Fast Motif Matching in R. *R package version 1.4.0* (2018).
79 Kolde, R. *Pheatmap: pretty heatmaps. R*, <https://cran.r-project.org/web/packages/pheatmap/index.html> (2012).
80 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).
81 Gittelman, R. M. *et al.* Comprehensive identification and analysis of human accelerated regulatory DNA. *Genome Res* **25**, 1245-1255, doi:10.1101/gr.192591.115 (2015).
82 Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**, 708-715, doi:10.1101/gr.1933104 (2004).
83 Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief Bioinform* **12**, 41-51, doi:10.1093/bib/bbq072 (2011).
84 Mafessoni, F. *et al.* A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc Natl Acad Sci U S A* **117**, 15132-15136, doi:10.1073/pnas.2004944117 (2020).
85 Prufer, K. *et al.* A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655-658, doi:10.1126/science.aao1887 (2017).
86 Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).
87 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).
88 Ghandi, M. *et al.* gkmSVM: an R package for gapped-kmer SVM. *Bioinformatics* **32**, 2205-2207, doi:10.1093/bioinformatics/btw203 (2016).

Methods

Specific details of all analyses can be found in our github page: https://github.com/konopkalab/Comparative_snATAC_snRNA

Sampling strategy for snRNA-seq and snATAC-seq

All human tissue was obtained from the University of Texas Neuropsychiatry Research Program (Dallas Brain Collection). Chimpanzee and macaque tissue were obtained from Yerkes National Primate Research Center. Brodmann area 23 (BA23, part of the posterior cingulate cortex) was dissected from frozen post-mortem tissue slabs. Humanized age (calculated as described before[18]) and sex were matched between species to minimize the effect of demographics. In total, 4 individuals were sequenced from each species **(Extended Figure 1A, Supplementary Table 1)**.

Single-nuclei RNA-seq library Preparation

Nuclei for snRNA-seq were isolated from human, chimpanzee, and macaque BA23 brain tissue. Briefly, the tissue was homogenized using a glass Dounce homogenizer in 2 ml of ice-cold lysis buffer (10 mM Tris-HCl, 10 mM NaCl, 3 mM MgCl2, and 0.1% Nonidet™ P40 Substitute) and was incubated on ice for 5 min. Nuclei were centrifuged at 500 × g for 5 min at 4 °C, washed with 4 ml ice-cold lysis buffer and, incubated on ice for 5 min. Nuclei were centrifuged at 500 × g for 5 min at 4 °C. After centrifugation, the nuclei were resuspended in 500 µl of nuclei suspension buffer (NSB) containing 1XPBS, 1%BSA (#AM2618, Thermo Fisher Scientific) and 0.2U/ul RNAse inhibitor (#AM2694, Thermo Fisher Scientific). Nuclei suspension was filtered through a 70-µm Flowmi Cell Strainer (#H13680-0070, Bel-Art). Debris was removed with a density gradient centrifugation using the Nuclei PURE 2M Sucrose Cushion Solution and Nuclei PURE Sucrose Cushion Buffer from Nuclei PURE Prep Isolation Kit (#NUC201-1KT, Sigma Aldrich). Nuclei PURE 2M Sucrose Cushion Solution and Nuclei PURE Sucrose Cushion Buffer were first mixed in a 9:1 ratio. 500 µl of the resulting sucrose solution was added to a 2 ml Eppendorf tube. 900 µl of the sucrose buffer was added to 500 µl of isolated nuclei in NSB. 1400 µl nuclei suspension was layered to the top of the sucrose buffer. This gradient was centrifuged at 13, 000 x g for 45 min at 4 °C. Nuclei pellet was resuspended, washed once in NSB and, filtered through a 70-µm Flowmi Cell Strainer (#H13680-0070, Bel-Art). Nuclei concentration was determined using 0.4% Trypan Blue (#15250061, Thermo Fisher Scientific). A final concentration of 1000 nuclei/µl was adjusted with NSB.

Droplet-based single-nuclei RNA-seq libraries were prepared using the Chromium Single Cell 3' v3.1 (1000121, 10x Genomics) according to the manufacturer's protocol[51]. Libraries were sequenced using an Illumina NovaSeq 6000.

Single-nuclei ATAC-seq library Preparation

For snATAC-seq, nuclei were isolated from human, chimpanzee, and macaque BA23 tissue as previously described (https://www.protocols.io/view/isolation-of-nuclei-from-frozen-tissue-for-atac-se-6t8herw). Briefly, tissue pieces neighboring to the tissue used for snRNA-seq were cut and homogenized using a glass Dounce homogenizer in ATAC-seq homogenization buffer (0.25M sucrose, 25mM KCl, 5mM $MgCl_2$, 20mM Tricine-KOH

(pH7.8), 1 mM DTT, 0.5mM spermidine, 0.15mM spermine, 0.3% NP40, protease inhibitors). The nuclei filtered through a 70-µm Flowmi Cell Strainer (#H13680-0070, Bel-Art) and were pelleted by centrifugation for 5 min at 4 °C at 350 x g in a 2 ml Eppendorf tube. The supernatant was discarded, and the nuclei were resuspended in 400 µl of homogenization buffer. 400 µl of 50% iodixanol solution was added to the nuclei suspension and was mixed by pipetting. 600 µl of 30% Iodixanol solution was layered under the 25% mixture. 600 µl of 40% Iodixanol solution was then layered under the 30% mixture. This gradient then centrifuged for 20 minutes at 4°C at 3,000 x g. After centrifugation the nuclei were recovered at the 30%-40% interface. Transfer the nuclei in a new Eppendorf tube and resuspend in 200 µl ATAC-RSC-Tween buffer (10mM Tris-HCl pH7.5, 10mM NaCl, 3mM $MgCl_2$, 0.1% Tween-20). Nuclei concentration was determined using 0.4% Trypan Blue (#15250061, Thermo Fisher Scientific). Nuclei integrity was tested by staining with Ethidium Homodimer-1 staining (Cat#E1169, Invitrogen). Droplet-based single-nuclei ATAC-seq libraries were prepared using the Chromium Single Cell ATAC Library kit (1000110, 10x Genomics) according to the manufacturer's protocols. Libraries were sequenced using an Illumina NovaSeq 6000.

Single-nuclei RNA-seq preprocessing and annotation

Bcl files were converted to fastq using *cellranger mkfastq*. Barcode correction and reference genome alignment were done using *cellranger count* with default parameters (Software: 10x Genomics Cell Ranger 3.1.0). For the alignment, genome builds GRCh38, panTro5 (Pan_tro 3.0), rheMac10 (Mmul_10) were used as reference genomes for human, chimpanzee and macaque, respectively. The BAM output from *cellranger count* was further processed to keep only uniquely mapped reads using samtools (*-q 255*)[52]. Since chimpanzee and macaque gene annotation files (gtf) are less accurate than human, chimpanzee and macaque reads were then mapped to human coordinates using CrossMap[53]. *featureCount* was used to count reads mapping to gene body[54], and *umi_tools*[55] was used to create the count matrix (Gene by cell barcode. Per sample, the top 50,000 cell barcodes with highest UMI count were pre-filtered for faster computation).

To remove ambient RNA contamination, we used CellBender on the un-normalized count matrix per sample[56]. We note that without ambient RNA removal, glial cells were shown to be conspicuously contaminated with neuronal ambient RNAs[21].

Empty-droplet filtered output from CellBender was further processed to retain only the protein coding and orthologous genes (between *H. sapiens*, *P. troglodytes*, *M. mulatta*) similar to Berto et al[18]. An orthologous gene list was obtained from Ensembl version 103[57]. For quality control, we only kept nuclei with >200 UMI and percentage of reads mapping to mitochondria < 5. We then clustered nuclei for further analysis. The following methods from Seurat v3[58] were used to perform and visualize clustering (a similar approach was followed for each new clustering; details are available in the publicly available code): normalization (*SCTransform*), dimensionality reduction (*RunPCA*), batch correction (*RunHarmony*, default parameters), k-nearest neighbors (*FindNeighbors)* on batch corrected dimensions and clusters identification by shared nearest neighbors (*FindClusters*). UMAP embedding was then computed for visualization in 2D space (*RunUMAP*). We removed clusters with an unusually high number of detected genes accompanied with high expression of at least two typically distinct marker genes as potential nuclei doublets. We re-clustered the nuclei and repeated this process if needed until no such clusters were found. We then used canonical marker genes (e.g *GAD1* for inhibitory neurons) and a reference dataset[6] (using label transfer, see next paragraph) to broadly annotate nuclei in each species. Major cell types were defined as: excitatory neurons, inhibitory neurons, astrocytes, oligodendrocytes, oligodendrocyte progenitor cells, and microglia. After broad annotation, we extracted each broad category (e.g. excitatory neurons) from all species and integrated them across species using the default approach in Seurat v3 across all samples (*SelectIntegrationFeatures, PrepSCTIntegration, FindIntegrationAnchors*). We then clustered the nuclei on the integrated matrix per cell type and further removed potential doublets with the same criteria as above. We additionally removed clusters with high enrichment in previously identified ambient RNA markers as previously described[21].

To annotate neuronal subtypes, we used a previous study[6] as reference to annotate our clusters via label transfer (*FindTransferAnchors*, *TransferData*). We assigned each cluster an annotation label based on the layer and marker gene of predominant predicted annotations per cluster. These annotations were also verified with known marker genes that separate certain neuronal subtypes **(Extended Figures 3)**.

We note that endothelial cells were removed from the analysis since we did not detect a distinct cluster of endothelial cells in snATAC-seq.

Single-nuclei ATAC-seq preprocessing and annotation

Bcl files were converted to fastq using *cellranger mkfastq*. Barcode correction and reference genome alignment were done using *cellranger-atac count* with default parameters (Software: 10x Genomics Cell Ranger ATAC 1.1.0). For the alignment, genome builds GRCh38, panTro5 (Pan_tro 3.0) and rheMac10 (Mmul_10) were used as reference genomes for human, chimpanzee and macaque, respectively. The BAM output from *cellranger count* was further processed to keep only uniquely mapped and properly paired reads using samtools[52]. Read duplicates were removed using *MarkDuplicates* from Picard tools[59]. Peak calling was performed using *macs2*[60] with following parameters: *--nomodel, --keep-dup all, extsize 200, --shift –100* to enrich for the cut sites. To obtain peaks concordant across samples, peak calling was performed by pooling all samples from each species as well as per each sample. Peaks from pooled samples were kept for further analysis only if they overlap >50% with peaks per sample in 3/4 of samples. This yielded a list of consensus peaks for each species.

To obtain a final set of peaks from consensus peaks, chimpanzee and macaque peaks were converted to human coordinates using *liftOver*[61]. All peaks were then merged using bedtools[62], resulting in merged peaks with a minimum distance of 200 between them (-d 200). To keep peaks with a reliable level of conservation across all species, merged peaks were reciprocally mapped to chimpanzee or macaque genomes and any peaks with more than 2-fold change in size, multi-mapped or less than 50% conserved (-minMatch=0.5) were discarded in each liftOver operation. Merged peaks were then filtered for the peaks that reciprocally mapped to both chimpanzee and macaque by requiring >50% overlap between a peak in the merged peak set and reciprocally mapped peak set (bedtools intersect -f 0.5 -F 0.5). Despite being conservative, this approach kept >93% of the initial peaks, indicating that sequence identity of most open chromatin peaks are reliably conserved across species and allow direct comparisons between species in downstream analysis (e.g differential accessibility). After this stage, peaks were also referred to as CREs (cis-regulatory regions).

To obtain peak-cell count matrix, reads were counted in each species' own coordinates using custom functions on bed files. To keep high quality cells, only the barcodes with >3000 reads in peaks, <100000 reads in peaks and >15% fraction of reads in peaks were kept for further analysis. Barcode multiplets[63] were additionally removed using cellranger's, *clean_barcode_multiplets_1.1.py* tool. Resulting matrices were processed separately in each species. Following methods from Seurat v3[58] were used to perform and visualize each clustering (please find details in the publicly available code): Dimensionality reduction was performed with latent

semantic indexing (LSI, using functions *RunTFIDF* and *RunSVD* in *Signac*[64]). Batch correction was achieved with harmony on LSI dimensions (*RunHarmony*[65]). Batch corrected dimensions were then used to compute k-nearest neighbors (*FindNeighbors)* and identify clusters by shared nearest neighbors (*FindClusters*). UMAP embedding was computed for visualization in 2D space (*RunUMAP*).

To annotate snATAC-seq cells, correspondence between gene accessibility and gene expression is required. To achieve this, gene activity matrix was calculated using *Cicero*[66] for each species. Only the CREs with more than 1% accessibility were retained for analysis, and CREs in protein coding genes (gene body +3kb upstream) were used to annotate CREs to genes (*annotate_cds_by_site*) which was further processed to build the unnormalized gene activity matrix (*build_gene_activity_matrix*). Both major cell types (e.g. Excitatory) and subtypes (e.g. L2-3_1) in snATAC-seq were annotated via label transfer with the corresponding snRNA-seq dataset as reference. All snRNA-seq to snATAC-seq label transfers were done separately for each species. Clusters with mixed annotation accompanied with unusually high number of reads in peaks and mixed marker gene activity (typically distinct marker genes highly accessible in the same cluster) were removed as potential doublets. Annotation label was assigned per cluster depending on the dominant annotation for each cluster. All cell types found in snRNA-seq were distinctly found in snATAC-seq and thus annotated with the same names.

Cell Type Fraction Comparisons

For comparison of cell type ratios, we calculated the fraction of glial cell types within all glia, the fraction of excitatory subtypes within all excitatory cells and the fraction of inhibitory subtypes within all inhibitor cells for each individual in both snRNA-seq and snATAC-seq. To determine whether the fraction differences were significant between species, we calculated the p-value using a log likelihood ratio on two nested models:

H0: Fraction ~ Assay (snRNA-seq or snATAC-seq)

H1: Fraction ~ Assay (snRNA-seq or snATAC-seq) + Species (e.g human and chimpanzee).

This was done for each pairwise species comparison per cell type. The statistics are available in **Supplementary Table 2**.

Single molecule fluorescent in situ hybridization

See **Supplementary Table 1** for sample demographics. Cortical BA23 (posterior cingulate cortex) and anterior cingulate cortex (ACC) samples from all species were postmortem, flash-frozen tissues that were embedded in OCT (optimal cutting temperature) compound. The tissue was sectioned at -20C to 20µm on Superfrost Plus Microscope slides. Single molecule fluorescent in situ hybridization (smFISH) was performed using RNAScope Multiplex v2 Fluorescent assays. Protease was applied for 30 minutes and all subsequent steps including probe application, tyramide signal amplification, channel development, and fluorophore application were performed according to the manufacturer's instructions for fresh frozen tissue except with the addition of Sudan Black B. 0.05% Sudan Black B was added to the tissue after application of DAPI to quench autofluorescence. Probes for *MOG* (human: 543181-C2, chimpanzee: 1076431-C2 Advanced Cell Diagnostics), *PDGFRA* (Advanced Cell Diagnostics, human: 604488, chimpanzee: 1120031), *THEMIS* (Advanced Cell Diagnostics, human: 407261), and *FOXP2* (Advanced Cell Diagnostics, human: 551661-C2) were incubated with the tissue and hybridized with their target genes. Opal fluorophores 570 (NC1601878, Akoya Biosciences, 1:750) and 620 (NC1612059, Akoya Biosciences, 1:750) were used to label the gene-specific probes after signal amplification. A 3-plex human (320861, Advanced Cell Diagnostics), and nonhuman primate (320901, Advanced Cell Diagnostics) positive control probe was used for each species alongside a primate negative control probe (320871, Advanced Cell Diagnostics).

To separate fluorophore signals, multispectral imaging was performed on a Zeiss LSM 880 in UT Southwestern's Quantitative Light Microscopy Core. Final imaging was performed on the Zeiss LSM 710 and Zeiss LSM 880 confocal microscope at x20 magnification in the UT Southwestern Neuroscience Microscopy Facility on chimpanzee and human samples.

To determine the composition of OPCs and MOLs in both BA23 (human: n=2, chimpanzee: n=3) and ACC (human: n=3, chimpanzee: n=3), We sampled 2-4 vertical bins (layer 1-6) of cortex from each individual and evenly divided each bin from the apical to basal boundary into 5 sections, and then we randomly selected 2-4 subareas (456x456 pixels) in each section to quantify the number of cells (DAPI 405 nm), OPCs (*PDGFRA*, 488 nm), and MOLs (*MOG*, 555 nm) by using self-generated ImageJ Macro code and R script in Fiji and R respectively (https://github.com/konopkalab/Comparative_snATAC_snRNA). Maximum intensity projection images were generated from 13 slices of Z stack. OPCs were defined as *PDGFRA* and DAPI double-positive cells, while MOLs were defined as *MOG* and DAPI double-positive cells. Data were analyzed using a linear

mixed model with species as the fixed factor and individual as the random factor per comparison (*lme4*[67] package in R, with REML = F).

To compare the expression of *FOXP2* in *THEMIS*+ neurons between human (n=3) and chimpanzee (n=3), we quantified the fraction of *FOXP2*+ neurons in *THEMIS*+ neurons, and the number of fluorescent puncta as a proxy for *FOXP2* expression levels in BA23. We sampled 2-3 images from the deep layers of each individual, and then we randomly selected 2-3 subareas of each image to quantify the fraction of DAPI (405 nm), *FOXP2* (488 nm), and *THEMIS* (555 nm) triple positive neurons in DAPI and *THEMIS* double positive neurons. Data were analyzed using a mixed linear model using species, image, and subarea as the fixed factors. For the puncta quantification, we used the same images as the fraction quantification but selected only the cells with individually distinguishable puncta. This resulted in the quantification of 3-11 THEMIS+ neurons and 3-9 THEMIS- neurons per image. Data were analyzed using a linear mixed model with species as the fixed factor and individual as the random factor per comparison (*lme4*[67] package in R, with REML = F).

Single-nuclei RNA-seq differential gene expression and identification of species-specific gene expression

We performed differential gene expression (DGE) using two approaches: a single-cell based DGE approach and a pseudobulk based DGE approach. We retained the pseudobulk DGE results for all analyses as both the HS-Genes and CS-Genes were more reproducible with previous studies[12,18] compared to the single-cell based DGE method **(Extended Figure 10C-E)**.

For the pseudobulk DGE method, we aggregated all cells per cell type and species using sumCountsAcrossCells from scuttle[68] and only retained the genes that were detected in all samples (UMI > 0) of at least one species. DGE analysis was performed using edgeR QLRT approach[69] and differentially expressed genes (DEGs) were determined with FDR (< 0.05) and logFC (logFC >0.3 or logFC < -0.3) cutoffs. DGE analysis was performed with the following covariates: humanized age, sex, and library batch. Humanized age was calculated as described before by linear modeling of life traits between species[70]. Genes with species-specific expression were determined as before[18]. Briefly, HS-Genes were determined as DEGs that are H > C = M or H < C = M (C = M was determined if FDR > 0.1, H: human, C: chimpanzee, M: rhesus macaque). The same criteria were used for

CS-Genes. Genes that are consistently different between macaque-human and macaque-chimpanzee were referred to as macaque versus human-chimpanzee genes.

For the single-cell DGE method, genes were tested for differential gene expression using *MAST*[68]. The same covariates were used as the pseudobulk method, except for *cngeneson* as recommended by the MAST approach[68]. Genes with FDR < 0.05 and absolute average log (ln) fold change > 0.25 were considered significant. Genes with species-specific expression were determined as described for the pseudobulk method above.

Single-nuclei ATAC-seq differential CRE accessibility and identification of species-specifically accessible CREs

Similar to DGE, we performed differential CRE accessibility using two approaches: a single-cell based approach and a pseudobulk based approach. We retained the pseudobulk method results for all analyses as both the HS-CREs and CS-CREs were more reproducible with the previous study[12] compared to the single-cell based method **(Extended Figure 10F)**.

For the pseudobulk method, we used the edgeR QLRT approach, which is widely used for differential accessibility analysis[71], similar to the DGE analysis. We aggregated all cells per cell type and species and only retained the CREs that were detected in all samples (total detected reads > 3) of at least one species, and among the top 100,000 CREs by accessibility per cell type. Differentially accessible CREs were determined with FDR (< 0.05) and logFC (logFC >0.3 or logFC < -0.3) cutoffs. Differentially accessible CRE analysis was performed with the following covariates: humanized age and sex. Species specifically accessible CREs were determined in the same manner as the species-specifically expressed genes described above.

For the single-cell method, CRE accessibility was used as the response variable and logistic regression was used to fit two models of covariates with or without species identity per comparison. Then, a log-likelihood ratio test was used to determine the p-value which was later adjusted with FDR correction. The covariates were: humanized age, sex, and total gene activity (as a measure of cell-depth and quality, calculated using Cicero[66]). To determine an effect size cutoff, we first calculated a mean accessibility ratio among tested CREs per pairwise comparison (MeanAccChimp / MeanAccHuman) and used this to normalize accessibility of one species to another. This was then used to compute delta accessibility (HumanAcc - ChimpAccNormalized) for all CREs,

which followed a normal distribution around zero. This calculation was done for each pairwise comparison per cell-type and 1.5 standard deviation (sd) away from the mean was used as cutoff. Therefore, only FDR < 0.05 and sd > 1.5 CREs were considered significant. Species-specifically differential CREs were determined with the same criteria used for species-specifically expressed genes as described above.

Analysis of previously published datasets

Khrameeva et al.[11] was analyzed from publicly available fastq files (GEO accession: GSE127898). Preprocessing (until count matrix) was done similar to our own dataset, including ambient RNA correction by CellBender[56]. Since species were mixed in the same library, we assigned cell barcodes to a given species (human, chimpanzee, bonobo, rhesus macaque) by counting reads with no mismatch (done for each species) and assigning the cell barcode to the species with the most counts. Our annotation corresponded with the original publication for >99.9% of the cell barcodes annotated in the original study[11]. We then used 200 UMI as cutoff, rather than 500 UMI in the original study, as we are interested in the ratios of OPC and MOLs, and glial cells have overall lower number of UMI **(Extended Figure 1E)**. We then used canonical markers to identify the major cell type and define the ratio of OPC and MOL nuclei per sample.

Kozlenkov et al[12]. was analyzed from supplementary tables. The overlap of CREs was tested for statistical significance with a Fisher's exact test. For overlap of species-specifically accessible CREs, we used the number of all CREs as background.

Berto et al.[18] OLIG2 dataset was deconvoluted using MuSiC[72] as previously done[73] except that the reference single cell study was used from this dataset (human, chimpanzee and macaque were used separately for comparisons).

Velmeshev et al.[23] raw count matrix was filtered to contain only L5/6 CC *THEMIS*+ neurons from the healthy controls. L5/6 CC was further subclustered and filtered to only contain *C1QL3*+ subclusters. We also only retained the orthologous protein coding genes initially identified for the original comparative analyses. Differential gene expression was performed between posterior cingulate cortex – prefrontal cortex and posterior cingulate cortex – anterior cingulate cortex using the pseudobulk DGE (edgeR QLRT) as described before.

Bakken et al. was analyzed for the proportional changes in the oligodendrocyte lineage. We obtained the metadata associated with the final count (NEMO identifier: dat-ek5dbmu) and computed fraction of OPC and MOL in all glia per individual. For species with both Snare-seq and single-nuclei transcriptome (human and marmoset), both datasets were utilized.

Epigenome-Transcriptome Associations

To test overlap of epigenomic and transcriptomic changes, we expanded the cis-regulatory elements on both sides of the TSS (transcription start site) with either 1) increasing distance **(Extended Figure 7A),** or 2) for 500kb to identify potential HS-CREs associated with HS-Genes. We used 500kb as most physical interactions between enhancers and promoters are within 500kb distance[74]. We determined that a HS-CRE and HS-Gene are associated if the following conditions are true:

1- They are both found in the same cell type (neuronal subtypes are treated as different cell types).
2- They are altered in the same direction (e.g HS-Open-CRE and HS-Up-Gene).
3- The HS-CRE is within 500kb on either side of the TSSs per HS-Gene.

Gene Set Enrichment Analyses

Gene ontology (GO) enrichment for HS-Genes was done using the clusterProfiler package in R[75]. HS-Up-Genes and HS-Down-Genes were tested separately with all genes tested for differential expression used as the background. Background was calculated separately for each cell type. Only the GO enrichments with FDR < 0.05 and fold change > 1.3 were considered significant.

Modern variant associated HS-CREs were first divided into HS-Open-CREs and HS-Closed-CREs. Then the nearby genes were annotated using *annotatr*[76]. Background genes were similarly identified by annotating all accessible CREs to their genes with *annotr*[76]. Similar to HS-Genes, GO enrichment was performed using the *clusterProfiler* package in *R*. Since only the L2-3_2 subtype showed enrichment, we performed modern variant GO enrichment only for this subtype.

Motif Enrichment Analysis

Non-redundant motifs for human were downloaded from the JASPAR 2018 database[77]. A binary CRE motif matrix (CREs in the rows, motifs in the columns) was created using *Signac*, which calculates the motif matrix using motifmatchr[78]. We then tested the enrichment of motifs in HS-Open-CREs per cell type using a log likelihood ratio test on two nested binomial linear regression models (Evolution: HS-Open-CRE or not):

H0: Evolution ~ CRE length

H1: Evolution ~ CRE length + Motif occurrence

CRE length was added as a covariate since longer CREs will include more motifs. To avoid capturing the motifs divergent between species in general, and to highlight the motifs only divergent in the human-evolution, the background was selected as all evolutionarily divergent CREs for the given cell type. Motifs with FDR < 0.05 and logFC > 0 were considered as significantly enriched. Motif enrichments were clustered and visualized using pheatmap[79].

Visualization of CREs

To visualize CREs across all species, we converted raw chimpanzee and rhesus macaque snATAC-seq reads to human coordinates using CrossMap[53]. This was done separately for reads counted in each cell type. For more accurate comparisons of the track plots between the species, reads were randomly down-sampled to the lowest number of read detected per species for the given subtype. The reads were then converted to bigwig format and visualized using IGV[80] (Integrated Genome Viewer). Tracks were log transformed and presented at the same scale for all comparisons.

Identification of cortical HARs

To identify HARs within the CREs our dataset, we followed a similar approach to a previous study[81]. We first segmented each CRE into 150bp (the size of the smallest CRE in our dataset). We then used the following 15 primate species from a 30-species alignment from UCSC[82]: *Homo sapiens, Pan troglodytes, Gorilla gorilla,*

Nomascus leucogenys, Macaca mulatta, Macaca fascicularis, Macaca nemestrina, Cercocebus atys, Chlorocebus sabaeus, Mandrillus leucophaeus, Colobus angolensis, Callithrix jacchus, Saimiri boliviensis, Cebus capucinus, Aotus nancymaae, and retained the CRE segments that have no gaps in at least 8 species and no gaps between human, chimpanzee and rhesus macaque (the species used in our single-cell genomics experiments) using the *rphast* package[83].

To estimate the neutral substitutions per CRE, we padded each CRE by 25kb upstream and 25kb downstream, and ran the *phyloFit* function with the following parameters on the phylogenetic tree of all species: subst.mod="SSREV", EM = T, nrates = 4. To test acceleration in human lineage, we then used phyloP function on each CRE segment using its corresponding neutral model with the following parameters: method = 'LRT', mode="ACC", branches = 'hg38'. The final list of HARs was determined using the cutoff p-value < 0.001.

Identification of modern human variants

The original publication of modern human variants lists 321,820 human-specific substitutions that contain an ancestral allele either in the Altai Neanderthal or in the Altai Denisovan genome[33]. Since the original publication, two additional high quality Neanderthal genomes have been reported[84,85]. We have therefore updated this original list of human-specific substitutions and only retained the substitutions that are different than in all reported high quality archaic genomes (3 Neanderthals and 1 Denisovan). This resulted in 98,550 human-specific substitutions. The original publication had only retained the substitutions that are present in >90% of present-day humans using the human polymorphism dataset[33]. Since then, the human polymorphism dataset expanded from 1092 individuals from 14 populations[86] to 2504 individuals from 26 populations[87]. Therefore, we updated this cutoff with the most recent 1000 genomes phase 3 dataset[87], which reduced the number of human-specific substitutions to 91,488. Since we were mainly interested in assessing the modern variant enrichment in HS-CREs compared to all CREs, we further filtered for the modern human variants overlapping the CREs in our dataset, which resulted in 12,161 variants. Out of 12,161 variants, 1920 variants (15.7%) overlapped HS-CREs.

Identification of human-specific substitutions

Our main objective to identify human-specific substitutions was to compare them with the modern human variants. Since modern variant analysis only used chimpanzee and gorilla as outgroup species[33], we also limited our comparison to apes. We used the 30 species genome-wide alignment and extracted the alignment for human, chimpanzee, gorilla and gibbon. We excluded orangutan because its alignment was not based on synteny (it was based on reciprocal blast) and showed more missing elements in the alignment compared to other species. Using this 4-way alignment, we then identified single-nucleotides that are only different in humans and map to the CREs identified in this dataset and referred to them as human-specific substitutions. Similar to the modern human variants, we further filtered human-specific substitutions for presence in at least 90% of modern day humans according to the 1000 genomes project phase 3 database[87].

Analyses of HS-CRE enrichment in HARs and modern variants

For a full list of published HARs, we merged the bed files of a compendium of HARs[27] and another HAR study based on accessibility patterns of chromatin[81] using bedtools[62]. To test the overlap of all CREs with published HARs, we generated background genomic regions of similar GC content and length using *genNullSeqs* from the gkmSVM package with default parameters[88]. We then randomly selected the same number of regions as the entire CRE list (n=100) and tested for significantly higher overlap of HARs with the observed CREs compared to the randomized background using an empirical p-value.

Enrichment of HARs in HS-CRE were tested by logistic regression. Predictor variables were CRE length and CRE evolution (HS or NS (non-significant)), and the response variable was whether the CRE contains a HAR or not. The effect of CRE evolution was tested with a likelihood ratio test. The test was performed for each major cell type separately (excitatory neurons, inhibitory neurons, MOLs, OPCs, astrocytes, microglia).

To test which cell-types evolved more recently after the split of modern humans from other ancient human species (Neanderthals and Denisovans)[33], we performed a negative binomial regression. Predictor variables were CRE length and CRE evolution and the response variable was the number of overlapping modern variants. The effect of CRE evolution was tested with a likelihood ratio test. The test was performed for each major cell type separately (excitatory neurons, inhibitory neurons, MOLs, OPCs, astrocytes, microglia). We also tested the

overall enrichment of modern variants by considering HS-CREs as a CRE that is a HS-CRE in at least one cell type.

Reported p-values were FDR adjusted in all enrichments. CRE length was used as a covariate in both enrichments since larger CREs tend to have more variants and a better chance to overlap HARs and modern variants.

Analysis of surgically resected human snATAC-seq

Raw fastq files were downloaded from the GEO database (accession number: GSE139914, brain region: BA38, middle temporal gyrus)[37]. We pre-processed the snATAC-seq similar to the original publication; however, instead of performing peak calling to generate the peak-cell matrix, we counted the reads in the CREs identified in our dataset for direct comparison of accessibility on the same CREs. We extracted the excitatory cells as they were annotated in the original study and annotated the subtypes by co-clustering with the human snATAC-seq excitatory subtypes in this study. We then performed differential accessibility analyses as described above, this time comparing the surgical human tissue and chimpanzee / rhesus macaque samples per excitatory subtype. Motif enrichment analyses were also performed as before on the HS-Open-CREs per excitatory subtype.

Data availability: Raw and processed data are available at NCBI GEO under the accession number GSE192774.Processed data associated with Bakken et al. was accessed from https://assets.nemoarchive.org/dat-ek5dbmu. Other datasets were obtained using their GEO accession numbers (Khrameeva et al: GSE127774, Berto et al: GSE107638, GSE123936, Franjic et al: GSE18653).

Code availability: All analysis scripts are available in our github page: https://github.com/konopkalab/Comparative_snATAC_snRNA

51 Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol Cell* **71**, 858-871 e858, doi:10.1016/j.molcel.2018.06.044 (2018).

52 Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049, doi:10.1038/ncomms14049 (2017).
53 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
54 Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006-1007, doi:10.1093/bioinformatics/btt730 (2014).
55 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2014).
56 Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491-499, doi:10.1101/gr.209601.116 (2017).
57 Fleming, S. J., Marioni, J. C. & Babadi, M. CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets. 791699, doi:10.1101/791699 %J bioRxiv (2019).
58 Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res* **49**, D884-D891, doi:10.1093/nar/gkaa942 (2021).
59 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821, doi:10.1016/j.cell.2019.05.031 (2019).
60 *Picard Toolkit*, <http://broadinstitute.github.io/picard/> (2019).
61 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).
62 Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief Bioinform* **14**, 144-161, doi:10.1093/bib/bbs038 (2013).
63 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
64 Lareau, C. A., Ma, S., Duarte, F. M. & Buenrostro, J. D. Inference and effects of barcode multiplets in droplet-based single-cell assays. *Nat Commun* **11**, 866, doi:10.1038/s41467-020-14667-5 (2020).
65 Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat Methods* **18**, 1333-1341, doi:10.1038/s41592-021-01282-5 (2021).
66 Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289-1296, doi:10.1038/s41592-019-0619-0 (2019).
67 Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67**, 1 - 48, doi:10.18637/jss.v067.i01 (2015).
68 Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16**, 278, doi:10.1186/s13059-015-0844-5 (2015).
69 Chen, Y., Lun, A. T. & Smyth, G. K. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Res* **5**, 1438, doi:10.12688/f1000research.8987.2 (2016).
70 McCarthy, D. J., Campbell, K. R., Lun, A. T. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179-1186, doi:10.1093/bioinformatics/btw777 (2017).
71 Gontarz, P. *et al.* Comparison of differential accessibility analysis strategies for ATAC-seq data. *Sci Rep* **10**, 10150, doi:10.1038/s41598-020-66998-4 (2020).
72 Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun* **10**, 380, doi:10.1038/s41467-018-08023-x (2019).
73 Mendizabal, I. *et al.* Cell type-specific epigenetic links to schizophrenia risk in the brain. *Genome Biol* **20**, 135, doi:10.1186/s13059-019-1747-7 (2019).
74 van Arensbergen, J., van Steensel, B. & Bussemaker, H. J. In search of the determinants of enhancer-promoter interaction specificity. *Trends Cell Biol* **24**, 695-702, doi:10.1016/j.tcb.2014.07.004 (2014).
75 Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287, doi:10.1089/omi.2011.0118 (2012).
76 Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics* **33**, 2381-2383, doi:10.1093/bioinformatics/btx183 (2017).
77 Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res* **46**, D260-D266, doi:10.1093/nar/gkx1126 (2018).

78 A, S. motifmatchr: Fast Motif Matching in R. *R package version 1.4.0* (2018).
79 Kolde, R. *Pheatmap: pretty heatmaps. R*, <https://cran.r-project.org/web/packages/pheatmap/index.html> (2012).
80 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).
81 Gittelman, R. M. *et al.* Comprehensive identification and analysis of human accelerated regulatory DNA. *Genome Res* **25**, 1245-1255, doi:10.1101/gr.192591.115 (2015).
82 Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**, 708-715, doi:10.1101/gr.1933104 (2004).
83 Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief Bioinform* **12**, 41-51, doi:10.1093/bib/bbq072 (2011).
84 Mafessoni, F. *et al.* A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc Natl Acad Sci U S A* **117**, 15132-15136, doi:10.1073/pnas.2004944117 (2020).
85 Prufer, K. *et al.* A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655-658, doi:10.1126/science.aao1887 (2017).
86 Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).
87 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).
88 Ghandi, M. *et al.* gkmSVM: an R package for gapped-kmer SVM. *Bioinformatics* **32**, 2205-2207, doi:10.1093/bioinformatics/btw203 (2016).

Acknowledgments

The authors thank Dr. Carol Tamminga and Kelly Gleason at UT Southwestern for postmortem human tissues. We also thank Dr. Hume Stroud and Dr. Maria Chahrour for their critical comments on the manuscript. We also thank Dr. Kate Luby-Phelps from UTSW LCIF (live cell imaging facility) and Dr. Shin Yamazaki from UTSW Neuroscience Microscopy Facility for their help with imaging. G.K. is a Jon Heighten Scholar in Autism Research and Townsend Distinguished Chair in Research on Autism Spectrum Disorders at UT Southwestern. E.C. is a Neural Scientist Training Program Fellow in the Peter O'Donnell Brain Institute at UT Southwestern. This work was partially supported by the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition Scholar Award to G.K.; NHGRI (HG011641) to G.K., S.V.Y., and C.C.S.; National Science Foundation (SBE-131719 and EF-2021635) to S.V.Y. and C.C.S.; the NIMH (MH103517) to T.M.P., G.K., and S.V.Y; NIH grants T32DA007290 and T32HL139438 to F.A., American Heart Association Postdoctoral Fellowship (915654) to Y.L and NIMH grant MH126481 to R.V. and G.K. The National Chimpanzee Brain Resource was supported by NINDS (R24NS092988). Macaque tissue collection and archiving was supported by the NIH National Center for Research Resources (P51RR165; superseded by the Office of Research Infrastructure Programs (OD P51OD11132)) to the Yerkes National Primate Research Center. The Zeiss LSM880 with Airyscan was supported by NIH grant 1S10OD021684-01 to Katherine Luby-Phelps.

Author contributions: S.V.Y., G.K. and E.C. designed the study. T.M.P. and C.C.S. performed tissue dissections. S.V.Y., T.M.P., C.C.S., and G.K. obtained funding. F.A. collected snRNA-seq and snATAC-seq data. Y.L., R.V. and E.O. performed smFISH experiments. Y.L. performed image quantification. E.C. performed all analyses. T.M.P. edited the manuscript. E.C., S.V.Y, and G.K. wrote the manuscript.

Competing interests: The authors declare no competing interests.

Additional information: Supplementary information is available for this paper at https://doi.org/10.1038/s41586-0XXXXX

Correspondence and requests for materials should be addressed to G.K. or S.V.Y.

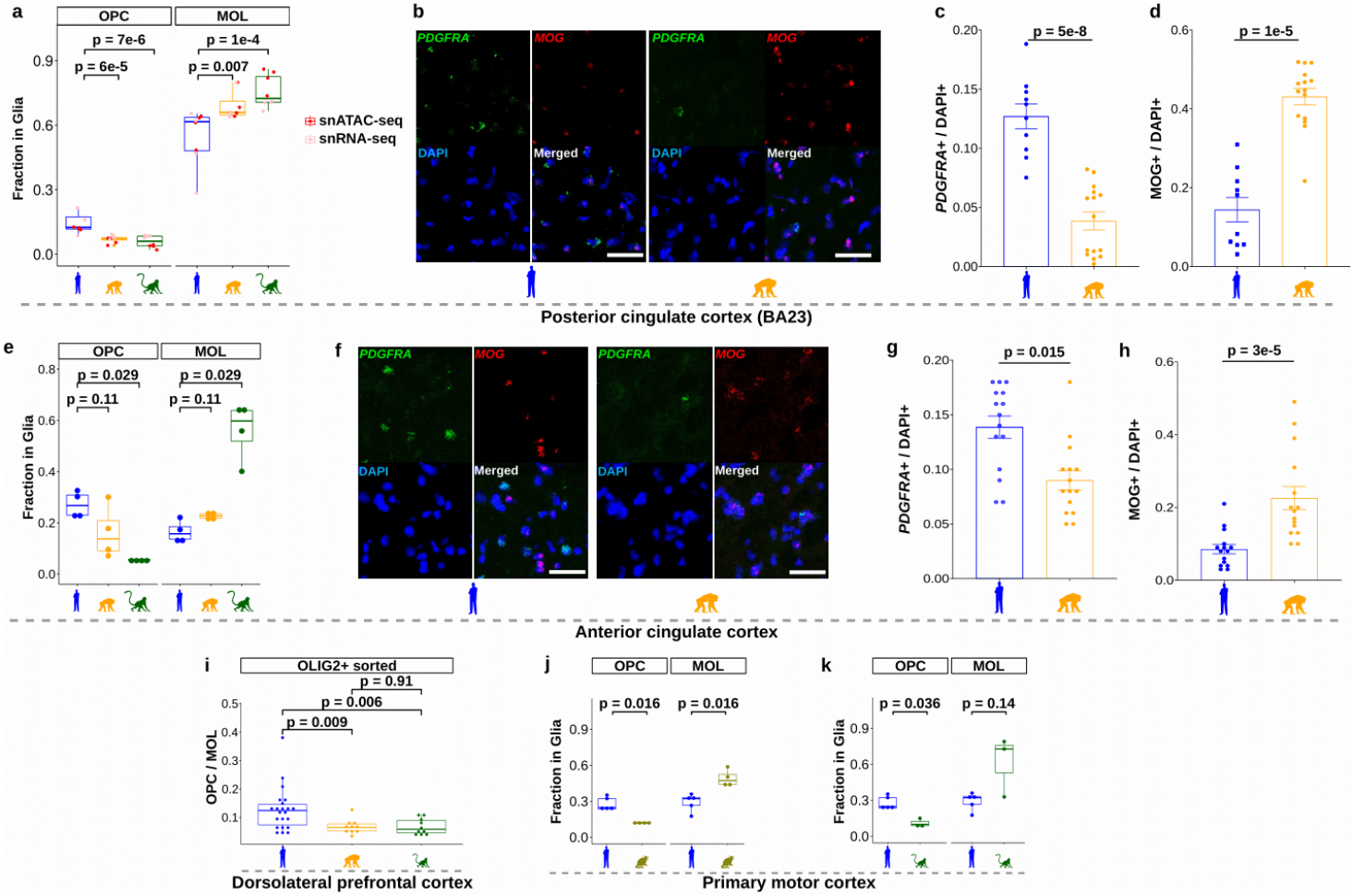


Figure 1: The oligodendrocyte lineage is proportionally altered in human evolution. (A) The fraction of OPC and MOL (mature oligodendrocytes) in glia are altered in humans. Each dot represents a sample (red: snATAC-seq, pink: snRNA-seq. N = 4 individuals per species per assay. P-value: likelihood ratio test, two-sided. See methods). **(B-D)** smFISH shows increased *PDGFRA* (OPCs) and decreased *MOG* (MOLs) signals in humans compared to chimpanzees (region: posterior cingulate cortex). **(B)** A representative image, scale bar is 100 µm. **(C-D)** Quantification of the fraction of OPCs and MOLs. Each data point is the average of all subareas in a section (2-4 subareas/5 sections/individual. Human: 10 sections. Chimpanzee:15 sections, see Methods). The p value is the main effect of species from a linear mixed model (random effect: individual, two-sided). Error bars represent SEM. **(E)** The fraction of OPCs or MOLs in glia based on snRNAseq from anterior cingulate cortex (n = 4 individuals per species, P-value: Wilcoxon rank sum test, two-sided). **(F)** smFISH similar to **(B)**, but for anterior cingulate cortex. **(G-H)** Quantification of the fraction of OPCs mOLs as in **(C-D)** Human: 15 sections, chimpanzee:15 sections, see Methods. **(I)** Deconvoluted proportions of cells from OLIG2 expressing bulk RNA-seq (reference dataset: human snRNA-seq from this study, n = 22 (human), 10 (chimpanzee), 10 (rhesus macaque) individuals. P-value: Wilcoxon rank sum test, two-sided). **(J-K)** Fraction of OPC or MOLs in glia per species in the primary motor cortex. **(J)** Human – marmoset comparison, **(K)** human – rhesus macaque comparison. N = 5 (human), 4 (marmoset), 3 (rhesus macaque) individuals. P-value: Wilcoxon rank sum test, two-sided. Boxplots represent median and interquartile range in panels **A**, **E**, **I-K**.

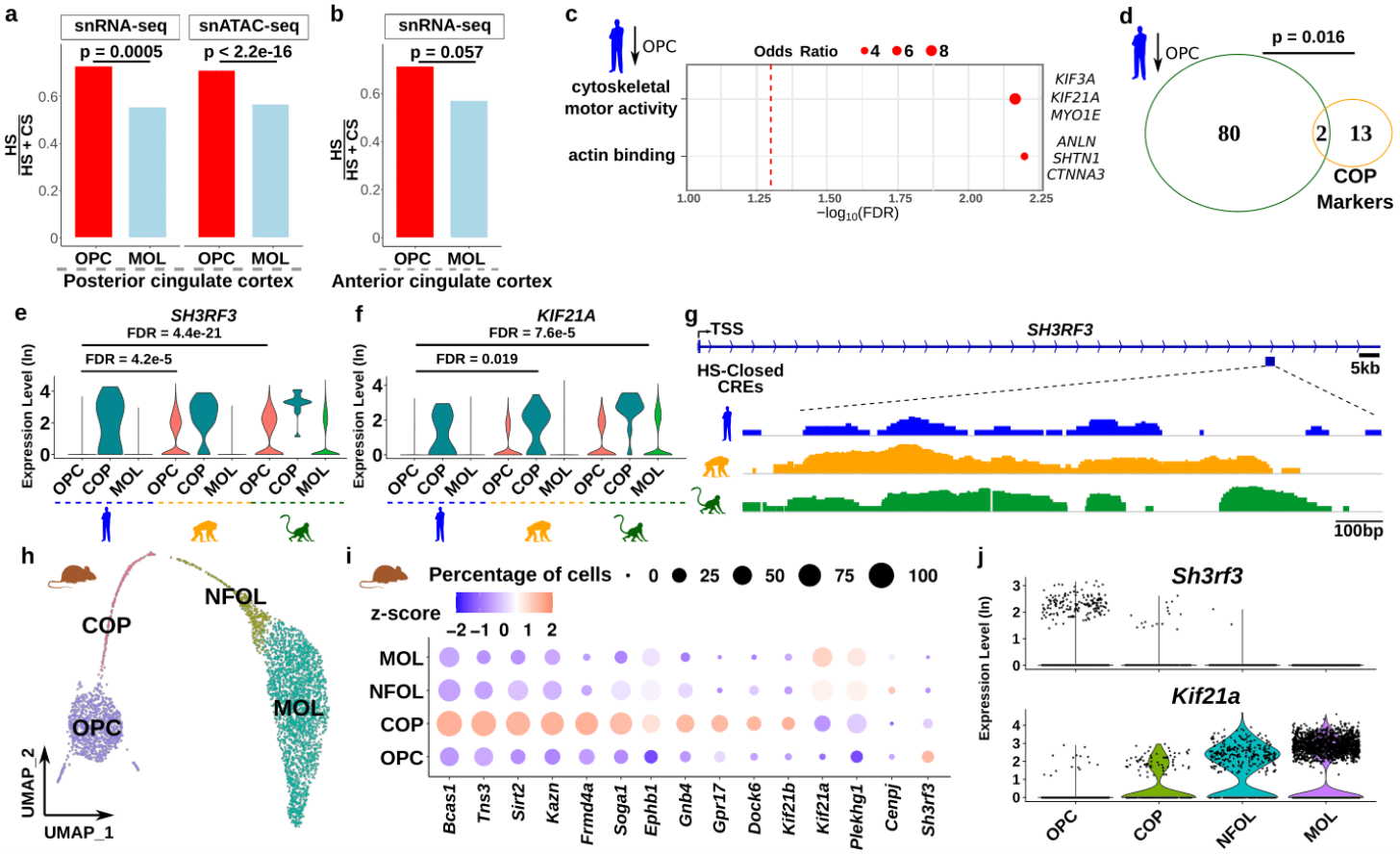


Figure 2: Human-specific gene regulatory changes in the oligodendrocyte lineage. (A) The proportions of HS (human-specific) regulatory changes to HS + CS (chimpanzee-specific) regulatory changes are higher in OPCs than MOLs (left: snRNA-seq, right: snATAC-seq). P value: chi-square test, two-sided. **(B)** Same as in **(A)**, except using the anterior cingulate cortex snRNA-seq dataset. **(C)** Gene ontology enrichment for HS-Down-Genes in OPCs highlights altered cytoskeletal function (p-value: Fisher's exact test, one-sided). **(D)** Overlap of HS-Down-Genes in OPCs and primate-conserved COP (committed oligodendrocyte progenitor) markers reveal two COP markers with loss-of-function in human OPCs (p-value: Fisher's exact test, one-sided). **(E-F)** Expression levels of **(E)** *SH3RF3* and **(F)** *KIF21A1* across cell types in human, chimpanzee, rhesus macaque. FDR corrected p-values compare the expression levels in OPCs between species (see Supplementary Table 3). **(G)** snATAC-seq coverage plots of the Human-DOWN-CRE near *SH3RF3* in OPCs. TSS: transcription start site. Track scales are the same in all species. **(H)** UMAP plot of oligodendrocyte lineage cells in mouse adult frontal cortex dataset. NFOL: newly formed oligodendrocytes. **(I)** Expression pattern of primate-conserved COP markers across mouse oligodendrocyte lineage cell types. Only *Sh3rf3* expression is decreased in COPs or NFOLs compared to OPCs. **(J)** Violin plots of *Sh3rf3* and *Kif21a* expression in mouse oligodendrocyte lineage cell types.

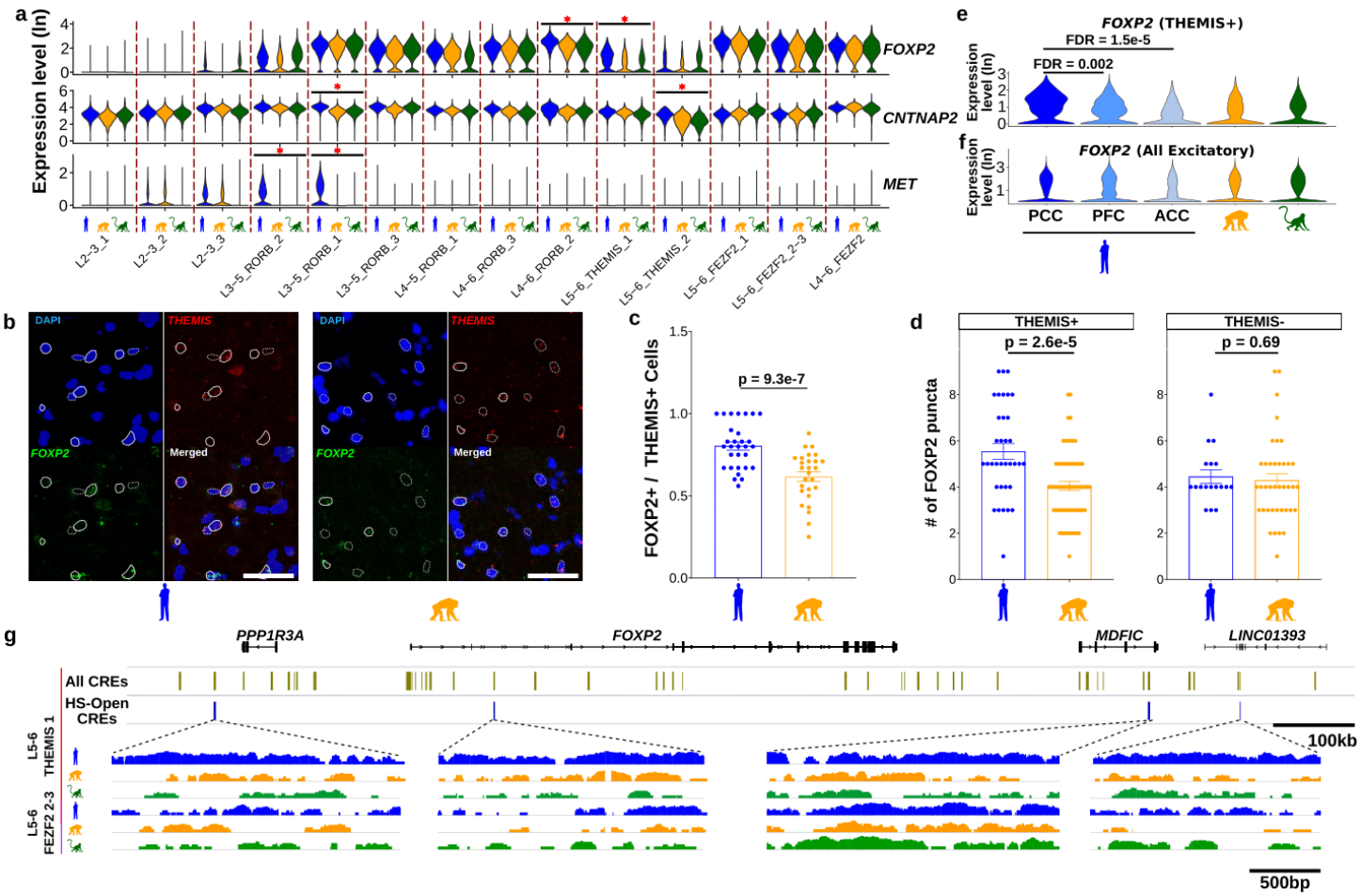


Figure 3: Subtype and cortical region-specific upregulation of *FOXP2* in human neurons. (A) Expression levels of *FOXP2, CNTNAP2 and MET* in the posterior cingulate cortex (PCC) show subtype-specific expression changes in the human brain. Human-specific expression is labeled with a red asterisk, x-axis denotes species and excitatory subtypes. **(B-D)** smFISH of *FOXP2* and *THEMIS* in anterior cingulate cortex (ACC) shows greater number of *FOXP2/THEMIS*+ cells in human compared to chimpanzee. **(B)** A representative image. Solid circles show *FOXP2* and *THEMIS* overlapping cells; dashed circles show *THEMIS*+ cells without *FOXP2* expression. Scale bar is 50 µm. **(C-D)** Quantification of the *FOXP2*+ cells **(C)** and **(D)** *FOXP2*+ puncta per cell in (left: *THEMIS*+ cells, right: *THEMIS*- cells). The p value is the main effect of species from a linear mixed model (random effect: individual, two-sided) with each data point representing a subarea/image/individual. Error bars represent SEM. N = 3 individuals per species. **(E-F)** *FOXP2* is upregulated in the PCC compared to pre-frontal cortex (PFC) and ACC in THEMIS+ neurons **(E)** but not among all excitatory neurons **(F)**. Y-axis: normalized and log-transformed expression levels. **(G)** snATAC-seq coverage plots of HS-Open-CREs near *FOXP2* in L5-6_THEMIS_1 and L5-6_FEZF2 2-3 neurons. The HS-Open-CREs shown have human-specific chromatin accessibility in L5-6_THEMIS_1 neurons but not in L5-6_FEZF2_2-3 neurons. Track scales are the same in all species.

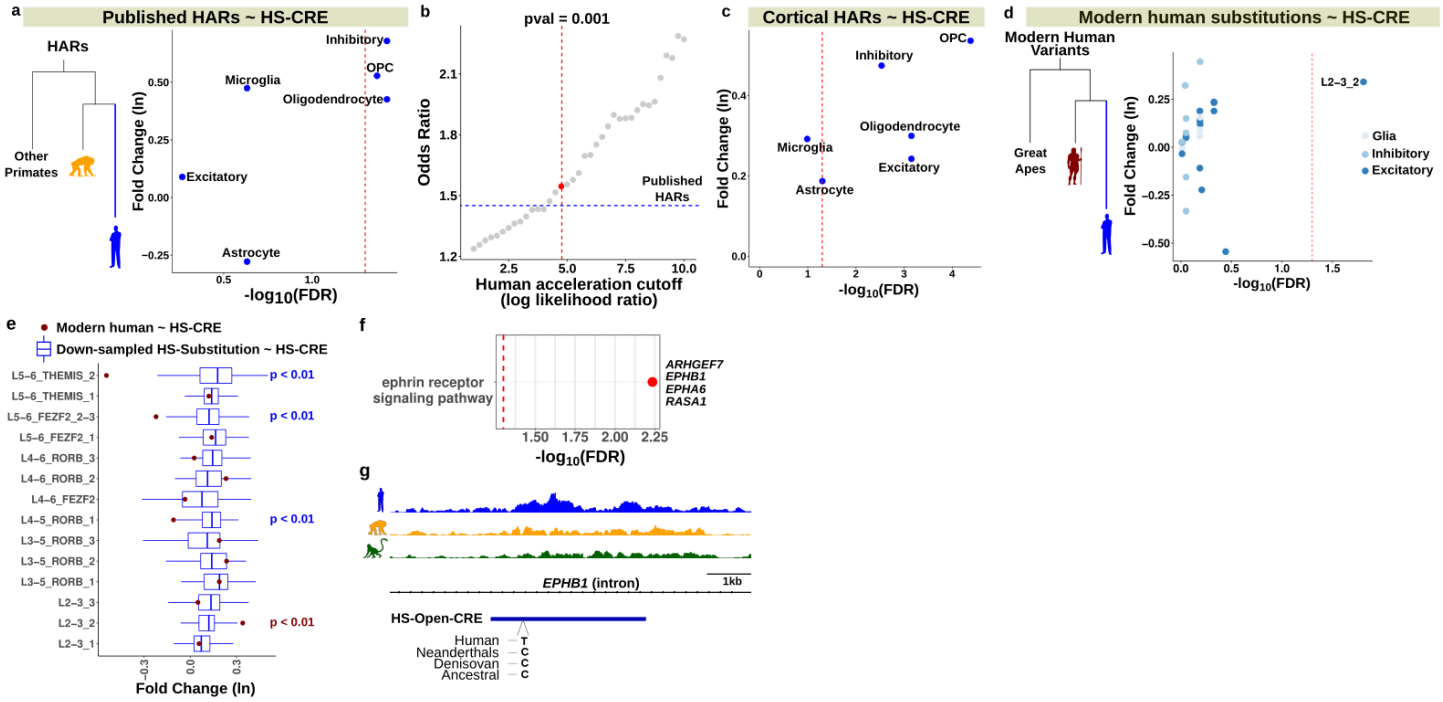


Figure 4: Significant association of chromatin accessibility changes with sequence divergence. (A) Enrichment of publicly available human accelerated regions (HARs) within the HS-CREs. Enrichment is tested by a logistic regression model with CRE length and evolution of the CRE as the predictor variables (HS-CRE or not HS-CRE) and HAR as the response variable (HAR or not HAR, P-value: likelihood ratio test, two-sided). **(B)** Odds ratio of HAR and HS-CRE association compared to HAR and NS-CRE (non-significant CREs; all CREs that are not HS-CREs) for published HARs (dashed blue line) and HARs identified in this study per log likelihood cutoff (x-axis). Red line indicates the log likelihood cutoff that corresponds to p=0.001 (one-sided). Red dot indicates the odds ratio that corresponds to p=0.001 cutoff. **(C)** Cortical HAR analysis reveals stronger association with HS-CREs (same enrichment analysis as **(A)** between cortical HARs and HS-CREs). **(D)** Modern human-specific variant enrichment within the HS-CREs. Enrichment is tested by a negative binomial regression model with CRE length and evolution of the CRE as the predictor variables (HS-CRE or not HS-CRE) and the number of modern human-specific variants as a response variable (P-value: likelihood ratio test, two-sided). **(E)** Log fold changes of substitution and HS-CRE association for substitutions on the human (blue) and modern human lineage (tile red) per excitatory subtype. Human lineage-specific substitutions were randomly down sampled for 100 times to 12,161 (the number of modern human-specific variants) for comparison. Empirical p-value (two-sided) is reported for conserved (in blue) and accelerated (in tile red) subtypes in modern humans. Boxplots represent median and interquartile range. **(F)** Gene ontology enrichment of HS-Open-CREs with modern human-specific variants in L2-3_2 subtype. **(G)** snATAC-seq coverage plots of Human-Open-CREs near *EPHB1* in L2-3_2 neurons which has a modern human variant. Track scales are the same in all species.

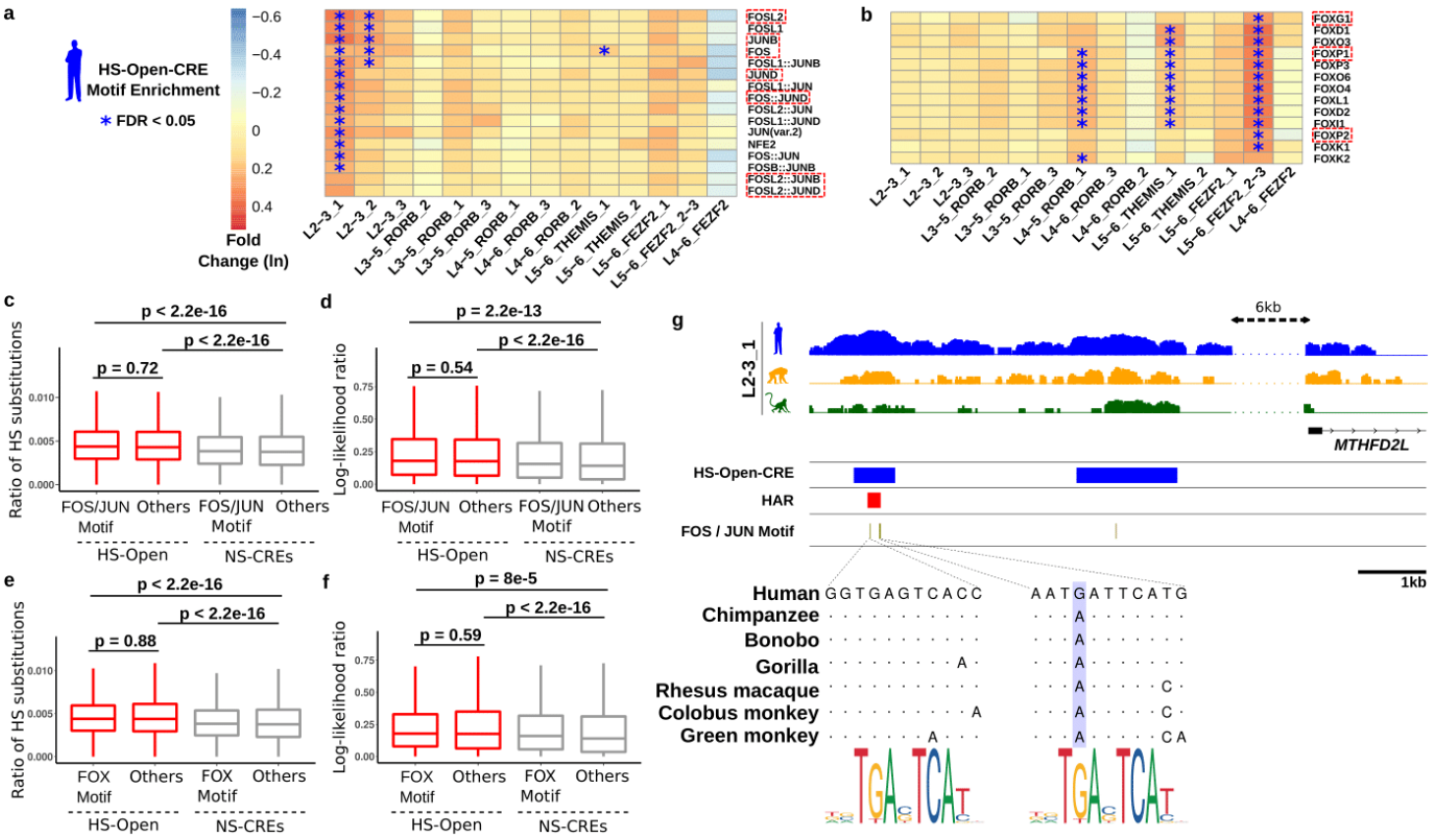
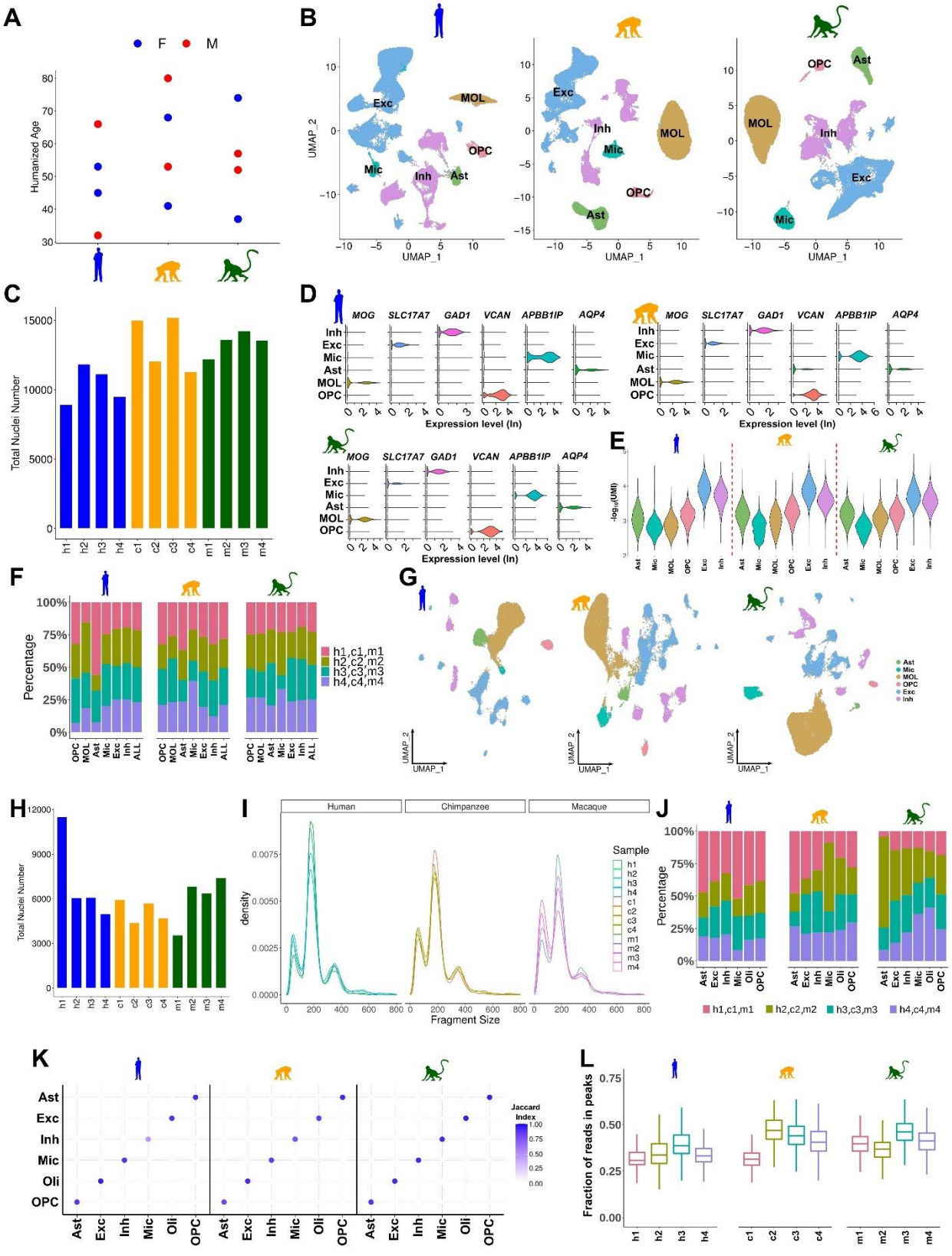
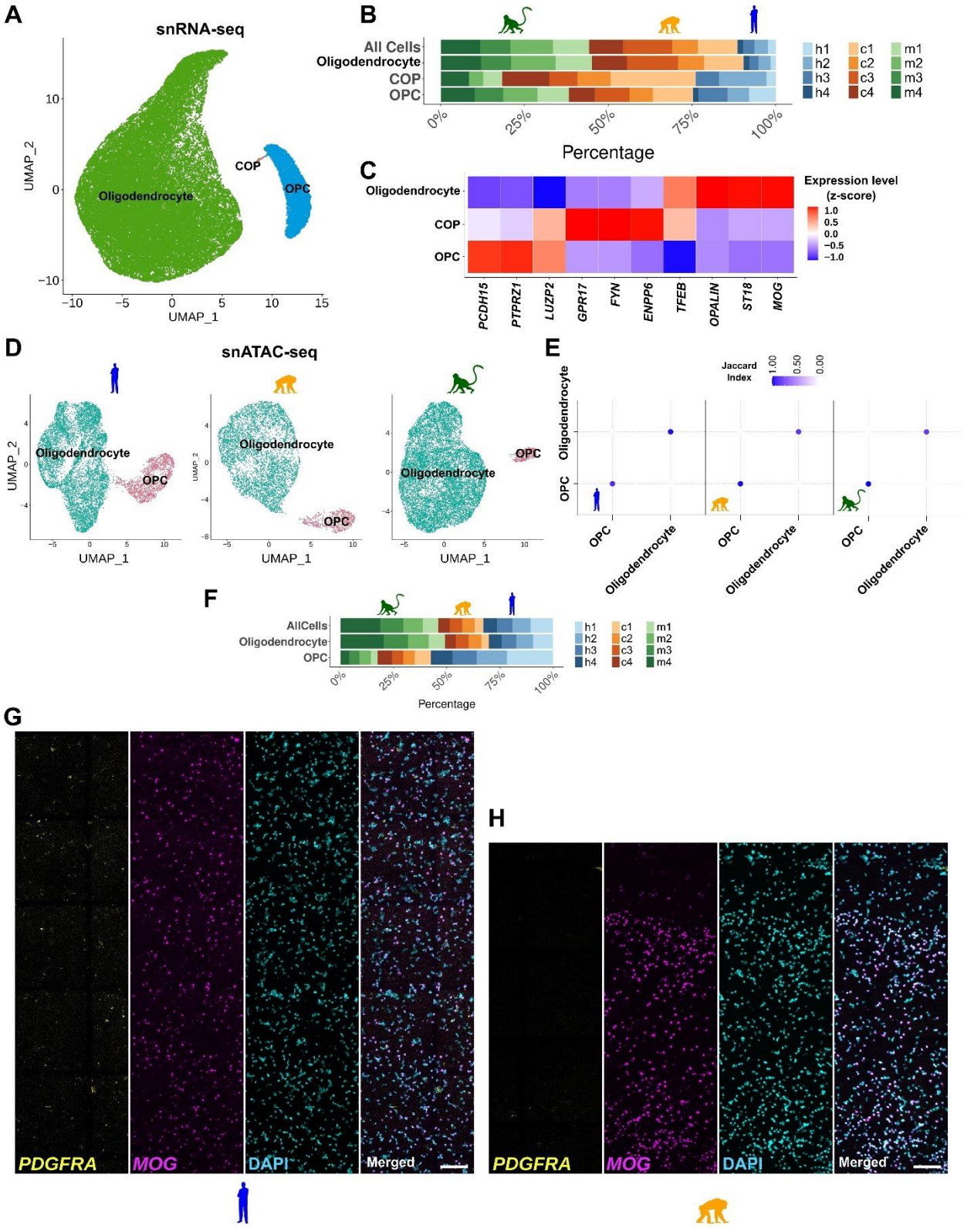


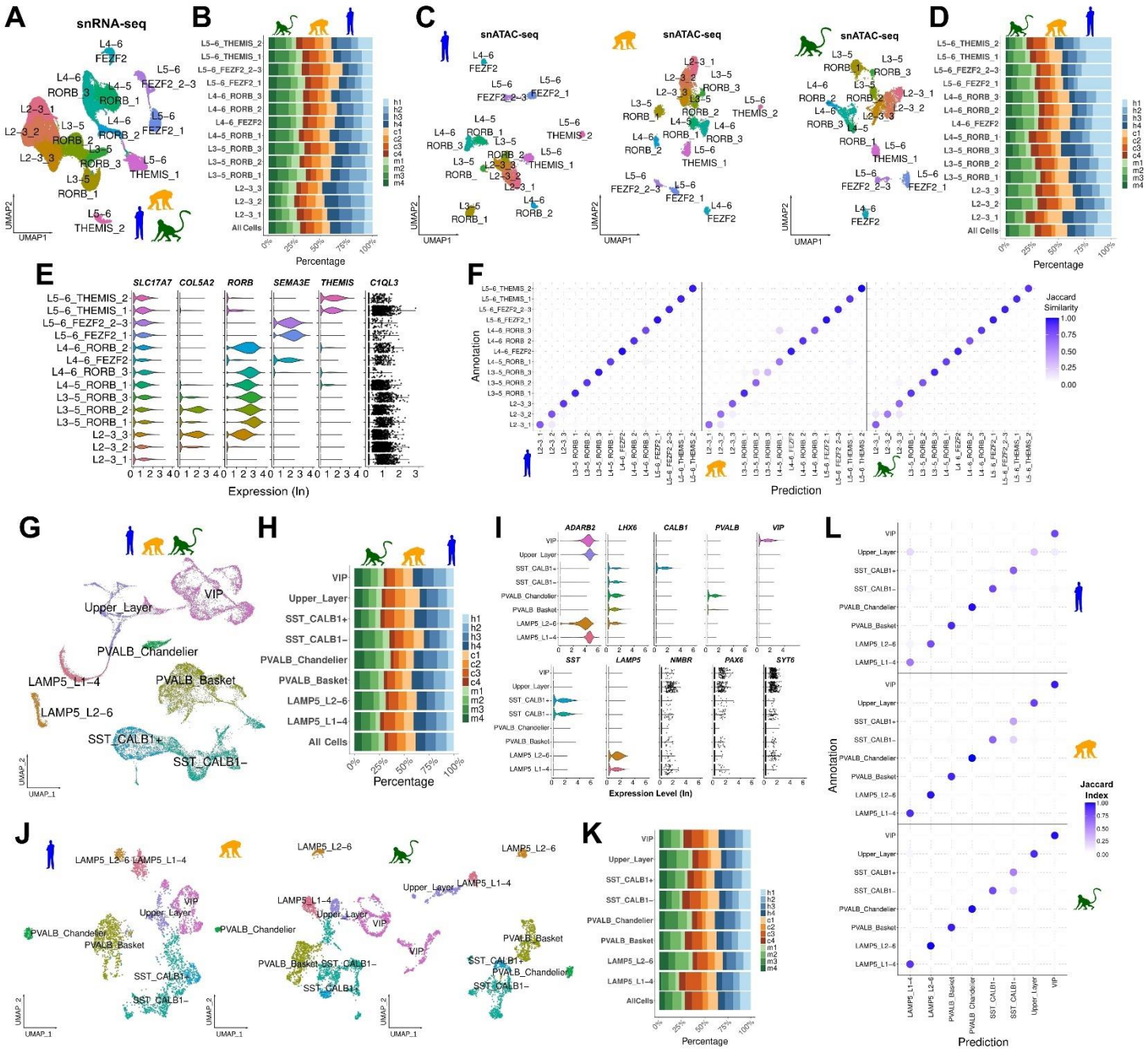
Figure 5: Accessibility changes highlight subtype-specific transcription factor target evolution in the human brain. (A-B) FOS / JUN **(A)** and FOX **(B)** family transcription factors (TFs) enrichments in HS-Open-CREs. Heatmap shows the log fold changes of TF motif enrichment within HS-Open-CREs per TF motif and per subtype (Blue asterisks: FDR < 0.05. Red rectangles: distinctly more accessible TFs within these subtypes). **(C)** Ratio of HS substitutions by CRE length per group of CREs. Groups from left to right: HS-Open-CREs that contain at least one motif, HS-Open-CREs that do not contain a motif, NS-CREs (non-significant CREs) that contain a motif, NS-CREs that do not contain a motif. Only the enriched subtypes per TF group and the highlighted TFs -in panel A- per TF family were used for these comparisons. (N = 1519, 2894, 38486, 109452 CREs left to right). Boxplots represent median and interquartile range. P-value: Wilcoxon rank sum test (two-sided). **(D)** Same CREs as **(C)** except using the mean log likelihood ratios per CREs computed in the HAR analysis. **(E-F)** Same comparison as **(C-D)** except using FOX motifs. (N = 1338, 3075, 38774, 109164 CREs left to right.). **(G)** Track plot of an HS-Open-CRE with a HAR and contains a human-specific gain of a FOS/JUN motif. Identical sequences with respect to the human sequence are shown with dots. The motif representation on the bottom is a FOS motif.


A
F
M
Humanized Age
B
UMAP_1
UMAP_2
Exc
MOL
OPC
Mic
Inh
Ast
C
Total Nuclei Number
h1
h2
h3
h4
c1
c2
c3
c4
m1
m2
m3
m4
D
MOG
SLC17A7
GAD1
VCAN
APBB1IP
AQP4
Expression level (ln)
E
F
Percentage
h1,c1,m1
h2,c2,m2
h3,c3,m3
h4,c4,m4
G
H
I
Human
Chimpanzee
Macaque
density
Fragment Size
Sample
J
K
Jaccard Index
Oli
L
Fraction of reads in peaks

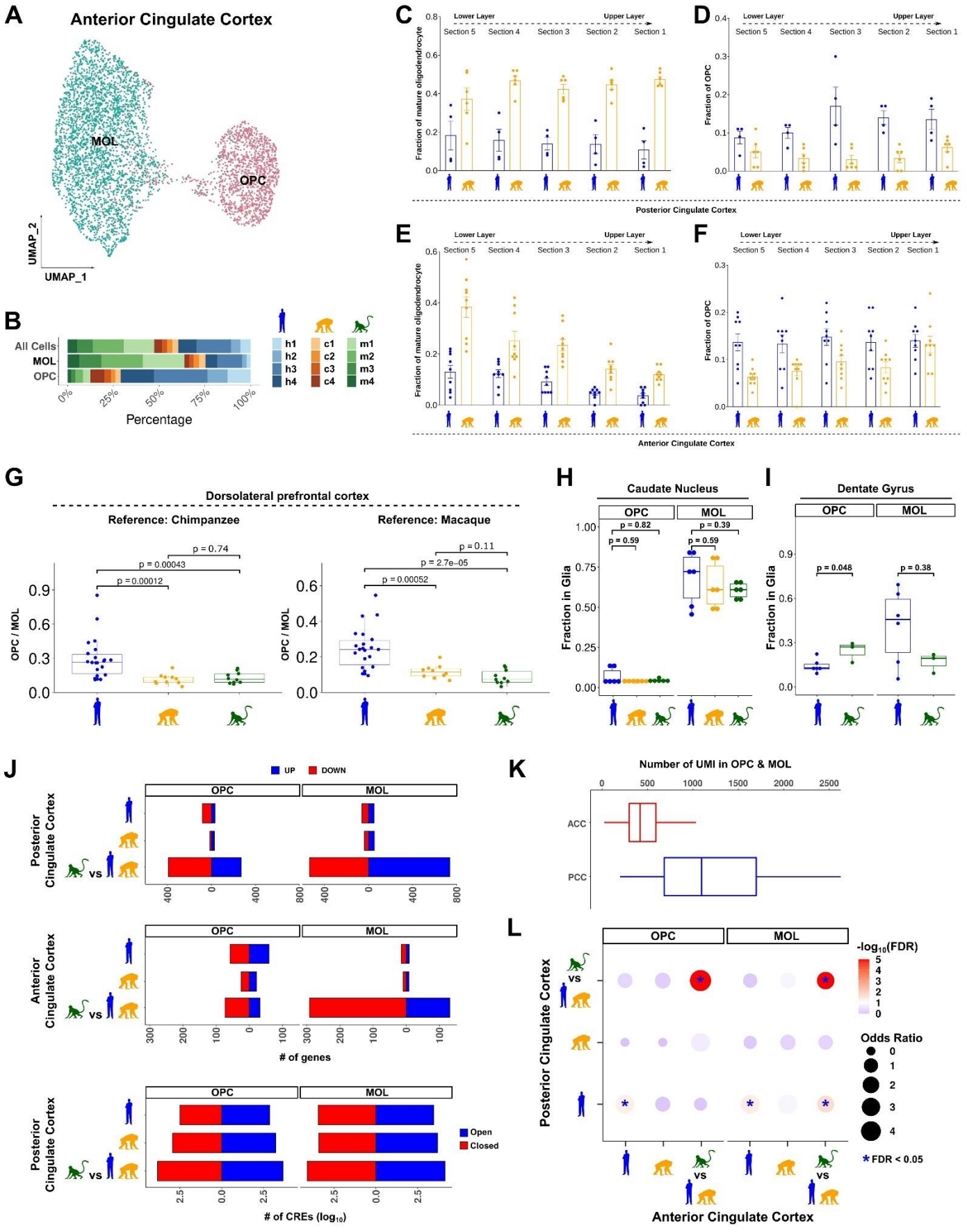
Extended Figure 1: Annotation and quality control of single-nuclei RNA-seq and single-nuclei ATAC-seq. (A) Distribution of sex and humanized age of samples. **(B)** Broadly annotated UMAP of nuclei per species. **(C)** Total nuclei number per sample after filtering. **(D)** Normalized, log (ln) transformed expression values of major cell type markers. **(E)** Violin plots of number of detected UMIs (log_{10} transformed) per major cell type. **(F)** Percentage of cells contributed per individual per species per major cell type. **(G)** Broad annotation of snATAC-seq data per species. **(H)** Total nuclei number per sample after quality control. **(I)** Nucleosome band pattern per sample; each line represents one sample. First, second and third peaks represent nucleosome free, mononucleosome and dinucleosome fractions, respectively. **(J)** Percentage of cells contributed per individual per species per major cell type. **(K)** Clarity of annotation transfer from snRNA-seq to snATAC-seq as displayed by Jaccard similarity index, which is the number of nuclei with the same final annotation and prediction (intersection) divided by the total number of nuclei with a given annotation or prediction (union). y-axis represents final annotation; x-axis represents the prediction which was assigned by label transfer per nucleus. Higher values indicate more similarity between final annotation and initial prediction. **(L)** Fraction of reads in peaks per sample (N = 9280, 5383, 5657, 4655, 5941, 4381, 5691, 4690, 3321, 6426, 5984, 6793 left to right). Boxplots represent median and interquartile range.


A
snRNA-seq
UMAP_2
UMAP_1
Oligodendrocyte
COP
OPC
B
All Cells
Oligodendrocyte
COP
OPC
0%
25%
50%
75%
100%
Percentage
h1
h2
h3
h4
c1
c2
c3
c4
m1
m2
m3
m4
C
Oligodendrocyte
COP
OPC
PCDH15
PTPRZ1
LUZP2
GPR17
FYN
ENPP6
TFEB
OPALIN
ST18
MOG
Expression level
(z-score)
1.0
0.5
0.0
−0.5
−1.0
D
snATAC-seq
Oligodendrocyte
OPC
UMAP_1
UMAP_2
E
Jaccard
Index
1.00
0.50
0.00
Oligodendrocyte
OPC
F
AllCells
Oligodendrocyte
OPC
Percentage
G
PDGFRA
MOG
DAPI
Merged
H
PDGFRA
MOG
DAPI
Merged

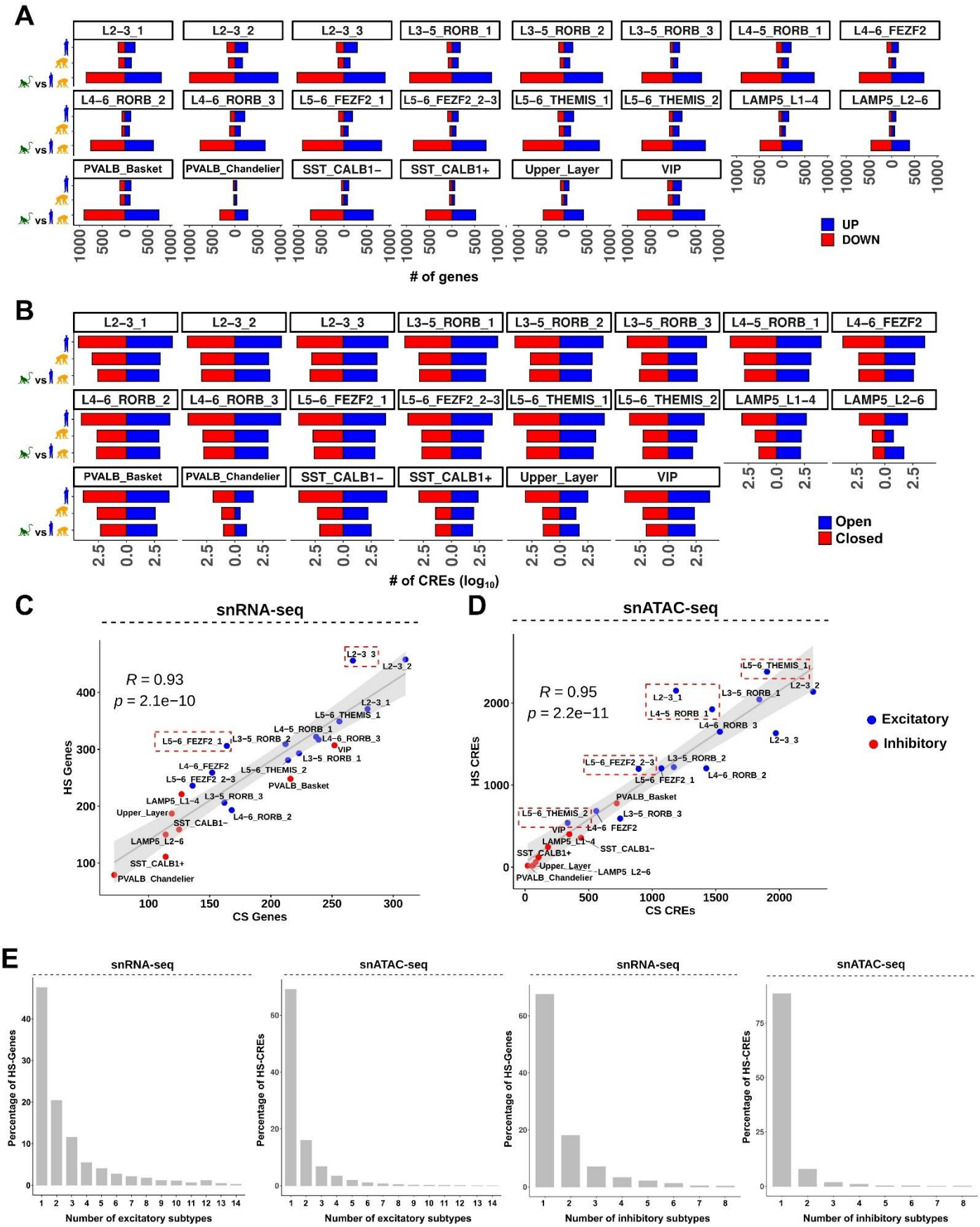
Extended Figure 2: Annotation of oligodendrocyte lineage cells. (A) UMAP visualization of integrated and annotated oligodendrocyte lineage nuclei in snRNA-seq. Oligodendrocyte: mature oligodendrocytes, COP: committer oligodendrocyte progenitor cells, OPC: oligodendrocyte progenitor cells. **(B)** Percentage of nuclei per sample for each subtype in snRNA-seq. **(C)** Normalized and scaled (z-scored) expression values of major oligodendrocyte lineage cell type markers. **(D)** UMAP visualization of annotated oligodendrocyte lineage nuclei in snATAC-seq per species. **(E)** Clarity of annotation transfer from snRNA-seq to snATAC-seq as displayed by Jaccard similarity index (similar to Extended Figure 1K). **(F)** Percentage of cells contributed per individual per species per major cell type. **(G-H)** smFISH of *PDGFRA* (OPC) and *MOG* (MOL) in humans **(G)** and chimpanzees **(H)** (region: posterior cingulate cortex. Images span all cortical layers in both species. Scale bar is 100 μm). Similar results have been obtained for 4 bins across 2 humans and for 6 bins across 3 chimpanzees (see Extended Figure 4C-D).



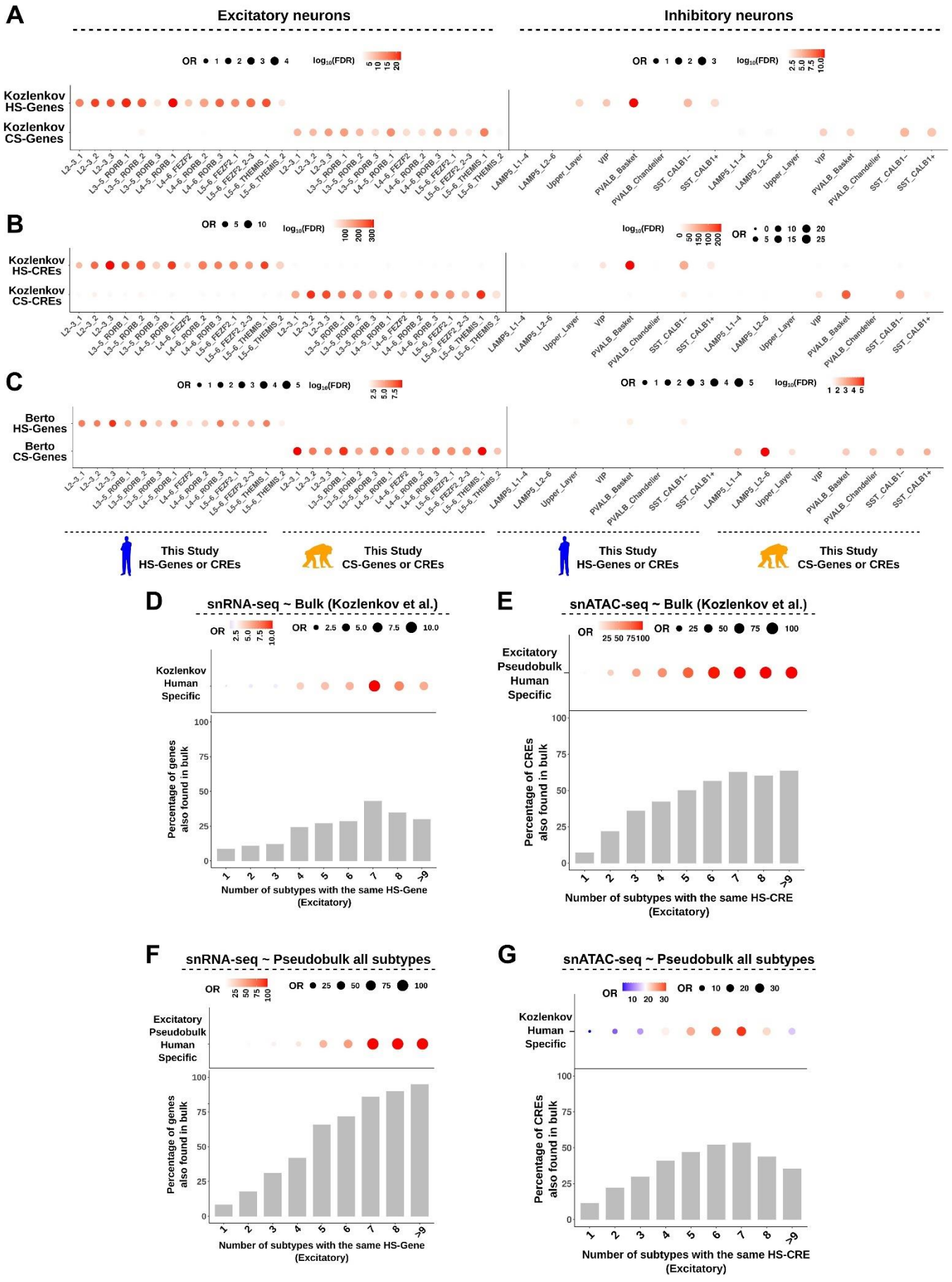
Extended Figure 3: Integration and annotation of neurons. (A) Annotated UMAP of excitatory neurons integrated across all species in snRNA-seq. **(B)** Percentage of nuclei per sample for each excitatory subtype in snRNA-seq. **(C)** Annotated snATAC-seq per species in the UMAP space. All 14 subtypes identified in snRNA-seq are also distinctly annotated in snATAC-seq for all species. **(D)** Percentage of nuclei per sample for each excitatory subtype in snATAC-seq. **(E)** Excitatory subtype markers for validation of annotation (expressions are normalized and log transformed). Note that the individual cells are plotted for *C1QL3* since the expression level is not sufficient for a violin plot. **(F)** Clarity of annotation transfer from snRNA-seq to snATAC-seq as displayed by Jaccard similarity index (similar to Extended Figure 1K). **(G)** Annotated UMAP of inhibitory neurons integrated across all species in snRNA-seq. **(H)** Percentage of nuclei per sample for each inhibitory subtype in snRNA-seq. **(I)** Known inhibitory subtype markers for validation of annotation. Expression levels are normalized and log transformed. Note that the individual cells are plotted for *NMBR, PAX6, SYT6* since the expression level is not sufficient for a violin plot. **(J)** Annotated snATAC-seq per species in the UMAP space. All 8 subtypes identified in snRNA-seq are also distinctly annotated in snATAC-seq for all species. **(K)** Percentage of nuclei per sample for each inhibitory subtype in snATAC-seq. **(L)** Clarity of annotation transfer from snRNA-seq to snATAC-seq as displayed by Jaccard similarity index (similar to Extended Figure 1K).


A
Anterior Cingulate Cortex
MOL
OPC
UMAP_2
UMAP_1
B
All Cells
MOL
OPC
0%
25%
50%
75%
100%
Percentage
h1
h2
h3
h4
c1
c2
c3
c4
m1
m2
m3
m4
C
Lower Layer
Upper Layer
Section 5
Section 4
Section 3
Section 2
Section 1
Fraction of mature oligodendrocyte
D
Fraction of OPC
Posterior Cingulate Cortex
E
F
Anterior Cingulate Cortex
G
Dorsolateral prefrontal cortex
Reference: Chimpanzee
p = 0.74
p = 0.00043
p = 0.00012
OPC / MOL
Reference: Macaque
p = 0.11
p = 2.7e−05
p = 0.00052
H
Caudate Nucleus
OPC
MOL
p = 0.82
p = 0.59
p = 0.39
p = 0.59
Fraction in Glia
I
Dentate Gyrus
p = 0.048
p = 0.38
J
UP
DOWN
Posterior Cingulate Cortex
vs
Anterior Cingulate Cortex
# of genes
Open
Closed
# of CREs (log10)
K
Number of UMI in OPC & MOL
ACC
PCC
L
-log10(FDR)
Odds Ratio
* FDR < 0.05

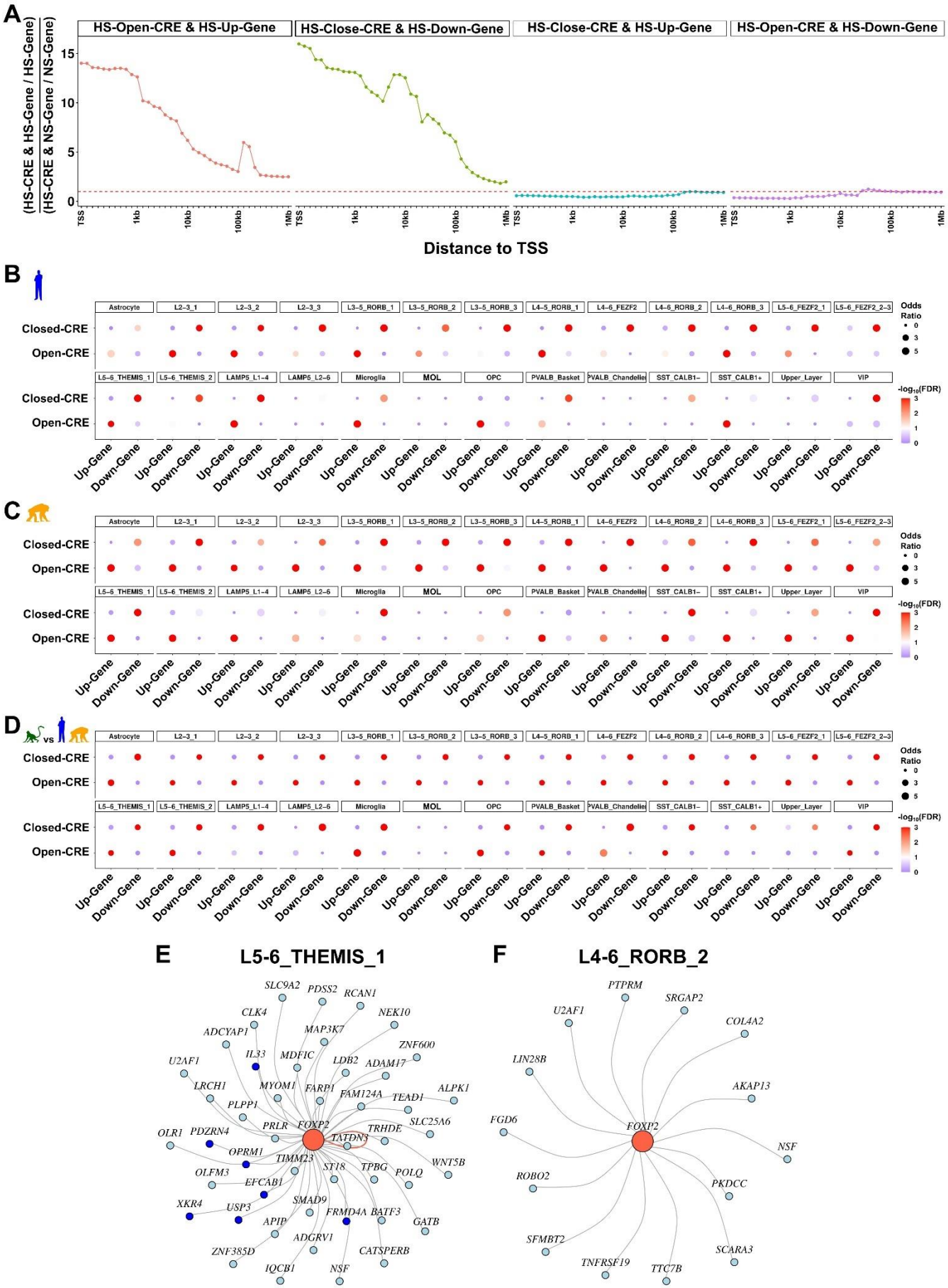
Extended Figure 4: Additional analyses on the oligodendrocyte lineage. (A) UMAP of MOLs and OPCs in the anterior cingulate cortex (ACC). **(B)** Percentage of cells contributed per individual per species per cell type. **(C-F)** Fractions of MOLs and OPCs in smFISH experiments per section (see Figure 1). Stitched column images encompassing all layers were divided into 5 equal parts from upper (Section 1) to lower layers (Section 5) in all images from human and chimpanzee. **(C-D)** are data from posterior cingulate cortex (PCC), and **(E-F)** are data from ACC. Each data point is a bin that contains sections from all layers. **C-D:** 4 bins from 2 humans, 6 bins from 3 chimpanzees. **E-F:** 9 bins from 3 humans and 3 chimpanzees. Data are represented as mean values +/- SEM. **(G)** Deconvoluted proportions from OLIG2+ bulk RNA-seq dataset (reference datasets: (left) chimpanzee, (right) rhesus macaque from this study). N = 22 (human), 10 (chimpanzee), 10 (rhesus macaque) individuals. P-value: Wilcoxon rank sum test, two-sided). **(H-I)** Fraction of OPCs or MOLs in glia in **(H)** caudate nucleus and **(I)** dentate gyrus per species. Each dot represents a sample (p-value: Wilcoxon rank sum test, two-sided. Caudate nucleus: N = 6 per species. Dentate gyrus: N = 6 for human, 3 for rhesus macaque). Box plots represent median and interquartile range in panels **G-I**. **(J)** Number of species-specific regulatory changes (PCC snRNA-seq (top), ACC snRNA-seq (middle), and PCC snATAC-seq (bottom, $\log_{10}$ transformed for better readability). **(K)** Distributions of UMIs (unique molecular identifiers) in ACC and PCC oligodendrocyte lineage nuclei (N=12 individuals both for PCC and ACC). Box plots represent median and interquartile range. **(L)** Enrichment results between species-specifically expressed genes in ACC (x-axis) and PCC (this study, y-axis). Blue asterisk indicates a significant overlap(FDR < 0.05, Fisher's exact test, one-sided).


A
L2–3_1
L2–3_2
L2–3_3
L3–5_RORB_1
L3–5_RORB_2
L3–5_RORB_3
L4–5_RORB_1
L4–6_FEZF2
L4–6_RORB_2
L4–6_RORB_3
L5–6_FEZF2_1
L5–6_FEZF2_2–3
L5–6_THEMIS_1
L5–6_THEMIS_2
LAMP5_L1–4
LAMP5_L2–6
PVALB_Basket
PVALB_Chandelier
SST_CALB1–
SST_CALB1+
Upper_Layer
VIP
vs
1000 500 0 500 1000
UP
DOWN
# of genes
B
2.5 0.0 2.5
Open
Closed
# of CREs (log10)
C
snRNA-seq
R = 0.93
p = 2.1e−10
HS Genes
CS Genes
D
snATAC-seq
R = 0.95
p = 2.2e−11
HS CREs
CS CREs
Excitatory
Inhibitory
E
snRNA-seq
snATAC-seq
Percentage of HS-Genes
Percentage of HS-CREs
Number of excitatory subtypes
Number of inhibitory subtypes

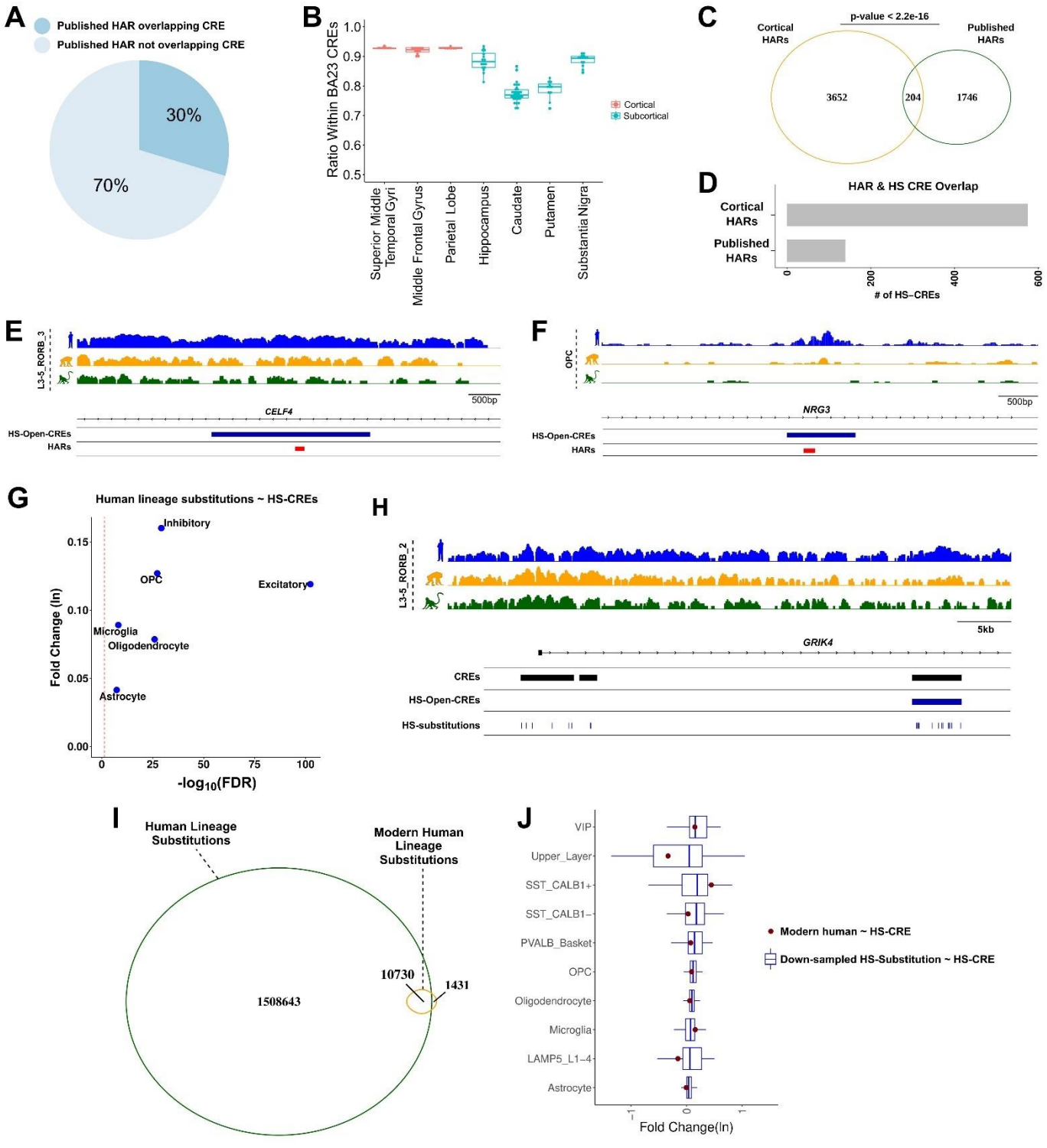
Extended Figure 5: Additional analyses of the regulatory changes in neuronal subtypes. (A-B) Number of regulatory changes that are human-specific, chimpanzee-specific or differential between rhesus macaque - human and rhesus macaque – chimpanzee in **(A)** snRNA-seq or **(B)** snATAC-seq (log_{10} transformed for better readability). **(C)** Scatter plots of number of HS-Genes and CS-Genes per neuronal subtype. Dashed rectangles indicate the subtypes with an excess number of human-specific regulatory gene expression changes (Two-sided chi-square test, FDR < 0.05). Shaded area indicates 95% confidence interval around the best fit (R indicates Spearman's rank correlation coefficient). **(D)** Same as **(C)** for HS-CREs and CS-CREs identified in snATAC-seq data. **(E)** Percentage distribution of excitatory HS-Genes that are found in only one subtype or shared among increasing number of subtypes (x-axis). Sum of all percentages equal 100. From left to right: in excitatory snRNA-seq, excitatory snATAC-seq, inhibitory snRNA-seq, inhibitory snATAC-seq.


A
Excitatory neurons
Inhibitory neurons
Kozlenkov HS-Genes
Kozlenkov CS-Genes
B
Kozlenkov HS-CREs
Kozlenkov CS-CREs
C
Berto HS-Genes
Berto CS-Genes
This Study HS-Genes or CREs
This Study CS-Genes or CREs
This Study HS-Genes or CREs
This Study CS-Genes or CREs
D
snRNA-seq ~ Bulk (Kozlenkov et al.)
Kozlenkov Human Specific
Percentage of genes also found in bulk
Number of subtypes with the same HS-Gene (Excitatory)
E
snATAC-seq ~ Bulk (Kozlenkov et al.)
Excitatory Pseudobulk Human Specific
Percentage of CREs also found in bulk
Number of subtypes with the same HS-CRE (Excitatory)
F
snRNA-seq ~ Pseudobulk all subtypes
Excitatory Pseudobulk Human Specific
Percentage of genes also found in bulk
Number of subtypes with the same HS-Gene (Excitatory)
G
snATAC-seq ~ Pseudobulk all subtypes
Kozlenkov Human Specific
Percentage of CREs also found in bulk
Number of subtypes with the same HS-CRE (Excitatory)

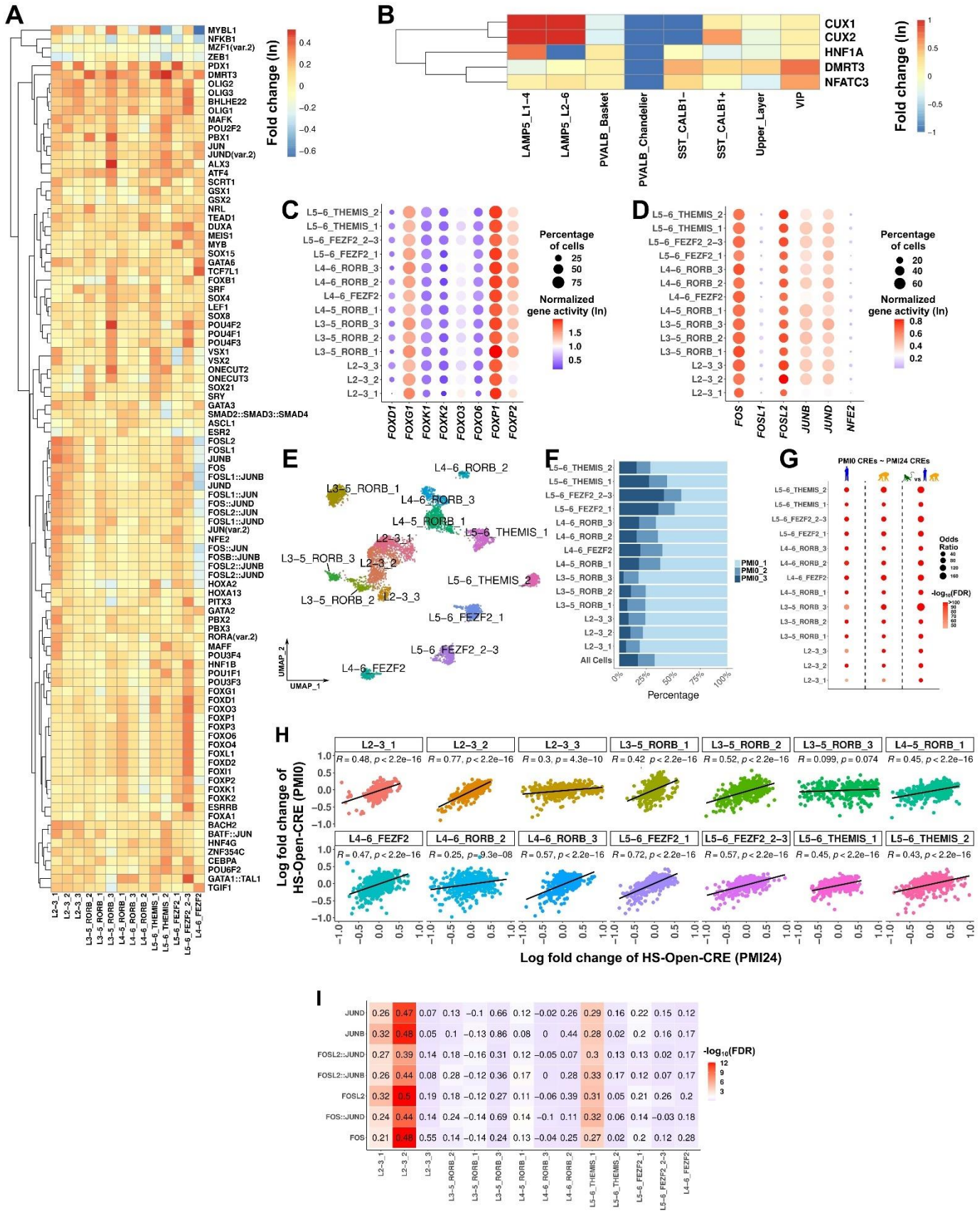
Extended Figure 6: Comparisons of neuronal expression patterns between this dataset and previous comparative bulk datasets. (A-C) Enrichments of species-specific expression patterns between this study and previous bulk studies between excitatory neurons (left) and inhibitory neurons (right). **(A)** Transcriptomic changes between the Kozlenkov et al. dataset and this dataset, **(B)** epigenomic changes between the Kozlenkov et al. dataset and this dataset, **(C)** transcriptomic changes between the Berto et al. dataset and this dataset. FDR values are from a Fisher's exact test with multiple testing correction. **(D-E)** Subtype-specific changes are captured less in the bulk RNA-seq datasets. **(D)** Comparison of excitatory HS-Genes between a previous bulk analysis and this dataset. Top: odds ratio between the bulk dataset and this dataset with increasing subtype specificity of HS- Genes (from right to left). Bottom: percentage of HS- Genes that were also found in the bulk dataset. **(E)** Same comparison as **(D)** with HS-CREs. **(F-G)** Subtype-specific changes are captured less when the subtypes are combined within the same dataset. **(F)** Same comparison as **(D)** with HS-Genes but this time pseudobulk data results are obtained by combining the subtypes in this study. **(G)** Same comparison as **(F)** with HS-CREs.


A
HS-Open-CRE & HS-Up-Gene
HS-Close-CRE & HS-Down-Gene
HS-Close-CRE & HS-Up-Gene
HS-Open-CRE & HS-Down-Gene
(HS-CRE & HS-Gene / HS-Gene)
(HS-CRE & NS-Gene / NS-Gene)
Distance to TSS
B
C
D
Astrocyte
L2–3_1
L2–3_2
L2–3_3
L3–5_RORB_1
L3–5_RORB_2
L3–5_RORB_3
L4–5_RORB_1
L4–6_FEZF2
L4–6_RORB_2
L4–6_RORB_3
L5–6_FEZF2_1
L5–6_FEZF2_2–3
L5–6_THEMIS_1
L5–6_THEMIS_2
LAMP5_L1–4
LAMP5_L2–6
Microglia
MOL
OPC
PVALB_Basket
PVALB_Chandelier
SST_CALB1–
SST_CALB1+
Upper_Layer
VIP
Closed-CRE
Open-CRE
Up-Gene
Down-Gene
Odds Ratio
-log10(FDR)
E
L5-6_THEMIS_1
SLC9A2
PDSS2
RCAN1
CLK4
NEK10
ADCYAP1
MAP3K7
ZNF600
U2AF1
IL33
MDFIC
LDB2
ADAM17
LRCH1
MYOM1
FARP1
FAM124A
TEAD1
ALPK1
PLPP1
PRLR
FOXP2
SLC25A6
TRHDE
OLR1
PDZRN4
TATDN3
OPRM1
TIMM23
WNT5B
OLFM3
ST18
TPBG
POLQ
EFCAB1
XKR4
USP3
SMAD9
FRMD4A
BATF3
GATB
APIP
ADGRV1
ZNF385D
CATSPERB
IQCB1
NSF
F
L4-6_RORB_2
PTPRM
SRGAP2
U2AF1
COL4A2
LIN28B
AKAP13
FGD6
FOXP2
NSF
ROBO2
PKDCC
SFMBT2
SCARA3
TNFRSF19
TTC7B

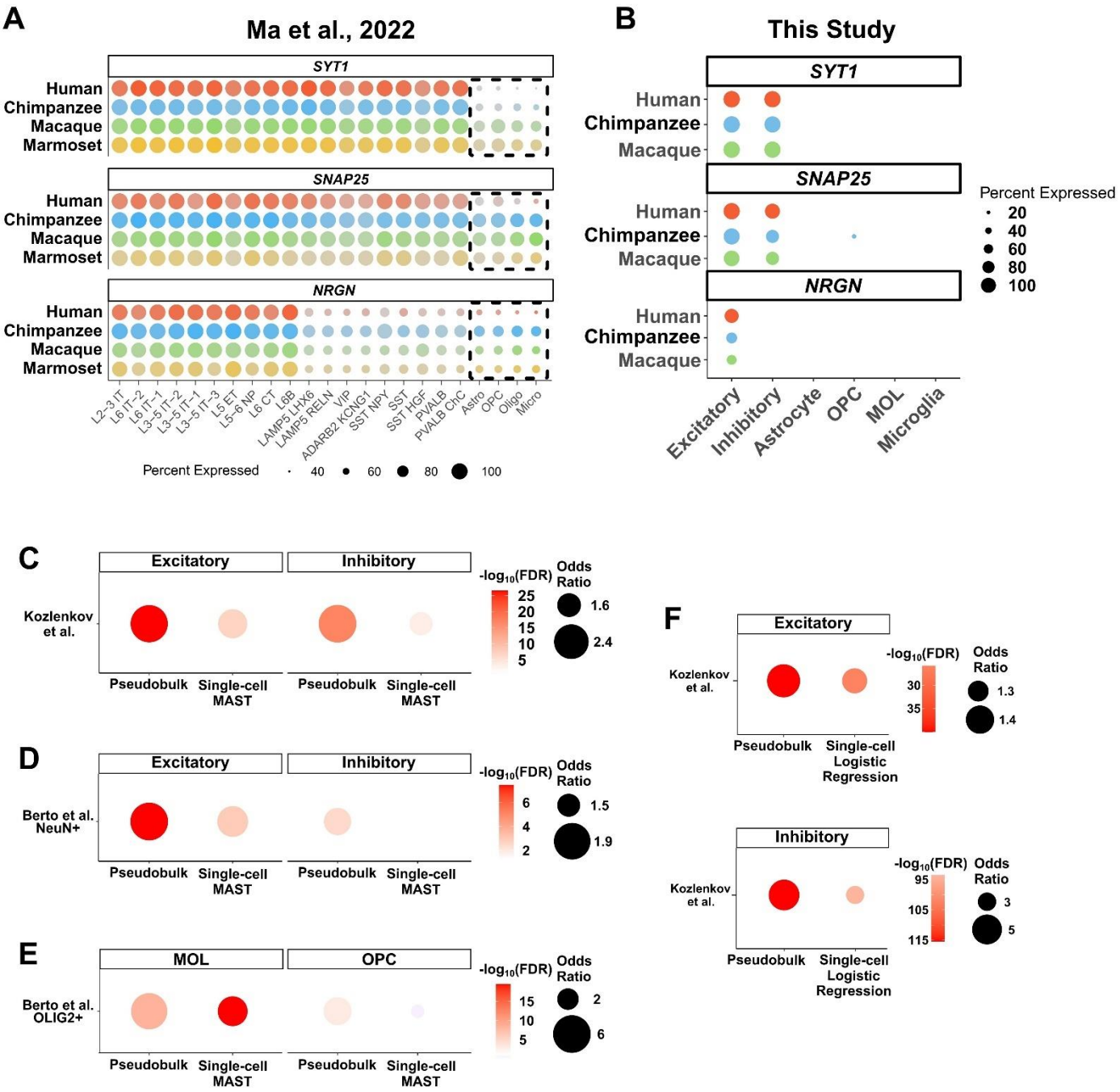
Extended Figure 7: Associations between HS-Genes and HS-CREs. (A) The specificity of association between HS-Genes and HS-CREs decreases with increasing distance from the transcription start site (TSS). Y-axis shows the odds ratio, which is defined by the ratio of HS-Genes associated with HS-CREs divided by the ratio of not significant genes (NS-Genes) associated with HS-CREs. We calculated the odds ratio for increasing the distance from the TSS in both directions for four different associations (from left to right): HS-Open-CRE & HS-Up-Gene, HS-Open-CRE & HS-Down-Gene, HS-Close-CRE & HS-Up-Gene, HS-Open-CRE & HS-Down-Gene. The value for each observation was obtained by taking the mean across all cell types. **(B-D)** Enrichments between HS-CRE associated genes within a 10kb window from the TSS and HS-Genes per cell type. **(E-F)** Putative target genes of human-specific *FOXP2* upregulation in **(E)** L5-6_THEMIS_1 and **(F)** L4-6_RORB_2 cells. All genes show human-specific up / down regulation in their respective subtype and reside within 500kb of at least one human-specific chromatin accessibility change that has a FOXP2 motif. Dark blue circles indicate the genes that are not altered in the other 12 excitatory subtypes (similar to *FOXP2* itself). Red loop in (**A**) indicates that *FOXP2* itself is also identified with this analysis in the L5-6 THEMIS 1 subtype.


A
Published HAR overlapping CRE
Published HAR not overlapping CRE
30%
70%
B
Ratio Within BA23 CREs
1.0
0.9
0.8
0.7
0.6
0.5
Cortical
Subcortical
Superior Temporal Gyri
Middle Frontal Gyrus
Parietal Lobe
Hippocampus
Caudate
Putamen
Substantia Nigra
C
p-value < 2.2e-16
Cortical HARs
Published HARs
3652
204
1746
D
HAR & HS CRE Overlap
Cortical HARs
Published HARs
0
200
400
600
# of HS−CREs
E
L3−5_RORB_3
500bp
CELF4
HS-Open-CREs
HARs
F
OPC
500bp
NRG3
HS-Open-CREs
HARs
G
Human lineage substitutions ~ HS-CREs
Inhibitory
OPC
Excitatory
Microglia
Oligodendrocyte
Astrocyte
Fold Change (ln)
0.15
0.10
0.05
0.00
0
25
50
75
100
-log10(FDR)
H
L3−5_RORB_2
5kb
GRIK4
CREs
HS-Open-CREs
HS-substitutions
I
Human Lineage Substitutions
Modern Human Lineage Substitutions
1508643
10730
1431
J
VIP
Upper_Layer
SST_CALB1+
SST_CALB1−
PVALB_Basket
OPC
Oligodendrocyte
Microglia
LAMP5_L1−4
Astrocyte
−1
0
1
Fold Change(ln)
Modern human ~ HS-CRE
Down-sampled HS-Substitution ~ HS-CRE

Extended Figure 8: Further associations between human-specific substitutions and human-specific chromatin accessibility changes. (A) Pie-chart distribution of published HARs overlapping CREs in this dataset. **(B)** Ratio of non-BA23 CREs overlapping BA23 CREs (denominator: all CREs in BA23). Each dot represents an independent library prep. Red datasets indicate cortical regions, blue datasets indicate sub cortical regions. (Sample sizes; Superior Middle Temporal Gyri: 8, Middle Frontal Gyrus: 12, Parietal Lobe: 7, Hippocampus: 16, Caudate: 32, Putamen: 11, Substantia Nigra: 14. Box plots represent median and interquartile range). **(C)** Overlap between cortical HARs (identified in this study) and published HARs (p-value: One-sided chi-square test). **(D)** Number of HS-CREs associated with a cortical HAR or a published HAR. **(E-F)** Examples of HS-Open-CRE associated HARs. Bottom panel shows the multi-species alignment for *CELF4* HAR. Dots represent no change from the human (hg38) sequence. Human-specific changes conserved in other lineages are highlighted in shaded blue. **(G)** Enrichment of human-specific substitutions within the HS-CREs per major cell type. Enrichment is tested by a negative binomial regression model with CRE length and evolution of the CRE as the predictor variables (HS-CRE or not HS-CRE) and number of human-specific substitutions as a response variable(Significance: likelihood ratio test). **(H)** Example of an HS-Open-CRE with many human specific substitutions. **(I)** Overlap of substitutions that are specific to the human lineage (in comparison to chimpanzee, gorilla and gibbon) and previously identified modern human substitutions. **(J)** Log fold changes of substitution and HS-CRE association for substitutions on the human (blue boxplots) and modern human lineage (tile red dots) per cell type (except for excitatory cells). Human lineage-specific substitutions were randomly down sampled for 100 times to 12,161 (the number of modern human-specific variants) for comparison. Box plots represent median and interquartile range.


A
B
C
D
E
F
G
H
I
Fold change (ln)
Percentage of cells
Normalized gene activity (ln)
Percentage
PMI0 CREs ~ PMI24 CREs
Odds Ratio
-log10(FDR)
Log fold change of HS-Open-CRE (PMI0)
Log fold change of HS-Open-CRE (PMI24)
UMAP_1
UMAP_2
LAMP5_L1−4
LAMP5_L2−6
PVALB_Basket
PVALB_Chandelier
SST_CALB1−
SST_CALB1+
Upper_Layer
VIP
CUX1
CUX2
HNF1A
DMRT3
NFATC3

Extended Figure 9: Supplementary motif enrichment results. (A-B) Hierarchical clustering of motif enrichments (log-fold change) in HS-Open-CREs across **(A)** excitatory and **(B)** inhibitory neuronal subtypes. Transcription factors (TFs) associated with each motif enrichment are displayed in rows and the neuronal subtypes are displayed in columns. Only the motifs enriched in at least one subtype are displayed. **(C-D)** Accessibility of **(C)** FOX and **(D)** FOS / JUN family TFs. Accessibility is assessed by the normalized gene activity scores (calculated using Cicero[51]) per gene per subtype. **(E)** Annotated UMAP of excitatory neurons in snRNA-seq of surgically resected samples (referred to as PMI0 compared to postmortem BA23 human samples that are referred to as PMI24 in this figure). **(F)** Percentage of nuclei per sample for each excitatory subtype in snRNA-seq. **(G)** Enrichments of species-specifically expressed genes when PMI0 or PMI24 datasets were used as the human dataset in the comparative analyses. **(H)** Pearson correlations (test for p-value is two-sided) between the log fold changes of HS-Open-CRE motif enrichments when PMI0 or PMI24 datasets were used as the human dataset in the comparative analyses. **(I)** Heatmap of motif FOS / JUN motif enrichments per excitatory subtype in HS-Open-CREs. Colors correspond to $-\log_{10}$(FDR); numbers correspond to log fold change of enrichment.



Extended Figure 10: Comparisons with external datasets (A) Expression levels of three ambient RNA markers highly expressed in neurons (*SYT1*, *SNAP25* and *NRGN*[21]) in the Ma et al. dataset[14]. The dot plot is generated through the interactive web tool linked to the original publication. Dashed square brackets indicate glial cell types, which show exceptionally low levels in the human dataset. Note that the smallest dot shows the presence of a transcript in 40% of the cells. **(B)** Same as **(A)** using this PCC dataset. Neuronal ambient RNA markers are detected at very low levels in glial cells across species after ambient RNA removal. **(C-E)** Enrichment of HS-Genes between the previous study (y-axis) and the current study (x-axis) with two alternative methods. **(F)** Enrichment of HS-CREs between the previous study (y-axis) and the current study (x-axis) with two alternative methods. For simplicity, we combined all HS-Genes from the subtypes of a major cell type (e.g. all excitatory neuronal subtypes were combined for the excitatory cell type comparisons). P-values were computed using a Fisher's exact test (one-sided) and false discovery rate (FDR) was calculated per panel.

CHAPTER 3: Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets

Published as:

Caglayan, E.**, Liu, Y., Konopka, G.** 2022. Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. *Neuron*. DOI: 10.1016/j.neuron.2022.09.010

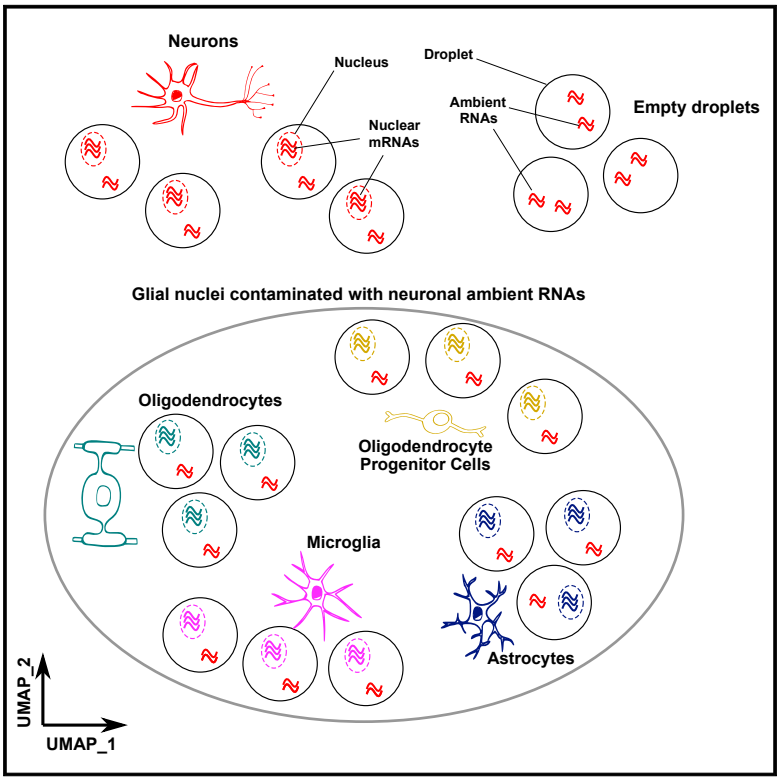
**co-corresponding authors

Neuron

NeuroResource

Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets

Graphical abstract



Highlights

- Ambient RNA in single-nuclei genomic data causes conspicuous contamination in glia
- Ambient RNA contamination can be mitigated by physical sorting or *in silico* methods
- Previously annotated immature oligodendrocytes are ambient RNA-contaminated glia
- Ambient RNA removal reveals unannotated COPs in all adult human brain datasets

Authors

Emre Caglayan, Yuxiang Liu, Genevieve Konopka

Correspondence

emre.caglayan@utsouthwestern.edu (E.C.), genevieve.konopka@utsouthwestern.edu (G.K.)

In brief

Caglayan et al. examine brain single-nuclei datasets and uncover signatures of neuronal contamination in glia that mask rare cell types and lead to improper cell annotation. The authors show how to reduce this contamination by using either *in silico* approaches or physical separation of cell types.

Caglayan et al., 2022, Neuron *110*, 4043–4056
December 21, 2022
https://doi.org/10.1016/j.neuron.2022.09.010

CellPress

Neuron

CellPress

NeuroResource

Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets

Emre Caglayan,[1,2,*] Yuxiang Liu,[1,2] and Genevieve Konopka[1,2,3,*]
[1]Department of Neuroscience, UT Southwestern Medical Center, Dallas, TX 75390, USA
[2]Peter O'Donnell Jr. Brain Institute, UT Southwestern Medical Center, Dallas, TX 75390, USA
[3]Lead contact
*Correspondence: emre.caglayan@utsouthwestern.edu (E.C.), genevieve.konopka@utsouthwestern.edu (G.K.)
https://doi.org/10.1016/j.neuron.2022.09.010

SUMMARY

Ambient RNA contamination in single-cell and single-nuclei RNA sequencing (snRNA-seq) is a significant problem, but its consequences are poorly understood. Here, we show that ambient RNAs in brain snRNA-seq datasets have a nuclear or non-nuclear origin with distinct gene set signatures. Both ambient RNA signatures are predominantly neuronal, and we find that some previously annotated neuronal cell types are distinguished by ambient RNA contamination. We detect pervasive neuronal ambient RNA contamination in all glial cell types unless glia and neurons are physically separated prior to sequencing. We demonstrate that this contamination can be removed *in silico* and show that previous single-nuclei RNA-seq-based annotations of immature oligodendrocytes are glial nuclei contaminated with ambient RNAs. After ambient RNA removal, we detect rare, committed oligodendrocyte progenitor cells not annotated in most previous adult human brain datasets. Together, these results provide an in-depth analysis of ambient RNA contamination in brain single-nuclei datasets.

INTRODUCTION

Single-nuclei RNA sequencing (snRNA-seq) experiments involve nuclei isolation and subsequent capture of each nucleus in a single droplet containing a unique cell barcode. However, this capture process can also encapsulate freely floating transcripts, resulting in contamination of the endogenous expression profile. These extraneous transcripts have been previously referred to as "ambient RNAs" (Luecken and Theis, 2019; Young and Behjati, 2020). Because ambient RNAs are expected to be primarily derived from more abundant cell types, ambient RNA contamination in less abundant cell types can result in a considerably skewed endogenous expression profile. Thus, failure to account for ambient RNA contamination can result in biological misinterpretation, especially in cell types with less abundant transcripts. Many previous studies have not removed ambient RNA contamination from the endogenous expression profiles of their datasets; therefore, it is possible that ambient RNA contamination constitutes an important problem that has led to misinterpretations in the downstream analyses. In our study, we address this possibility by reanalyzing several previously published datasets.

In addition to contaminating the endogenous expression profiles of nuclei, ambient RNAs are also captured in droplets that do not capture nuclei. These droplets are referred to as "empty droplets" (Lun et al., 2019; Macosko et al., 2015). Interestingly, empty droplets do not always show a clear separation from the non-empty droplets in terms of unique read counts, underscoring the high levels of ambient transcripts that are captured in most single-nuclei (and single-cell) preparations (Lun et al., 2019). The lack of clear separation between empty and non-empty droplets based on read counts makes it difficult to justify using a read-count-based cutoff (also called UMI—unique molecular identifier—cutoff). A UMI cutoff can lead to the miscalling of empty droplets as non-empty droplets, or the miscalling of certain cell types that contain fewer transcripts than others as empty droplets (Luecken and Theis, 2019). Recent tools have addressed this problem by distinguishing non-empty from empty droplets by using other metrics, such as expression profile and nuclear fraction (Heiser et al., 2021; Lun et al., 2019; Muskovic and Powell, 2021). However, the composition of empty droplets is tissue-dependent and the specific composition of empty droplets for a given tissue is often not explored. Without proper understanding of the transcriptional composition of empty droplets, it can be difficult to decide whether a given cluster of cell barcodes are empty droplets or non-empty droplets that captured real nuclei/cells.

Taken together, ambient RNAs pose two challenges: contamination of the endogenous profile of real nuclei/cells and ambiguous separation of empty and non-empty droplets. With respect to these two issues, contamination in real nuclei/cells has

received less attention. Several recently developed tools are specifically designed to remove ambient RNA contamination from real nuclei/cells; however, their utilization has been limited (Fleming et al., 2019; Yang et al., 2020; Young and Behjati, 2020). Additionally, while these tools aim to be effective across different tissues, the genetic signatures of ambient RNA contamination differ between tissue types. Thus, it is also important to characterize the overrepresented genes in the ambient RNA population of a given tissue type. This will aid both the interpretation of previously published datasets as well as assess the effects of ambient RNA contamination removal tools for specific tissues by evaluating the levels of ambient RNA population before and after the removal of contamination. In this study, we focus on snRNA-seq studies from brain tissue, one of the most frequently profiled tissues using this technique due to its high cellular heterogeneity (Bakken et al., 2021; Jäkel et al., 2019; Nagy et al., 2020; Velmeshev et al., 2019). Brain tissue is also a good model to understand the effects of ambient RNA contamination, as neurons are more abundant and contain more transcripts than glia in the adult mammalian cortex (Ruzicka et al., 2020). Therefore, ambient RNA profiles from snRNA-seq studies of brain should be biased for neurons. We hypothesize that this bias may contribute to both empty droplets with neuronal signatures as well as distinctive neuronal read contamination in non-neuronal cell types.

Here, we analyzed ambient RNA signatures from human brain snRNA-seq datasets (Table S1) by retaining additional cell barcodes that are typically removed due to low UMI counts. We found two types of ambient RNAs separated by their intronic read ratio: non-nuclear ambient RNAs with low intronic reads and nuclear ambient RNAs with high intronic reads. Comparisons with nuclei-sorted datasets (SDs) revealed that non-nuclear ambient RNA can be cleared by physical nuclei sorting. We show that the ambient RNA signature is predominantly neuronal in origin and that all glia are contaminated with ambient RNAs. Ambient RNA contamination in glia was removed in datasets depleted of neurons (NeuN− SDs) before droplet capture (Hodge et al., 2019). As an *in silico* alternative, we used CellBender (Fleming et al., 2019) and subsequent subcluster cleaning, which also removed detectable ambient RNA contamination from glia. Re-analysis of the oligodendrocyte (OL) lineage trajectory after ambient RNA removal revealed that previously annotated immature OLs are likely glial nuclei contaminated with ambient RNA. Instead, we found a rare, transient cell type named COPs (committed oligodendrocyte progenitor cells), which were previously described in an adolescent mouse dataset (Marques et al., 2016) but not described in most human snRNA-seq datasets, with few exceptions (Jäkel et al., 2019; Perlman et al., 2020). Together, our results provide an in-depth analysis of ambient RNA contamination in single-nuclei brain datasets and reveal misinterpreted results that can be explained by ambient RNA contamination.

RESULTS

Both nuclear and non-nuclear ambient RNAs confound cell-type annotation

Studies of adult brain tissue have repeatedly shown a greater number of transcripts present in neurons compared with glia (Ruzicka et al., 2020). Interestingly, several snRNA-seq studies reported some neuronal cell types (e.g., Neu-NRGN and Neu-mat in Figure 1A) that have fewer transcript counts than other neurons (Ruzicka et al., 2020; Velmeshev et al., 2019). We also noticed that Neu-NRGNs had higher mitochondrial reads than other neuronal cell types (Velmeshev et al., 2019) (Figure 1B). Because this dataset was generated by nuclei isolation but not nuclei sorting (purification of DAPI+ nuclei with flow cytometry), we refer to it as non-SD 1 (NSD1). Considering the possibility of non-nuclear transcript contamination in nuclei-based sequencing, cell barcodes with high non-nuclear contamination should contain lower intronic read ratios because intronic reads will not be present in non-nuclear transcripts. We thus hypothesized that Neu-NRGNs and Neu-mat may be represented by cell barcodes with a high amount of such non-nuclear transcript contamination in NSD1. To test this hypothesis, we calculated the intronic read ratio per cell barcode (see STAR Methods) and found that Neu-NRGN but not Neu-mat displayed markedly lower intronic read ratios compared with other cell types (Figure 1C). To test whether these cell types are similar to normally discarded cell barcodes with low UMI counts, we then clustered an excess number of cell barcodes that contained low UMI counts together with the annotated cell barcodes from the original publication. To identify annotated cell barcodes that cluster with low UMI cell barcodes, we focused on the clusters that were largely, but not fully, filtered out (>75% filtered) in the original publication and named them "ambient clusters" (Figures 1D and 1E). We found that Neu-NRGN cell barcodes predominantly clustered with the ambient clusters while other cell types were almost absent in ambient clusters, except for Neu-mat (Figure 1F). If Neu-NRGN cell barcodes are indeed highly contaminated with non-nuclear ambient RNA, they should also be depleted of long non-coding RNAs (lncRNAs), which are retained in the nucleus (Guo et al., 2020). Indeed, Neu-NRGN barcodes contained fewer lncRNAs than other neurons (Figure 1G). An interesting example is *MALAT1*, which has an elevated expression in the brain (Bernard et al., 2010) and was depleted among Neu-NRGN cell barcodes relative to other cell types (Figure 1H). Together, these results indicate that Neu-NRGN cell barcodes contain high non-nuclear ambient RNA contamination and are unlikely to represent intact nuclei.

We hypothesized that non-nuclear ambient RNA can be removed by fluorescence activated nuclei sorting (FANS). To test this, we analyzed another cortical snRNA-seq dataset, named SD1, in which the authors performed nuclei sorting (purification of DAPI+ nuclei with flow cytometry) (Lake et al., 2018). This sample preparation contrasts with the previous dataset (NSD1) we discussed that performed nuclei isolation but not nuclei sorting by flow cytometry (Figure 1). Because a low intronic read ratio indicates the presence of non-nuclear transcripts, we then determined the intronic read ratio in both datasets to assess whether non-nuclear transcripts are removed by nuclei sorting. As expected, NSD1 displayed a low intronic read ratio in cell barcodes with low UMI counts, as these cell barcodes are primarily associated with ambient RNAs (Figure 2A). In contrast, we found that the intronic read ratio did not markedly change with increasing UMI counts in SD1 (Figure 2B). We then assessed ambient RNA signatures after the

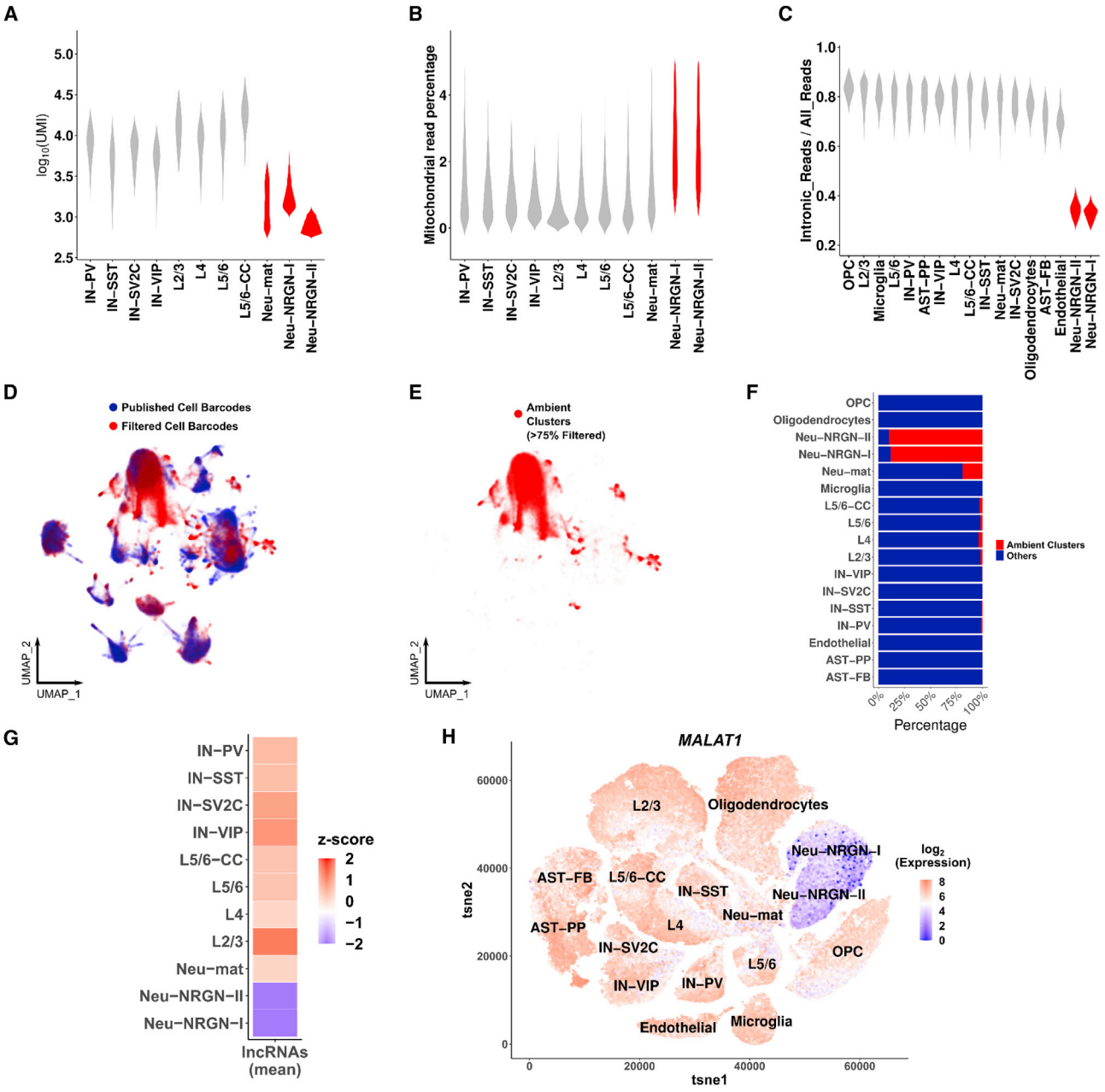


Figure 1. Neu-NRGNs are comprised of non-nuclear ambient RNAs

(A) Log10 transformed UMI counts per neuronal cell type in NSD1. Cell types with an unusually low UMI count are shown in red.

(B) Mitochondrial read percentage per neuronal cell type in NSD1. Cell types with significantly higher mitochondrial read percentage are shown in red (Wilcoxon rank sum test, p value < 0.05).

(C) Intronic read ratios of all cell types in NSD1.

(D) UMAP representation after co-embedding of same numbers of published (blue) and filtered cell barcodes (red) (dataset: NSD1).

(E) Clusters that are >75% composed of filtered cell barcodes are highlighted and named ambient clusters (dataset: NSD1).

(F) Bar plot of the percentage of cell barcodes in ambient clusters per cell type. Red: ambient clusters, blue: other clusters.

(G) Heatmap of normalized, log-transformed, and *Z* scored expression levels of lncRNAs across cell types. The means of all lncRNAs were taken before calculating *Z* scores.

(H) T-distributed stochastic neighbor embedding (tSNE) plot of the normalized and log-transformed expression level of *MALAT1* in all nuclei in NSD1.

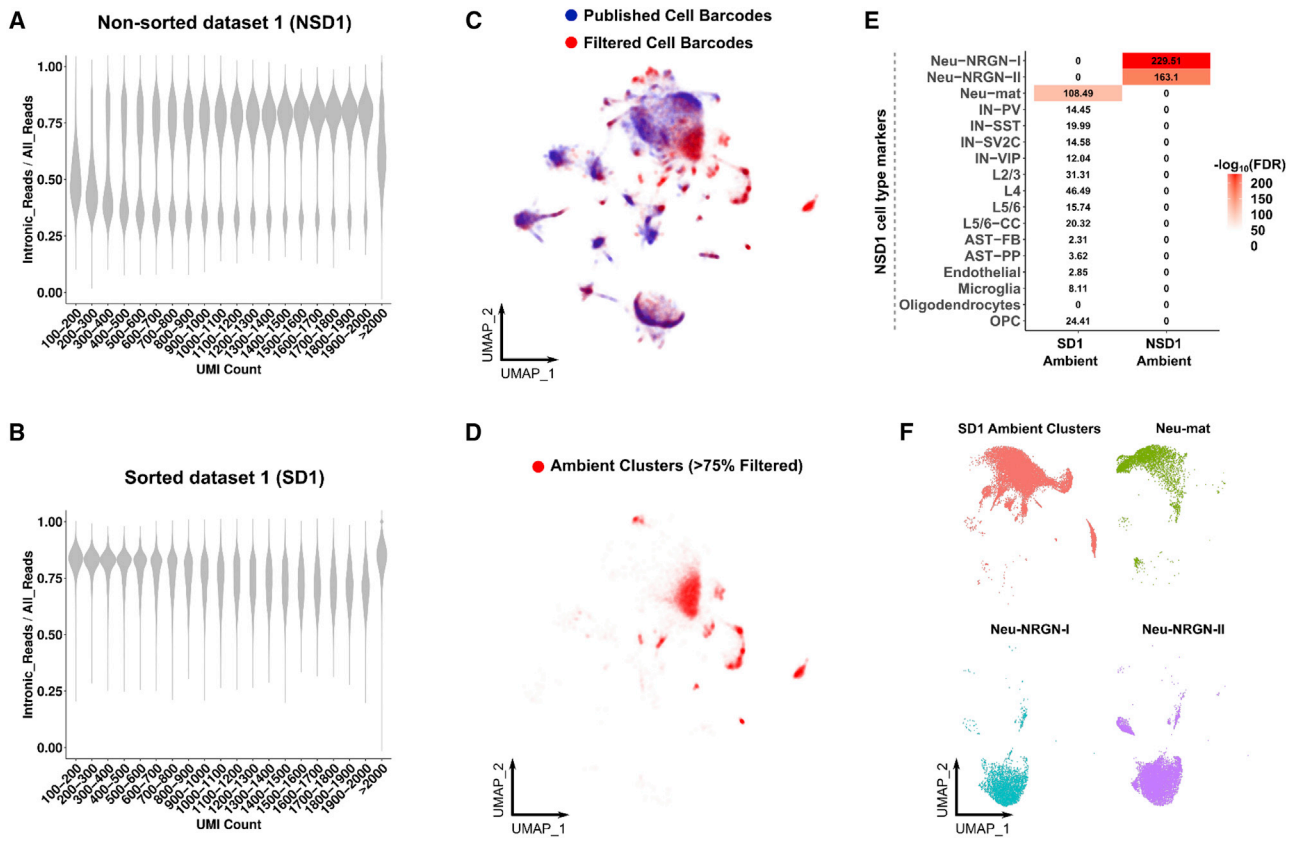


Figure 2. Non-nuclear and nuclear ambient RNAs are distinct from each other

(A and B) Intronic read ratio across increasing UMI count in (A) NSD1 and (B) SD1. UMI counts are divided into intervals of 100, from 100 to 2,000.

(C) UMAP representation after co-embedding of the same numbers of published (blue) and filtered cell barcodes (red) (dataset: SD1).

(D) Clusters that are >75% composed of filtered cell barcodes are highlighted and named ambient clusters (dataset: SD1).

(E) Heatmap of enrichments between ambient RNA markers and Neu-mat or Neu-NRGN cell types in NSD1; Fisher's exact test, $\log_{10}$(FDR).

(F) Co-embedding of Neu-NRGNs, Neu-mat, and SD1 ambient clusters.

See also Figures S1 and S2 and Tables S2 and S3.

removal of non-nuclear ambient RNA and found ambient clusters in SD1 (Figures 2C and 2D). Interestingly, SD1 ambient cluster markers were highly and distinctly enriched in the Neu-mat markers from NSD1 (Figure 2E). We also observed significant enrichment of SD1 ambient cluster markers in other neuronal cell-type markers, although it was markedly less compared with Neu-mat. Co-clustering of SD1 ambient clusters and NSD1 cell types also grouped SD1 ambient clusters and Neu-mat cell barcodes together (Figure 2F). Overall, these results indicate that Neu-mat cell barcodes carry an unusually high nuclear ambient RNA signature, and Neu-NRGN cell barcodes are distinctly contaminated with non-nuclear ambient RNA. These results also reveal that nuclei sorting, but not nuclei isolation alone, effectively removes non-nuclear ambient RNA; however, nuclei sorting cannot remove nuclear ambient RNA. Therefore, we named NSD1 ambient RNA markers "non-nuclear ambient markers" and SD1 ambient RNA markers "nuclear ambient markers" (Table S2). We selected a combination of the top 500 nuclear ambient markers and the top 500 non-nuclear ambient markers (logFC > 1 and false discovery rate [FDR] < 0.05) for the subsequent enrichment analyses.

To provide independent support for these results, we then analyzed a second snRNA-seq cortical human dataset that did not include nuclei sorting (NSD2, see STAR Methods and Table S1) and a cortical human dataset that was generated after nuclei sorting (SD2) (Tran et al., 2021). We reproduced a similar pattern of increasing intronic read ratio with increasing UMI in the NSD2, whereas this was not observed in SD2 (Figures S1A and S1B). We then similarly identified ambient clusters from the NSD2 dataset (Figure S1C). The intronic read ratio distribution in ambient clusters was bimodal, and we divided the clusters into two categories: cell barcodes with high intronic read ratio (High-Intron-CB) and cell barcodes with low intronic read ratio (Low-Intron-CB) (Figure S1D). In line with the previous results, genes overrepresented in Low-Intron-CB were highly enriched in non-nuclear ambient RNA markers, whereas High-Intron-CB were enriched in nuclear ambient RNA markers (Figure S1E).

Signatures and sources of non-nuclear and nuclear ambient RNAs

To better understand the ambient RNA marker genes, we performed gene ontology enrichment and found that non-nuclear

ambient RNA markers are enriched for genes involved in ribosomal, mitochondrial, and synaptic functions, whereas nuclear ambient RNA markers are enriched for genes related to synaptic function (Figures S2A and S2B; Table S3). As previously hypothesized (Thrupp et al., 2020), we asked whether non-nuclear ambient RNAs are enriched in genes comprising synaptosomes, which are also marked by ribosomal, mitochondrial, and synaptic activity (Hafner et al., 2019). Using the top 500 markers in each ambient RNA group, we found that non-nuclear ambient RNAs are more enriched than nuclear ambient RNAs for transcripts of both vGLUT1-depleted (originating from postsynapse + soma) and vGLUT1-enriched (originating from presynapse) synaptosome markers (Hafner et al., 2019) (Figure S2C). Because nuclear ambient RNAs also showed significant association and were enriched in overall synaptic function but not ribosomal and mitochondrial function (Table S3), we then hypothesized that nuclear ambient RNAs are derived from highly expressed genes captured in neuronal nuclei. Indeed, nuclear ambient RNAs largely overlapped with genes that are highly expressed in neurons (Figure S2C). We note that this also explains the significant enrichments between neuronal cell-type markers and nuclear ambient RNA markers that we observed (Figure 2E). We also observed that the top non-nuclear ambient RNA and nuclear ambient RNA markers were also distinct, further underscoring the hypothesis that ambient RNAs are derived from different sources (Figure S2C).

Ambient RNA contamination of glial nuclei can be removed *in silico*

Given that neuronal genes are overrepresented in both ambient RNA types, we then hypothesized that ambient RNA contamination can make the transcriptomic profile of glial cell types appear more neuronal-like. As sorting for nuclei that do not express the neuronal marker NeuN (referred to as NeuN−) prior to droplet capture should remove neuronal ambient RNAs, we compared SDs and NSDs with three NeuN− sorted snRNA-seq datasets (NeuN− SD) (Bakken et al., 2021; Hodge et al., 2019; Sadick et al., 2022). We identified genes significantly overrepresented in the four exemplar datasets (SD1, SD2, NSD1, NSD2) compared with the three NeuN− SDs in 6 major cell types: excitatory and inhibitory neurons, OLs, OPCs (OL progenitor cells), astrocytes (AST), and microglia (MIC) (Table S4). We called these genes "NeuN− depleted genes." For each comparison, we then selected the top 500 NeuN− depleted genes and performed enrichment for ambient RNA markers. Despite the neuronal signature of ambient RNA markers, their enrichments within the NeuN− depleted genes were consistently significant in all glial cell types across the studies (Figures 3A and 3B). Notably, the association of ambient RNA markers and NeuN− depleted genes was consistently less in neuronal cell types than in glia (Figures 3A and 3B). Together, these results show that glial nuclei in cortical snRNA-seq are likely contaminated with neuronal ambient RNAs.

Our findings suggest that neuronal ambient RNAs contaminate glial nuclei unless samples are sorted to remove neuronal nuclei prior to droplet capture. To further test our finding, we used a dataset that employed two sorting strategies: one that depleted neurons (NeuN− and LHX2+ sorting also referred to as NeuN− SD3 in our comparisons) and one that did not deplete neurons (SOX9+ sorting) (Sadick et al., 2022) (Figures S3A–S3C). In line with our hypothesis, ambient RNA markers appeared less "expressed" across glial nuclei in the dataset with neuron depletion (Figure S3D), and genes less represented in the neuron depleted dataset were significantly enriched in ambient RNA markers (Figure S3E). We note that two ambient RNA markers we exemplify (*CSMD1* and *RALYL*) showed similar expression patterns between the two datasets. These genes are likely endogenously expressed in OPCs as high expression levels are detectable across the other NeuN− SDs (Bakken et al., 2021; Hodge et al., 2019), underscoring the importance of distinguishing ambient RNA contamination from endogenous transcripts per cell type. Together, these results provide further proof of neuronal ambient RNA contamination in the cortical snRNA-seq datasets.

Ambient RNA contamination within droplets that contain real nuclei is a general problem in snRNA-seq experiments, and various tools exist to remove ambient RNA contamination (Fleming et al., 2019; Yang et al., 2020; Young and Behjati, 2020). To assess the performance of these tools in the analyzed datasets, we used NeuN− SDs as the ground truth and asked which tool would lower the percentage of reads explained by ambient RNA markers to the levels observed in NeuN− SDs in glial cell types. We applied SoupX (Young and Behjati, 2020), DecontX (Yang et al., 2020), and CellBender (Fleming et al., 2019), with default parameters on each dataset. Overall, we observed a lower percentage of ambient RNA in NeuN− SDs compared with other datasets where no removal was performed (Figure S4). Among the three ambient RNA removal tools, CellBender was consistently better at reducing the ambient RNA contamination levels across the datasets (Figures S4A–S4D). Comparison of all NeuN− SDs and ambient RNA removal tools in glial cell types showed that there was no significant difference between the NeuN− SDs and CellBender results in terms of the percentage of reads explained by ambient RNA markers (Figure S4E). We highlight this in OLs from SD1 that show low levels of *SYT1*, *CSMD1*, and *KCNIP4* after CellBender, similar to NeuN− SDs, while DecontX- and SoupX-treated datasets display substantial levels of contamination (Figure S4F). Together, these results show that CellBender performs better than DecontX and SoupX to remove neuronal ambient RNA contamination from glial nuclei.

We next asked whether CellBender could fully remove ambient RNA contamination. For each dataset, we calculated the enrichments between ambient RNA markers and genes depleted in NeuN− SD1 (chosen as these genes have the lowest ambient RNA percentage in glial cell types among all NeuN− SDs) before and after applying CellBender (Figures S4A–S4D). Focusing on the OLs (from SD1), we found that CellBender substantially reduced ambient RNA contamination (Figure 3C). However, enrichment was still significant, indicating that ambient RNA contamination was not fully removed. To investigate this, we next subclustered OLs and found that markers of a small subcluster were highly enriched in ambient RNAs (Figures 3D and 3E). Removing this subcluster fully removed detectable ambient RNA contamination from OLs (Figures 3C and 3F) and increased correlation with NeuN− sorted OLs (Figure 3G). We then applied

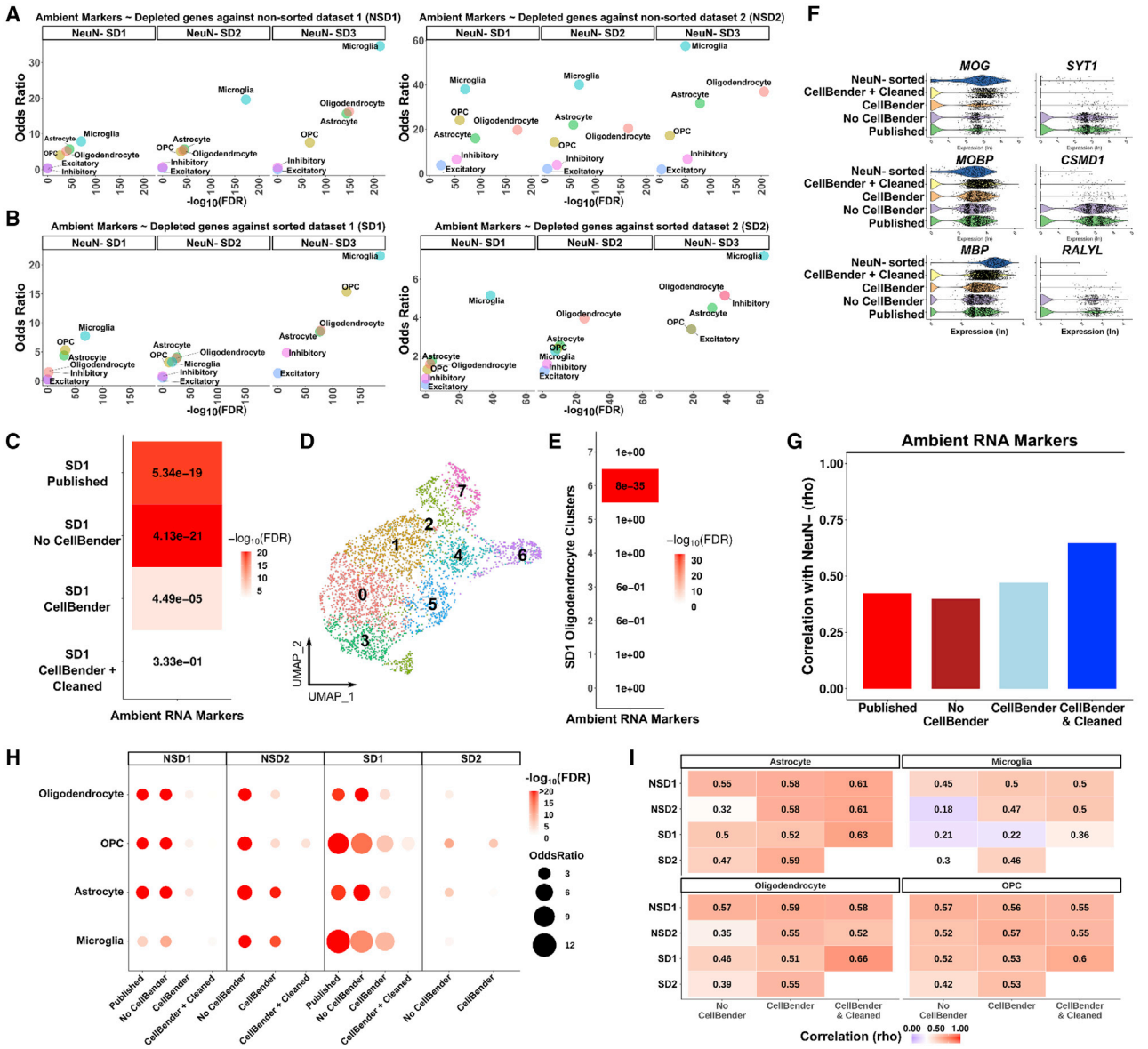


Figure 3. Ambient RNAs contaminate glia expression profiles

(A and B) Dot plots using odds ratio and FDR adjusted p values as measurements of ambient RNA enrichment of genes depleted in NeuN− sorted datasets compared with other datasets that did not perform NeuN− sorting; comparisons include: (A) between NSDs and NeuN− SDs; and (B) between SDs and NeuN− SDs (per major cell type using a Fisher's exact test).

(C) The same enrichment as in (A and B) after each analysis (in y axis as rows) performed in oligodendrocytes from the SD1 dataset. Numbers: FDR value; colors scale: $-\log_{10}$(FDR).

(D) UMAP plot of SD1 oligodendrocytes after CellBender.

(E) Heatmap of enrichment between oligodendrocyte cluster markers and ambient RNA markers.

(F) Violin plots of gene expression (log transformed) in oligodendrocytes after each analysis. Left column: oligodendrocyte markers; right column: ambient RNA markers. NeuN− sorted, NeuN− SD1.

(G) Spearman rank correlations of all genes between SD1 oligodendrocytes and NeuN− sorted oligodendrocytes after each analysis (x axis).

(H) The same enrichment as in (C) performed in all datasets and glial cell types after each analysis.

(I) Spearman rank correlations of all genes with the NeuN− sorted dataset. Correlations were performed per cell type per dataset (y axis) after each analysis (x axis). The numbers and color of the heatmaps indicate the correlation coefficient.

See also Figures S3–S9 and Table S4.

this procedure to each glial cell type per dataset and found that there was little to no contamination after CellBender and additional subcluster cleaning (Figure 3H). The removed subclusters had a consistently lower intronic read ratio in datasets that did not undergo nuclei-sorting, in line with the expectation that ambient RNA contamination contains non-nuclear reads unless nuclei are physically sorted (Figure S5A). Indeed, nuclei-SDs contained similar intronic read ratios between removed subclusters and other nuclei. (Figure S5B). We note that SD2 contained fewer nuclei compared with the other datasets and did not demonstrate robust subclusters; thus, we omitted subcluster cleaning for SD2.

Although neuronal nuclei are also expected to be contaminated with ambient RNAs, NeuN-based sorting is not helpful in revealing ambient RNA contamination in neuronal nuclei, as ambient RNAs are dominated by neuronal signatures. To assess the levels of ambient RNA contamination in neuronal nuclei, we leveraged the lower intronic read ratio of relatively more contaminated nuclei in the NSDs (Figure S5A). To reveal this association for all nuclei, we calculated a non-nuclear ambient RNA percentage and assessed its association with a non-intronic read ratio. Both measures are expected to be higher in ambient RNA-contaminated nuclei. Indeed, we observed high correlations between the two metrics for all cell types (Figures S6A–S6C and S6E). In line with the previous results, ambient RNA contamination was decreased by CellBender and further removed by subcluster cleaning in all glial cell types (Figures S6A and S6B). Strikingly, CellBender did not reduce ambient RNA contamination from the neurons (Figures S6C and S6E). As expected, non-nuclear ambient RNA markers *NRGN* and *CHN1* levels were higher in more contaminated neuronal nuclei, whereas nuclear-retained *MALAT1* levels were lower (Figures S6D and S6F). Contamination patterns were similar among the previously annotated neuronal subtypes of NSD1, indicating that ambient RNA contamination in neurons is a cell-type agnostic problem and, unlike glia, is not accounted for by CellBender (Figures S7A and S7B). Other ambient RNA contamination removal tools were also more effective in glia than neurons, indicating a general deficiency in the current methods to remove ambient RNA contamination from the dominant cell type in the tissue (Figures S7C and S7D).

To test the effect of ambient RNA removal on all genes, we then assessed the correlation of all expressed genes between a given dataset and a NeuN− SD. We found that ambient RNA removal consistently increased overall correlations, indicating that ambient RNA removal results in better reproducibility between datasets (Figure 3I). Neurons were also similarly correlated with the NeuN+ SD before and after CellBender (Figure S5C). These results indicate that ambient RNA contamination in glia can be effectively removed with CellBender and subcluster cleaning without undesired effects on the overall gene expression profile.

Ambient RNA contamination is also detected in a mouse brain snRNA-seq dataset

To assess whether ambient RNA contamination is similar in mouse cortical snRNA-seq data, we generated snRNA-seq datasets from the frontal cortex of four P56 (postnatal day 56) mice. Similar to human datasets, the intronic read ratio was less in cell barcodes with low UMI counts (Figure S8A) and ambient clusters were distributed bimodally with Low-Intron-CB and High-Intron-CB (Figures S8B and S8C). Low-Intron-CB markers were also enriched in non-nuclear ambient RNAs, whereas High-Intron-CB markers were enriched in nuclear ambient RNAs (Figure S8D). We similarly ran CellBender and performed subcluster cleaning on glial cell types. Both steps selectively removed the ambient RNA signature from all glial cell types (Figure S8E). Thus, we conclude that ambient RNA types and the contamination of glial cell types by neuronal ambient RNAs are not specific to human brain datasets.

In situ hybridization reveals no overlap of ambient RNA markers and glia

Because ambient RNA contamination arises from the RNAs released from dissociated cells and nuclei, we hypothesized that assays carried out using intact tissue should reveal considerably lower expression of neuronal ambient markers in glia. To test this, we used probes to an OL marker (*Mog*) and three ambient RNA markers (*Rbfox1*, *Snap25*, *Syt1*) and performed pairwise single molecule fluorescent *in situ* hybridization (smFISH) on cortical slices from adult mice. Indeed, we found almost no overlap between any of the three ambient RNA markers and *Mog* (Figures S9A–S9D; Table S5). In contrast, the snRNA-seq from the mouse frontal cortex indicated that >75% of nuclei contained reads from all three markers in the OLs (Figure S9E). This prevalent contamination was abolished after the ambient RNA removal process (CellBender + subcluster cleaning) (Figure S9E). These results provide further support for the prevalence of neuronal ambient RNA contamination in glial snRNA-seq nuclei.

Previously annotated immature oligodendrocytes are glia contaminated with ambient RNAs

Glia can express genes that are typically associated with neuronal function. For example, OPCs can make synapse-like contacts with axons and express glutamatergic receptors that bind to neurotransmitters secreted by neurons, affecting OL maturation in vitro (Fields, 2015; Luse and Korey, 1959; Wake et al., 2011). We also found that glutamatergic receptors functionally studied in OL maturation (e.g., *GRIA2*, *GRIA4*, and *GRM5* in OPCs [Fields, 2015; Kougioumtzidou et al., 2017; Wake et al., 2011]) remain present in our analysis after ambient RNA removal (Figure S10A). However, such ambiguity in cell-type expression patterns raises the possibility that neuronal ambient RNA contamination in glia might have been implicated with biological function in previous snRNA-seq studies. For example, the snRNA-seq study that generated SD1 identified "immature OLs," which were marked by greater expression of many neuronal genes, but many marker genes of this cell-type annotation were not independently validated (Lake et al., 2018). Based on our findings, we considered the alternative possibility that this excessive neuronal gene expression signature is ambient RNA contamination. In line with this interpretation, we found that ~80% (46 out of 58) of immature OL markers overlapped with the top 200 most abundant ambient RNA markers (Figure 4A). Using the gene-cell matrix from the original publication, we reconstructed the lineage trajectory between OPC and OL

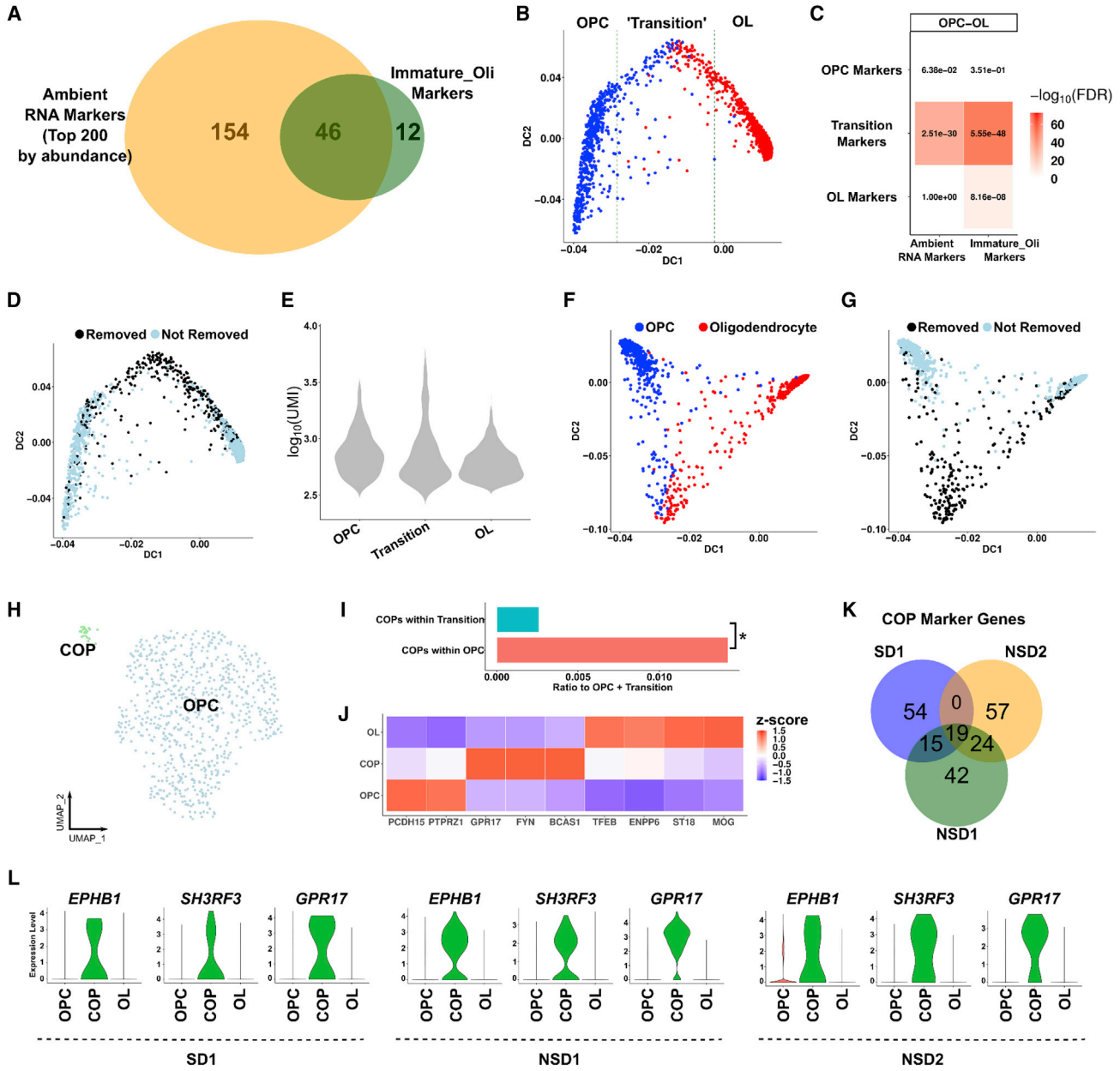


Figure 4. Ambient RNA contamination causes misinterpretation of transitioning oligodendrocytes in the human brain

(A) Overlap of the immature oligodendrocyte markers in SD1 and the top 200 most abundant ambient RNA markers.

(B) The oligodendrocyte lineage trajectory as reconstructed with *destiny*. The "transition" zone: the 400 cell barcodes around the middle cell barcode based on DC1.

(C) Heatmap enrichments between the trajectory zones (OPC, transition, and OL) and either ambient RNA or immature oligodendrocyte markers using a Fisher's exact test. Numbers: FDR; color scale: $-\log_{10}$(FDR).

(D) The same lineage trajectory as (B) with cell barcodes removed after subcluster cleaning highlighted.

(E) UMI counts of cell barcodes within the OPC, transition, or OL zones.

(F) The oligodendrocyte lineage trajectory after CellBender.

(G) The same lineage trajectory as (F) with cell barcodes removed after subcluster cleaning highlighted.

(H) UMAP of OPC subclustering. COP, committed OPCs.

(I) The ratio of COPs within OPCs or within the transitioning cells to the total number of OPCs and transitioning cells. Asterisk: p value < 0.05, chi-square test.

(J) Heatmap of oligodendrocyte lineage markers (*Z* scored across cell types per marker gene).

(K) Overlap of COP markers (compared with OPCs) across datasets. The top 100 markers were selected (FDR < 0.05).

(L) Violin plots of the expression levels of the top COP markers in three datasets.

See also Figures S10–S12 and Tables S5 and S6.

(Figure 4B). Cell barcodes between OPC and OL (“transitioning cells”) showed high enrichment for both immature OL and ambient RNA markers (Figures 4B and 4C), and these cell barcodes were removed during our subcluster cleaning procedure (Figure 4D). These cell barcodes displayed similar UMI counts compared with OPC and OL, indicating that they are also not glia-neuron doublets (Figure 4E). Because CellBender reduces ambient RNA contamination, we also assessed the OPC-OL trajectory after CellBender, which revealed a less continuous trajectory compared with the original dataset (Figure 4F). Similar to the original dataset, cell barcodes between OPC and OL were removed during subcluster cleaning (Figure 4G). We performed smFISH experiments to examine whether cells expressing the annotated immature OL markers (*GRIN2A* and *SYT1*) also express an OL lineage marker (*OLIG2*). We found essentially no overlap of expression of these genes in human cortical samples (Figures S9F and S9G; Table S5). In contrast, there was ~10% overlap of these markers in the SD1 snRNA-seq dataset (Figure S9H). Together, these results indicate that the OPC-OL pseudotime trajectory is driven by ambient RNA contamination rather than the biological differentiation of OLs.

We hypothesized that ambient RNA-contaminated nuclei can be detected as transitioning cells between any two glial cell types in pseudotime analysis, as all glial nuclei are expected to contain neuronal ambient RNA contamination in a brain gray matter preparation. Therefore, we generated a pseudotime analysis between OPC and AST and found similar “transitioning cells” (Figures S10B and S10D). Because OPCs can achieve multipotency under certain conditions (Chamling et al., 2021; Sim et al., 2011; van Bruggen et al., 2017), we could not exclude the possibility that this could be a real biological function (i.e., OPCs differentiating into AST). For a definitive answer, we tested two non-OL lineage cell types, AST and MIC, which also revealed similar “transitioning cells” that were highly enriched in both ambient RNA and immature OL markers and were effectively removed by our ambient RNA removal process (Figures S10E–S10G). Given that immature OLs also lack known markers of COPs or premyelinating OLs (Pre-OLs) (e.g., *BCAS1*, *ENPP6*, and *GPR17* [Hughes and Stockton, 2021]) (Figure S10H), our results indicate that nuclei previously annotated as immature OLs in several snRNA-seq studies are not transitioning cells but rather glia with a high contamination of neuronal ambient RNA.

Ambient RNA removal reveals rare cell type in adult human brain snRNA-seq datasets

Initial single-cell studies on the adolescent mouse OL lineage identified COPs and NFOLs (newly formed OLs) as transitioning OL cells (Marques et al., 2016). This work established marker genes, including *Gpr17*, which peaked in COPs, was reduced in NFOLs, and was absent in mature OLs (Marques et al., 2016). A study in human-induced pluripotent stem cell-derived OPC culture also showed that GPR17 regulated OL maturation in human cells (Merten et al., 2018). However, few single-cell RNA-seq studies in the adult human brain have identified these populations (Fernandes et al., 2021; Jäkel et al., 2019). Because these studies used different annotation labels and marker genes, it is also unclear whether transitioning OLs are consistent across human datasets. Robust annotation of these cells in human datasets is crucial to understand the role of the OL lineage in neurological diseases (Akay et al., 2021; Jäkel et al., 2019; Nagy et al., 2020; Phan et al., 2020).

To determine whether we can identify COPs after ambient RNA removal, we subclustered OPCs. *GPR17*+ COPs were detectable and clustered separately from OPCs in NSD1, NSD2, and SD1 (Figures 4H, S11A, and S11C). In SD2, plotting of COP markers in the Uniform Manifold Approximation and Projection (UMAP) space revealed a small population of nuclei with high expression of COP markers, although they did not cluster separately due to the low number of nuclei in this dataset (Figure S11E). Importantly, COPs were significantly depleted within the transition zone of the pseudotime plot (Figure 4I), further indicating that previous pseudotime analyses were driven by ambient RNA contamination rather than demonstrating biological underpinnings of OL maturation. To validate the identity of COPs, we then plotted genes known to be associated with COPs or Pre-OLs (*BCAS1* [Fard et al., 2017], *GPR17* [Chen et al., 2009], *FYN* [Sperber and McMorris, 2001]) as well as genes that are upregulated in Pre-OLs but are also expressed in OLs (*TFEB* [Sun et al., 2018], *ENPP6* [Xiao et al., 2016]). We found that *BCAS1*, *GPR17*, and *FYN* selectively marked COPs, and *TFEB* and *ENPP6* were upregulated in COPs compared with OPCs consistently across the datasets (Figures 4J, S11B, S11D, and S11E). Overall, 58 out of the 211 (27%) top markers of COPs were shared between at least two datasets (Figure 4K). To highlight previously undescribed markers for COPs in the adult human brain, we then found the most specific COP markers compared with both OPCs and OLs which—in addition to *GPR17*, *BCAS1*, and *FYN*—revealed *TNS3* (Marques et al., 2016), *SH3RF3*, *EPHB1*, *CRB1*, *SIRT2*, and *ARHGAP5* as additional potential markers for future studies of OL biology in the human brain (Figure 4L; Table S6). Given that we could detect similar COP populations in all datasets, we also re-analyzed a previous study that identified COPs in the adult human brain white matter (Jäkel et al., 2019). Surprisingly, in the original annotation, COP markers did not have a higher expression in COPs than OPCs (Figure S12A). Clustering OPCs and COPs revealed a subpopulation of nuclei that were very similar to COPs in other human datasets on account of their marker gene expression levels (“COPs-New”) (Figures S12B and S12C). To assess whether previously annotated COPs (“COPs-Old”) could be ambient RNA contamination, we also checked expression levels of neuronal genes. This revealed high expression of both ambient and non-ambient neuronal genes, indicating “COPs_Old” might be OL-neuron doublets rather than ambient RNA contamination (Figure S12C). Indeed, COPs-Old displayed similar UMI count levels to neuronal cell types, in contrast to ambient RNA driven clusters which contained lower UMI count levels (Figures 1A, 4E, and S12D). These results provide further evidence of the extreme rarity of COPs in human brain datasets, which can be masked by technical artifacts.

Stepwise guideline for detection and removal of ambient RNAs

Our results show that a combination of existing tools and careful analysis can remove ambient RNA contamination and improve the biological relevance of results. To illustrate our approach in a more direct way, we present a stepwise guideline that outlines

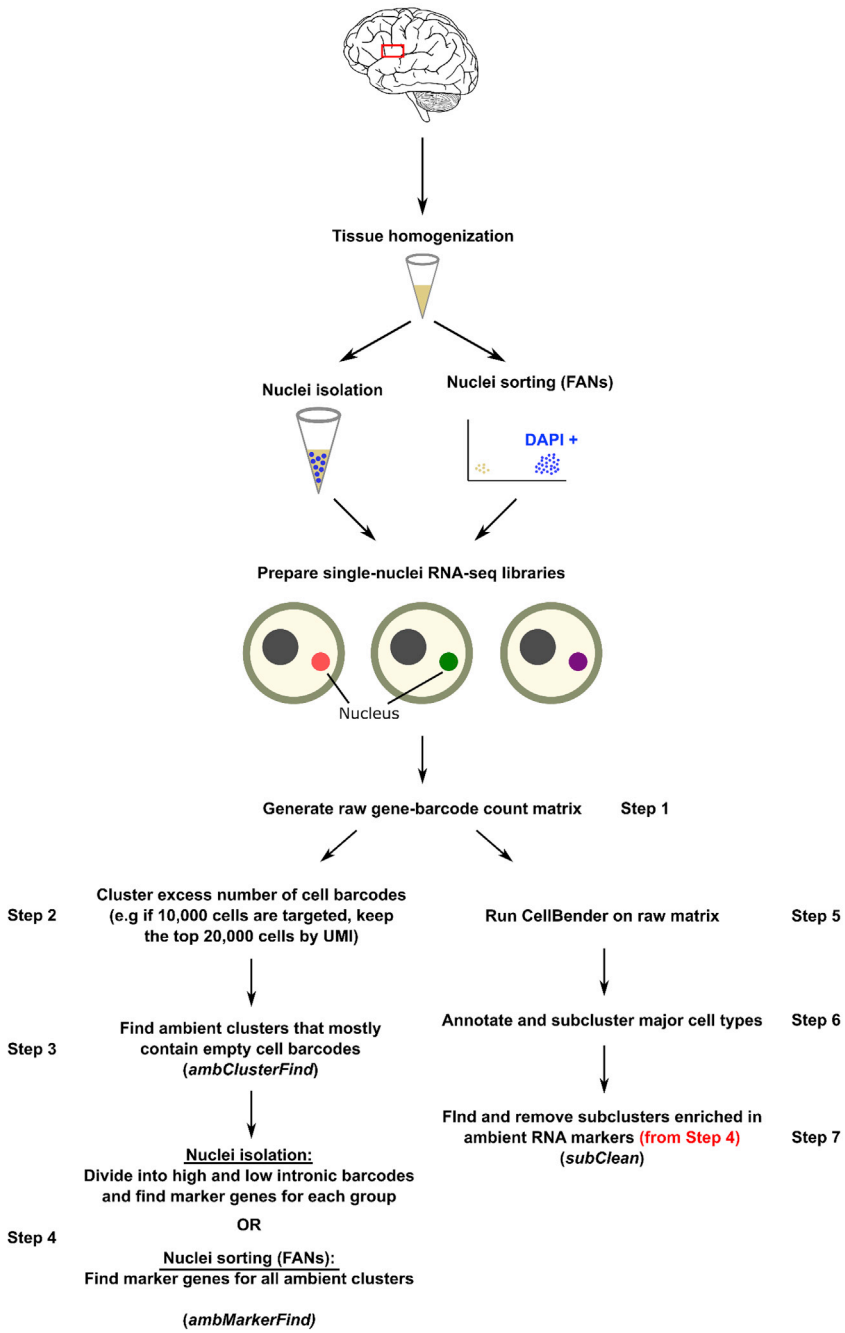


Figure 5. Stepwise guidelines of ambient RNA marker detection and ambient RNA removal

Steps 1–4 describe how to identify ambient RNA markers in the given dataset. Steps 5–7 describe how to use this information to further remove ambient RNA-contaminated cell barcodes after a formal ambient RNA contamination removal tool such as CellBender is applied. Left, for non-sorted datasets; right, for nuclei-sorted datasets.

the major steps important in our analysis (Figure 5). Although ambient RNA removal tools aim to be a one-step solution for this problem, we advise researchers to identify ambient RNA populations and their markers genes in their own dataset, which is achievable using common methods (Figure 5, steps 1–4). This can then be used to assess whether a specific cell population is marked by high ambient RNA contamination, which may not have been removed or cleaned of ambient RNAs by the specialized tools (e.g., CellBender, Figure 5, steps 5–7). Taken together, we show pervasive contamination of glia by neuronal ambient RNAs and successfully remove them using available methods, which reveals the underappreciated biology of transitioning OLs in the adult human brain. We also provide a stepwise guideline outlining our integrated approach to tackle ambient RNA contamination in single-nuclei datasets from brain tissue.

DISCUSSION

Here, we provide an in-depth examination of ambient RNAs in brain snRNA-seq datasets. We identify nuclear and non-nuclear ambient RNAs with different gene signatures and find that previously annotated neuronal cell types have a high contamination of ambient RNAs (Ruzicka et al., 2020; Velmeshev et al., 2019). We then show that the high prevalence of neuronal reads in ambient RNAs contaminate glia but can be effectively removed using CellBender and additional subcluster cleaning. These results are not unique to the human brain and are reproducible in mouse cortical snRNA-seq data. We also show that immature OLs previously identified in snRNA-seq datasets are artifacts of neuronal ambient RNA contamination. After ambient RNA removal, we can identify populations of COPs in all human brain snRNA-seq datasets and highlight both known and previously undescribed markers of COPs. Finally, we provide a stepwise guideline of ambient RNA marker identification and removal.

Our findings suggest that single-nuclei isolation does not entirely remove non-nuclear reads. The presence of cell barcodes with high proportions of non-nuclear reads indicates that cytoplasmic mature RNAs also contribute to contamination during nuclei isolation. We found that marker genes of non-nuclear reads significantly overlap with mRNAs that localize to synaptosomes (Figure S2) and that the non-nuclear ambient RNAs are largely abolished when the intact nuclei are physically sorted by FANS (Figures 2B and S1B). Although these results indicate that non-nuclear reads are likely derived from all cell types, it is

also possible that mature mRNAs can be carried over into droplets by the endoplasmic reticulum that is still attached to the nucleus after isolation or sorting. Therefore, some non-nuclear reads may be derived from the same cell as the captured nucleus.

We leveraged the intronic read ratio difference between the empty and non-empty droplets to reveal that previously annotated cell clusters (Neu-NRGNs) contain high levels of non-nuclear ambient RNA contamination and are likely empty droplets (Figure 1C). However, non-nuclear contamination measured by intronic read ratio is not sufficient to identify all healthy nuclei/cells. For example, a recent study highlighted the distinction of damaged cells and empty droplets that only contain ambient RNA (Muskovic and Powell, 2021). The authors noted that damaged cells contain a similar intronic read ratio (i.e., nuclear fraction) to real cells, but they display lower UMI counts compared with other cells with similar annotation. Similarly, we find that the Neu-mat cluster has lower UMI counts compared with other neurons despite having a similar intronic read ratio, indicating that this cluster likely contains damaged nuclei (Figures 1A and 1C). Additionally, we observe that while empty droplets have lower intronic read ratios, these ratios are still substantially higher than zero, indicating that nuclear reads also contribute to ambient RNAs (Figures 1C and 2A). Finally, intronic read ratio is not an indicator of empty droplets in datasets that underwent nuclei sorting by flow cytometry, as this procedure only removes non-nuclear ambient RNAs (Figure 2B). In nuclei-SDs (e.g., SDs in this study), empty droplets can be better identified by assessing a given cluster's enrichment for the ambient RNA markers (Figure 2E). We offer several functions to find ambient RNA markers for this purpose (Figure 5, steps 1–4).

In addition to empty droplets, ambient RNAs can contaminate all non-empty droplets. We focused on contamination in glial nuclei as they contain fewer transcripts than other cell types in the brain. We found that ambient RNA markers were underrepresented in the glial nuclei from studies that physically separated neurons and glia, indicating that some reads mapping to neuronal genes in datasets without neuron-glia separation are not representative of neuronal endogenous expression (Figures 3A and 3B). To remove the neuronal ambient RNA contamination in glia, we utilized CellBender (Fleming et al., 2019) and subsequent detection of subclusters with ambient RNA contamination. Together, these methods removed neuronal ambient RNA contamination from glial nuclei and improved correspondence with NeuN− SDs (Figures 3C–3I). Based on these results, we recommend two approaches for cortical snRNA-seq experiments: (1) physical separation of glia from nuclei (e.g., by FANS) or (2) *in silico* cleaning of neuronal ambient RNA contamination. Our approach for the *in silico* cleaning involves two steps: using a formal ambient RNA removal tool (CellBender) and subsequent removal of contaminated subclusters that are not successfully cleaned of ambient RNAs after CellBender (Figures 3C–3E). We show that failure to remove ambient RNA contamination can have important consequences, such as the misannotation of contaminated glial nuclei as immature OLs (Lake et al., 2018). Importantly, CellBender alone did not remove all ambient RNA contamination, and the remaining contaminated nuclei were positioned between OPCs and OLs in the pseudotime trajectory (Figures 4B–4G). We thus recommend utilizing both CellBender and subsequent subcluster cleaning to account for ambient RNA contamination. We provide a stepwise guideline for the *in silico* approach we have taken to remove ambient RNA contamination from glial cell types (Figure 5).

Neuronal reads are abundant within the ambient RNA population, making ambient RNA contamination in glial cell types distinct from the endogenous gene expression of glial nuclei. In contrast, ambient RNA contamination in neurons is difficult to separate from the endogenous neuronal gene expression, and cell barcodes with a higher percentage of ambient RNA markers may be biologically relevant. As an unbiased method, we used the positive correlation of non-intronic read ratio and ambient RNA percentage across cell barcodes as a measure of contamination in NSDs. This revealed that neither CellBender nor other tools (DecontX and SoupX) could substantially reduce ambient RNA contamination in neuronal cell types (Figures S6C–S6F and S7). Although the cell barcodes with high contamination can be manually removed, determining a threshold for removal would be arbitrary and could reduce the number of nuclei retained for analysis. Currently, we suggest caution in interpreting "novel" neuronal cell types and cell states even if the common ambient RNA removal tools are applied. Our study shows that ambient RNA contamination in neurons can be assessed by using metrics such as intronic read ratio (if the dataset has non-nuclear ambient RNAs) and the percentage of ambient RNA markers identified from the same dataset or similarly prepared datasets from similar tissues.

We showed that analysis of nuclei from the OL lineage after ambient RNA removal revealed COPs in all independent datasets. Although COPs have been identified before (Marques et al., 2016; Perlman et al., 2020), many studies did not annotate them (Nagy et al., 2020; Ruzicka et al., 2020; Sadick et al., 2022; Tran et al., 2021; Velmeshev et al., 2019). This could be hindered by both ambient RNA contamination and the rarity of COPs in the adult brain. COPs are ~0.04% of cells in the adult brain (NSD2 samples from 30 to 80 years old) and ~0.3% of cells in the adolescent brain (NSD1 samples from 4 to 22 years old). We also showed that previously annotated COPs in adult human brain white matter are likely OL-neuron doublets and that real COPs are detectable and similarly rare (~0.1% of all cells) (Figure S11). In line with this, live cell imaging in the mouse brain showed that ~80% of transitioning OLs rapidly undergo cell death, which should result in a transient and rare cell population (Hughes et al., 2018). A carbon dating study of human genomic DNA in OLs also showed low levels of oligodendrogenesis in adulthood, which further supports the rarity of transient cells in the adult human brain (Yeung et al., 2014). Despite being rare, COPs are critical to examine because OL maturation is altered in both neurological diseases (de Faria et al., 2021; Phan et al., 2020) and human evolution (Miller et al., 2012; Zhu et al., 2018). Thus, ambient RNA removal is important for accurate analysis of underrepresented cell types. Another recent study also uncovered that glial cell types respond to enzymatic dissociation during single-cell and single-nucleus library preparation and confound the transcriptomic profile (Marsh et al., 2022). Here, we show that all glial cell types are also contaminated

with neuronal ambient RNA transcripts, causing misinterpretation of glial single-cell analysis. Together, these results indicate that both data generation and data analysis of glial cell types should be revisited and updated.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human cortical tissue
  - Mouse cortical tissue
- METHOD DETAILS
  - Preprocessing and count matrix generation
  - Single-nuclei library preparation
  - Ambient cluster analysis
  - Comparison with NeuN- datasets
  - Ambient RNA removal with CellBender, DecontX and SoupX
  - Subcluster cleaning of glia after CellBender
  - Assessment of ambient RNA contamination signatures in glia
  - *In-situ* hybridization and image quantification
  - Pseudotime analysis
  - OPC subcluster analysis
  - Identification of COP Marker Genes
  - Other enrichments
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2022.09.010.

ACKNOWLEDGMENTS

The authors thank Dr. Dmitry Velmeshev, Dr. Arnold Kriegstein, Dr. Tao Wang, and Dr. Lu Sun for their critical comments on the manuscript. We also thank Dr. Shin Yamazaki and the UTSW Neuroscience Microscopy Facility for their help with imaging and Dr. Shane A. Liddelow and Michael O'Dea for additional information about their sorted dataset. The authors thank the NIH NeuroBioBank for providing human brain tissue. G.K. is a Jon Heighten Scholar in Autism Research and Townsend Distinguished Chair in Research on Autism Spectrum Disorders at UT Southwestern. E.C. is a Neural Scientist Training Program fellow in the Peter O'Donnell Brain Institute at UT Southwestern. This work was partially supported by the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition Scholar award, NHGRI (HG011641), NINDS (NS115821), and NIMH (MH126481, MH103517) to G.K., and an American Heart Association Postdoctoral Fellowship (915654) to Y.L.

AUTHOR CONTRIBUTIONS

E.C. and G.K. conceptualized the study. Y.L. collected snRNA-seq data and performed smFISH. E.C. performed all analyses. Y.L. edited the manuscript. E.C. and G.K. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 23, 2022
Revised: April 20, 2022
Accepted: September 8, 2022
Published: October 13, 2022

REFERENCES

Akay, L.A., Effenberger, A.H., and Tsai, L.H. (2021). Cell of all trades: oligodendrocyte precursor cells in synaptic, vascular, and immune function. Genes Dev. *35*, 180–198. https://doi.org/10.1101/gad.344218.120.

Angerer, P., Haghverdi, L., Büttner, M., Theis, F.J., Marr, C., and Buettner, F. (2016). destiny: diffusion maps for large-scale single-cell data in R. Bioinformatics *32*, 1241–1243. https://doi.org/10.1093/bioinformatics/btv715.

Ayhan, F., Douglas, C., Lega, B.C., and Konopka, G. (2021). Nuclei isolation from surgically resected human hippocampus. Star Protoc. *2*, 100844. https://doi.org/10.1016/j.xpro.2021.100844.

Bakken, T.E., Jorstad, N.L., Hu, Q., Lake, B.B., Tian, W., Kalmbach, B.E., Crow, M., Hodge, R.D., Krienen, F.M., Sorensen, S.A., et al. (2021). Comparative cellular analysis of motor cortex in human, marmoset and mouse. Nature *598*, 111–119. https://doi.org/10.1038/s41586-021-03465-8.

Bernard, D., Prasanth, K.V., Tripathi, V., Colasse, S., Nakamura, T., Xuan, Z., Zhang, M.Q., Sedel, F., Jourdren, L., Coulpier, F., et al. (2010). A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. EMBO J. *29*, 3082–3093. https://doi.org/10.1038/emboj.2010.199.

Chamling, X., Kallman, A., Fang, W., Berlinicke, C.A., Mertz, J.L., Devkota, P., Pantoja, I.E.M., Smith, M.D., Ji, Z., Chang, C., et al. (2021). Single-cell transcriptomic reveals molecular diversity and developmental heterogeneity of human stem cell-derived oligodendrocyte lineage cells. Nat. Commun. *12*, 652. https://doi.org/10.1038/s41467-021-20892-3.

Chen, Y., Lun, A.T., and Smyth, G.K. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000Res *5*, 1438. https://doi.org/10.12688/f1000research.8987.2.

Chen, Y., Wu, H., Wang, S., Koito, H., Li, J., Ye, F., Hoang, J., Escobar, S.S., Gow, A., Arnett, H.A., et al. (2009). The oligodendrocyte-specific G protein-coupled receptor GPR17 is a cell-intrinsic timer of myelination. Nat. Neurosci. *12*, 1398–1406. https://doi.org/10.1038/nn.2410.

de Faria, O., Jr., Pivonkova, H., Varga, B., Timmler, S., Evans, K.A., and Káradóttir, R.T. (2021). Periods of synchronized myelin changes shape brain function and plasticity. Nat. Neurosci. *24*, 1508–1521. https://doi.org/10.1038/s41593-021-00917-2.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Fard, M.K., van der Meer, F., Sánchez, P., Cantuti-Castelvetri, L., Mandad, S., Jäkel, S., Fornasiero, E.F., Schmitt, S., Ehrlich, M., Starost, L., et al. (2017). BCAS1 expression defines a population of early myelinating oligodendrocytes in multiple sclerosis lesions. Sci. Transl. Med. *9*. https://doi.org/10.1126/scitranslmed.aam7816.

Fernandes, M.G.F., Luo, J.X.X., Cui, Q.L., Perlman, K., Pernin, F., Yaqubi, M., Hall, J.A., Dudley, R., Srour, M., Couturier, C.P., et al. (2021). Age-related injury responses of human oligodendrocytes to metabolic insults: link to BCL-2 and autophagy pathways. Commun. Biol. *4*, 20. https://doi.org/10.1038/s42003-020-01557-1.

Fields, R.D. (2015). A new mechanism of nervous system plasticity: activity-dependent myelination. Nat. Rev. Neurosci. *16*, 756–767. https://doi.org/10.1038/nrn4023.

Fleming, S.J., Marioni, J.C., and Babadi, M. (2019). CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets. Preprint at bioRxiv. https://doi.org/10.1101/791699.

Guo, C.J., Xu, G., and Chen, L.L. (2020). Mechanisms of long noncoding RNA nuclear retention. Trends Biochem. Sci. *45*, 947–960. https://doi.org/10.1016/j.tibs.2020.07.001.

Hafner, A.S., Donlin-Asp, P.G., Leitch, B., Herzog, E., and Schuman, E.M. (2019). Local protein synthesis is a ubiquitous feature of neuronal pre- and postsynaptic compartments. Science *364*, eaau3644. https://doi.org/10.1126/science.aau3644.

Heiser, C.N., Wang, V.M., Chen, B., Hughey, J.J., and Lau, K.S. (2021). Automated quality control and cell identification of droplet-based single-cell data using dropkick. Genome Res. *31*, 1742–1752. https://doi.org/10.1101/gr.271908.120.

Hodge, R.D., Bakken, T.E., Miller, J.A., Smith, K.A., Barkan, E.R., Graybuck, L.T., Close, J.L., Long, B., Johansen, N., Penn, O., et al. (2019). Conserved cell types with divergent features in human versus mouse cortex. Nature *573*, 61–68. https://doi.org/10.1038/s41586-019-1506-7.

Hughes, E.G., Orthmann-Murphy, J.L., Langseth, A.J., and Bergles, D.E. (2018). Myelin remodeling through experience-dependent oligodendrogenesis in the adult somatosensory cortex. Nat. Neurosci. *21*, 696–706. https://doi.org/10.1038/s41593-018-0121-5.

Hughes, E.G., and Stockton, M.E. (2021). Premyelinating oligodendrocytes: mechanisms underlying cell survival and integration. Front. Cell Dev. Biol. *9*, 714169. https://doi.org/10.3389/fcell.2021.714169.

Jäkel, S., Agirre, E., Mendanha Falcão, A., van Bruggen, D., Lee, K.W., Knuesel, I., Malhotra, D., Ffrench-Constant, C., Williams, A., and Castelo-Branco, G. (2019). Altered human oligodendrocyte heterogeneity in multiple sclerosis. Nature *566*, 543–547. https://doi.org/10.1038/s41586-019-0903-2.

Koopmans, F., van Nierop, P., Andres-Alonso, M., Byrnes, A., Cijsouw, T., Coba, M.P., Cornelisse, L.N., Farrell, R.J., Goldschmidt, H.L., Howrigan, D.P., et al. (2019). Syngo: an evidence-based, expert-curated knowledge base for the synapse. Neuron *103*, 217–234.e4. https://doi.org/10.1016/j.neuron.2019.05.002.

Kougioumtzidou, E., Shimizu, T., Hamilton, N.B., Tohyama, K., Sprengel, R., Monyer, H., Attwell, D., and Richardson, W.D. (2017). Signalling through AMPA receptors on oligodendrocyte precursors promotes myelination by enhancing oligodendrocyte survival. eLife *6*, e2808. https://doi.org/10.7554/eLife.28080.

Lake, B.B., Chen, S., Sos, B.C., Fan, J., Kaeser, G.E., Yung, Y.C., Duong, T.E., Gao, D., Chun, J., Kharchenko, P.V., and Zhang, K. (2018). Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat. Biotechnol. *36*, 70–80. https://doi.org/10.1038/nbt.4038.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.

Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. *15*, e8746. https://doi.org/10.15252/msb.20188746.

Lun, A.T., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with bioconductor. F1000Res *5*, 2122. https://doi.org/10.12688/f1000research.9501.2.

Lun, A.T.L., Riesenfeld, S., Andrews, T., Dao, T.P., Gomes, T., participants in the 1st Human Cell Atlas Jamboree, and Marioni, J.C. (2019). EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol. *20*, 63. https://doi.org/10.1186/s13059-019-1662-y.

Luse, S., and Korey, S.J.H.-H. (1959). The Biology of Myelin (New York: Hoefer-Harper), pp. 59–81.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002.

Marques, S., Zeisel, A., Codeluppi, S., van Bruggen, D., Mendanha Falcão, A., Xiao, L., Li, H., Häring, M., Hochgerner, H., Romanov, R.A., et al. (2016). Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. Science *352*, 1326–1329. https://doi.org/10.1126/science.aaf6463.

Marsh, S.E., Walker, A.J., Kamath, T., Dissing-Olesen, L., Hammond, T.R., de Soysa, T.Y., Young, A.M.H., Murphy, S., Abdulraouf, A., Nadaf, N., et al. (2022). Dissection of artifactual and confounding glial signatures by single-cell sequencing of mouse and human brain. Nat. Neurosci. *25*, 306–316. https://doi.org/10.1038/s41593-022-01022-8.

Merten, N., Fischer, J., Simon, K., Zhang, L., Schröder, R., Peters, L., Letombe, A.G., Hennen, S., Schrage, R., Bödefeld, T., et al. (2018). Repurposing HAMI3379 to block GPR17 and promote rodent and human oligodendrocyte differentiation. Cell Chem. Biol. *25*, 775–786.e5. https://doi.org/10.1016/j.chembiol.2018.03.012.

Miller, D.J., Duka, T., Stimpson, C.D., Schapiro, S.J., Baze, W.B., McArthur, M.J., Fobbs, A.J., Sousa, A.M., Sestan, N., Wildman, D.E., et al. (2012). Prolonged myelination in human neocortical evolution. Proc. Natl. Acad. Sci. USA *109*, 16480–16485. https://doi.org/10.1073/pnas.1117943109.

Muskovic, W., and Powell, J.E. (2021). DropletQC: improved identification of empty droplets and damaged cells in single-cell RNA-seq data. Genome Biol. *22*, 329. https://doi.org/10.1186/s13059-021-02547-0.

Nagy, C., Maitra, M., Tanti, A., Suderman, M., Théroux, J.F., Davoli, M.A., Perlman, K., Yerko, V., Wang, Y.C., Tripathy, S.J., et al. (2020). Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. Nat. Neurosci. *23*, 771–781. https://doi.org/10.1038/s41593-020-0621-y.

Perlman, K., Couturier, C.P., Yaqubi, M., Tanti, A., Cui, Q.L., Pernin, F., Stratton, J.A., Ragoussis, J., Healy, L., Petrecca, K., et al. (2020). Developmental trajectory of oligodendrocyte progenitor cells in the human brain revealed by single cell RNA sequencing. Glia *68*, 1291–1303. https://doi.org/10.1002/glia.23777.

Phan, B.N., Bohlen, J.F., Davis, B.A., Ye, Z., Chen, H.Y., Mayfield, B., Sripathy, S.R., Cerceo Page, S., Campbell, M.N., Smith, H.L., et al. (2020). A myelin-related transcriptomic profile is shared by Pitt-Hopkins syndrome models and human autism spectrum disorder. Nat. Neurosci. *23*, 375–385. https://doi.org/10.1038/s41593-019-0578-x.

Ruzicka, W.B., Mohammadi, S., Davila-Velderrain, J., Subburaju, S., Tso, D.R., Hourihan, M., and Kellis, M. (2020). Single-cell dissection of schizophrenia reveals neurodevelopmental-synaptic axis and transcriptional resilience. Preprint at medRxiv. https://doi.org/10.1101/2020.11.06.20225342.

Sadick, J.S., O'Dea, M.R., Hasel, P., Dykstra, T., Faustin, A., and Liddelow, S.A. (2022). Astrocytes and oligodendrocytes undergo subtype-specific transcriptional changes in Alzheimer's disease. Neuron *110*, 1788–1805.e10. https://doi.org/10.1016/j.neuron.2022.03.008.

Shen, L. (2020). GeneOverlap: An R package to test and visualize gene overlaps. http://shenlab-sinai.github.io/shenlab-sinai/.

Sim, F.J., McClain, C.R., Schanz, S.J., Protack, T.L., Windrem, M.S., and Goldman, S.A. (2011). CD140a identifies a population of highly myelinogenic, migration-competent and efficiently engrafting human oligodendrocyte progenitor cells. Nat. Biotechnol. *29*, 934–941. https://doi.org/10.1038/nbt.1972.

Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. Genome Res. *27*, 491–499. https://doi.org/10.1101/gr.209601.116.

Sperber, B.R., and McMorris, F.A. (2001). Fyn tyrosine kinase regulates oligodendroglial cell development but is not required for morphological differentiation of oligodendrocytes. J. Neurosci. Res. *63*, 303–312. https://doi.org/10.1002/1097-4547(20010215)63:4<303::AID-JNR1024>3.0.CO;2-A.

Squair, J.W., Gautier, M., Kathe, C., Anderson, M.A., James, N.D., Hutson, T.H., Hudelle, R., Qaiser, T., Matson, K.J.E., Barraud, Q., et al. (2021). Confronting false discoveries in single-cell differential expression. Nat. Commun. *12*, 5692. https://doi.org/10.1038/s41467-021-25960-2.

Srinivasan, A. (2016). gread: fast reading and processing of common gene annotation and next generation sequencing format files. https://rdrr.io/github/asrinivasan-oa/gread/.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell *177*, 1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.

Sun, L.O., Mulinyawe, S.B., Collins, H.Y., Ibrahim, A., Li, Q., Simon, D.J., Tessier-Lavigne, M., and Barres, B.A. (2018). Spatiotemporal control of CNS myelination by oligodendrocyte programmed cell death through the TFEB-PUMA axis. Cell *175*, 1811–1826.e21. https://doi.org/10.1016/j.cell.2018.10.044.

Thrupp, N., Sala Frigerio, C., Wolfs, L., Skene, N.G., Fattorelli, N., Poovathingal, S., Fourne, Y., Matthews, P.M., Theys, T., Mancuso, R., et al. (2020). Single-nucleus RNA-seq is not suitable for detection of microglial activation genes in humans. Cell Rep. *32*, 108189. https://doi.org/10.1016/j.celrep.2020.108189.

Tran, M.N., Maynard, K.R., Spangler, A., Huuki, L.A., Montgomery, K.D., Sadashivaiah, V., Tippani, M., Barry, B.K., Hancock, D.B., Hicks, S.C., et al. (2021). Single-nucleus transcriptome analysis reveals cell-type-specific molecular signatures across reward circuitry in the human brain. Neuron *109*, 3088–3103.e5. https://doi.org/10.1016/j.neuron.2021.09.001.

van Bruggen, D., Agirre, E., and Castelo-Branco, G. (2017). Single-cell transcriptomic analysis of oligodendrocyte lineage cells. Curr. Opin. Neurobiol. *47*, 168–175. https://doi.org/10.1016/j.conb.2017.10.005.

Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., Bhaduri, A., Goyal, N., Rowitch, D.H., and Kriegstein, A.R. (2019). Single-cell genomics identifies cell type-specific molecular changes in autism. Science *364*, 685–689. https://doi.org/10.1126/science.aav8130.

Wake, H., Lee, P.R., and Fields, R.D. (2011). Control of local protein synthesis and initial events in myelination by action potentials. Science *333*, 1647–1651. https://doi.org/10.1126/science.1206998.

Xiao, L., Ohayon, D., McKenzie, I.A., Sinclair-Wilson, A., Wright, J.L., Fudge, A.D., Emery, B., Li, H., and Richardson, W.D. (2016). Rapid production of new oligodendrocytes is required in the earliest stages of motor-skill learning. Nat. Neurosci. *19*, 1210–1217. https://doi.org/10.1038/nn.4351.

Yang, S., Corbett, S.E., Koga, Y., Wang, Z., Johnson, W.E., Yajima, M., and Campbell, J.D. (2020). Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol. *21*, 57. https://doi.org/10.1186/s13059-020-1950-6.

Yeung, M.S., Zdunek, S., Bergmann, O., Bernard, S., Salehpour, M., Alkass, K., Perl, S., Tisdale, J., Possnert, G., Brundin, L., et al. (2014). Dynamics of oligodendrocyte generation and myelination in the human brain. Cell *159*, 766–774. https://doi.org/10.1016/j.cell.2014.10.011.

Young, M.D., and Behjati, S. (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. GigaScience *9*, giaa151. https://doi.org/10.1093/gigascience/giaa151.

Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. Omics *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.

Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049. https://doi.org/10.1038/ncomms14049.

Zhu, Y., Sousa, A.M.M., Gao, T., Skarica, M., Li, M., Santpere, G., Esteller-Cucala, P., Juan, D., Ferrández-Peral, L., Gulden, F.O., et al. (2018). Spatiotemporal transcriptomic divergence across human and macaque brain development. Science *362*, eaat8077. https://doi.org/10.1126/science.aat8077.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Adult mouse (P56) frontal cortex specimens	Table S1	N/A
Adult human posterior cingulate cortex specimens	Table S1	N/A
Deposited Data		
Mouse snRNA-seq data	This paper	GEO: GSE198640
Human nuclei-sorted snRNA-seq data (SD1)	Lake et al., 2018	GEO: GSE97930
Human nuclei-sorted snRNA-seq data (SD2)	Tran et al., 2021	https://research.libd.org/globus (endpoint: jhpce#tran2021)
Human non-sorted snRNA-seq data (NSD1)	Velmeshev et al., 2019	BioProject: PRJNA434002
Human non-sorted snRNA-seq data (NSD2)	This paper	GEO: GSE198951
NeuN- sorted dataset 1	Hodge et al., 2019	https://portal.brain-map.org/atlases-and-data/rnaseq/human-mtg-smart-seq
NeuN- sorted dataset 2	Bakken et al., 2021	https://assets.nemoarchive.org/dat-ek5dbmu
NeuN- sorted dataset 3	Sadick et al., 2022	GEO: GSE167494
Human cortical oligodendrocyte data	Jäkel et al., 2019	GEO: GSE118257
Software and Algorithms		
CellRanger v.3.0.2	10x Genomics	https://www.10xgenomics.com/products/single-cell-gene-expression/
R version 4.1.2	The R Project	https://www.r-project.org/
Seurat_3.0.1	Stuart et al., 2019	https://github.com/satijalab/seurat
Scran_1.18.7	Chen et al., 2016	https://bioconductor.org/packages/release/bioc/html/scran.html
GeneOverlap_1.30.0	Shen, 2020	https://bioconductor.org/packages/release/bioc/html/GeneOverlap.html
CellBender_0.2.0	Fleming et al., 2019	https://github.com/broadinstitute/CellBender
DecontX	Yang et al., 2020	https://github.com/campbio/celda
SoupX	Young and Behjati, 2020	https://github.com/constantAmateur/SoupX
Destiny_3.8.1	Angerer et al., 2016	https://bioconductor.org/packages/release/bioc/html/destiny.html
clusterProfiler_4.2.2	Yu et al., 2012	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
SynGO	Koopmans et al., 2019	https://www.syngoportal.org/
Umi-tools_1.1.1	Smith et al., 2017	https://github.com/CGATOxford/UMI-tools
STAR_2.7.10	Dobin et al., 2013	https://github.com/alexdobin/STAR
Subread_2.0.1	Liao et al., 2014	https://sourceforge.net/projects/subread/files/
Gread_0.99.3	Srinivasan et al., 2016	https://rdrr.io/github/asrinivasan-oa/gread/
Critical commercial assays		
RNAscope® Multiplex Fluorescent Reagent Kit v2	ACD Bio-techne	Catalog #: 323100
Chromium Single Cell 3′ v3	10x Genomics	Cat#1000153
Chemicals, peptides, and recombinant proteins		
RNAscope® Probe-Hs-OLIG2-C2- mRNA	ACD Bio-techne	Catalog #: 424191-C2
RNAscope® Probe Hs-SYT1-C3- mRNA	ACD Bio-techne	Catalog #: 525791-C3
RNAscope® Probe-Hs-GRIN2A- mRNA	ACD Bio-techne	Catalog #: 485841
RNAscope® Probe-Mm-Mog-C2 mRNA	ACD Bio-techne	Catalog #: 492981-C2
RNAscope® Probe-Mm-Syt1- mRNA	ACD Bio-techne	Catalog #: 491831
RNAscope® Probe-Mm-Rbfox1- mRNA	ACD Bio-techne	Catalog #: 519911
RNAscope® Probe-Mm-Snap25- mRNA	ACD Bio-techne	Catalog #: 516471

RESOURCE AVAILABILITY

Lead contact

Further requests for resources should be directed to and will be fulfilled by the lead contact, Genevieve Konopka (genevieve.konopka@utsouthwestern.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Raw fastq files of mouse single-nuclei RNA-seq dataset are accessible in GEO with accession number: GSE198640. Re-analyzed processed matrices are accessible in GEO with accession number: GSE198951. All analysis codes are available in our github page: https://github.com/konopkalab/Ambient_RNA_In_Brain_snRNAseq.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human cortical tissue

Human postmortem posterior cingulate cortex samples were provided by the NIH NeuroBioBank. 17μm tissue sections were sectioned for the in-situ hybridization experiment detailed in the methods. Tissues from both adult (age range 45-75) males and females were used in the experiments. Demographic information of the samples is listed in Tables S1 and S5.

Mouse cortical tissue

All experiments were performed according to procedures approved by the UT Southwestern Institutional Animal Care and Use Committee. Mouse frontal cortex samples were collected for the single-nuclei RNA-sequencing and in-situ hybridization as detailed in the methods. Adult (postnatal day 56) male and female wildtype C57BL/6J mice were used in the experiments. Mice were maintained on a 12-hr light on/off schedule. Detailed information of the samples is listed in Table S1.

METHOD DETAILS

Preprocessing and count matrix generation

Datasets were downloaded from NCBI-GEO database (Table S1). Within the datasets, we only used the cells generated from cortical brain tissue. Specifically, NSD1 was from prefrontal and anterior cingulate cortex (n = 41 samples from anterior cingulate cortex and prefrontal cortex), NSD2 was from posterior cingulate cortex (n = 4 samples), SD1 was from prefrontal (n = 13) and visual cortex (n = 24), and SD2 was from anterior cingulate cortex (see below for final sample size). Barcode correction and filtering was done using *umi_tools whitelist* (retained top 20,000-80,000 cell barcodes per sample depending on the dataset to keep the ambient cell barcode population) and *umi_tools extract* (Smith et al., 2017). Alignment was done using STAR aligner (Dobin et al., 2013) with the reference genomes of GRCh38 (for the human datasets) or GRCm38 (for the mouse dataset). *featureCount* was used to count reads mapping to gene body only for uniquely mapping reads (Liao et al., 2014), and *umi_tools count* was used to create the count matrix. Count matrix using only intronic reads were similarly obtained using featureCount on a custom gtf that only contained introns (created using *construct_introns* from *gread* package in R (Srinivasan, 2016)). Intronic read ratios were then calculated per cell barcode by taking the ratio of the number of UMI counts mapping to introns and the number of UMI counts mapping to the gene body.

For DAPI+ sorted datasets, a Spearman's rank correlation between UMI counts (log_{10}) and intronic read ratio was computed for each sample. Only the samples with correlation lower than the correlation coefficient of 0.05 were considered sorted and used for further analysis. This criterion retained all samples in the SD1 study (Lake et al., 2018) and one sample (Br5400_sACC) in the SD2 study (Tran et al., 2021).

For ambient RNA cleanup, CellBender was used on the raw matrix of gene counts with default parameters (Fleming et al., 2019).

Single-nuclei library preparation

We processed 4 c57BL/6J P56 mice (2 males and 2 females). Mice were rapidly decapitated and brains were quickly removed. The isolated brain was quickly transferred to an ice-cold coronal brain section mold (Braintree Scientific, BS-A 5000) and washed with ice-cold 1X PBS (Cytiva, SH30256.01). The boundary of the olfactory bulb and frontal cortex was aligned to the first indentation where the first razor blade (Fisher Scientific, 12-640) was inserted to remove the olfactory bulb. The second razor blade was inserted into the third indentation. The coronal sections matched with coronal numbers 22-36 in the Allen Brain Atlas: Mouse Reference Atlas, Version 2 (2011). We then removed the subcortical region from this section and separated the left and right hemisphere samples into different Eppendorf tubes. The tubes were flash frozen in liquid nitrogen.

The nuclei isolation procedure was modified from our previous work (Ayhan et al., 2021). The frozen section from the left hemisphere was transferred to a Dounce homogenizer with 2ml of ice-cold Nuclei EZ lysis buffer (Sigma-Aldrich, NUC101). We then inserted pestle A for 23 strokes followed by pestle B for 23 strokes on ice. The homogenized sample was transferred to a 15ml conical

tube. We added 2ml of ice-cold Nuclei EZ lysis buffer and incubated on ice for 5min. Nuclei were collected by centrifuging at 500 g for 5min at 4C. We discarded the supernatant and added 4ml of ice-cold Nuclei EZ lysis buffer to resuspend nuclei. We then repeated the incubation and centrifuge steps and resuspended the nuclei in 200ul of nuclei suspension buffer: 1X PBS, 1% BSA (ThermoFisher, AM2618), and 0.2 U/ul RNAse inhibitor (ThermoFisher, AM2696). Finally, the nuclei suspension was filtered twice through Flowmi Cell Strainers (Bel-Art, H13680-0040). We mixed 10ul of nuclei suspension with 10ul of 0.4% Trypan Blue (Gibco, 15-250-061) and loaded this suspension on a hemocytometer (SKC, DHC-N015) to determine the concentration. 10,000 nuclei/sample were used to prepare snRNA-seq libraries using 10X Genomics Single Cell 3′Reagent Kits v3 (Zheng et al., 2017). Libraries were sequenced by the McDermott Sequencing Core at UT Southwestern on a NovaSeq 6000.

Tissue processing, single-nuclei RNA-seq library preparation and sequencing for the NSD2 dataset was performed as previously described (Ayhan et al., 2021).

Ambient cluster analysis

To retain cell barcodes that predominantly contain ambient RNAs, we kept two times more cell barcodes than the original publication per sample. For datasets generated in this study, we retained two times more cell barcodes than the number of nuclei targeted. Therefore, the final count matrix included both the cell barcodes that mostly represented real nuclei (and were annotated as real cell types in the published datasets) and newly retained cell barcodes that mostly represented empty droplets. Since not all newly retained cell barcodes are predominantly ambient RNAs (e.g. they could be doublets, or low quality nuclei of various cell types), we then performed clustering to identify clusters that contained high numbers of newly retained cell barcodes and clustered distinctly compared to annotated cell types per dataset. The following methods from Seurat v3 (Stuart et al., 2019) were used to perform and visualize clustering: normalization (*SCTransform*), dimensionality reduction (*RunPCA*), batch correction (*RunHarmony*, default parameters), k-nearest neighbors (*FindNeighbors*) on batch corrected dimensions and clusters identification by shared nearest neighbors (*FindClusters*). UMAP embedding was then computed for visualization in 2D space (*RunUMAP*). Clusters that were largely composed of newly retained cell barcodes (>75%) were annotated as ambient clusters. We note that 75% is unusually high since only 50% of the newly retained barcodes were originally filtered out in the previous publication.

To identify ambient cluster marker genes, we ran DGE (differential gene expression) analysis using pseudobulk edgeR (Chen et al., 2016). Briefly, counts were aggregated per sample and pseudobulk DGE was run with *pseudoBulkDGE* function *(method = 'edgeR')* in the *scran* package (Lun et al., 2016). Ambient cluster markers were identified with logFC > 0.3 and FDR < 0.05 cutoffs.

Enrichment of ambient cluster markers with annotated cell types was done using a Fisher's exact test from the *GeneOverlap* package (Shen, 2020). The total number of expressed genes were used as background (we followed this strategy for all Fisher's exact test enrichments).

Comparison with NeuN- datasets

To find differentially expressed genes in each dataset compared to the NeuN- sorted datasets, we first identified 6 major cell types in all datasets: excitatory neurons, inhibitory neurons, oligodendrocytes, OPCs, microglia and astrocytes. Using pseudobulk DGE (see above) in matched cell types, we then identified differentially expressed genes with significantly higher number of reads than in the given NeuN- sorted dataset. This was performed separately for each NeuN- sorted and other datasets. For the NeuN- sorted dataset and SD2 comparisons we used Wilcoxon rank sum test (*FindMarkers* function from Seurat) since we retained only one sample from this dataset (see Preprocessing and Count Matrix Generation).

For enrichment with ambient cluster markers, we selected the top 500 differentially expressed genes (ranked by logFC) among the NeuN- depleted genes in each comparison (logFC > 1 and FDR < 0.05). Similarly, the top 500 ambient RNA markers were selected from both nuclear ambient RNA and non-nuclear ambient RNA markers. Enrichment analyses were performed as above.

Ambient RNA removal with CellBender, DecontX and SoupX

All ambient RNA removal tools were run with the default parameters and according to the instructions. For CellBender (Fleming et al., 2019), the input was the raw gene – cell barcode count matrix. For DecontX (Yang et al., 2020), the input was the filtered matrix as recommended (Yang et al., 2020). For SoupX, both the filtered and raw matrices were given as input (Young and Behjati, 2020).

Subcluster cleaning of glia after CellBender

To subcluster glia after CellBender, we used the annotation provided in the original publication and processed each glia cell type separately per study. For the datasets generated in this study, we performed clustering as described and annotated glia based on established marker genes (e.g. *MBP*, *PCDH15*, *APBB1IP*, *SLC1A3*). Clustering was done similarly as above and marker genes of subclusters (identified using the default parameters in Seurat's *FindAllMarkers* (Stuart et al., 2019) function were tested for enrichment of ambient RNA markers using a Fisher's exact test. For this, we selected the top 500 (by logFC) ambient RNA markers from both nuclear and non-nuclear ambient RNA lists and combined them. We removed the subclusters with distinctly high levels of enrichment of ambient RNAs (FDR < 0.001 and odds ratio > 3) compared to other subclusters. All steps of ambient RNA contamination removal are outlined in Figure 5. We also showcase our mouse snRNA-seq dataset and provide analysis scripts that match each step in the stepwise guideline in our github page: https://github.com/konopkalab/Ambient_RNA_In_Brain_snRNAseq.

Assessment of ambient RNA contamination signatures in glia

To compare ambient RNA contamination in glia after each type of analyses, we first found cell barcodes common between annotated cell barcodes in each original publication and the retained cell barcodes after CellBender. We then only retained common cell barcodes in all downstream analyses that compared the original dataset and analyses that included ambient RNA contamination removal (Figure 3). This was to ensure that only gene expression levels were different and the enrichments are not driven by cell barcode differences between analyses. However, we note that the analysis with CellBender + subcluster cleaning contained fewer cell barcodes as ambient RNA rich subclusters were removed after CellBender. To test ambient RNA marker enrichment, we first found differentially expressed genes with significantly greater number of reads than in the NeuN- sorted dataset per cell type per dataset (logFC > 1 and FDR < 0.05). These gene lists were then tested for enrichment of ambient RNA markers (the combined top 500 genes were used as before) using a Fisher's exact test.

To test whether the global gene expression profile is altered after these different analysis methods, we found genes expressed in at least 5% of cells per cell type per dataset to remove lowly expressed genes. We then kept the genes that survive this threshold in all datasets and performed a Spearman rank correlation between each dataset and the NeuN- sorted dataset using the average log expression of genes in the normalized matrix.

In-situ hybridization and image quantification

Flash-frozen human postmortem cortical BA23 samples (n=3) and mouse whole brains (n=2) were embedded in Tissue-Tek CRYO-OCT Compound (#14-373-65, Thermo Fisher Scientific). We sectioned tissue at -20C to 17μm on Superfrost Plus Microscope slides (#12-550-15, Thermo Fisher Scientific). Fluorescent in situ hybridization (FISH) was performed using RNAScope® Multiplex Fluorescent Reagent Kit v2 assay for fresh frozen tissue (#323100, Advanced Cell Diagnostics) with the additional step of 0.05% Sudan Black B incubation at room temperature for 10 minutes after application of DAPI to quench autofluorescence. Species-specific probes were used for human: Hs-GRIN2A-C1 (485841), Hs-OLIG2-C2 (424191-C2), Hs-SYT1-C3 (525791-C3) and mouse: Mm-Syt1-C1 (491831), Mm-Rbfox1-C1 (519911), Mm-Snap25-C1 (516471), Mm-Mog-C2 (492981-C2) respectively. Opal fluorophores 520 (FP1487001KT, Akoya Biosciences), 570 (FP1488001KT, Akoya Biosciences) and 620 (FP1495001KT, Akoya Biosciences) were used to label C1, C2, and C3 channel respectively for the gene-specific probes after signal amplification.

We captured images from cortical areas of human and mouse by using a Zeiss LSM 710 at x20 magnification in the UT Southwestern Neuroscience Microscopy Facility. Maximum intensity projection images were generated from 13 slices of a Z stack. We randomly sampled 2-4 cortical areas (488x488 μm) from each brain section for both human and mouse to manually quantify the number of cells (DAPI 405 nm), neurons (*GRIN2A*, *Syt1*, *Rbfox1*, and *Snap25*, 488 nm; *SYT1*, 594 nm), and oligodendrocytes (*OLIG2* and *Mog*, 555 nm). We then calculated the fraction of neurons, oligodendrocytes, and overlap between the two cell types.

Pseudotime analysis

To be consistent with SD1 (Lake et al., 2018), we used the *DiffusionMap* function from *destiny* (Angerer et al., 2016) only on the visual cortex samples to build pseudotime trajectories between OPC-OL or between other pairs of glial cell types using the matrix provided by the authors. Diffusion maps were created with parameters *n_pcs=100* and *k=100*. The first two eigenvectors of diffusion maps were plotted for visualization. To identify markers of 'transitioning' cell barcodes, we found the middle cell barcode based on the first eigenvector (DM1) and labeled 200 cell barcodes around the middle cell barcode as 'transitioning cells'. The remaining two groups of cell barcodes were labeled by their original annotation label (e.g. OPC). We then found marker genes for each of these pseudotime groups (FDR < 0.05 and logFC > 0.25 using *FindMarkers* in *Seurat* (Stuart et al., 2019)) and ran enrichment with ambient RNA markers and immature oligodendrocyte markers identified in SD1 using a Fisher's exact test.

OPC subcluster analysis

To identify potential transitioning OPCs, we separately subclustered OPCs from three different datasets: SD1 (Lake et al., 2018), NSD1 (Velmeshev et al., 2019), and NSD2 (GEO accession: GSE198951) after CellBender and subcluster cleaning based on high ambient RNA contamination. We further removed subclusters with high expression of markers from two distinct major cell types as potential doublets. Committed oligodendrocyte progenitors (COPs) were identified by high expression of *GPR17* (as previously established (Marques et al., 2016)) among other markers (e.g. *BCAS1*, *FYN*).

To identify subclusters of Jäkel et al. (Jäkel et al., 2019), we performed dimensionality reduction and clustering on cells with the annotation of 'OPCs' and 'COPs' using Seurat v3 as described above. We then identified 'COPs-New' by the established marker genes (*BCAS1, FYN, GPR17*). For the heatmaps, all mature oligodendrocytes were combined and annotated as 'OL'. Normalized and log transformed expression levels for each gene was then z-transformed across 4 cell type annotations (OPC, COP-New, COP-Old, OL). For the UMI counts plots, we retained the original labels for neuronal cell types. Both control and multiple sclerosis samples were used and no additional cell filtering (other than subsetting by annotation) was applied for all analyses.

Identification of COP Marker Genes

Genes upregulated in COPs compared to OPCs were identified using the FindMarkers function from Seurat. Significant genes (FDR < 0.05 and expressed in >10% of COPs) were ranked by their avg_logFC and the top 100 genes per dataset were selected.

To highlight genes specific to COPs compared to OPCs and OLs, we found the percentage of nuclei that expressed at least one read of each significant gene. We then computed the difference of percentages between both COPs-OPCs and COPs-OLs. We then took the intersection of the top 20 genes with the greatest difference in favor of COPs in both comparisons. This was repeated for all three datasets. Genes that marked COPs in at least two datasets were reported as COP markers within the oligodendrocyte lineage in human brain (Table S5).

Other enrichments

Gene ontology (GO) enrichment of ambient RNA signatures was done using the clusterProfiler package in R (Yu et al., 2012) with all expressed genes used as the background. The full table of GO results is available in Table S3.

To test enrichment of ambient RNA markers with vGLUT1-Depleted and vGLUT1-Enriched genes from Hafner et al. (Hafner et al., 2019), we first converted the mouse gene symbols to human gene symbols using SynGO (Koopmans et al., 2019). Fisher's exact tests were performed as before.

To overlap ambient RNAs with highly represented genes in neurons in snRNA-seq datasets, we identified the top-represented genes among all neurons by taking the mean of each gene across all cell barcodes annotated as neurons separately in both SD1 and NSD1 (except for Neu-NRGNs and Neu-mat). The intersection of the top 500 genes in both datasets (403 genes) was used to overlap with ambient RNA markers.

QUANTIFICATION AND STATISTICAL ANALYSIS

All analysis-specific quantifications and statistics can be found in their corresponding method section. Individual statistics (e.g adjusted p-value, odds ratio, fold change) for each comparison can be found in the figure legends and on the figures. Sample sizes of the snRNA-seq dataset can be found in the methods. Unless otherwise stated, all samples from the associated publication were retained for the analyses of this study. Sample sizes of the in-situ hybridization experiment can also be found in the methods and the figure legends. We did not conduct a separate benchmarking for the selection of the statistical analyses, however we strived to select the most up to date and benchmarked methods (e.g we favored pseudobulk methods instead of the single-cell based methods for the differential gene expression analyses (Squair et al., 2021)).

Neuron, Volume *110*

Supplemental information

Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets

Emre Caglayan, Yuxiang Liu, and Genevieve Konopka

Supplemental Information

Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets

Emre Caglayan, Yuxiang Liu and Genevieve Konopka

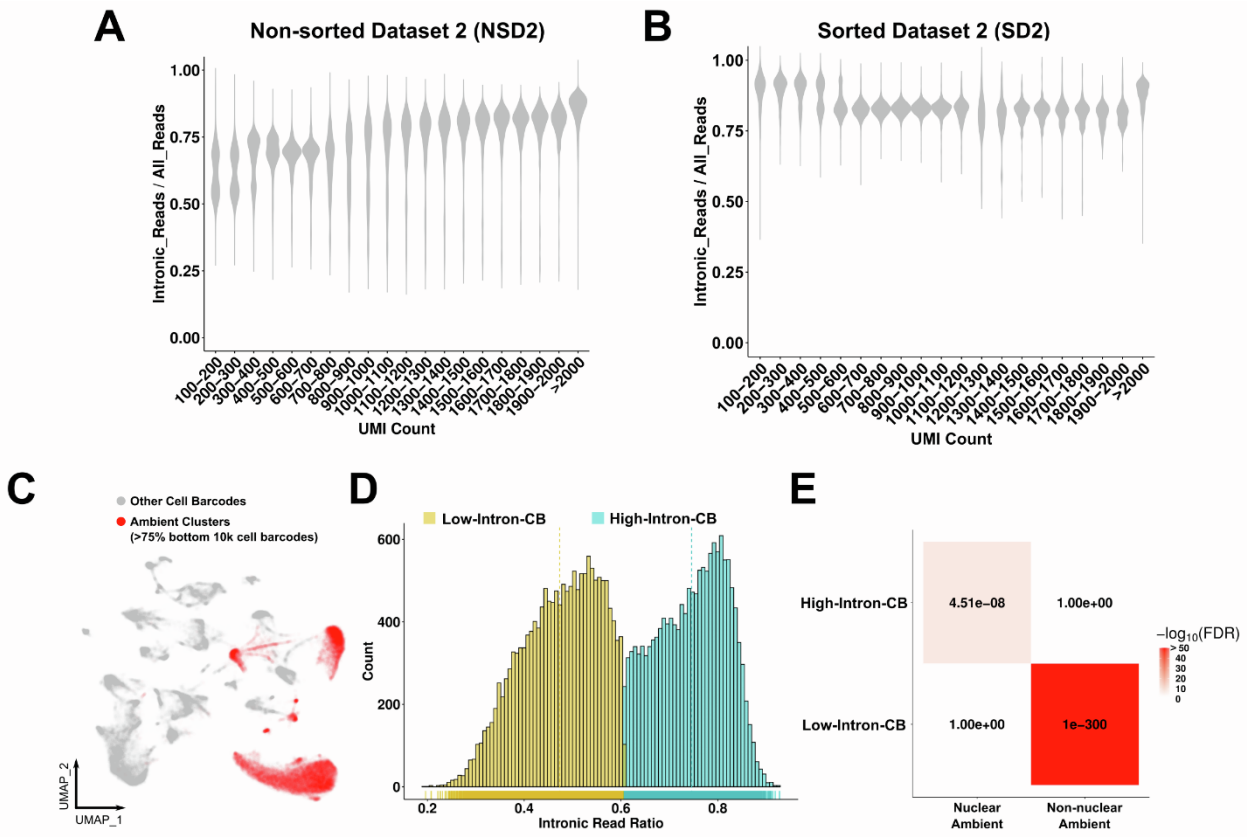


Figure S1 (Related to Figure 2). Independent confirmation of nuclear and non-nuclear ambient RNA types. (A-B) Plots of intronic read ratios across increasing UMI counts in **(A)** non-sorted dataset 2 (NSD2) and **(B)** sorted dataset 2 (SD2). UMI counts are divided into intervals of 100 from 100-2000. **(C)** The clusters that contain >75% of filtered cell barcodes are highlighted and named ambient clusters (dataset: NSD2). **(D)**. Plot of the distribution of intronic read ratios within ambient clusters. Yellow: Low-Intron-CB (CB: Cell Barcodes), blue: High-Intron-CB. **(E)** Heatmap of enrichments between ambient RNA types. Nuclear ambient RNA and non-nuclear ambient RNAs from SD1 and NSD1 were compared (see Figures 1-2).

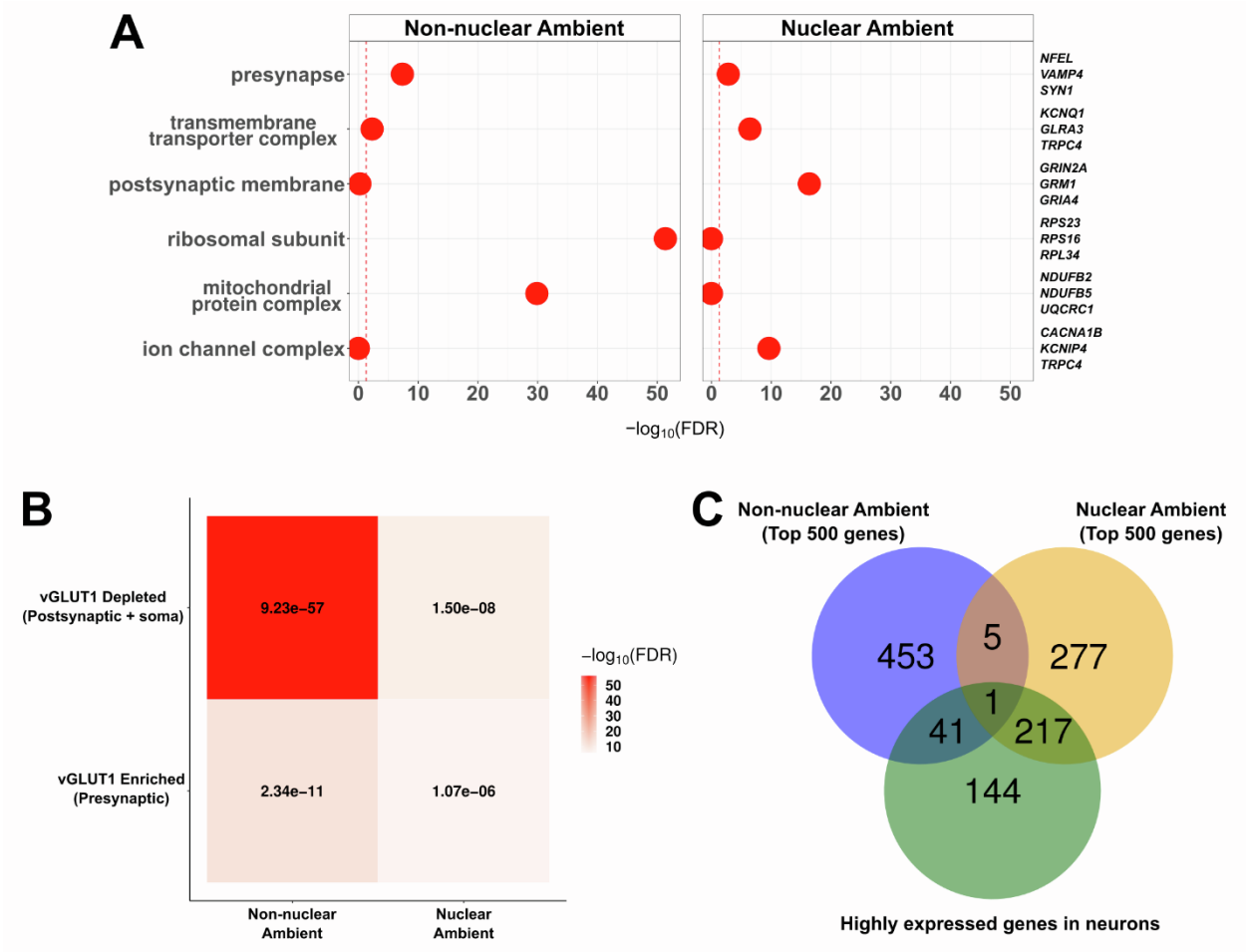


Figure S2 (Related to Figure 2). Characterization of ambient RNA markers. (A) Gene ontology (GO) enrichments for non-nuclear and nuclear ambient RNA markers. Example genes per GO term are shown on the right. **(B)** Heatmap of enrichments between presynaptic synaptosomes (vGLUT1 Enriched) and others (vGLUT1 Depleted) (Fisher's exact test; numbers indicate FDR; heatmap color indicates -$\log_{10}$FDR). Genes from the synaptosome dataset were converted from mouse to human symbols prior to enrichment. **(C)** Overlaps between the top 500 most distinct ambient RNA markers and the top 500 highly expressed genes in neurons.

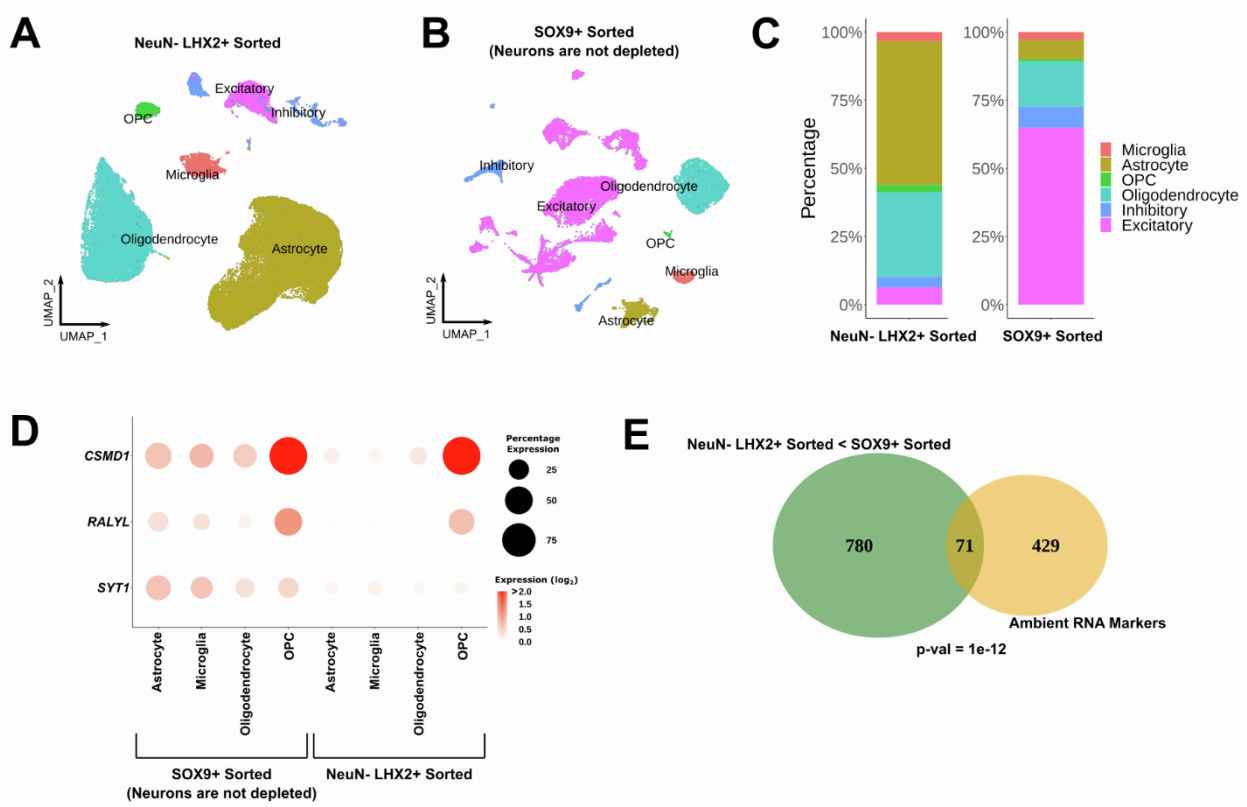


Figure S3 (Related to Figure 3). Comparison of neuron depleted and non-depleted snRNA-seq datasets from the same study. (A-B) UMAP plots of sorted datasets with neuron depletion (NeuN- LHX2+ sorted, also used as NeuN- SD3) **(A)** and without neuron depletion (SOX9+ sorted) **(B)**. **(C)** Stacked bar plots of both datasets that show cell type composition by percentage. **(D)** Dot plot of expression levels (normalized, log2 transformed) of selected ambient RNA marker genes across glial cell types. **(E)** Overlap between genes overrepresented in the SOX9+ sorted dataset and the top 500 ambient RNA markers. Only nuclear ambient RNA markers were used since non-nuclear ambient RNAs were removed in the sorted datasets.

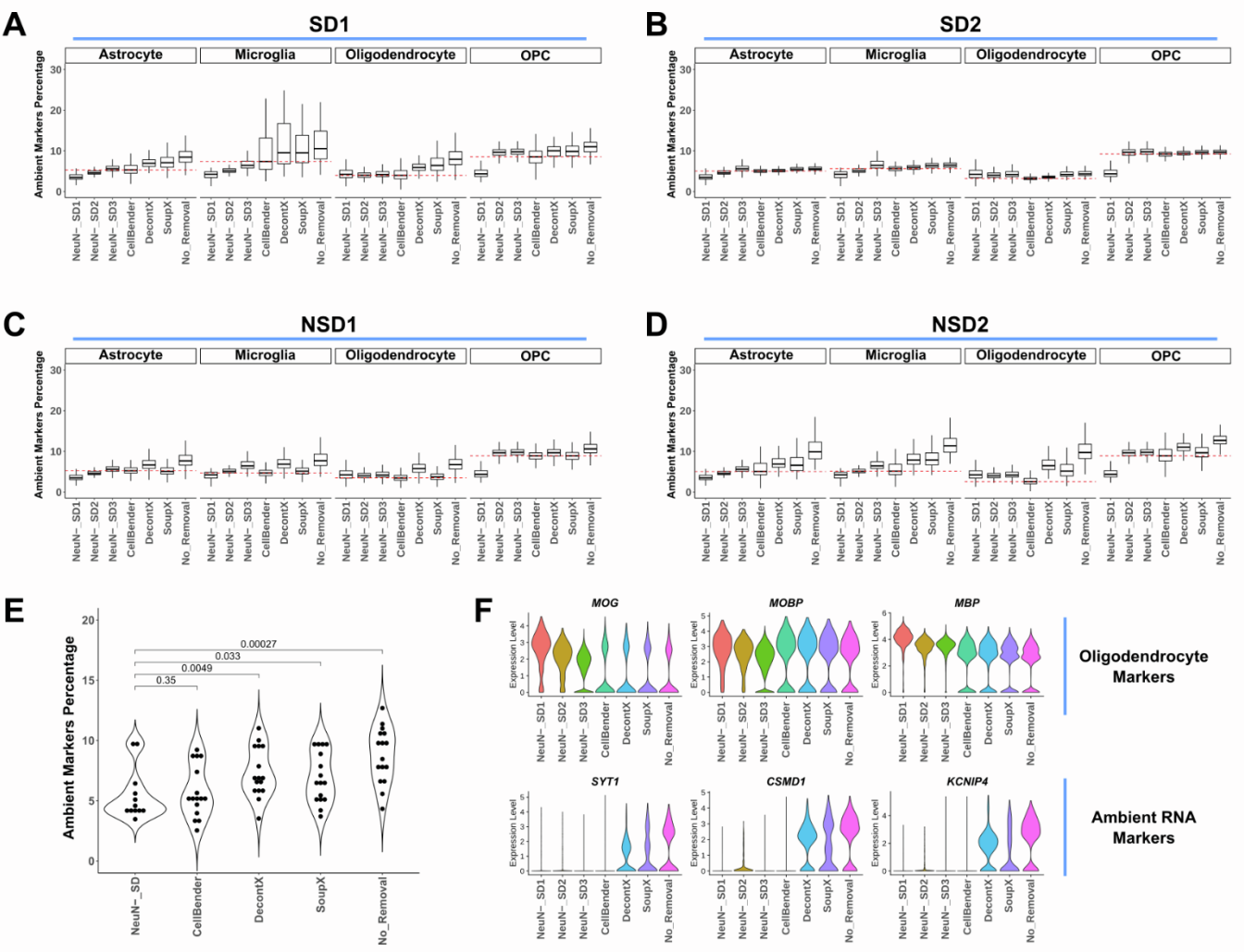


Figure S4 (Related to Figure 3). Comparison of ambient RNA removal tools. All tools were evaluated based on the percentage of reads explained by ambient RNA markers in glial cell types. These percentages were then compared to the percentages of ambient RNA markers in NeuN- sorted datasets (NeuN- SDs). **(A-D)** Comparison of ambient RNA marker percentages in SD1 **(A)**, SD2 **(B)**, NSD1 **(C)** and NSD2 **(D)**. Dashed lines in red correspond to the median values of the CellBender result. **(E)** Summary of comparisons in **A-D**. Each dot represents a median value of the boxplots in **A-D**. Numbers indicate p-values from one-sided Wilcoxon rank sum tests between the NeuN- SD results and the results from each ambient RNA removal tool or no removal. **(F)** Violin plots of the expression levels of selected ambient RNAs after implementation of CellBender, DecontX or SoupX. The plot contains the expression values of oligodendrocytes from SD1.

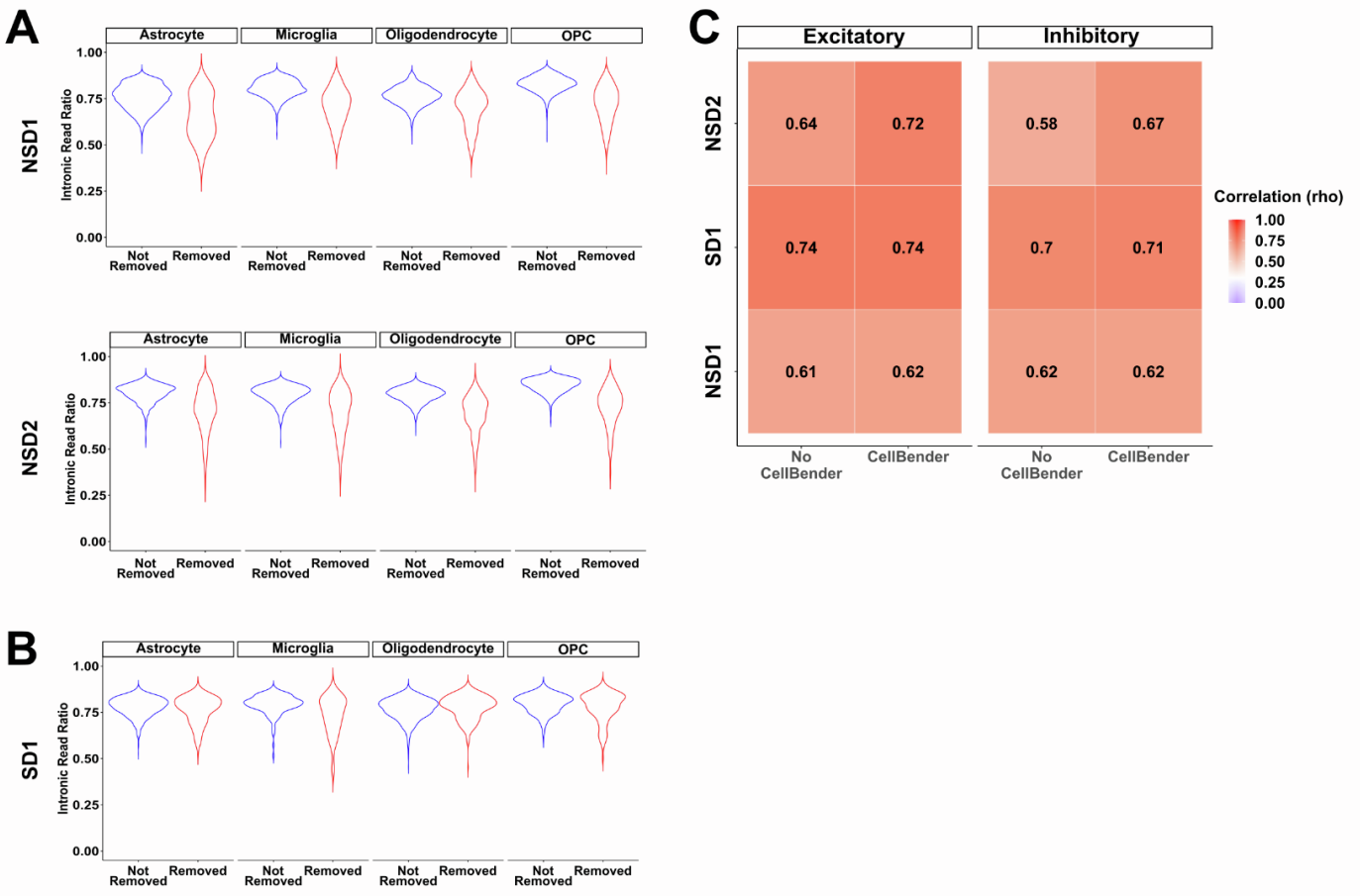


Figure S5 (Related to Figure 3). Supplementary analyses of CellBender adjustment and subcluster cleaning. (A) Intronic read ratios per cell barcode in subclusters that were removed due to ambient RNA contamination (red) or not removed (blue) per glial cell type in datasets that did not perform nuclei sorting (NSD1 and NSD2). **(B)** Same as in **(A)** but in a dataset that performed nuclei sorting (SD1). **(C)** Heatmap of Spearman rank correlations of all genes with the NeuN+ sorted dataset (SD2). Correlations were performed per cell type per dataset (y-axis) after each analysis (x-axis). Both numbers and heatmaps indicate the magnitude of correlation coefficient.

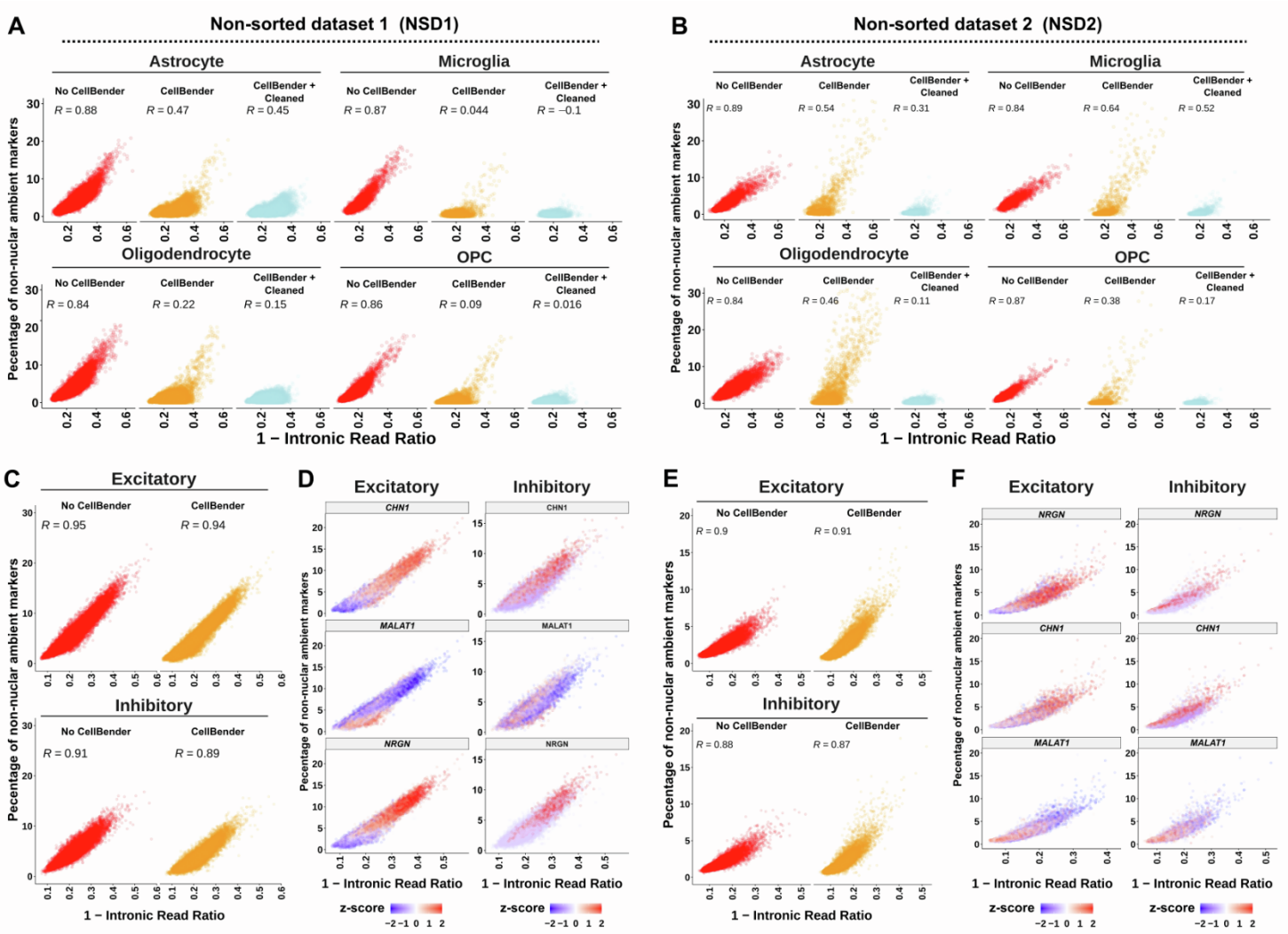


Figure S6 (Related to Figure 3). **Association of non-intronic read ratios and the percentage of non-nuclear ambient markers across nuclei per cell type. (A-B)** Scatter plots of non-intronic read ratios (x-axis) versus the percentage of non-nuclear ambient markers (y-axis) in glial cell types either before CellBender (red), after CellBender (orange), or after CellBender + subcluster cleaning (lightblue) using either the NSD1 **(A)** or NSD2 dataset **(B)**. **(C, E)** Scatter plots of non-intronic read ratios (x-axis) versus the percentage of non-nuclear ambient markers (y-axis) in excitatory and inhibitory neurons using either the NSD1 **(C)** or NSD2 **(E)** dataset. **(D, F)** Normalized and z-transformed expression levels of non-nuclear ambient markers *NRGN*, *CHN1* and nuclear-retained non-coding gene *MALAT1* in either the NSD1 **(D)** or NSD2 **(F)** dataset. R corresponds to the Spearman's rank correlation coefficient.

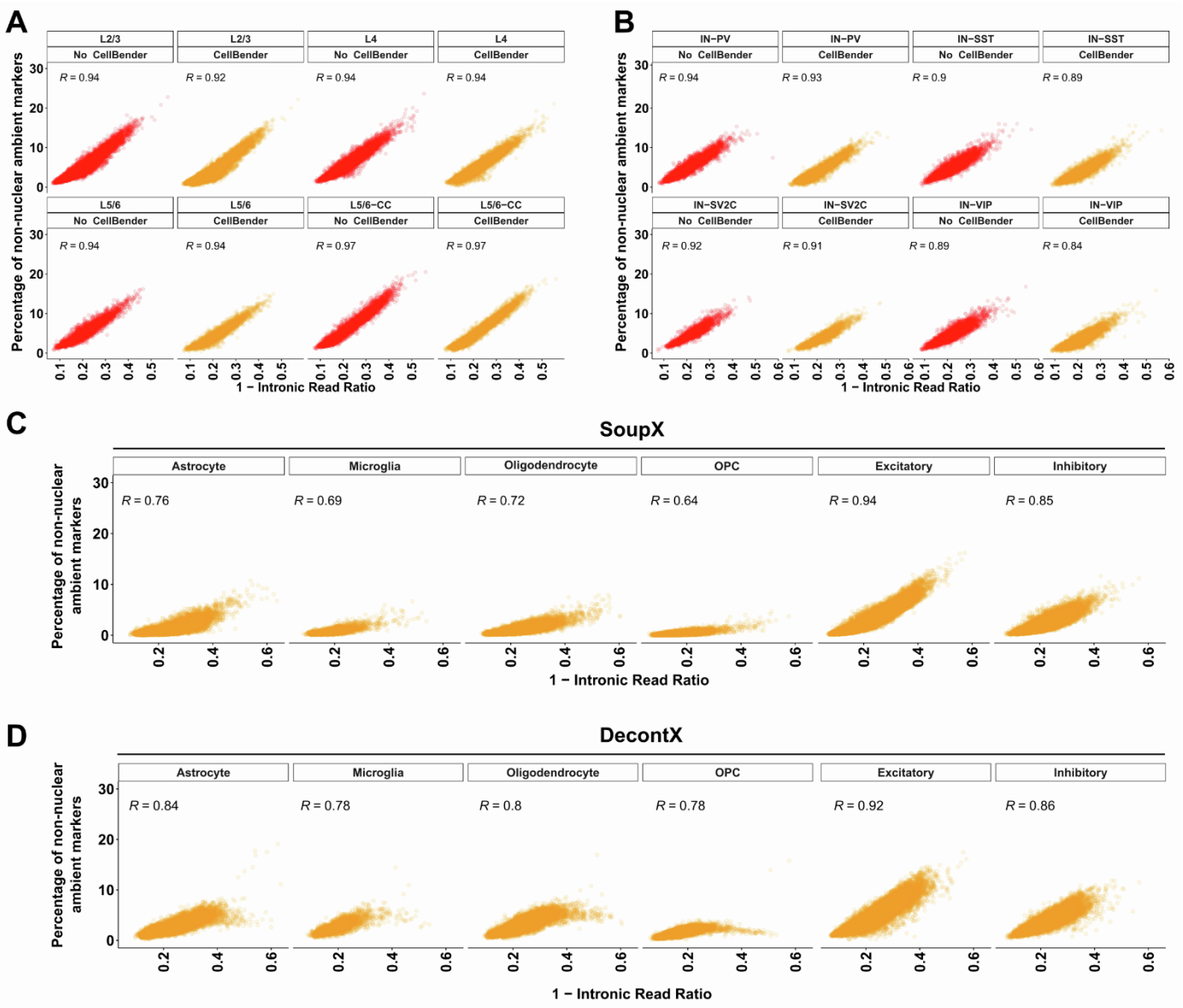


Figure S7 (Related to Figure 3). Association of non-intronic read ratios and the percentage of non-nuclear ambient markers across nuclei in subtypes or using other tools. (A) Scatter plots of non-intronic read ratios (x-axis) versus the percentage of non-nuclear ambient markers (y-axis) per annotated excitatory subtype before CellBender (red) or after CellBender (orange). **(B)** Same as **A**, but for inhibitory subtypes. **(C)** Scatter plots of the non-intronic read ratios and percentage of non-nuclear ambient markers per annotated broad cell type after SoupX. **(D)** Scatter plots of the non-intronic read ratios (x-axis) versus the percentage of non-nuclear ambient markers per annotated broad cell type after DecontX. All plots are from the NSD1 dataset. R corresponds to the Spearman's rank correlation coefficient.

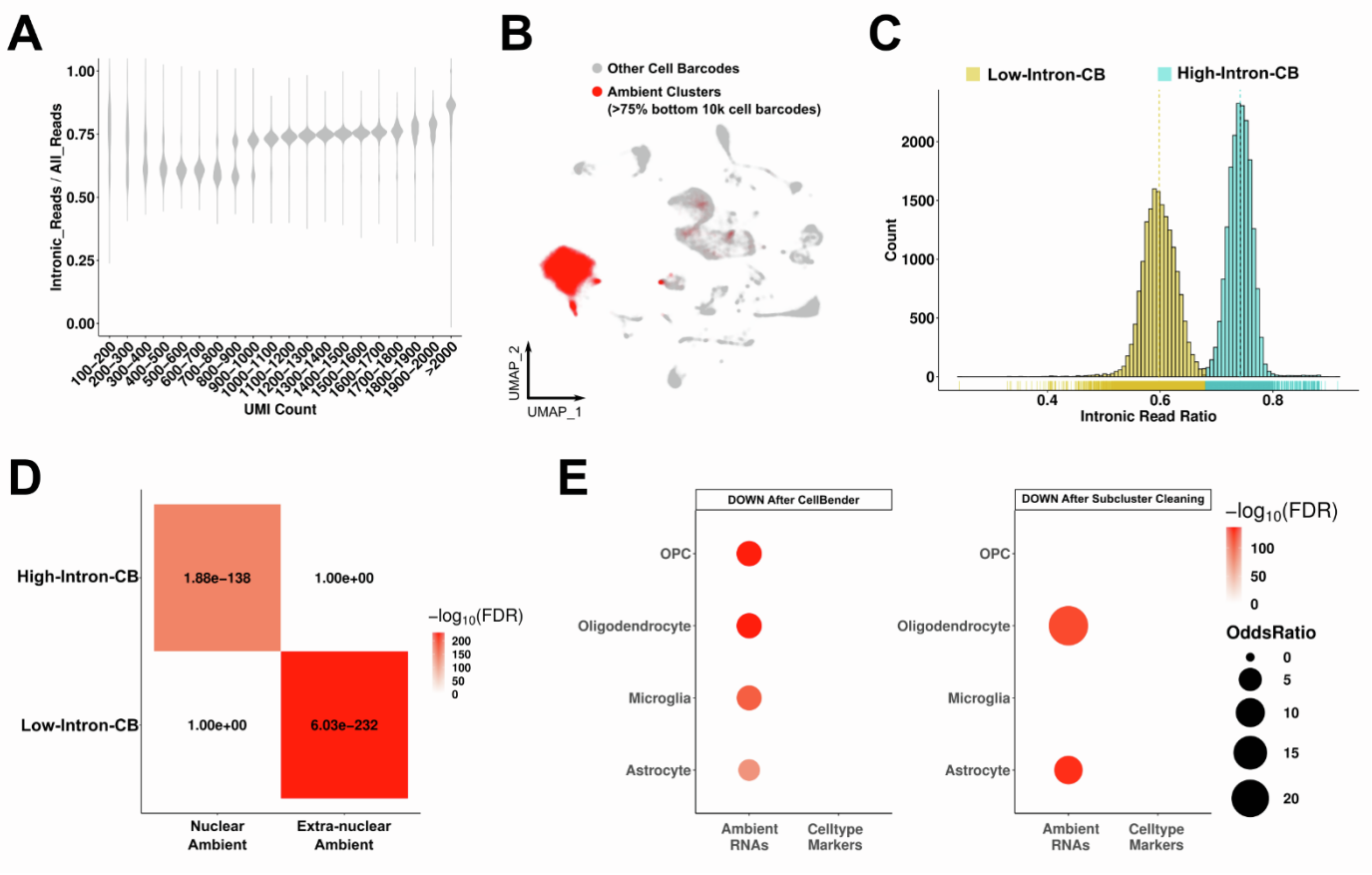


Figure S8 (Related to Figure 3). Ambient RNAs in a mouse brain snRNA-seq dataset. (A) The intronic read ratio across increasing UMI counts in a mouse brain snRNA-seq dataset with no nuclei-sorting. UMI counts are divided into intervals of 100 from 100-2000. **(B)** The clusters that contain greater than >75% of filtered cell barcodes are highlighted and named ambient clusters. **(C)** The distribution of intronic read ratios within ambient clusters. Yellow: Low-Intron-CB (CB: Cell Barcodes), blue: High-Intron-CB. **(D)** Heatmap of enrichments between ambient RNA types. Nuclear ambient RNA and non-nuclear ambient RNAs are identified from SD1 and NSD1 (see Figures 1-2). **(E)** Dot plot enrichments between genes with significantly lower expression (DOWN) after CellBender (left) or subclustering steps (right) with ambient RNA markers or cell type markers. Cell types are indicated in the y-axis.

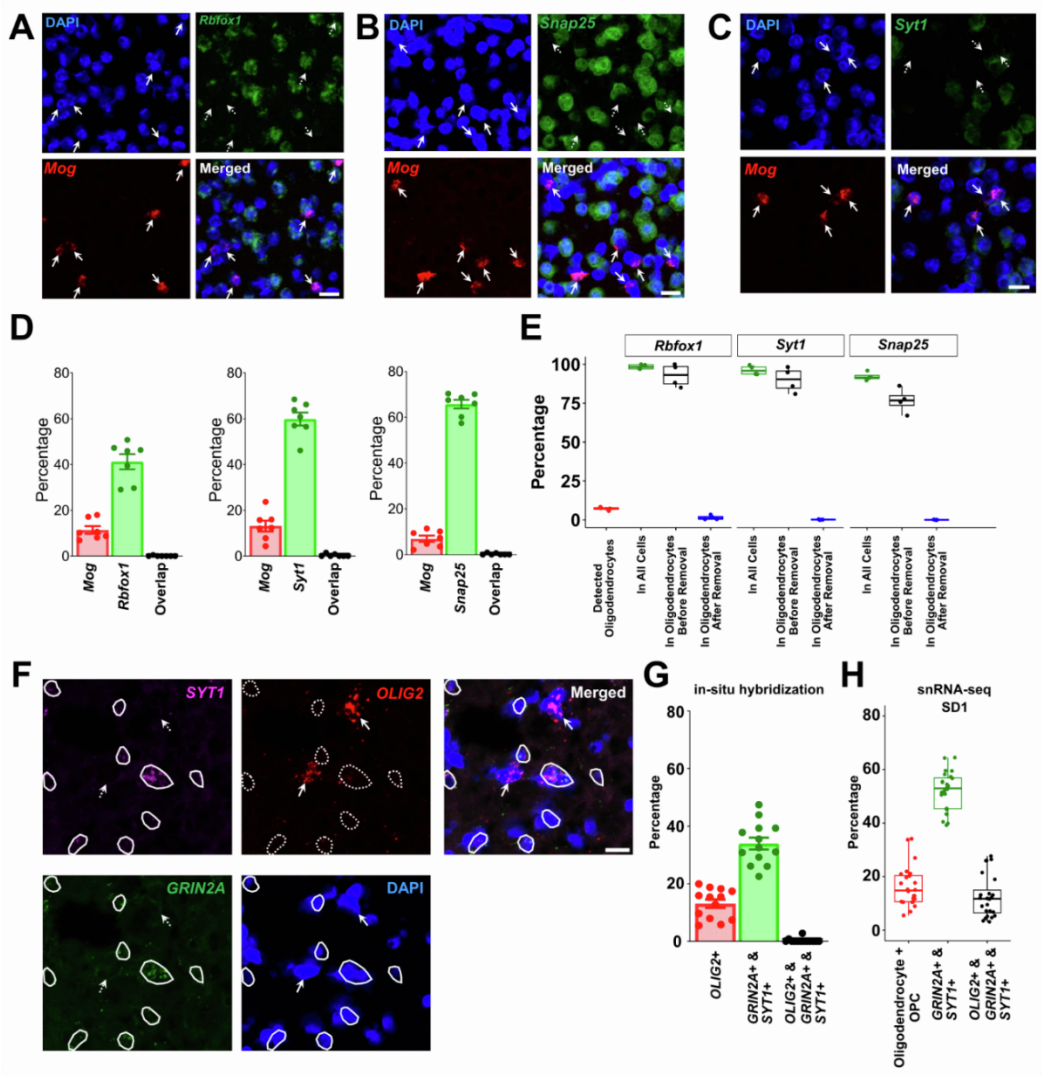


Figure S9 (Related to Figure 3): In situ hybridization does not detect ambient RNA markers in oligodendrocytes. (**A-C**) Representative images of smFISH for a marker of mature oligodendrocytes (*Mog*+ cells, solid arrows) and 3 markers of ambient RNAs (*Rbfox1*, *Snap25*, *Syt1*) reveal no overlap (dashed arrows) in adult mouse frontal cortex. (**D**) Quantification of smFISH experiments to indicate the percentage of cells positive for each gene relative to the number of DAPI+ cells (7 sagittal images obtained from 2 mice were quantified for each experiment). Data are represented as mean ± SEM (**E**) The percentage of cells that contain at least one read of each gene per given population for each specified analysis in the adult mouse snRNA-seq dataset. The dataset is the one from Supplementary Figure 8. (**F**) Representative images of smFISH for a marker of mature oligodendrocytes (*OLIG2*+ cells, arrows) or genes that mark neurons (*SYT1* and *GRIN2A*, positive cells are highlighted in circles) in adult human posterior cingulate cortex. (**G**) Quantification of smFISH experiments to indicate the percentage of cells positive for each gene relative to the number of DAPI+ cells (13 images from 6 tissue sections obtained from a total of 3 individuals were quantified for each experiment). Data are represented as mean ± SEM. (**H**) Percentage of cells that were annotated as belonging to the oligodendrocyte lineage or containing at least one read from the given gene in the original SD1 dataset.

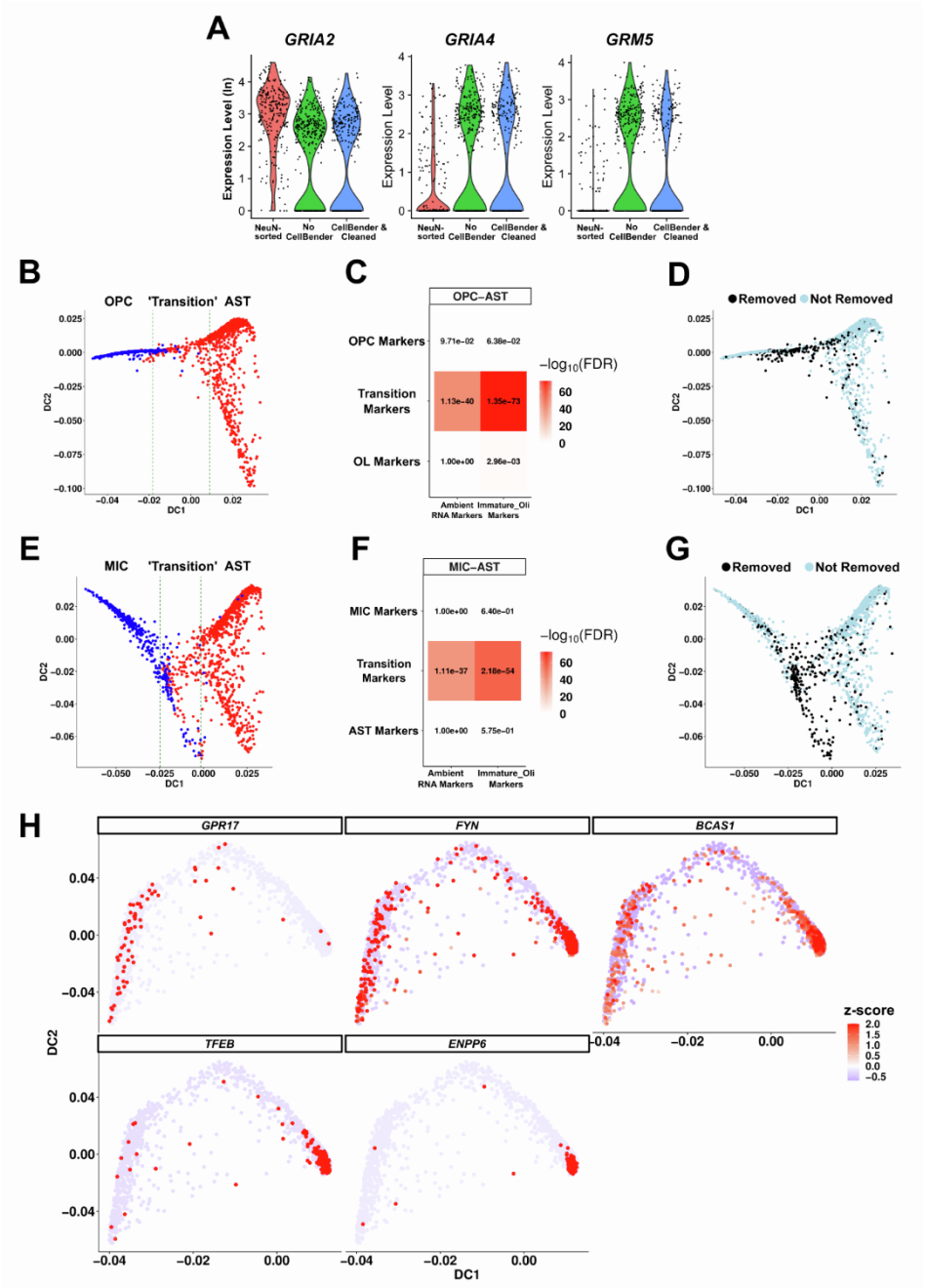


Figure S10 (Related to Figure 4). Immature oligodendrocytes are explained by ambient RNA contamination. (A) Violin plots of expression of glutamatergic receptors in NeuN- sorted OPCs and SD1 OPCs with or without ambient RNA removal. **(B)** Pseudotime trajectory of SD1 as reconstructed with *destiny* between OPCs and AST (astrocytes). The 'transition' zone was defined as the 400 nuclei around the middle nucleus based on DC1. **(C)** Heatmaps of enrichments between trajectory zones (OPC, Transition, AST) and ambient RNA or immature oligodendrocyte markers (Fisher's exact test; numbers indicate FDR; color scale is -$\log_{10}$(FDR)). **(D)** The same lineage trajectory as **(B)** with the nuclei removed after subcluster cleaning highlighted. **(E-G)** The same trajectory approach used in **(B-D)** but instead using MIC (microglia) and AST. **(H)** Z-transformed gene expression of COP or Pre-OL (premyelinating oligodendrocyte) markers in the OPC-OL lineage trajectory.

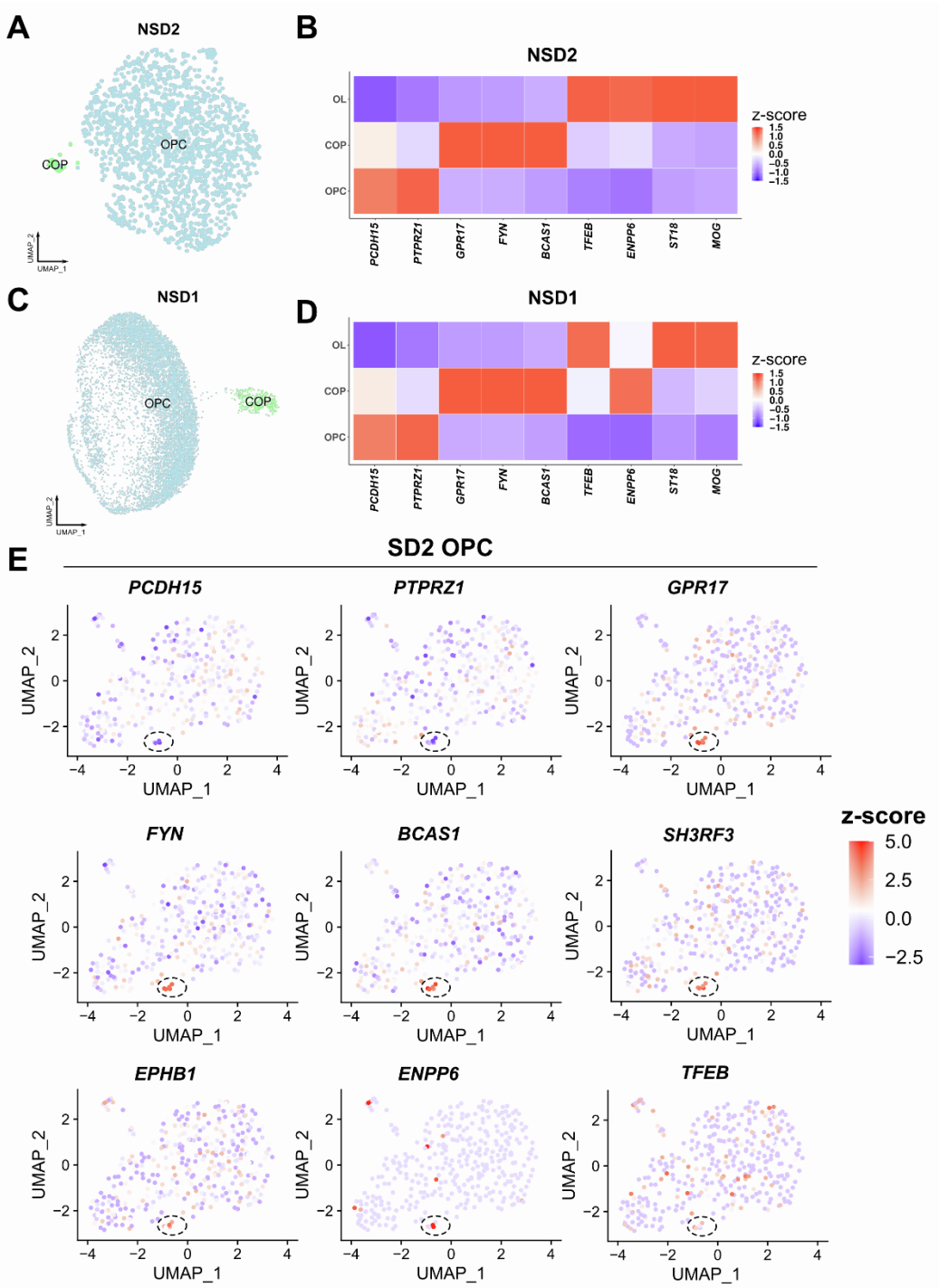


Figure S11 (Related to Figure 4). Committed progenitor cells in additional adult human brain snRNA-seq datasets. (A, C) UMAP of OPC subclustering from **(A)** NSD2 or **(C)** NSD1 datasets. COP: committed OPCs. **(B, D)** Heatmap of z-transformed gene expression of oligodendrocyte lineage marker genes (z-scored across cell types per marker gene) from **(B)** NSD2 or **(D)** NSD1 datasets. OPC markers: *PCDH15*, *PTPRZ1*. COP markers: *GPR17*, *FYN*, *BCAS1*. COP and OL markers: *ENPP6*, *TFEB*. **(E)** Feature plots of oligodendrocyte lineage markers genes for OPCs from SD2. The color scheme reflects the z-score per the expression of each gene across cell types (legend on the right side of the plot).

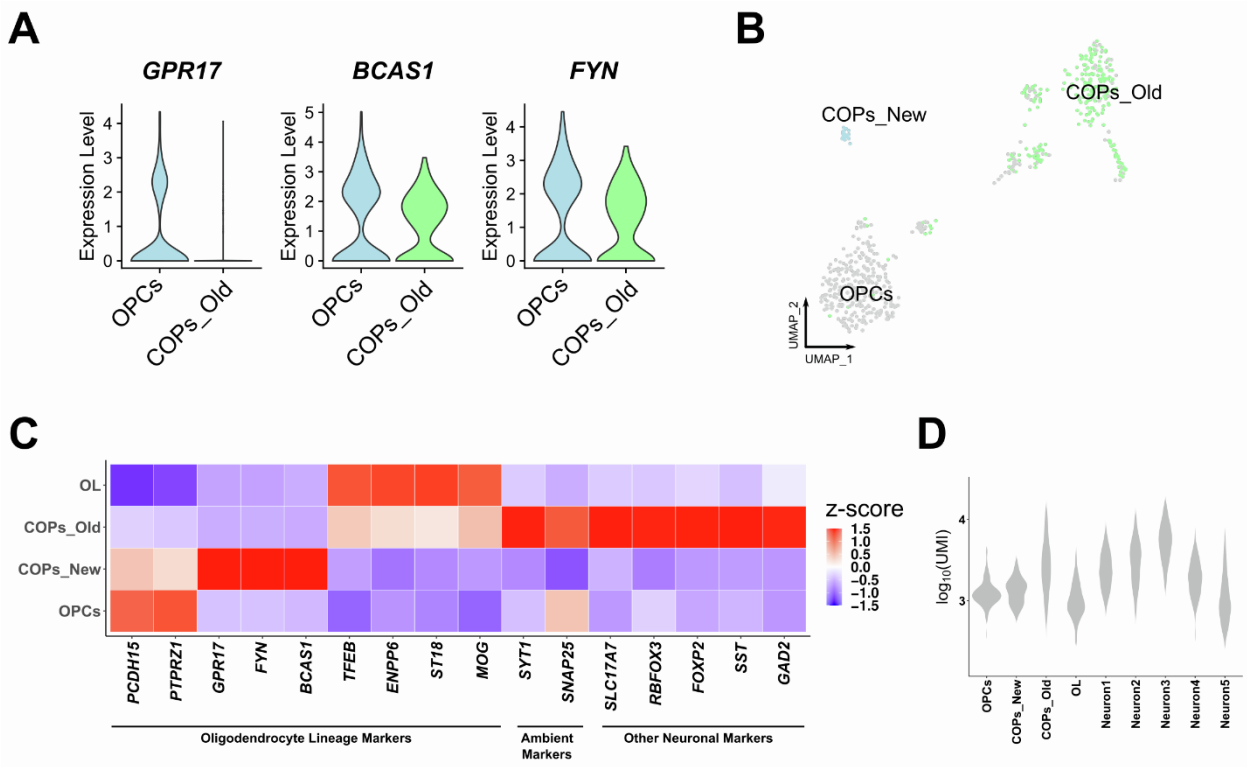


Figure S12 (Related to Figure 4). Re-assessment of previous COP annotations in human brain snRNA-seq white matter dataset. (A) Violin plots of expression levels (normalized, log transformed) of COP markers in the original annotation of OPCs and COPs ('COPs_Old') versus expression in nuclei we hypothesize to be true COPs ('COPs_New'). **(B)** UMAP plot of OPCs and COPs. The small subpopulation suspected to be real COPs is indicated as 'COPs_New' whereas the nuclei previously annotated as COPs are shown as 'COPs_Old'. **(C)** Heatmap of z-scored gene expression of oligodendrocyte lineage markers, two top ambient RNA markers (*SYT1*: Nuclear, *SNAP25*: Extra-nuclear) and other neuronal markers in the same dataset. The colors indicate the z-scored expression across cell types per marker gene. Note that z-scored expression only shows the relative expression levels among the four cell type annotations. **(D)** Log_{10} transformed UMI count values per cell type. OL: oligodendrocytes. Neuronal cell types are annotated as in the original publication.

CHAPTER 4: Association between resting-state functional brain connectivity and gene expression is altered in autism spectrum disorder

Published as:

Berto, S.*, Treacher, A. H.*, **Caglayan, E.***, Luo, D., Haney, J. R., Gandal, M. J., Geschwind, D. H., Montillo, A. A., & Konopka, G. 2022. Association between resting-state functional brain connectivity and gene expression is altered in autism spectrum disorder. *Nature Communications*. DOI: 10.1038/s41467-022-31053-5

*co-first authors

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-31053-5 OPEN

Association between resting-state functional brain connectivity and gene expression is altered in autism spectrum disorder

Stefano Berto[1,9], Alex H. Treacher[2,9], Emre Caglayan[1,9], Danni Luo[2], Jillian R. Haney[3,4,5], Michael J. Gandal[3,4,5,6], Daniel H. Geschwind[3,4,5,6], Albert A. Montillo[2,7,8]✉ & Genevieve Konopka[1]✉

Gene expression covaries with brain activity as measured by resting state functional magnetic resonance imaging (MRI). However, it is unclear how genomic differences driven by disease state can affect this relationship. Here, we integrate from the ABIDE I and II imaging cohorts with datasets of gene expression in brains of neurotypical individuals and individuals with autism spectrum disorder (ASD) with regionally matched brain activity measurements from fMRI datasets. We identify genes linked with brain activity whose association is disrupted in ASD. We identified a subset of genes that showed a differential developmental trajectory in individuals with ASD compared with controls. These genes are enriched in voltage-gated ion channels and inhibitory neurons, pointing to excitation-inhibition imbalance in ASD. We further assessed differences at the regional level showing that the primary visual cortex is the most affected region in ASD. Our results link disrupted brain expression patterns of individuals with ASD to brain activity and show developmental, cell type, and regional enrichment of activity linked genes.

[1] Department of Neuroscience, UT Southwestern Medical Center, Dallas, TX 75390, USA. [2] Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center, Dallas, TX 75390, USA. [3] Program in Neurobehavioral Genetics, Department of Psychiatry, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA. [4] Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA. [5] Program in Neurogenetics, Department of Neurology, Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA. [6] Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA. [7] Department of Radiology, University of Texas Southwestern Medical Center, Dallas, TX, USA. [8] Advanced Imaging Research Center, University of Texas Southwestern Medical Center, Dallas, TX, USA. [9] These authors contributed equally: Stefano Berto, Alex H. Treacher, Emre Caglayan. ✉email: Albert.Montillo@utsouthwestern.edu; Genevieve.Konopka@utsouthwestern.edu

Brain architecture and activity are governed by gene regulatory mechanisms that can be captured using transcriptomic measures[1–3]. How these mechanisms are impacted in neuropsychiatric disorders such as autism spectrum disorder (ASD) remain incompletely understood. Recent advances in human brain imaging genomics have the translational potential to address the challenge of detecting genes associated with either structural or functional measurements[4–6]. For instance, several studies have highlighted the influence of genetic variants on brain imaging phenotypes, identifying common loci that affect brain morphology, structure, and connectivity[7–11]. However, despite this considerable progress in understanding the genetic influence on human brain phenotypes, the gene regulatory mechanisms supporting such functional measurements remain mostly unknown. Identification of such gene expression patterns that underlie functional measures of human brain activity is particularly compelling as such insights will provide opportunities for future modulation of normal or pathological behaviors.

To date, several studies using resting-state functional MRI (rs-fMRI) measurements across cortical regions have identified gene expression patterns that support functional signals in human brain[12–15]. Such studies were a pioneering first step to determine reliable sets of genes that correlate with functional brain network measurements. These studies also established methodologies that can also be applied to study the association between gene expression and functional measurements in neuropsychiatric disorders. For example, genomic perturbations associated with differences in brain activity in a neuropsychiatric disorder such as ASD can now be examined. Individuals with ASD have alterations in both brain activity[16–18], and gene expression patterns (including at the cell-type level)[19–22]; thus, integrating datasets of brain imaging phenotypes, transcriptional landscapes, and cell-type expression patterns should provide insight into ASD pathophysiology. Moreover, because several ASD-relevant genes are chromatin modifiers or involved in neuronal activity[23–26], we hypothesized that brain gene expression patterns that typically support functional brain activity in healthy individuals might be severely affected in ASD. Therefore, coupling measurements of brain gene expression and activity has the potential to identify genes whose expression underlies functional networks observed in rs-fMRI and how such relationships are altered in ASD.

Here, we apply an approach to understand the gene expression signals that may underlie human brain activity (as assessed by rs-fMRI) relevant to ASD. In contrast with previous studies that used a reference dataset from a small number of "control" brain donors[27–29] or blood[30], we use post-mortem brain gene expression datasets from a greater number of individuals who are characterized as either neurotypical or who were diagnosed with ASD. Because of the rarity of post-mortem tissue available from ASD brain donors, our study is restricted to a subset of cortical regions. Nonetheless, we identify genes with expression patterns in brains from individuals with ASD that are differentially correlated with rs-fMRI activity. We also identify a small number of cortical regions that display the greatest impact of gene expression on brain activity (e.g., primary visual cortex and inferior temporal cortex). Our analyses consider the developmental expression pattern of the genes we identify related to ASD status. We find that many of these genes have altered expression patterns over postnatal development into adulthood suggesting that these particular genes are indeed relevant for brain activity responsiveness. Together, our results provide key insights into both specific genes and cortical regions that are at risk in ASD. The coupling of two diverse measurements (transcriptome and rs-fMRI) facilitates the prioritization of specific ASD mechanisms that might be missed by using only one type of dataset.

Results

Integration of resting-state functional MRI and gene expression measures in individuals with ASD and controls. To identify differentially correlated genes, we determined the spatial similarity between rs-fMRI and gene expression changes in the human brain of subjects with ASD compared to controls across 11 matched cortical regions. We used rs-fMRI data from an imaging database containing individuals with ASD and matched controls (ABIDE I[31] and ABIDE II[32]) and cortical RNA-sequencing (RNA-seq) datasets from persons with ASD and matched controls across development into adulthood[33] (Fig. 1). We computed two extensively validated measures of brain activation to characterize brain function from rs-fMRI. The first brain measure, fractional Amplitude of Low-Frequency Fluctuations (fALFF)[34], quantifies a subset of brain activity within the low frequency band that form a fundamental feature of the resting brain, and that activity is vitally important whether at rest (daydreaming, musing) or attending to a specific task. The second brain measure, Regional Homogeneity (ReHo)[35], is a complementary measure of the similarity in the temporal activation pattern manifested by clusters of voxels rather than single voxels as in fALFF. This measure of local functional connectivity is itself a close derivative of the underlying brain activity[35]. We generated voxel-wise maps of fALFF and ReHo for a total of 1983 subjects from the ABIDE I and ABIDE II datasets (ASD = 916, CTL = 1067; Supplementary Fig. 1 and Supplementary Data 1), and analyzed a total of 11 regions of interest (ROIs) matching the transcriptomic data using Brodmann area (BA) designations: BA1/2/3/5 (somatosensory cortex), BA4/6 (premotor and primary motor cortex), BA7 (superior parietal gyrus), BA9 (dorsolateral prefrontal cortex), BA17 (primary visual cortex), BA20/37 (inferior temporal cortex), BA24 (dorsal anterior cingulate cortex), BA38 (temporal pole), BA39/40 (inferior parietal cortex), BA41/42/22 (superior temporal gyrus), BA44/45 (inferior frontal gyrus).

We first assessed differences between cases and controls for both fALFF and ReHo (Fig. 2a). We identified 4 ROIs with a significant difference for fALFF and 1 ROI for ReHo (Wilcoxon Rank Sum Test, $p < 0.05$; Supplementary Fig. 2a). BA20/37 was commonly different using either measurement. Even though effect sizes were small between cases and controls for all the ROIs analyzed (Cohen's d; $d < 0.3$) in agreement with other reports[36,37], we observed consistency between fALFF and ReHo (Spearman Rank Correlation, rho = 0.46; Fig. 2b). These data reflect subtle, yet replicable functional activity measurements linked to ASD calculated by two rs-fMRI measurements. However, because the differences between cases and controls using rs-fMRI were minimal with a small to null contribution to the analysis, we assessed the rs-fMRI—gene expression relationship using the control subject ReHo and fALFF values. We first assessed the complementarity of these two rs-fMRI measurements in controls. There was a significant correlation between fALFF and ReHo values across individuals in each singular ROI analyzed (Spearman's rho = 0.58, $p < 2.2 \times 10^{-16}$; Supplementary Fig. 3a, b). These data further confirmed the complementarity of these two distinct measurements of rs-fMRI values. To understand ASD pathophysiology in the context of brain activity and gene expression, we spatially matched RNA-seq data[33] from 11 cortical areas for a total of 360 tissue samples from cases (ASD) and 302 control samples (CTL) (Supplementary Fig. 4a). The variance explained by technical and biological covariates was accounted for and removed before further analyses (see "Methods" and Supplementary Fig. 4b).

Identification of genes differentially correlated with rs-fMRI between ASD and controls. We sought to identify genes with correlated expression to imaging measurements across regional

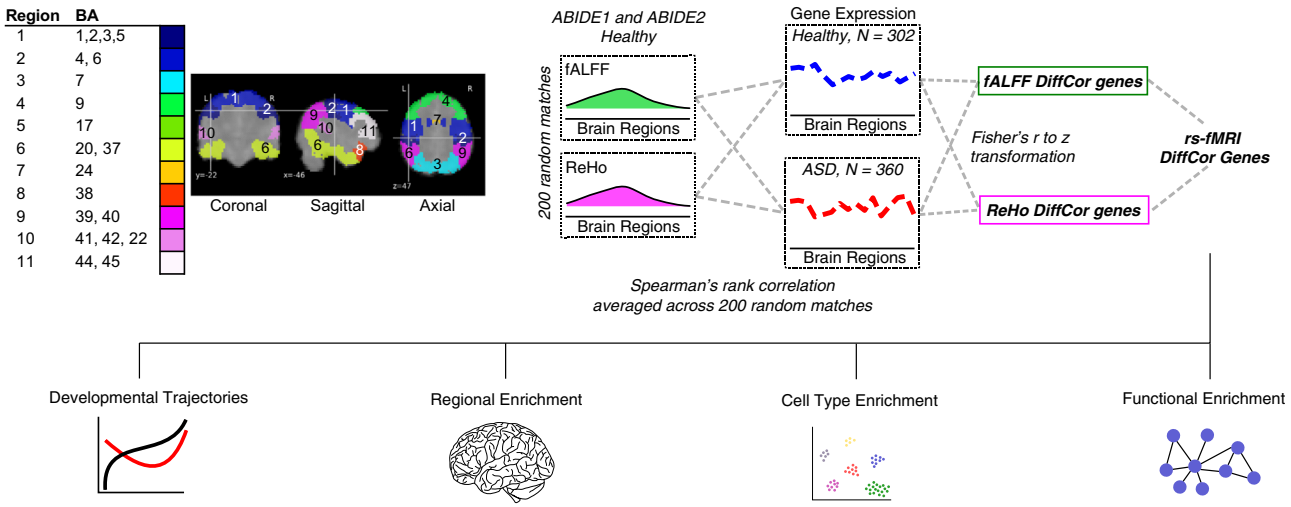


Fig. 1 Flowchart of the analytical framework and pipeline. Gene expression values were obtained for 11 cortical regions from individuals diagnosed with autism spectrum disorder (ASD) and demographically matched neurotypical individuals (CTL: control). For fMRI, the ABIDE I and II datasets were used to select ASD and CTL individuals that demographically matched with the gene expression cohort. Regional Homogeneity (ReHo) and fractional amplitude of low frequency oscillations (fALFF) were calculated from these individuals for the 11 cortical regions. 200 subsampled matches were selected from CTL individuals. Spearman's rank correlation was used to infer a correlation between gene expression and fMRI separately for both fMRI measurements. For differential correlation (DC genes) between ASD and CTL, the mean correlation values of matches were transformed to a z score (Fisher's r to z transformation) and statistics combined (Fisher's combined method). A gene was called differentially correlated if the differential correlation *p*-value was less than 0.01 (DiffCorP < 0.01) and FDR < 0.05 in CTL. The final list consisted of genes that are differentially correlated using both fMRI measurements.

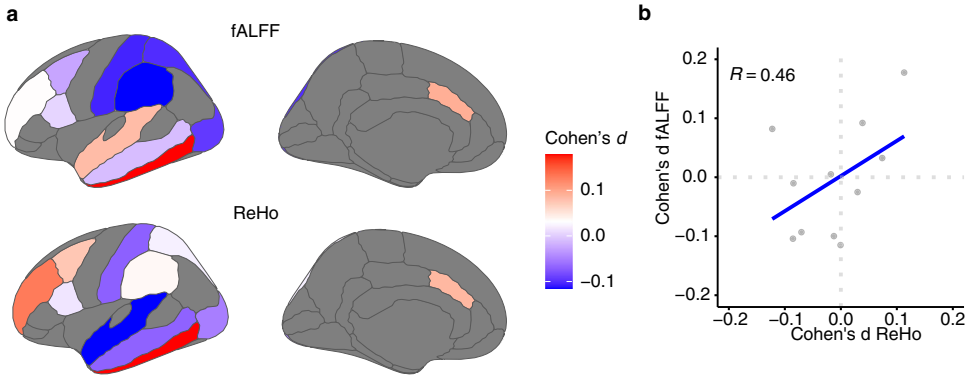


Fig. 2 Imaging differences between ASD and CTL. a Differences between ASD and CTL calculated by Cohen's d (effect sizes) derived from ASD–CTL comparison for both rs-fMRI measurements across the ROIs analyzed. **b** Scatter plot depicting the spatial correlation between Cohen's d values of fALFF and ReHo. Each dot corresponds to the ROI analyzed.

rs-fMRI values. To do this, we used Spearman's rank correlation between mean regional values of fALFF or ReHo and regional gene expression. To take advantage of the entire ABIDE dataset, we randomly sampled from the ABIDE dataset 200 times, and correlated each sample with the genomic data (see "Methods"). We defined genes correlated with ReHo and/or fALFF in both controls and ASD (Supplementary Fig. 5a and Supplementary Data 2). Using a Fisher r-to-z transformation, we assessed the significance of the difference between ASD and CTL correlations in both fALFF or ReHo values. We next used a Fisher's method to combine the resultant *p*-values defining 415 differentially correlated genes (DC genes; Diff Cor $P < 0.01$, CTL FDR < 0.05; Fig. 3a; "Methods"). DC genes showed a high proportion of positively correlated genes with similar correlation coefficients in both measurements (59.8%; Fig. 3b; Supplementary Fig. 5b). We next examined the effect sizes and the relationship between fALFF and ReHo values (Fig. 3c; Supplementary Fig. 5c). For a $P < 0.01$, DC genes showed an effect size larger than 1.8, resulting in ~3% of the gene expressed in our data (Fig. 3c). Among the genes with highest effect size, we found *FILIP1*, which encodes a filamin A binding protein important for cortical neuron migration and dendrite morphology[38–40], and *GABRQ*, a gene encoding a GABA receptor subunit highly expressed in von Economo neurons[41,42]. In addition, the effect sizes of the DC genes calculated with fALFF and ReHo strongly correlate (Spearman Rank Correlation, rho = 0.54, $p < 2.2 \times 10^{-16}$; Supplementary Fig. 5c), further confirming the reproducibility of the DC genes in two different rs-fMRI measurements.

Next, we compared the genes we identified with genes linked with rs-fMRI values from independent studies[14,15]. Because these earlier studies analyzed only healthy individuals, we first compared the genes correlated only in CTL with the ones previously reported. We found that previously fMRI-correlated genes were significantly enriched in CTL genes, revealing reproducibility of fMRI-correlated genes despite variation in cortical regions and type of fMRI measurements (Wang et al.:

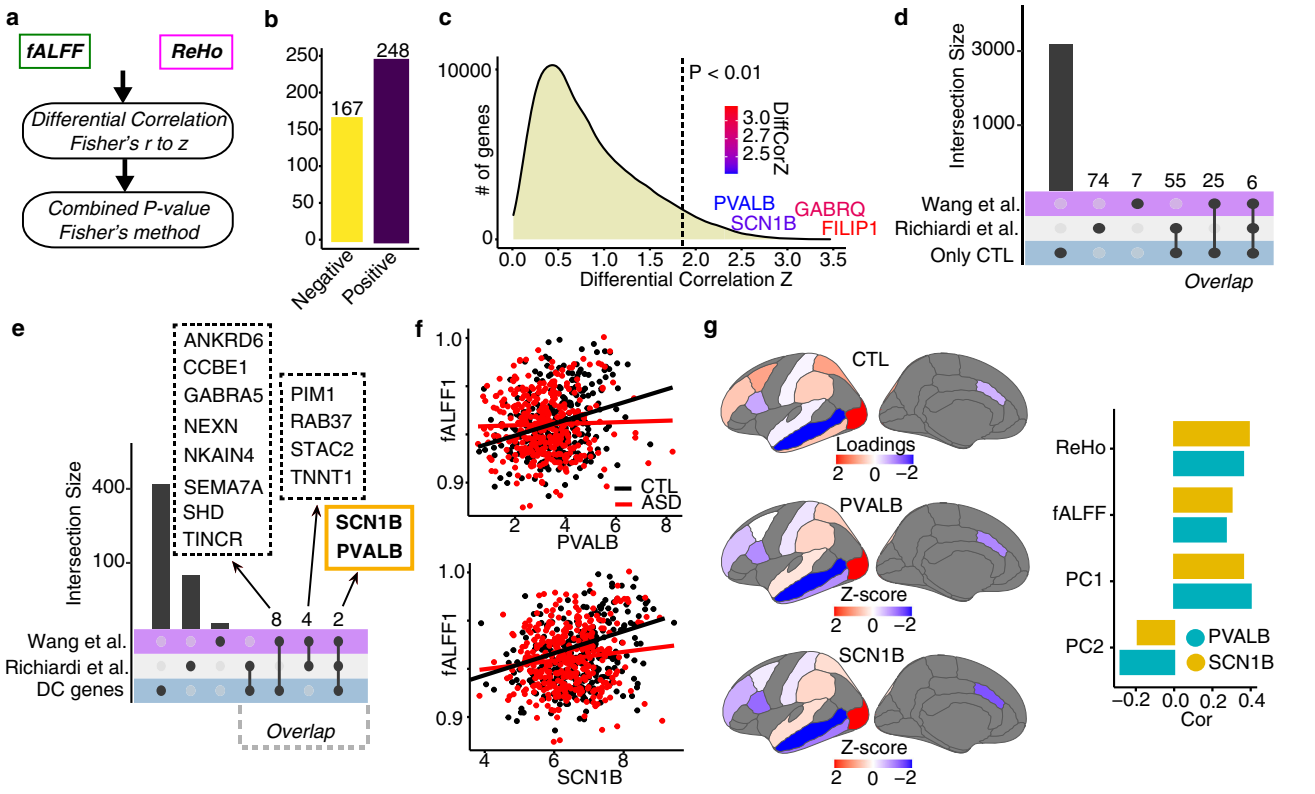


Fig. 3 Overview of differentially correlated genes. a Schematic workflow to define differentially correlated genes. **b** Barplot depicting the number of differentially correlated genes positively (purple) and negatively (yellow) correlated with CTL. **c** Density plot depicting the distribution of differential correlation effect sizes (z). Line corresponds to the *p* value cutoff used. Genes of interest are highlighted with a gradient color that reflects the relative effect size. *P*-value threshold for Differential Correlation analysis is shown based on Fisher's combined *P*-value. **d** Upset plot showing the intersection between the genes significantly correlated only in CTL for both fALFF and ReHo and two previous rs-fMRI—gene expression studies using only data from neurotypical individuals. **e** Upset plot showing the intersection between DC genes and two previous rs-fMRI—gene expression studies using only data from neurotypical individuals. **f** Scatterplots showing the relationship (Spearman's rank correlation) between rs-fMRI (*Y*-axis) and gene expression for two candidate genes (*X*-axis) in CTL and ASD. **g** Gradient of CTL expression (PC1), PVALB, and SCN1B gene expression. Bar plot depicts the correlation between PVALB and SCN1B gene expression with ReHo, fALFF, expression PC1, and expression PC2.

odds ratio (OR) = 17.3, FDR = 2.5×10^{-15}, Richiardi et al.: OR = 3.2, FDR = 1.5×10^{-10}; Fig. 3d, Supplementary Fig. 5d). Among them, 6 genes overlapped in all previous studies, and 2 out of the 6 genes were also among the DC genes between ASD-CTL (*PVALB and SCN1B*; Fig. 3d–f). These two genes are particularly compelling as *SCN1B*, which encodes a beta-1 subunit of voltage-gated sodium channel, is a highly expressed gene in fast-spiking parvalbumin (PVALB+) cortical interneurons, which play a key role in neuronal networks, and whose oscillations are linked with ASD[43–46]. Because PVALB gene expression has a rostrocaudal axis gradient[47], we next evaluated the spatial distribution of both candidates' gene expression in the ROIs. We found that both PC1 (principal component 1), as well as SCN1B and PVALB, displayed differences in the rostrocaudal axis (*PVALB* ~ PC1, *rho* = 0.41; *SCN1B* ~ PC1, *rho* = 0.37), with higher expression in caudal cortical regions (Fig. 3g). These genes were similarly correlated with rs-fMRI measurements, (*PVALB* ~ fALFF, *rho* = 0.32; *SCN1B* ~ fALFF, *rho* = 0.33; *PVALB* ~ ReHo, *rho* = 0.38; *SCN1B* ~ ReHo, *rho* = 0.42), but these correlations are affected by ASD status (Fig. 3f). Overall, these results identify many brain activity-related genes and imply that some of the high confidence genes such as *PVALB* and *SCN1B* support brain activity affected in ASD.

Differentially correlated genes have specific developmental trajectories. Although we identified DC genes across all samples with a median age of 22 years old, we asked how DC genes compare between CTL and ASD across development given that autism is a neurodevelopmental disorder. We leveraged the transcriptomic dataset from this study to detect whether DC genes follow a specific developmental trajectory in individuals with ASD compared with CTL subjects (see "Methods"). We identified three main clusters of DC genes: one highly expressed in adults (Adult), one highly expressed in early development (EarlyDev), and one with relatively stable trajectory throughout development (Stable) (Fig. 4a). Interestingly, genes in the Adult cluster are upregulated until adulthood in neurotypical individuals but this upregulation is delayed in individuals with ASD. In contrast, the genes in the Stable and EarlyDev clusters follow a similar trajectory in both groups (Fig. 4a and Supplementary Fig. 6a). Because each region differs by sample size, we used a subsample approach and recalculated the developmental trajectories. We found that differences in sample size between regions did not affect the overall result (Supplementary Fig. 6b). We additionally assessed gene expression patterns in BrainSpan dataset[48] generated using healthy brain tissue (0–40 years old) and found similar trajectory patterns with the Adult cluster displaying immediate upregulation until adulthood similar to CTL in our dataset (Supplementary Fig. 6c).

Next, we sought to understand the functional properties of the genes associated with these developmental trajectories. Overall, we found enrichments for transporter activity, ion channel activity, and DNA-binding activity which are crucial for proper

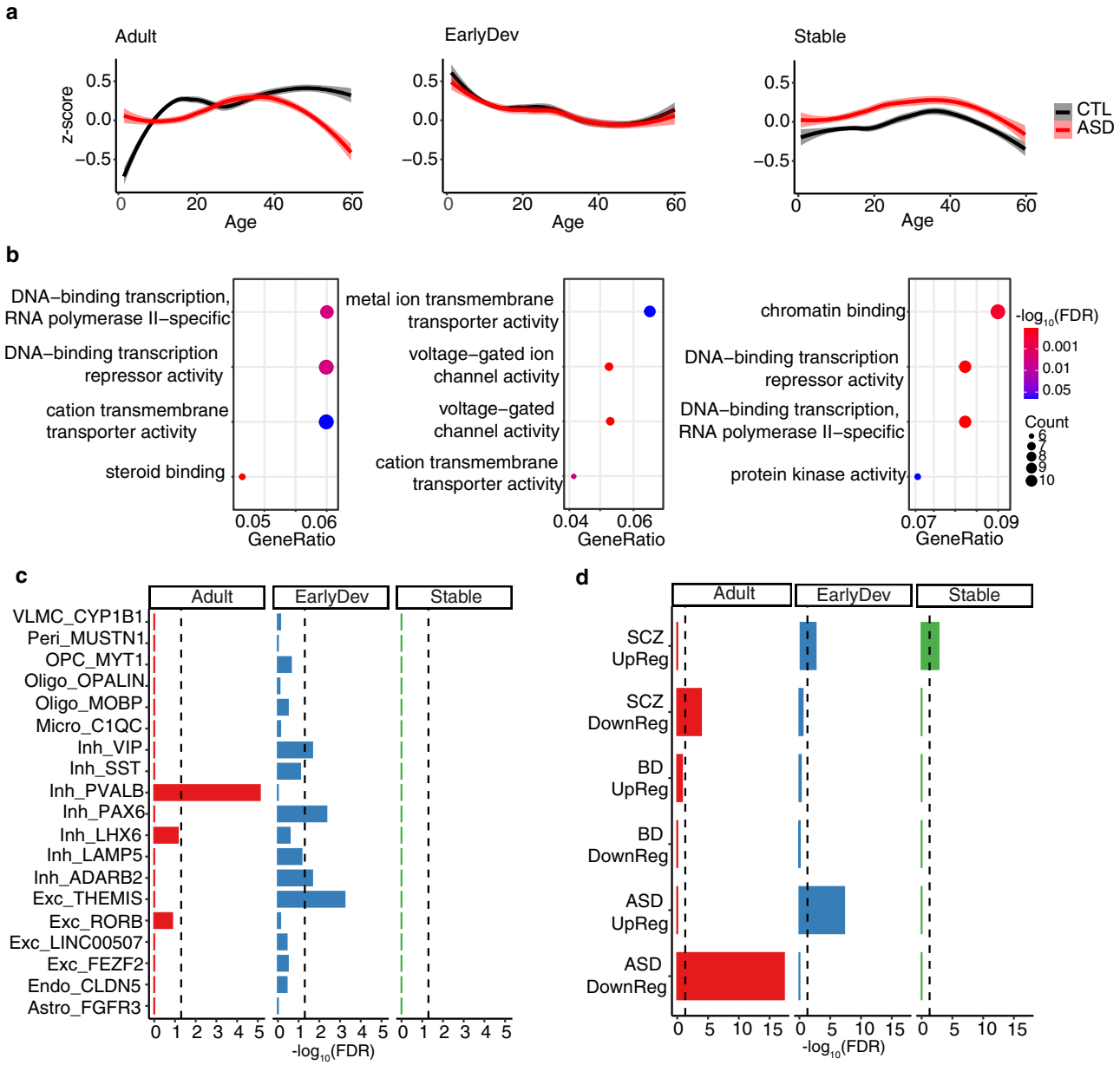


Fig. 4 Differentially correlated genes in individuals with ASD are important for brain development. a The 415 genes were clustered in three developmental time groups: adult (Adult), early development (EarlyDev), stable (Stable). *X*-axis represents developmental time. *Y*-axis represents the expression based on human brain developmental time. Loess regression with confidence intervals depicts the overall distribution. Smooth curves are shown with 95% confidence bands with relative trendlines. **b** Bubblechart representing the functional enrichment for modules associated with developmental time. *Y*-axis corresponds to the odds ratio, *X*-axis corresponds to the $-\log_{10}$(FDR). **c** Bar plot depicting the $-\log_{10}$(FDR) of the Fisher's exact test enrichment between developmental clusters and cell-type markers (*Y*-axis) from single-nuclei RNA-seq from multiple brain regions ("Methods"). VLMC=vascular leptomeningeal cell, Peri=pericytes, OPC=oligodendrocyte precursor cell, Micro=microglia, Inh=inhibitory neuron, Exc=excitatory neuron, Endo=endothelial cell, Astro=astrocyte. **d** Bar plot depicting the $-\log_{10}$(FDR) of the Fisher's exact test enrichment between developmental clusters and genes differentially regulated in ASD, schizophrenia (SCZ), and bipolar disorder (BD) (*Y*-axis) from an independent study ("Methods").

development and have been repeatedly implicated in ASD[25,49,50] (Fig. 4b and Supplementary Data 2). However, these enrichments were not distinct for a single developmental trajectory. In contrast, enrichment in steroid binding was only present in the Adult cluster with relatively high significance (Fig. 4b). Steroid binding was mainly driven by enrichment for estrogen receptor (*ESRRG*, *ESRRA*) and nuclear glucocorticoid receptor (*NR3C1*, *NR3C2*) genes. We find this intriguing given that steroid levels are altered in autistic individuals even in early development[51,52] and the ratio of sexes was very similar in our dataset (CTL female ratio: ~0.18, ASD female ratio: ~0.18). Thus, our results indicate that altered steroid biology in ASD is linked to brain activity changes across cortical regions.

To understand the cell type-specific properties of the rs-fMRI genes, we performed enrichment for gene expression data derived from single-cell RNA-seq studies[41] ("Methods"). We observed that the genes in the Adult cluster were highly enriched for parvalbumin (*PVALB*) expressing interneurons whereas EarlyDev genes were enriched for excitatory neurons. No cell-type enrichment was detected for the genes in the Stable cluster (Fig. 4c). Because PVALB expression follows an anterior to posterior regional gradient[47], we imputed $PVALB^{+}$ interneurons

abundance for each of the region analyzed (see "Methods"). We conducted a deconvolution analysis using MuSiC[53], which allows the inference of relative cell-type abundance in bulk data. Single-nuclei RNA-seq from a multi-cortical region data was used to infer cell-type proportions[54]. We estimated the relative cell-type abundance by subjects and brain regions. As expected, the *PVALB*+ interneurons fractional abundance was higher in posterior regions compared with anterior regions (Supplementary Fig. 6d; Supplementary Data 3). Notably, the relative abundance of these interneurons was significantly reduced in individuals with ASD in posterior regions such as BA7 and BA17. These data indicate that our results are potentially driven by *PVALB*+ interneurons regional abundance further demonstrating the important role of these interneurons in ASD pathology. We next investigated the association of developmental gene clusters with genomic data from brain disorders including ASD[55]. The Adult cluster is enriched for downregulated genes in individuals with ASD while the EarlyDev cluster is enriched for upregulated genes in individuals with ASD (Fig. 4d). This result was relatively specific to ASD as similar gene lists from individuals with schizophrenia or bipolar disorder showed little to no enrichment (Fig. 4d). This result was further confirmed using modules of co-expressed genes dysregulated in such disorders (Supplementary Fig. 6e). Together, these data extend the emerging picture of molecular pathways disrupted in ASD corresponding to rs-fMRI measurements[14,15,30,56].

The relationship of rs-fMRI and gene expression is altered at the brain region level. Due to the limited number of samples per ROI, we were not able to assess the association between brain activity and gene expression at a regional level. We overcame this limitation with a leave-one-region out (LoRo) approach inferring the contribution of each region in our results (see "Methods"). Briefly, by leaving one of the 11 regions out at a time, we were able to test whether the differential correlation was affected by one region or several specific regions. We calculated the z-score from the z-to-r Fisher transformation from each analysis and combined with the Fisher's method (Supplementary Data 4). We observed a significant contribution from the primary visual cortex (BA17), temporal cortex (BA20/37, BA38), parietal cortex (BA39/40), and motor cortex (BA4/6) (Fig. 5a). We next examined the enrichment of regional differential expressed genes (DEG; FDR < 0.05, $|\log_2(FC)| > 0.3$; see "Methods") between ASD-CTL in DC genes. We explored whether any of the developmental gene clusters were enriched for specific regional DEG. Interestingly, we found the highest enrichment of Adult and EarlyDev cluster genes in cortical areas associated with vision and proprioception (BA17 and BA7) (Fig. 5b). Taken together, these results support the emergent role of the visual cortex in ASD pathophysiology[57,58].

Discussion

Assessing gene expression in the brain permits a relevant examination of how biological pathways might be altered in the tissue of interest and connected to genetic predispositions. Moreover, functional imaging provides an important window into phenotypes associated with mental illness. Combining these approaches can help begin to bridge the gap between genes and behavior. Indeed, previous work has demonstrated a correspondence between human brain gene expression and functional connectivity as measured by fMRI[14,15,30]. However, the studies using brain gene expression only used neurotypical populations. Local brain activity measures such as ReHo and fALFF can assess neuronal connectivity and activity. When restricted to a specific image acquisition site and age range (e.g., children or adolescents), previous studies using ReHo and fALFF have found significant differences between CTL and individuals with ASD in cortical regions but in different brain regions and directions[59–63]. However, protocol variability across sites can induce inconsistent findings in functional connectivity[64]. A quantitative meta-analysis indicated that only connectivity between the dorsal posterior cingulate cortex and the right medial paracentral lobule consistently differs between individuals with ASD and CTL subjects across sites and ages;[65] however, these regions were not available for tissue sampling in this study. Structural imaging studies have also indicated the difficulty in finding differences between individuals with ASD and CTL subjects when no age restriction is imposed[66–68]. In contrast to these age and site-restricted reports, our study includes ages from 5 to 64 years and data from 37 sites whose differences are retrospectively normalized and such differences with previous studies likely underlie our finding of few significant differences in brain activity between cases and controls.

We speculated that the expression of genes and their association with brain activity may underscore their potential relevance for any functional brain activity that is disrupted in ASD. In line with this, our results suggest that genes typically associated with rs-fMRI lose their association when ASD genomics are included. These genes are important for brain development, regional differences, and excitatory/inhibitory identity. As previously reported, GABAergic signaling is disrupted across mouse models of ASD[69] and GABA interneurons have a key role for cortical circuitry and plasticity[70–72]. Interestingly, genes highly expressed in the Adult gene cluster that are significantly associated with brain activity are overrepresented in a subpopulation of inhibitory interneurons expressing parvalbumin. In contrast, genes highly expressed in early development are overrepresented in excitatory neurons. In line with the role of parvalbumin neurons in normal brain circuitry and oscillations[70,73,74], this distinct association might underscore the excitation-to-inhibition ratio imbalance in autism. Moreover, the relative abundance explained by the Adult genes of *PVALB*+ interneurons is significantly decreased in individuals with ASD. Because spatial *PVALB*+ expression covaries with rs-fMRI across regions[47], we hypothesized that the relative increased abundance of these interneurons in the visual cortices and conversely the reduction shown in individuals with ASD explains the differential correlation in the specific subset of Adult genes. Therefore, these results further underscore the important role of parvalbumin interneurons in autism.

We hypothesize that genes severely dysregulated in autism such as *SCN1B*, *KCNAB3, FMN1, or VAMP1* might additionally contribute to the excitation-to-inhibition ratio affecting normal network function and circuitry. Moreover, previous studies have shown that inhibitory neurons control visual response precision with increased activity leading to a sharpening of feature selectivity in mouse primary visual cortex[58]. Additionally, multiple lines of evidence have indicated that individuals with ASD show slower switching between images in binocular rivalry[57,75–77]. Here, we provide evidence that regional brain expression influences the association between rs-fMRI values and gene expression, with the visual cortex as the major contributor to the variance explaining the rs-fMRI—gene association. Therefore, these results contribute to a consistent emerging role of the visual cortex in ASD pathology. However, because subjects who underwent fMRI measurements might not have had uniform instructions (or resultant behavioral compliance) to keep their eyes open or closed, it is possible that the visual cortex data could be influenced by such behavior.

Finally, any functional interpretation of the genes identified should be made with caution. Here, we assessed the relationship between gene expression and rs-fMRI across cortical areas based on correlation, which is not necessarily evidence of causation.

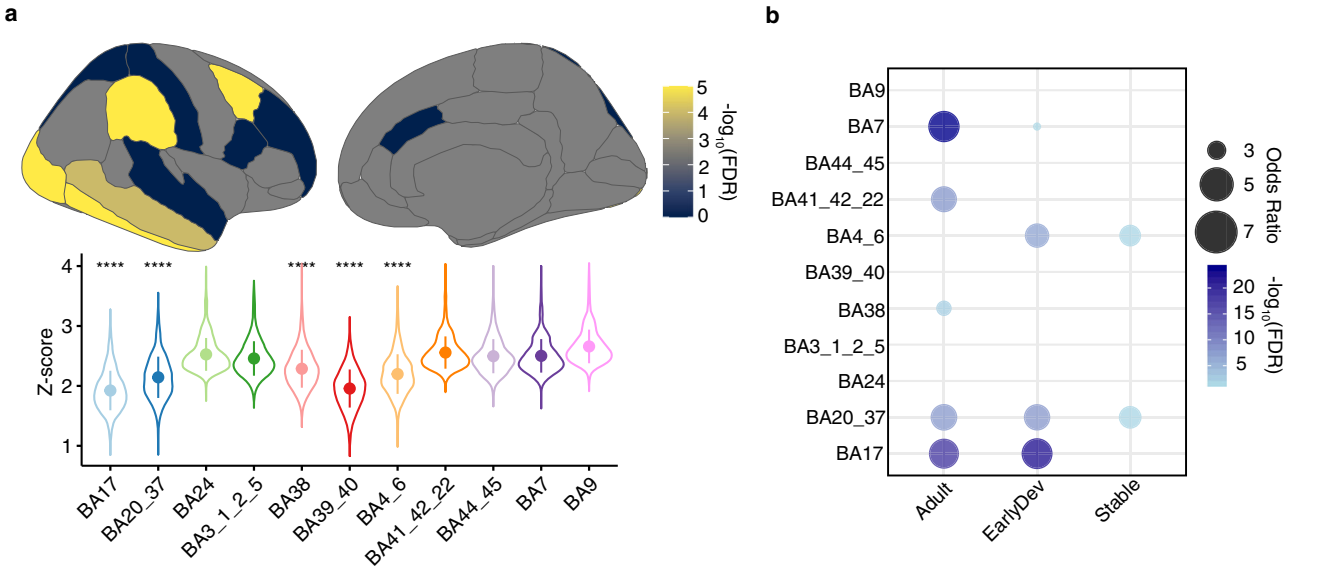


Fig. 5 Leave one region out (LoRo) analysis underscores the importance of specific brain regions. a Brain visualizations and violin plots depicting the contribution of each ROI in the differential correlation after the specific region was removed (LoRo). Brain visualizations represent the $-\log_{10}$(FDR) of the comparative analysis between ROIs after LoRo analysis (One-sided Wilcoxon's rank sum test). Violins represent the Z-score (*Y*-axis) of differential correlation after the specific region (*X*-axis) was removed. **** corresponds to a significant lower Z-score compared with other regions ($p < 0.001$; One-sided Wilcoxon's rank sum test). Dots represent the mean Z-score for the specific Brodmann area. Lines represent the standard deviation (SD). $N = 415$ genes from independent analysis. Exact *P*-value: BA17 $p < 2e^{-16}$, BA20_37 $p < 2e^{-16}$, BA38 $p = 6.8e^{-05}$, BA39_40 $p < 2e^{-16}$, BA4_6 $p < 2e^{-16}$. **b** Bubblechart with $-\log_{10}$(FDR) and Odds Ratio from Fisher's Exact test representing enrichment between developmental groups and genes differentially expressed in each region. *X*-axis shows abbreviations for each region. *Y*-axis represents each developmental cluster identified.

Moreover, functional imaging analysis of the ROIs used does not show large differences between neurotypical and individuals with autism. The primary limiting factor in spatial resolution and brain coverage is driven by the restricted tissue sampling/availability from postmortem disease cohorts. Larger sample sizes may in the future allow for a more detailed investigation of these genes at the regional level increasing both specificity and sensitivity. Additionally, candidate genes should be further analyzed in vivo using model systems to provide a basic understanding of their effects on brain activity. In conclusion, we have established that autism pathology significantly impacts the relationship between gene expression and functional brain activity. Our results uncovered genes that are important for excitation-to-inhibition ratio balance and visual cortex function. These results provide molecular mechanisms for future studies relevant to understanding brain activity in individuals with autism.

Methods

All research in this manuscript complies with all relevant ethical regulations. This study was approved by the UT Southwestern Medical Center Institutional Review Board.

fALFF and ReHo. To provide image-derived phenotypes (IDPs) for each subject in the ABIDE cohort, regional measures of brain function were computed including the fractional amplitude of low-frequency fluctuation (fALFF; https://fcp-indi.github.io/docs/latest/user/alff.html?highlight=falff) and regional homogeneity (ReHo; https://fcp-indi.github.io/docs/latest/user/reho). Supplementary Fig. 1 illustrates the main processing steps of the image analysis pipeline.

Imaging materials. This study used resting-state fMRI from the 916 ASD and 1067 CTL subjects of both ABIDE I and ABIDE II[32,78]. Details of the pulse sequence parameters used in this data acquisition are provided in Supplementary Data 1. After the removal of subjects with image artifacts, high head movement, or poor MNI152 coregistration, we analyzed the data from the remaining 710 ASD (79% male), and 606 CTL (87% male) subjects, whose age ranges from 5 to 64 years.

fMRI preprocessing. All data from each subject were preprocessed consistently—as described below—and are illustrated in Supplementary Fig. 1. The 3dSkullStrip method from the brain extraction tool (BET) was applied to remove skull and non-brain tissue[79]. The first 5 volumes were censored to allow for MRI scanner dynamic instability to settle. To correct for head movement, volume realignment was applied frame by frame, to register each volume to the mean volume with an affine transformation. Slice timing correction was applied to ensure volume slices align temporally.

Images were processed with a generalized linear model (GLM) to regress out: (1) global signal fluctuation, (2) physiological noise represented by white matter and CSF fluctuation, (3) fluctuation correlated with the 6 original affine head motion parameters (X/Y/Z/pitch/roll/yaw), (4) their first derivatives, squares, and squared derivatives, and 5) noise fluctuations captured from five components from aCompCor[80]. Scrubbing was applied to remove frames with a Jenkinson framewise displacement (FWD) > 0.5 mm, and subsequently replaced with an interpolated frame. ReHo was calculated with scrubbed data; however, ALFF and fALFF were not calculated with scrubbed data because the framewise removal and alteration disrupts the temporal structure precluding Fourier transform-based approaches[81].

For subjects with multiple fMRI scans, the scan with the lowest head motion, measured by mean FWD, was selected for analysis. For each resulting subject scan, a subject was excluded if their scan had excessive head motion. Specifically, scans meeting at least one of these three requirements were removed: (1) mean FWD > 0.30 mm, (2) greater than 50% of frames being scrubbed, or (3) scans with outlier mean, 1st, 2nd, or 3rd quantile DVARS values. DVARS was defined as the root mean square of the temporal change of the fMRI voxel-wise signal at each time point[82,83]. The package CPAC v1.8.0 was used for fMRI pre-processing including head motion correction, scrubbing, and nuisance regression.

Calculation of fALFF and ReHo. We computed fALFF and ReHo from the resting-state fMRI using C-PAC (v1.8.0)[84] in native subject space, resulting in a volumetric map of fALFF and a map of ReHo for each subject. fALFF[34] quantifies the slow oscillations in brain activity that form a fundamental feature of the resting brain. ALFF is defined as the total power within the low-frequency range (0.01–0.1 Hz) and forms an index of the intensity of the low-frequency oscillations. The normalized ALFF known as fALFF is defined as the power within that low-frequency range normalized by the total power in the entire detectable frequency range. fALFF characterizes the contribution of specific low-frequency oscillations to the entire frequency range[34]. To increase the signal to noise ratio by removing high-frequency information, we spatially smoothed each derivative map with a Gaussian kernel. ReHo[35] aims to detect complementary brain activity manifest by clusters of voxels rather than single voxels as in fALFF. ReHo evaluates the similarity of the activity time courses of a given voxel to those of neighboring voxels using Kendall's coefficient of concordance (KCC)[85] as the index of time series similarity. This

measure requires the cluster size as an input to define the size of the neighborhood. In this study, we used a cluster size of 27 voxels. The 26 neighbors of a voxel, *x*, are those within a 3 x 3 x 3 voxel cube centered on voxel *x*. The similarity of the activation time courses between each voxel, *x*, and its 26 nearest neighbors was calculated using: $W_x = (\sum(R_i)^2 - n(\bar{R})^2)/(\frac{1}{12}K^2(n^3 - n))$. W_x is the KCC for voxel *x* and ranges from 0 to 1, representing no concordance to complete concordance. R_i is the rank sum of the *i*th time point. R is the mean value over the R_i's. K is the cluster size for the voxel time series (here $K = 27$). n is the total number of ranks.

Registration. The mean processed fMRI image was nonlinearly registered directly to an EPI template in MNI152 space using the symmetric normalization (SyN) non-linear registration method of the ANTs (v2.3.5) package[86,87]. The resulting composite transform was then applied to both the fALFF and ReHo maps to provide derivative maps in normalized MNI152 space. We used EPInorm-based registration as it better accounts for nonlinear B0 field inhomogeneities at the air to tissue interfaces[88,89]. Supplementary Figure 7 illustrates the improvement EPInorm-based registration has over more the commonly applied T1norm based registration. In this study, EPInorm registration yielded more accurate spatial normalization of the brains to the standard atlas space in which regional values are computed. Regions of improved registration included the sinuses which present air/tissue interfaces that induce non-linear distortions which are properly handled through EPInorm co-registration. EPInorm registration also had a substantially lower standard deviation around the brain periphery across the 1316 subjects assessed.

Lastly, subjects with poor EPInorm registration[88] (discussed below) were removed. Specifically, mis-registration was identified through a combination of manual inspection and through the detection of scans with an outlier number of misaligned brain-masked voxels using the interquartile range (IQR) outlier test[90].

Segmentation. In this study, we adapted the Brodmann atlas publicly available through MRICron (v1.0.9) to form the 11 multi-area regions from which tissue samples were drawn from matched donor brains. Supplementary Data 1 illustrates how we combined Brodmann areas to generate 11 regions that correspond with the RNA sequence data. We used the resulting 11 region atlas to assign a region label (parcellate) to each voxel in the fALFF and ReHo maps to enable computation of the mean regional fALFF and ReHo values for all subjects.

Site correction. We accessed publicly available ABIDE data across 30 different sites. These sites used MRI devices from different manufacturers (Siemens, Philips, GE) and used different MRI pulse sequences and participant protocols, which can cause differences in the absolute value of the fMRI acquired and can affect fALFF and ReHo values (Supplementary Data 1). As the mean fALFF and ReHo varied between sites, we applied a correction to minimize site differences. To suppress site differences, the difference between the cohorts mean regional value and each site's mean regional value was calculated. This regional difference was then subtracted from each region value for all subjects belonging to the corresponding site.

Derivative map normalization. To provide better inter-subject comparisons, we normalized regional fALFF and ReHo values to the weighted mean, weighted by the number of voxels for each region, over all of the regional values for each subject. To reduce the impact of confounders, we regressed out age, site, and sex using a linear model.

RNA-seq processing and analysis. Quality control was performed using FastQC (v.0.11.9). Reads were aligned to the human hg38 reference genome using STAR[91] (v.2.5.2b). Picard tool was implemented to refine the quality control metrics (http://broadinstitute.github.io/picard/) and to calculate sequencing statistics. Gencode annotation for hg38 (v.25) was used for reference alignment annotation and downstream quantification. Gene level expression was calculated using RSEM[92]. Dup15q individuals were removed from the initial data[33]. Technical replicates were collapsed by the maximum expression value and maximum RNA integrity value. A total of 302 Control and 360 ASD were used for the final analysis. Supplementary Figure 8 represents the pairwise comparison of demographics from the RNA-seq and rsfMRI datasets. Supplementary Data 1 provides details on all the covariates. Only protein-coding genes were considered. Counts were normalized using counts per million reads (CPM) with the *edgeR* (v3.32.0) package in R[93]. Normalized data were log2 scaled with an offset of 1. Genes were considered expressed with $\log_2(\text{CPM} + 1) > 0.5$ in at least 80% of the subjects. Normalized data were assessed for effects from known biological covariates (*Sex*, *Age, Ancestry,* and *PMI*), technical variables related to sample processing (*Batch, BrainBank, RNA Integrity value (RIN)*) and technical variables related to sequencing processing based on PICARD statistics (https://broadinstitute.github.io/picard/).

We used the following sequencing covariates:
picard_gcbias.AT_DROPOUT, star.deletion_length, picard_rnaseq.PCT_INTERGENIC_BASES,
picard_insert.MEDIAN_INSERT_SIZE, picard_alignment.PCT_CHIMERA
Spicard_alignment.PCT_PF_READS_ALIGNED, star.multimapped_percent,
picard_rnaseq.MEDIAN_5PRIME_BIAS, star.unmapped_other_percent,
picard_rnaseq.PCT_USABLE_BASES, star.uniquely_mapped_percent.

Residualization was applied using a linear model. All covariates except Diagnosis, Subjects and Regions were taken into account:

mod <- lm(gene expression ~ Sex + Age + Ancestry + PMI + Batch + BrainBank + RIN + seqCovs).

This method allowed us to remove variation explained by biological and technical covariates.

Adjusted expression was calculated by extracting the residuals per each gene and adding the mean of the gene expression: *adjusted gene expression <− residuals(mod) + mean(gene expression)*

Adjusted CPM values were used for rs-fMRI—gene expression correlation and resultant visualization.

fMRI-gene expression correlation analysis. We performed Spearman's rank correlation between the mean regional values of fALFF and ReHo and the regional gene expression across the 11 cortical areas analyzed. To define fMRI-gene expression relationships, we used random subsampling (200 times) of neurotypical individuals from the ABIDE I and II datasets. We matched the number of subjects per each cortical area (e.g., 25 ASD subjects for BA17). We performed correlation across the regions using all 11 areas matching with the gene expression dataset and averaged Spearman's rank statistics over the 200 subsamples. *P*-values from Spearman's rank statistics were adjusted by Benjamini–Hochberg FDR. Differential Correlation analysis was performed comparing the resulting Rho from neurotypical individuals to individuals with ASD for each gene using the *psych* (v2.0.12) package in R. We combined the resultant Differential Correlation *p*-values and effect sizes using a Fischer's combination test in R. Significant results are reported at FDR < 0.05 for neurotypical individuals' statistics and *P*-value of combined differential correlation at $p < 0.01$.

Leave-one-region out (LoRo) analysis. We performed the same subsampling approach followed by differential correlation analysis as described above leaving one region out at the time. This method allowed us to determine the effect of each region in the resultant *z* from the differential correlation analysis between healthy individuals and autistic individuals. Next, we calculated the contribution of each region based on a principal component analysis using the resultant *z*-values. We visualized resultant contributions in a multi-dimensional plot.

Developmental analysis. The identification of gene clusters with different developmental trajectories was performed on DC genes using all subjects except for individuals above 60 yr as they were represented only in the ASD group.

We applied residualization as previously described removing the age from the covariates.

mod <− lm(gene expression ~ Sex + Ancestry + PMI + Batch + BrainBank + RIN + seqCovs).

Then, we scaled gene expression and divided genes into three clusters according to the scaled expression values of healthy subjects only, using the *Kmeans* function from the *amap* (v0.8) package in R. We plotted the developmental trajectories using the loess regression and *ggplot2* (v3.3.2) package in R. To make loess regression computationally possible, 8000 data points were randomly sampled. Repeated samplings yielded very similar patterns. We made no adjustments for developmental time points and the *x*-axis directly represents the age of the subjects. We annotated clusters based on visual inspection of their trajectory. To subsample diagnosis-region groups (e.g., ASD BA17 samples), we determined the diagnosis-region group with minimum number of samples and randomly subset other groups to that number. Then we plotted expression values with loess regression as before.

To assess the significance of trajectories, we compared gene expression between age brackets of 5 years using t-test (One-tailed. Greater expression for Adult (e.g Ha: 0–5 < 5–10) and less expression for EarlyDev (Ha: 0–5 > 5–10) with increasing age).

The BrainSpan dataset[48] was downloaded from www.brainspan.org (normalized matrix: "RNA-Seq Gencode v10 summarized to genes"). Data were then log2 transformed ($\log2(\text{data} + 1)$). To match with the current study, the following brain regions were removed: AMY, OFC, Ocx, URL, DTH, CB, CBC, MD, STR, and HIP. For each gene, the expression values were z-transformed across samples. To understand expression pattern across ages, samples were divided into age groups per 5 years. Only postnatal samples were kept to match with the current study.

Allen single nuclei RNA-seq analysis. Multi-Region snRNA-seq[41] (MTG, V1C, M1C, CgGr, S1C, A1C) was from the Allen Brain Map portal (https://portal.brain-map.org/atlases-and-data/rnaseq). Briefly, data was analyzed using Seurat[94] (v3.9.9). Data was subsetted by removing nuclei with >10,000 UMI and >5% of mitochondrial gene expressed. Published cell-type annotations included in the metadata were used for downstream analyses. We identified cell-type markers using *FindMarkers* function based on Wilcoxon-rank sum test statistics. Markers were defined by Percentage of Cells expressing the gene in the cluster >0.5, FDR < 0.05 and $|\log_2(\text{FC})| > 0.3$.

Functional enrichment. We performed the functional annotation of differentially expressed and co-expressed genes using ToppGene[95]. We used the GO and KEGG databases. Pathways containing between 5 and 2000 genes were retained. We applied a Benjamini–Hochberg FDR ($P < 0.05$) as a multiple comparisons adjustment. Brain expressed genes (Brainspan, $N = 15585$) were used as background.

Gene set enrichment. We performed gene set enrichment for neuropsychiatric DGE[55], neuropsychiatric modules[55], and cell-type markers[41] using a Fisher's exact test in R with the following parameters: alternative = "greater", conf.level = 0.95. We reported odds ratios (OR) and Benjamini–Hochberg adjusted *P*-value (FDR). Brain expressed genes (Brainspan, $N = 15585$) were used as background.

Deconvolution. Deconvolution was performed by *MuSiC* (v0.1.1)[53] in R. This method leverages transcriptomic signatures of cell-types considering cross-subject heterogeneity and gene expression stochasticity. Bulk RNA-seq data is deconvoluted to obtain proportions of cell-types in each sample. We used single-cell data that was downloaded from the Allen Brain Map portal (https://portal.brain-map.org/atlases-and-data/rnaseq). Published cell-type annotations included in the metadata were used as reference for cell-type abundance inference.

Statistical analysis and reproducibility. No statistical methods were used to predetermine sample sizes. Nevertheless, the data here reported is in line with the sample size of previous studies[96,97]. Samples were not randomized. ASD subjects with Chromosome 15q Duplication were excluded from the downstream analysis. Data collection and analysis were not performed blind to the conditions of the experiments. Findings were not replicated due to the limitation of the multi-region ASD transcriptome data. Nevertheless, we used two independent rs-fMRI measurement to refine and increase the confidence of our findings. For fALFF/ReHo rs-fMRI values and bulk RNA-seq transcriptomic data, distribution was assumed to be normal but this was not formally tested. Non-parametric tests have been used to avoid uncertainty when possible. Data collection and analysis were not performed blind to the conditions of the experiments.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The imaging data from ABIDE I and II are available to approved investigators who register with the NITRC (Neuroimaging Informatics Tools and Resources Clearinghouse) and 1000 Functional Connectomes Project to gain access. Details and access information are provided here: http://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html and here: http://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html.

The source bulk RNA-seq data generated in this manuscript are available via the PsychENCODE Knowledge Portal (https://psychencode.synapse.org/). The PsychENCODE Knowledge Portal is a platform for accessing data, analyses, and tools generated through grants funded by the National Institute of Mental Health (NIMH) PsychENCODE program. Data is available for general research use according to the following requirements for data access and data attribution: (https://psychencode.synapse.org/DataAccess). For access to content described in this manuscript see: https://doi.org/10.7303/syn4587615.

Code availability

Custom R code and data to support the data correction, correlation analysis, visualizations, functional, and gene set enrichments are available at https://github.com/konopkalab/AUTISM_rsFMRI_GeneExpressionCorrelations and https://github.com/DeepLearningForPrecisionHealthLab/AUTISM_rsfMRI_ProcessingConnectivityExtractionAndSubjectMatching.

Received: 29 January 2021; Accepted: 13 May 2022;
Published online: 09 June 2022

References

1. Hawrylycz, M. et al. Canonical genetic signatures of the adult human brain. *Nat. Neurosci.* **18**, 1832–1844 (2015).
2. Hawrylycz, M. J. et al. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391–399 (2012).
3. Lein, E. S., Belgard, T. G., Hawrylycz, M. & Molnar, Z. Transcriptomic perspectives on neocortical structure, development, evolution, and disease. *Annu Rev. Neurosci.* **40**, 629–652 (2017).
4. Anderson, K. M. et al. Heritability of individualized cortical network topography. *Proc Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.2016271118 (2021).
5. Grasby, K. L. et al. The genetic architecture of the human cerebral cortex. *Science* https://doi.org/10.1126/science.aay6690 (2020).
6. Anderson, K. M. et al. Gene expression links functional networks across cortex and striatum. *Nat. Commun.* **9**, 1428 (2018).
7. Yadav, S. K. et al. Genetic variations influence brain changes in patients with attention-deficit hyperactivity disorder. *Transl. Psychiatry* **11**, 349 (2021).
8. Moreau, C. A., Ching, C. R., Kumar, K., Jacquemont, S. & Bearden, C. E. Structural and functional brain alterations revealed by neuroimaging in CNV carriers. *Curr. Opin. Genet Dev.* **68**, 88–98 (2021).
9. Radonjic, N. V. et al. Structural brain imaging studies offer clues about the effects of the shared genetic etiology among neuropsychiatric disorders. *Mol. Psychiatry* **26**, 2101–2110 (2021).
10. Hashem, S. et al. Genetics of structural and functional brain changes in autism spectrum disorder. *Transl. Psychiatry* **10**, 229 (2020).
11. Fakhoury, M. Imaging genetics in autism spectrum disorders: Linking genetics and brain imaging in the pursuit of the underlying neurobiological mechanisms. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **80**, 101–114 (2018).
12. Hariri, A. R. & Weinberger, D. R. Imaging genomics. *Br. Med. Bull.* **65**, 259–270 (2003).
13. Konopka, G. Cognitive genomics: Linking genes to behavior in the human brain. *Netw. Neurosci.* **1**, 3–13 (2017).
14. Richiardi, J. et al. BRAIN NETWORKS. Correlated gene expression supports synchronous activity in brain networks. *Science* **348**, 1241–1244 (2015).
15. Wang, G. Z. et al. Correspondence between resting-state activity and brain gene expression. *Neuron* **88**, 659–666 (2015).
16. Nair, A., Treiber, J. M., Shukla, D. K., Shih, P. & Muller, R. A. Impaired thalamocortical connectivity in autism spectrum disorder: A study of functional and anatomical connectivity. *Brain* **136**, 1942–1955 (2013).
17. Buckner, R. L., Andrews-Hanna, J. R. & Schacter, D. L. The brain's default network: Anatomy, function, and relevance to disease. *Ann. N. Y Acad. Sci.* **1124**, 1–38 (2008).
18. Just, M. A., Cherkassky, V. L., Keller, T. A. & Minshew, N. J. Cortical activation and synchronization during sentence comprehension in high-functioning autism: Evidence of underconnectivity. *Brain* **127**, 1811–1821 (2004).
19. Velmeshev, D. et al. Single-cell genomics identifies cell type-specific molecular changes in autism. *Science* **364**, 685–689 (2019).
20. Parikshak, N. N. et al. Author correction: Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* **560**, E30 (2018).
21. Gandal, M. J. et al. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359**, 693–697 (2018).
22. Voineagu, I. et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380–384 (2011).
23. Vorstman, J. A. S. et al. Autism genetics: Opportunities and challenges for clinical translation. *Nat. Rev. Genet.* **18**, 362–376 (2017).
24. de la Torre-Ubieta, L., Won, H., Stein, J. L. & Geschwind, D. H. Advancing the understanding of autism disease mechanisms through genetics. *Nat. Med.* **22**, 345–361 (2016).
25. De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
26. Ebert, D. H. & Greenberg, M. E. Activity-dependent neuronal signalling and autism spectrum disorder. *Nature* **493**, 327–337 (2013).
27. Lombardo, M. V. et al. Atypical genomic cortical patterning in autism with poor early language outcome. *Sci. Adv.* **7**, eabh1663 (2021).
28. Xie, Y. et al. Brain mRNA expression associated with cortical volume alterations in autism spectrum disorder. *Cell Rep.* **32**, 108137 (2020).
29. Romero-Garcia, R., Warrier, V., Bullmore, E. T., Baron-Cohen, S. & Bethlehem, R. A. I. Synaptic and transcriptionally downregulated genes are associated with cortical thickness differences in autism. *Mol. Psychiatry* **24**, 1053–1064 (2019).
30. Lombardo, M. V. et al. Large-scale associations between the leukocyte transcriptome and BOLD responses to speech differ in autism early language outcome subtypes. *Nat. Neurosci.* **21**, 1680–1688 (2018).
31. Di Martino, A. et al. The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* **19**, 659–667 (2014).
32. Di Martino, A. et al. Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Sci. Data* **4**, 170010 (2017).
33. Gandal, M. J. et al. Broad transcriptomic dysregulation occurs across the cerebral cortex in ASD. *Nature* (2022).
34. Zou, Q. H. et al. An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: Fractional ALFF. *J. Neurosci. Methods* **172**, 137–141 (2008).

35. Zang, Y., Jiang, T., Lu, Y., He, Y. & Tian, L. Regional homogeneity approach to fMRI data analysis. *Neuroimage* **22**, 394–400 (2004).
36. King, J. B. et al. Generalizability and reproducibility of functional connectivity in autism. *Mol. Autism* **10**, 27 (2019).
37. Holiga, S. et al. Patients with autism spectrum disorders display reproducible functional connectivity alterations. *Sci. Transl. Med.* https://doi.org/10.1126/scitranslmed.aat9223 (2019).
38. Yagi, H. et al. Filamin A-interacting protein (FILIP) is a region-specific modulator of myosin 2b and controls spine morphology and NMDA receptor accumulation. *Sci. Rep.* **4**, 6353 (2014).
39. Nagano, T., Morikubo, S. & Sato, M. Filamin A and FILIP (Filamin A-Interacting Protein) regulate cell polarity and motility in neocortical subventricular and intermediate zones during radial migration. *J. Neurosci.* **24**, 9648–9657 (2004).
40. Nagano, T. et al. Filamin A-interacting protein (FILIP) regulates cortical cell migration out of the ventricular zone. *Nat. Cell Biol.* **4**, 495–501 (2002).
41. Hodge, R. D. et al. Transcriptomic evidence that von Economo neurons are regionally specialized extratelencephalic-projecting excitatory neurons. *Nat. Commun.* **11**, 1172 (2020).
42. Dijkstra, A. A., Lin, L. C., Nana, A. L., Gaus, S. E. & Seeley, W. W. Von economo neurons and fork cells: A neurochemical signature linked to monoaminergic function. *Cereb. Cortex* **28**, 131–144 (2018).
43. Filice, F., Schwaller, B., Michel, T. M. & Grunblatt, E. Profiling parvalbumin interneurons using iPSC: Challenges and perspectives for Autism Spectrum Disorder (ASD). *Mol. Autism* **11**, 10 (2020).
44. Hashemi, E., Ariza, J., Rogers, H., Noctor, S. C. & Martinez-Cerdeno, V. The number of parvalbumin-expressing interneurons is decreased in the prefrontal cortex in autism. *Cereb. Cortex* **27**, 1931–1943 (2017).
45. Wohr, M. et al. Lack of parvalbumin in mice leads to behavioral deficits relevant to all human autism core symptoms and related neural morphofunctional abnormalities. *Transl. Psychiatry* **5**, e525 (2015).
46. Kawaguchi, Y. & Kubota, Y. Neurochemical features and synaptic connections of large physiologically-identified GABAergic cells in the rat frontal cortex. *Neuroscience* **85**, 677–701 (1998).
47. Anderson, K. M. et al. Transcriptional and imaging-genetic association of cortical interneurons, brain function, and schizophrenia risk. *Nat. Commun.* **11**, 2889 (2020).
48. Miller, J. A. et al. Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199–206 (2014).
49. Satterstrom, F. K. et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584.e523 (2020).
50. Lord, C. et al. Autism spectrum disorder. *Nat. Rev. Dis. Prim.* **6**, 5 (2020).
51. Baron-Cohen, S. et al. Foetal oestrogens and autism. *Mol. Psychiatry* **25**, 2970–2978 (2020).
52. Baron-Cohen, S. et al. Elevated fetal steroidogenic activity in autism. *Mol. Psychiatry* **20**, 369–376 (2015).
53. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 380 (2019).
54. Tasic, B. et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).
55. Gandal, M. J. et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* https://doi.org/10.1126/science.aat8127 (2018).
56. Seidlitz, J. et al. Morphometric similarity networks detect microscale cortical organization and predict inter-individual cognitive variation. *Neuron* **97**, 231–247.e237 (2018).
57. Spiegel, A., Mentch, J., Haskins, A. J. & Robertson, C. E. Slower binocular rivalry in the autistic brain. *Curr. Biol.* **29**, 2948–2953.e2943 (2019).
58. Lee, S. H. et al. Activation of specific interneurons improves V1 feature selectivity and visual perception. *Nature* **488**, 379–383 (2012).
59. Zhang, Y., Miao, B., Guan, J. & Meng, Q. Fractional amplitude of low-frequency fluctuation and degree centrality in autistic children: A resting-state fMRI study. *SPIE* https://doi.org/10.1117/12.2501762 (2018)
60. Itahashi, T. et al. Alterations of local spontaneous brain activity and connectivity in adults with high-functioning autism spectrum disorder. *Mol. Autism* **6**, 30 (2015).
61. Dajani, D. R. & Uddin, L. Q. Local brain connectivity across development in autism spectrum disorder: A cross-sectional investigation. *Autism Res.* **9**, 43–54 (2016).
62. Maximo, J. O., Keown, C. L., Nair, A. & Muller, R. A. Approaches to local connectivity in autism using resting state functional connectivity MRI. *Front. Hum. Neurosci.* **7**, 605 (2013).
63. Paakki, J. J. et al. Alterations in regional homogeneity of resting-state brain activity in autism spectrum disorders. *Brain Res.* **1321**, 169–179 (2010).
64. Nair, S. et al. Local resting state functional connectivity in autism: Site and cohort variability and the effect of eye status. *Brain Imaging Behav.* **12**, 168–179 (2018).
65. Lau, W. K. W., Leung, M. K. & Lau, B. W. M. Resting-state abnormalities in Autism Spectrum Disorders: A meta-analysis. *Sci. Rep.* **9**, 3892 (2019).
66. Khundrakpam, B. S., Lewis, J. D., Kostopoulos, P., Carbonell, F. & Evans, A. C. Cortical thickness abnormalities in autism spectrum disorders through late childhood, adolescence, and adulthood: A large-scale MRI study. *Cereb. Cortex* **27**, 1721–1731 (2017).
67. Nomi, J. S. & Uddin, L. Q. Developmental changes in large-scale network connectivity in autism. *Neuroimage Clin.* **7**, 732–741 (2015).
68. Uddin, L. Q., Supekar, K. & Menon, V. Reconceptualizing functional brain connectivity in autism from a developmental perspective. *Front. Hum. Neurosci.* **7**, 458 (2013).
69. Gogolla, N. et al. Common circuit defect of excitatory-inhibitory balance in mouse models of autism. *J. Neurodev. Disord.* **1**, 172–181 (2009).
70. Tremblay, R., Lee, S. & Rudy, B. GABAergic interneurons in the neocortex: From cellular properties to circuits. *Neuron* **91**, 260–292 (2016).
71. Turkheimer, F. E., Leech, R., Expert, P., Lord, L. D. & Vernon, A. C. The brain's code and its canonical computational motifs. From sensory cortex to the default mode network: A multi-scale model of brain function in health and disease. *Neurosci. Biobehav. Rev.* **55**, 211–222 (2015).
72. Hensch, T. K. Critical period plasticity in local cortical circuits. *Nat. Rev. Neurosci.* **6**, 877–888 (2005).
73. Ferguson, B. R. & Gao, W. J. PV interneurons: Critical regulators of E/I balance for prefrontal cortex-dependent behavior and psychiatric disorders. *Front. Neural Circuits* **12**, 37 (2018).
74. Sohal, V. S., Zhang, F., Yizhar, O. & Deisseroth, K. Parvalbumin neurons and gamma rhythms enhance cortical circuit performance. *Nature* **459**, 698–702 (2009).
75. Mentch, J., Spiegel, A., Ricciardi, C. & Robertson, C. E. GABAergic inhibition gates perceptual awareness during binocular rivalry. *J. Neurosci.* **39**, 8398–8407 (2019).
76. Wykes, K. M., Hugrass, L. & Crewther, D. P. Autistic traits are not a strong predictor of binocular rivalry dynamics. *Front. Neurosci.* **12**, 338 (2018).
77. Karaminis, T., Lunghi, C., Neil, L., Burr, D. & Pellicano, E. Binocular rivalry in children on the autism spectrum. *Autism Res.* **10**, 1096–1106 (2017).
78. Di Martino, A. et al. The autism brain imaging data exchange: Towards large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* **19**, 659–667 (2013).
79. Smith, S. M. Fast robust automated brain extraction. *Hum. Brain Mapp.* **17**, 143–155 (2002).
80. Behzadi, Y., Restom, K., Liau, J. & Liu, T. T. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage* **37**, 90–101 (2007).
81. Kelly, C., Biswal, B. B., Craddock, R. C., Castellanos, F. X. & Milham, M. P. Characterizing variation in the functional connectome: Promise and pitfalls. *Trends Cogn. Sci.* **16**, 181–188 (2012).
82. Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L. & Petersen, S. E. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* **59**, 2142–2154 (2012).
83. Smyser, C. D. et al. Longitudinal analysis of neural network development in preterm infants. *Cereb. Cortex* **20**, 2852–2862 (2010).
84. Craddock, C. et al. Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (C-PAC). *Front. Neuroinform.* https://doi.org/10.3389/conf.fninf.2013.09.00042 (2013).
85. Kendall, M. G. & Gibbons, J. D. *Rank Correlation Methods* 5 edn (Oxford University Press, 1990).
86. Avants, B. B. et al. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage* **54**, 2033–2044 (2011).
87. Avants, B. B., Epstein, C. L., Grossman, M. & Gee, J. C. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* **12**, 26–41 (2008).
88. Calhoun, V. D. et al. The impact of T1 versus EPI spatial normalization templates for fMRI data analyses. *Hum. Brain Mapp.* **38**, 5331–5342 (2017).
89. Dohmatob, E., Varoquaux, G. & Thirion, B. Inter-subject registration of functional images: Do we need anatomical images? *Front. Neurosci.* **12**, 64 (2018).
90. Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P. & Meester, L. E. *A Modern Introduction to Probability and Statistics: Understanding Why and How* (Springer, 2005).
91. Dobin, A. et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
92. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **12**, 323 (2011).

93. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
94. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 e1821 (2019).
95. Chen, J., Xu, H., Aronow, B. J. & Jegga, A. G. Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinform.* **8**, 392 (2007).
96. Fromer, M. et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).
97. Li, M. et al. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science* https://doi.org/10.1126/science.aat7615 (2018).

Acknowledgements

G.K. is a Jon Heighten Scholar in Autism Research and Townsend Distinguished Chair in Research on Autism Spectrum Disorders at UT Southwestern Medical Center. E.C. is a Neural Scientist Training Program Fellow in the Peter O'Donnell Brain Institute at UT Southwestern. Data were generated as part of the PsychENCODE Consortium. Visit 10.7303/syn24240356 for a complete list of grants and PIs. Tissue specimens and/or data used in this research were obtained from the Autism BrainNet (formerly the Autism Tissue Program), which is sponsored by the Simons Foundation, and the University of Maryland Brain and Tissue Bank, which is a component of the NIH NeuroBiobank. We are grateful to the patients and families who participate in the tissue donation programs. Funding for this work was provided by grants to D.H.G. (NIMH R01MH110927, U01MH115746, P50-MH106438, and R01 MH-109912, R01 MH094714), grants to M.J.G. (SFARI Bridge to Independence Award, NIMH R01-MH121521, NIMH R01-MH123922, NICHD-P50-HD103557), and grants to J.R.H. (Achievement Rewards for College Scientists Foundation Los Angeles Founder Chapter, UCLA Neuroscience Interdepartmental Program). This work was also supported by the NIMH (MH102603, MH126481), NINDS (NS106447, NS115821), NHGRI (HG011641), the Simons Foundation (SFARI #573689), and the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition—Scholar Award (220020467) to G.K.

Author contributions

S.B., E.C., A.H.T., D.L., A.A.M., and G.K. analyzed the data and wrote the paper. J.R.H., M.J.G., and D.H.G. collected samples, processed RNA, and generated bulk RNA-seq libraries. A.H.T. analyzed the ABIDE I and ABIDE II data. A.A.M. and G.K. designed and supervised the study, and provided intellectual guidance. All authors discussed the results and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-31053-5.

Correspondence and requests for materials should be addressed to Albert A. Montillo or Genevieve Konopka.

Peer review information *Nature Communications* thanks Michael Hawrylycz and the other anonymous reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

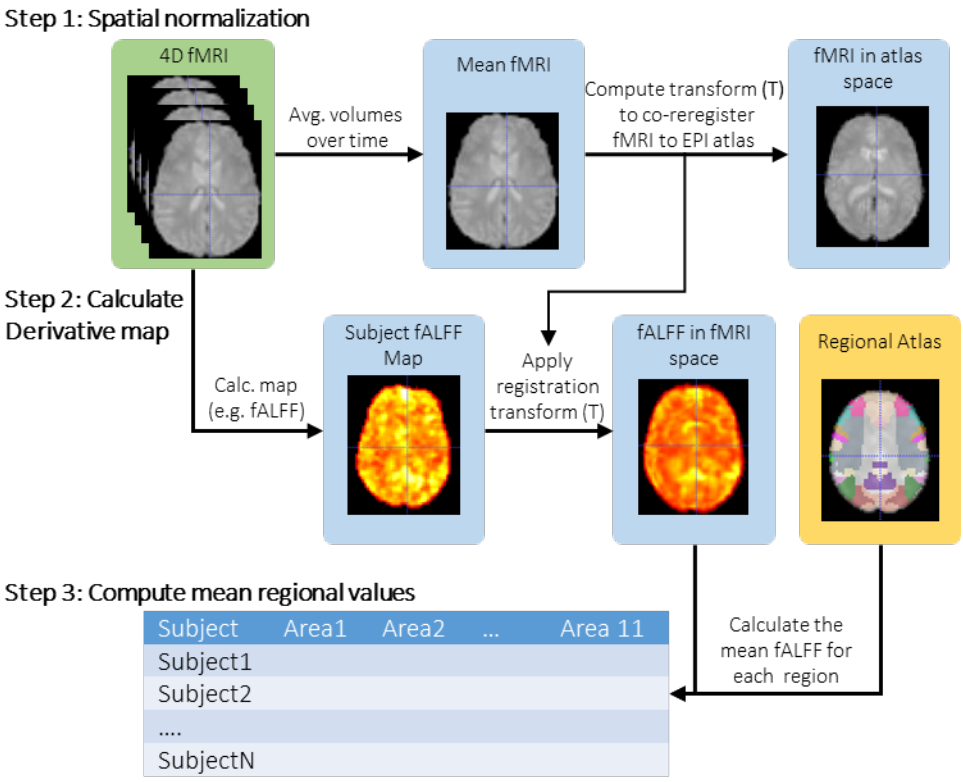
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Association between resting-state functional brain connectivity and gene expression is altered in autism spectrum disorder

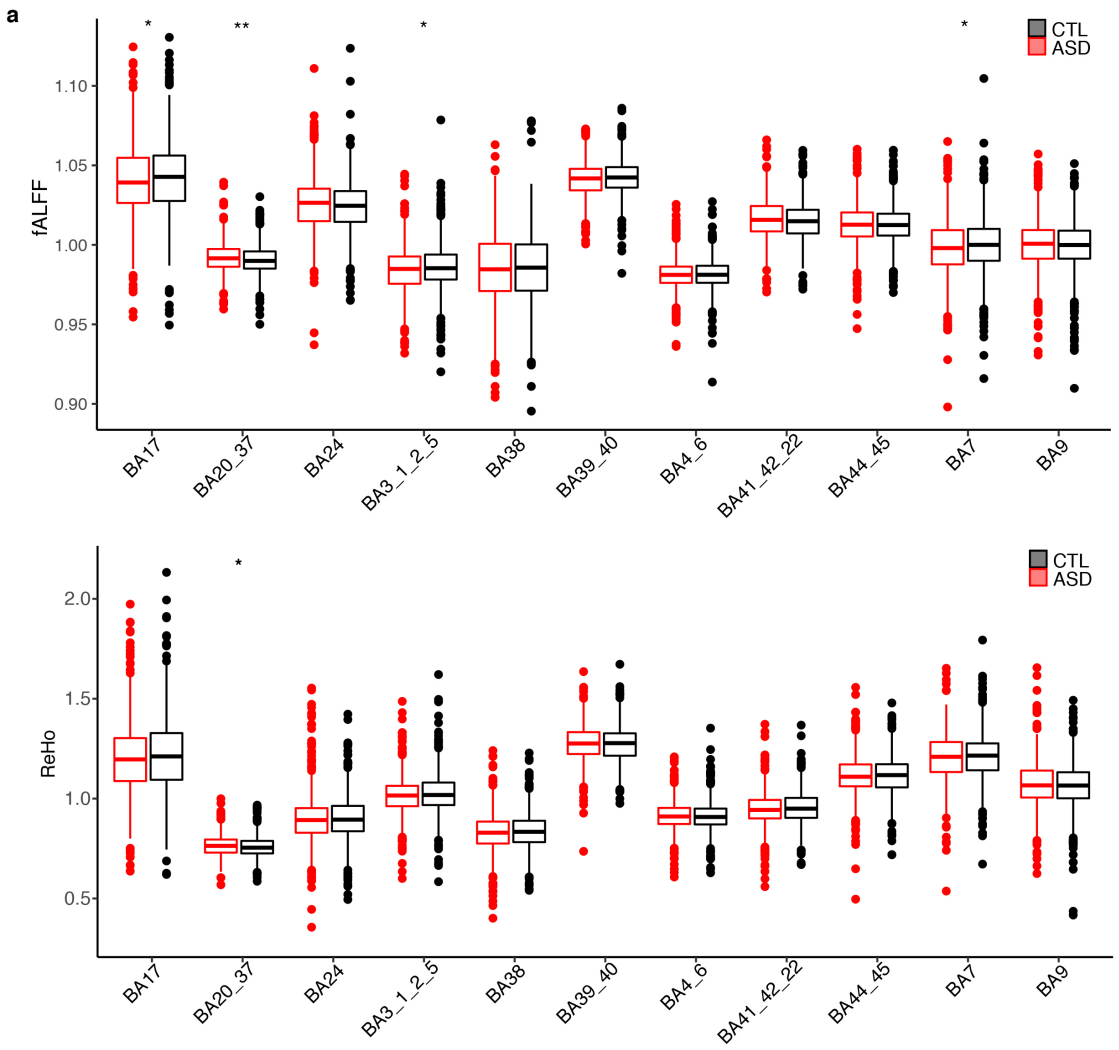
Stefano Berto[1], Alex Treacher[2], Emre Caglayan[1], Danni Luo[2], Jillian R. Haney[3,4,5], Michael J. Gandal[3,4,5,6], Daniel H. Geschwind[3,4,5,6], Albert Montillo[2,7,8] and Genevieve Konopka[1]

[1] Department of Neuroscience, UT Southwestern Medical Center, Dallas, TX 75390, USA.
[2] Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center, Dallas, TX 75390, USA.
[3] Program in Neurobehavioral Genetics, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA.
[4] Department of Neurology, Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA.
[5] Department of Psychiatry, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA.
[6] Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA.
[7] Department of Radiology, University of Texas Southwestern Medical Center, Texas, USA
[8] Advanced Imaging Research Center, University of Texas Southwestern Medical Center, Texas, USA

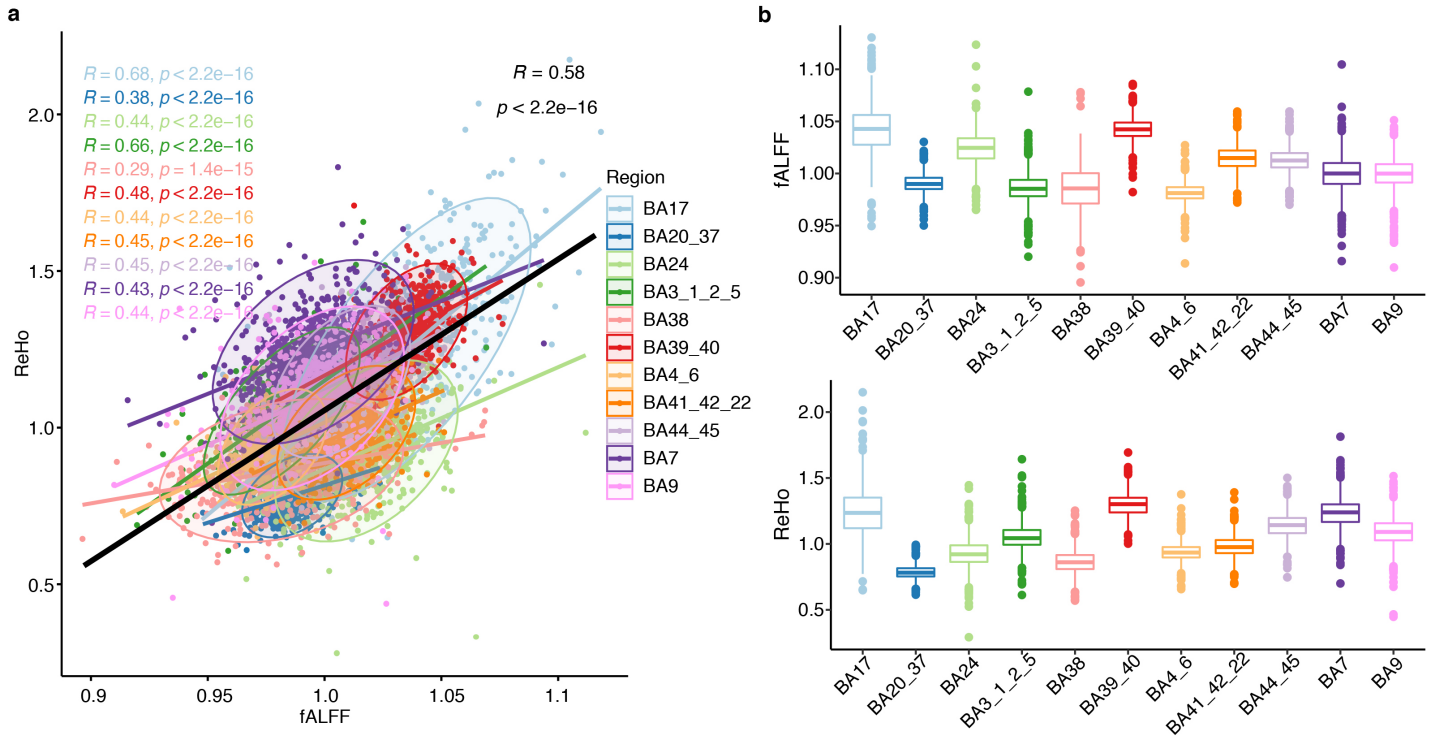
Supplementary Figures 1-8



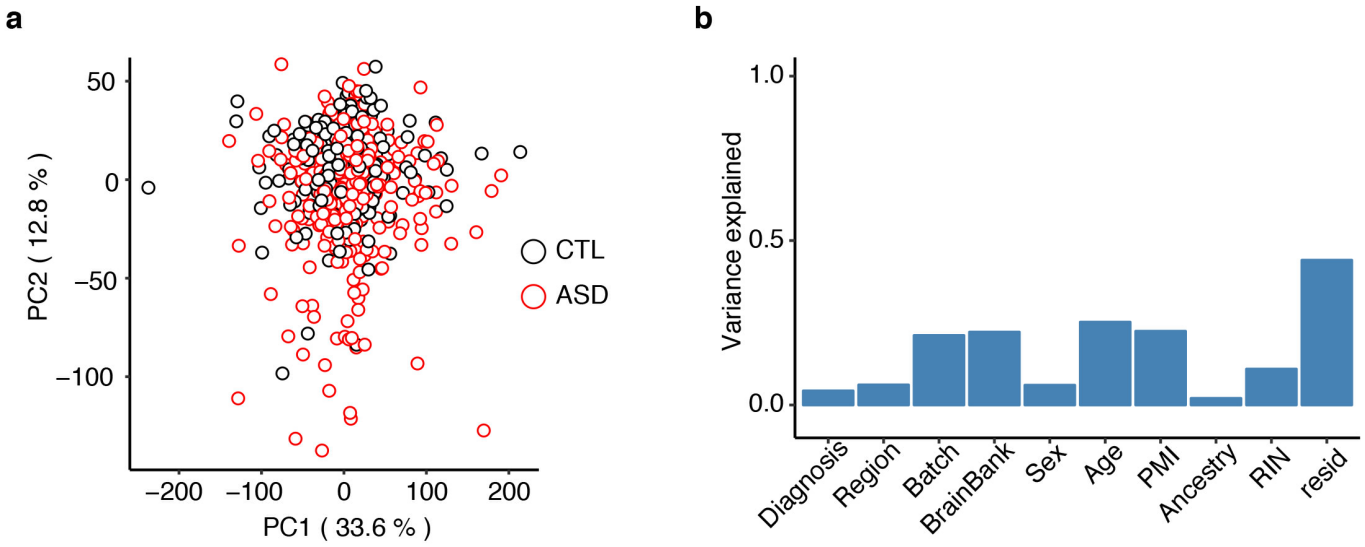
Supplementary Figure 1: Overview of image analysis pipeline. Three steps are used to calculate regional values from MRI. Step 1: Subject fMRI (green) is spatially normalized to atlas space. Step 2: Local functional activity measures are derived for each subject (e.g. fALFF) and coregistered to atlas space. Step 3: Mean regional measures are computed using the atlas (yellow) for each region for each subject.



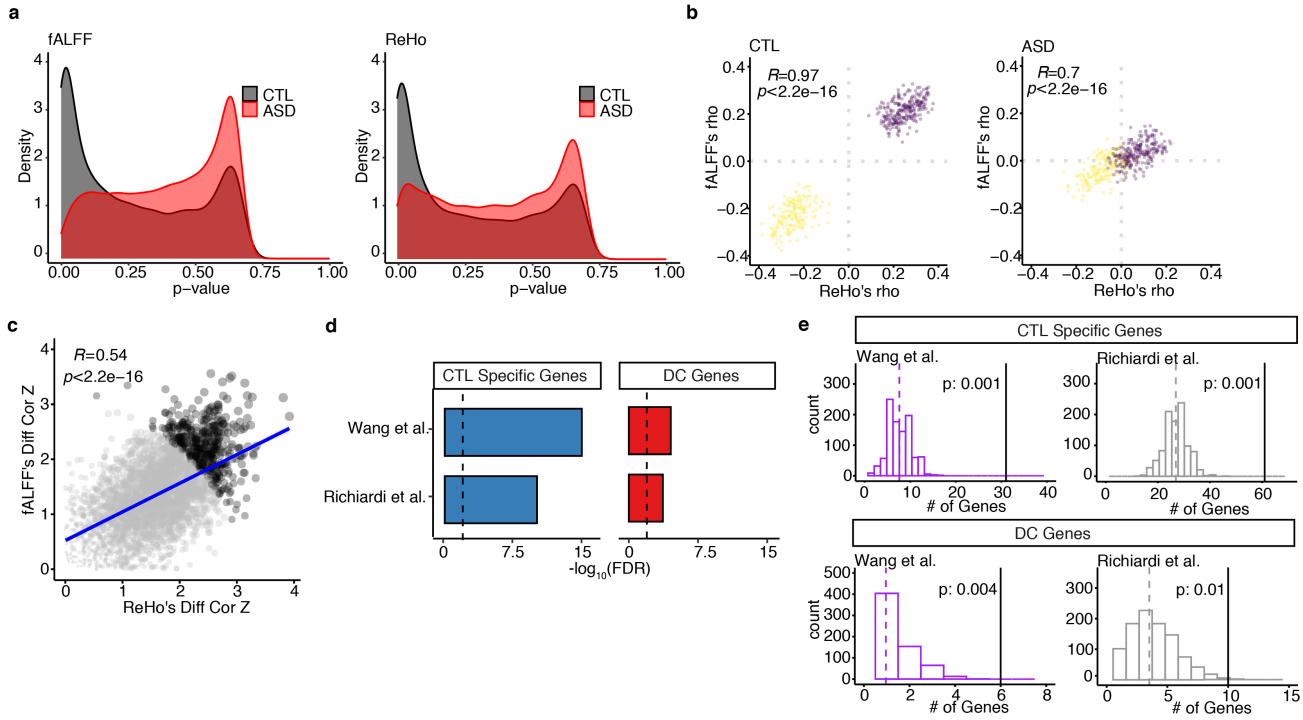
Supplementary Figure 2. rs-fMRI data quality control. a, Boxplots of fALFF and ReHo measurements comparison between ASD (red) and CTL (black) across multiple ROIs analyzed. Stars correspond to the significant differences between ASD and CTL based on one-sided Wilcoxon rank sum's test (** p = 0.001, * p = 0.05). Boxes extend from the 25th to the 75th percentiles, the center lines represent the median. ASD: N = 606 biologically independent samples, CTL: N = 710 biologically independent samples. Exact P-value fALFF: BA17 p = 0.048, BA20_37 p = 0.001, BA7 p = 0.04. Exact P-value ReHo: BA20_37 p = 0.026.



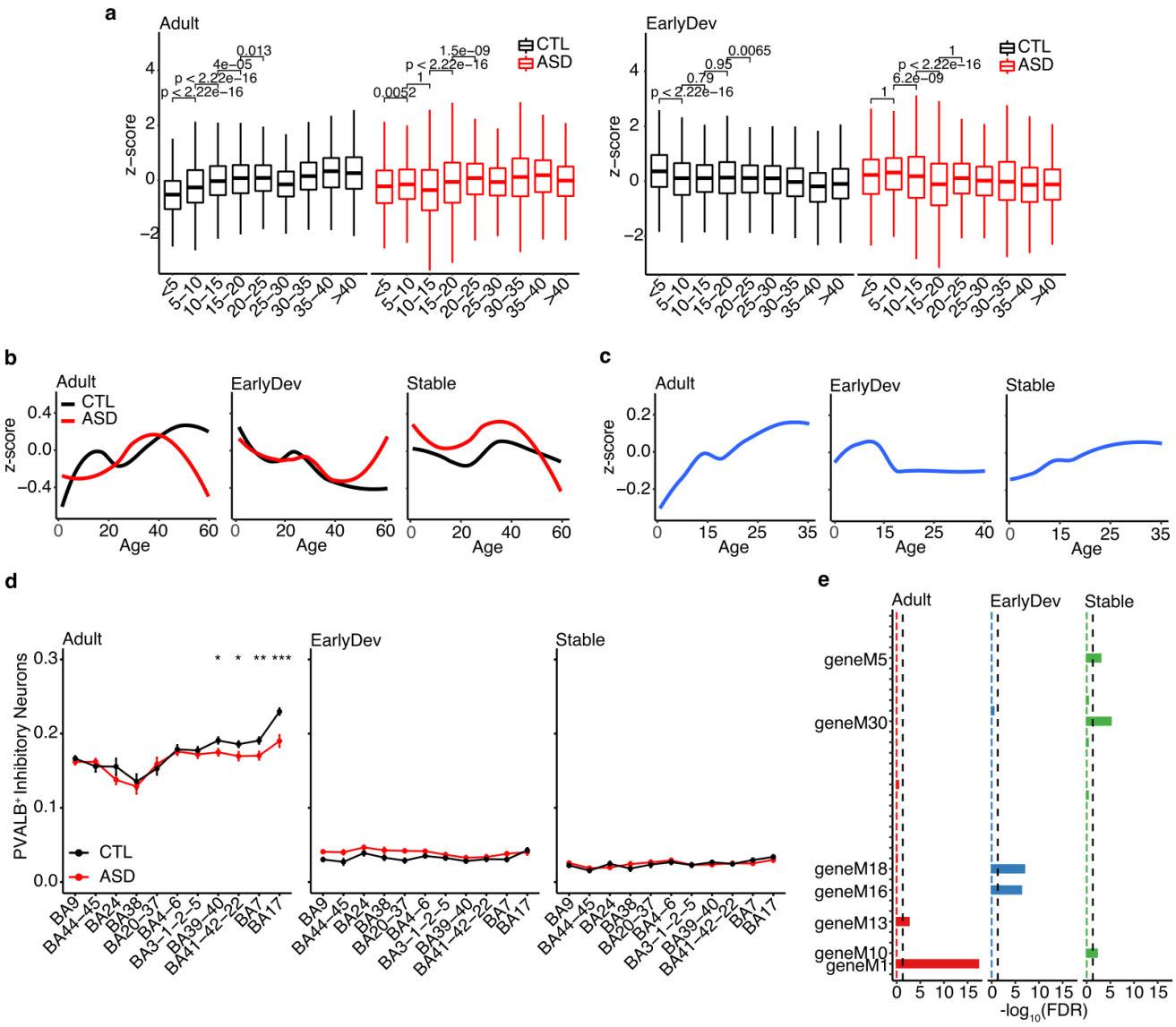
Supplementary Figure 3. Comparisons of the two types of rs-fMRI measurements. a, Scatter plot comparing fALFF (X-axis) and ReHo (Y-axis) between the 11 ROIs analyzed in CTL. Spearman rank *rho* values and associated p-values are shown colored by ROIs. Black line corresponds to the across-ROIs (pancortical) correlation (Spearman's rank correlation test, two-tailed). **b,** Distribution of fALFF and ReHo measurements in the 11 ROIs analyzed in CTL. Boxes extend from the 25th to the 75th percentiles, the center lines represent the median. ASD: N = 606 biologically independent samples, CTL: N = 710 biologically independent samples.



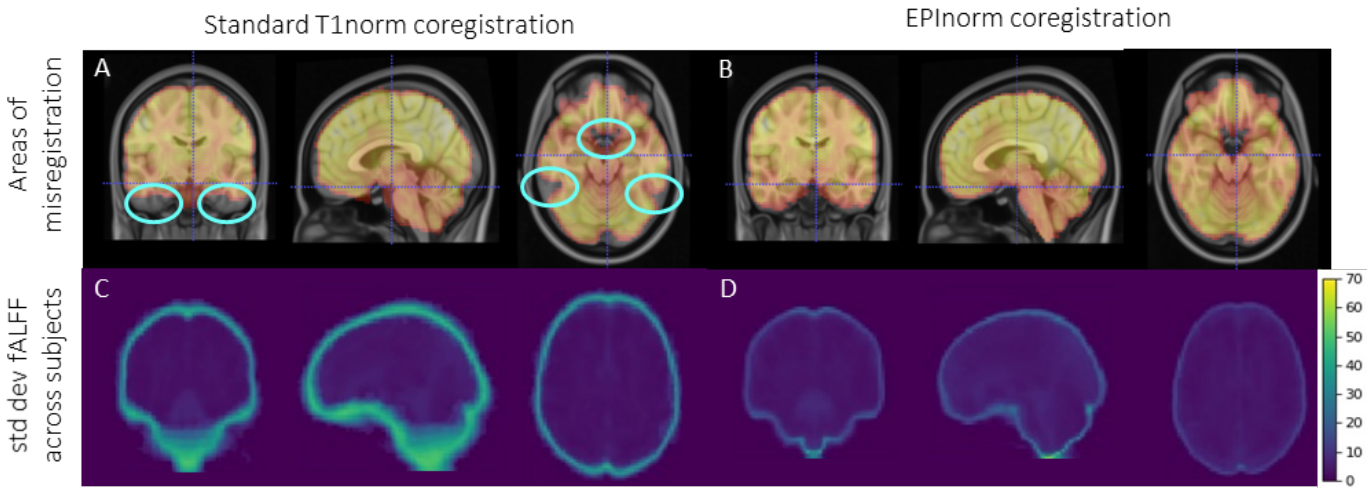
Supplementary Figure 4. **RNA-seq data quality control and covariate metrics. a**, Principal component analysis based on the RNA-seq data of all the subjects used in this study. **b**, Variance explained by each covariate adjusted across 10 principal components.



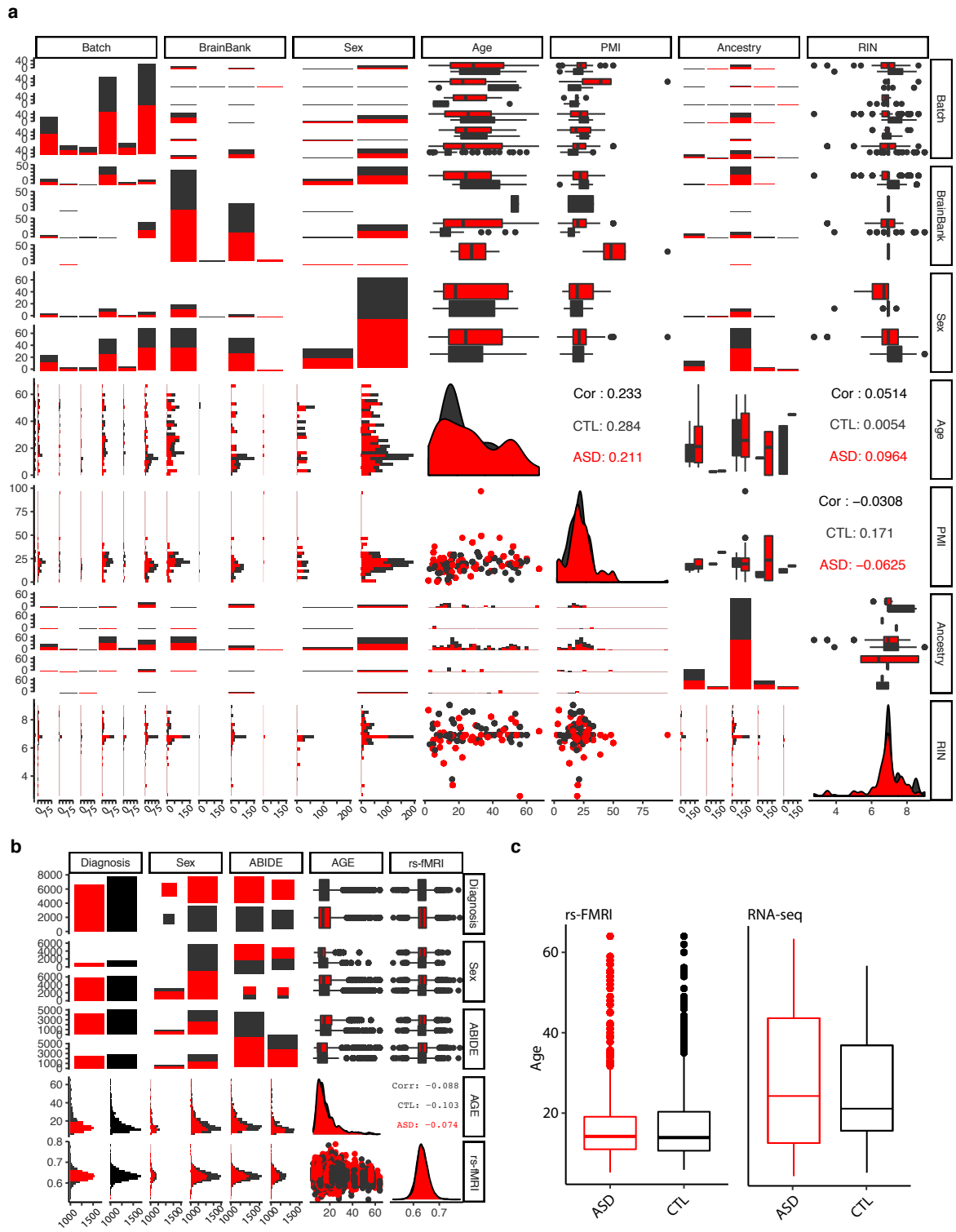
Supplementary Figure 5. **Comparison of differentially correlated genes using either imaging metric or with other published datasets. a**, Distribution of Spearman's rank p-values in both rs-fMRI measurements for ASD and CTL. **b**, Scatterplot depicting the correlation between CTL and ASD fALFF and ReHo *rho* values for the differentially correlated genes (DC genes) (Spearman's rank correlation test, two-tailed). **c**, Scatterplot depicting the correlation between fALFF and ReHo differentially correlated effect sizes (Spearman's rank correlation test, two-tailed). **d**, Barplot depicting the -$\log_{10}$(FDR) of the overlap between CTL specific and DC Genes with previously published rs-fMRI genes (Fisher's Exact Test). **e**, Permutation tests of the presented overlaps. Brain expressed genes were randomly permuted matching the number of rs-fMRI genes identified in the two independent studies. Histograms show the distribution of overlapped genes between permuted data and genes found in this study. Overlaps and p-values were obtained using 1000 random permutations. Black lines indicate the original overlap.



Supplementary Figure 6. **Gene expression comparisons of ASD and CTL across developmental trajectories. a**, Statistical comparison of developmental trajectories. Samples were divided into age brackets and age brackets were compared by one-sided t-test (alternative hypothesis: greater with increasing age in Adult, less with increasing age in EarlyDev). Numbers on graph are p-values, y-axis indicates Z-scored gene expression. Note that z-score spans a larger interval compared to Figure 4a. ASD: N = 360 biologically independent samples, CTL: N = 302 biologically independent samples. Boxes extend from the 25th to the 75th percentiles, the center lines represent the median. **b,** Developmental trajectories after equalizing diagnosis-region groups by random subsampling. Loess regression was used to fit smooth curve for the values of all genes per cluster across development. Smooth curves are shown with 95% confidence bands. Y-axis indicates Z-scored gene expression. **c**, Developmental trajectories of Adult, EarlyDev and Stable gene clusters using BrainSpan atlas (similar to Figure 4a). **d,** Line chart showing the median with standard error of *PVALB*+ interneurons imputed proportions across the 11 regions analyzed based on Adult, EarlyDev, Stable genes. Stars correspond to the significant differences between ASD and CTL based on one-sided Wilcoxon rank sum's test (** p = 0.01, * p = 0.05). Y-axis indicates cortical regions by anterior-to-posterior ordering. **e**, Barplot of enrichment between developmental clusters and modules associated with autism, bipolar disorder, and schizophrenia (Y-axis) from an independent study.



Supplementary Figure 7: **Quantitative comparison of EPInorm-based versus T1norm-based co-registration.** Each panel shows a mid-coronal image (left), mid-sagittal image (middle) and mid-axial image (right). Top panels show fMRI (yellow/red overlay) coregistered to MNI T1 anatomical atlas (underlay) using (A) T1norm-based co-registration and (B) EPI based registration. Both co-registrations are 3-dimensional. EPInorm based co-registration better aligns the derivative maps (fALFF and ReHo) to brain anatomy and reduces areas of misregistration (blue circles). Bottom panels show the standard deviation for fALFF values across subjects using (C) T1norm-based co-registration and (D) EPInorm-based co-registration.



Supplementary Figure 8. Averaged demographic information for ASD and CTL groups. **a**) Pairwise comparison of demographic information containing biological and technical covariates for RNA-seq. In red: ASD subjects; in black: control. **b**) Pairwise comparison of demographic information containing biological and technical covariates for rs-fMRI. In red: ASD subjects; in black: control. **c**) Distribution of the age of the individuals who provided data for either RNA-seq or rs-fMRI studies. RNA-seq: ASD: N = 360 biologically independent samples, CTL: N = 302 biologically independent samples. rs-fMRI: ASD: N = 606 biologically independent samples, CTL: N = 710 biologically independent samples. Boxes extend from the 25th to the 75th percentiles, the center lines represent the median.

CHAPTER 5: Gene-expression correlates of the oscillatory signatures supporting human episodic memory encoding

Published as:

Berto, S., Fontenot, M. R., Seger, S., Ayhan, F., **Caglayan, E.**, Kulkarni, A., Douglas, C., Tamminga, C. A., Lega, B. C., & Konopka, G. 2021. Gene-expression correlates of the oscillatory signatures supporting human episodic memory encoding. *Nature Neuroscience*. DOI: 10.1038/s41593-021-00803-x

ARTICLES
https://doi.org/10.1038/s41593-021-00803-x

nature neuroscience

Gene-expression correlates of the oscillatory signatures supporting human episodic memory encoding

Stefano Berto[1], Miles R. Fontenot[1], Sarah Seger[2], Fatma Ayhan[1], Emre Caglayan[1], Ashwinikumar Kulkarni[1], Connor Douglas[1], Carol A. Tamminga[3], Bradley C. Lega[2]✉ and Genevieve Konopka[1]✉

In humans, brain oscillations support critical features of memory formation. However, understanding the molecular mechanisms underlying this activity remains a major challenge. Here, we measured memory-sensitive oscillations using intracranial electroencephalography recordings from the temporal cortex of patients performing an episodic memory task. When these patients subsequently underwent resection, we employed transcriptomics on the temporal cortex to link gene expression with brain oscillations and identified genes correlated with oscillatory signatures of memory formation across six frequency bands. A co-expression analysis isolated oscillatory signature-specific modules associated with neuropsychiatric disorders and ion channel activity, with highly correlated genes exhibiting strong connectivity within these modules. Using single-nucleus transcriptomics, we further revealed that these modules are enriched for specific classes of both excitatory and inhibitory neurons, and immunohistochemistry confirmed expression of highly correlated genes. This unprecedented dataset of patient-specific brain oscillations coupled to genomics unlocks new insights into the genetic mechanisms that support memory encoding.

Genome-wide association studies (GWAS) and gene expression profiling of the human brain have unlocked the ability to investigate the genetic basis of complex brain phenomena. These datasets have principally been applied to noninvasive imaging studies, especially correlations with structural magnetic resonance imaging (MRI) or resting-state functional MRI[1–4]. Existing methods have relied on published datasets of gene expression from postmortem brains, which means that neurophysiological and behavioral data are not from the same individuals who contributed gene expression data[5,6]. This limits the potential impact of such approaches to determine how genes support key cognitive processes such as episodic memory and highlights the need to develop new datasets in which individuals contribute both neurophysiological and gene expression data[7]. Another issue affecting previous studies is that neurophysiological measurements such as resting-state functional MRI are not directly linked to cognitive phenomenon. Thus, we previously attempted to correlate gene expression levels with oscillatory signatures of successful memory encoding[8], as the fundamental role of these oscillations in supporting memory behavior has been well established in rodents and humans[9,10]. These oscillatory signatures are measures of the degree to which memory encoding success modulates oscillatory power in a given frequency band. They were quantified using intracranial electrodes implanted for seizure mapping purposes, with recordings made as participants performed an episodic memory task. We used a large database of intracranial electroencephalography (iEEG) recordings obtained over 10 years to piece together a distribution of these oscillatory signatures across brain regions. We identified genes correlated with these oscillatory signatures, including those previously linked to memory formation in rodent investigations, genes linked to cognitive disorders such as autism spectrum disorder (ASD) and novel genes that are prime targets for further investigation. However, as with other studies, this dataset did not have the benefit of both neurophysiological and gene expression information from the same individuals.

With the goal of explicating links between gene expression and brain oscillations and identifying propitious targets for neuromodulation to treat memory disorders, here, we compiled an unprecedented dataset from 16 human participants who first underwent iEEG during which we measured oscillatory signatures of episodic memory encoding using a well-refined signal-processing pipeline. These participants then underwent a temporal lobectomy, during which an en bloc resection of the lateral temporal lobe permitted the acquisition of high-quality tissue specimens that were processed immediately after removal from a common brain region (Brodmann area 38 (BA38)) from which in vivo recordings had been previously obtained. This approach allowed us to identify genes linked with mnemonic oscillatory signatures by correlating gene expression information with iEEG data obtained from the same individuals. Prioritization was performed using the following different steps: multivariate analyses (MVAs) followed by decomposition by brain oscillation using correlations; gene regulatory network connectivity and cell-type-specific expression and/or epigenomic state; and immunofluorescence staining confirmation. This robust analytical approach to combine human electrophysiological data by iEEG and genomic data from the same participants highlighted genes that might be relevant for mechanisms of episodic memory.

We made the a priori decision to focus on BA38 in this analysis for the following reasons: (1) the region has been shown to exhibit strong memory-related oscillatory signatures in multiple investiga-

[1]Department of Neuroscience, UT Southwestern Medical Center, Dallas, TX, USA. [2]Department of Neurosurgery, UT Southwestern Medical Center, Dallas, TX, USA. [3]Department of Psychiatry, UT Southwestern Medical Center, Dallas, TX, USA. ✉e-mail: Bradley.lega@utsouthwestern.edu; Genevieve.Konopka@utsouthwestern.edu

tions[11,12]; (2) resection of this region is standardized in an en bloc temporal lobectomy operation, thereby allowing preservation of blood supply to the region until a period of less than 5 min from procurement of tissue for processing and maximizing the quality of the specimen; and (3) iEEG investigations preceding temporal lobectomy in this particular population invariably include sampling from this region.

An inevitable feature of our dataset is that the participants suffered from intractable epilepsy, which presents an important caveat to the interpretations of the results. However, recent experiments have shown that blood-oxygenation-level-dependent patterns elicited during successful encoding in patients with epilepsy participating in cognitive studies do not show significant differences compared to healthy controls[13]. Moreover, since we examined gene–oscillatory signature correlations across these individuals rather than in comparison to an alternative cohort of data, we could institute control methodologies to partially account for this concern. These included strict artifact-rejection routines and the exclusion of data from regions of seizure onset, as well as using matched postmortem gene expression samples from both unaffected individuals and patients with epilepsy to adjust gene expression levels.

Results

Generation of within-individual memory oscillatory signatures and a gene expression dataset. To determine the relationship between memory-related brain oscillations and gene expression, we analyzed iEEG recorded as participants encoded episodic memories along with gene expression data from the same 16 individuals (Supplementary Table 1). Oscillatory signatures of successful memory encoding (subsequent memory effects (SMEs)) were calculated from recorded iEEG signals by comparing oscillatory patterns during successful versus unsuccessful memory encoding. We use the term "oscillations" to describe oscillatory power extracted in predefined frequency bands, but address issues related to the use of this term in the Discussion. We used the free-recall task, a standard test of episodic memory for which oscillatory patterns have been well described[14], and calculated oscillatory signatures utilizing our well-established signal-processing pipeline[8,15,16] (Fig. 1 and Methods). On average, participants remembered 24.3% of memory items, with a rate of list intrusion (erroneous recollection) of 5.4%. These characteristics are consistent with previous publications of the performance of participants undergoing iEEG during this task[11]. Further behavioral characteristics, including response probability curves by serial position and conditional response probability curves, are shown in Extended Data Fig. 1a,b. These revealed expected patterns for free recall, including primacy and recency effects and temporal contiguity for immediate lags.

SMEs were extracted from electrodes located in the temporal pole by first normalizing the iEEG signal following wavelet decomposition and statistically comparing oscillatory values between successful versus unsuccessful encoding events across 56 log-spaced frequencies from 2 to 120 Hz. This was done using a permutation procedure, whereby trial labels are shuffled 1,000 times within each recording electrode. We made the a priori decision to average oscillatory data for each individual across all electrodes localized to the anterior temporal pole (BA38) by expert neuroradiology review, as this seemed the most generalizable approach. Extended Data Fig. 1c shows that the variance of SME values across participants for all bands is greater than the variance within participants, which supports the validity of this approach. Data were averaged over a mean of 3.6 electrodes per participant. The resulting oscillatory signatures were averaged into six predefined frequency bands before entering these data into our model to estimate gene correlation values (Fig. 1d,f). The proportion of electrodes exhibiting significant differences in oscillatory power between successful and unsuccessful encoding demonstrated that significant memory-related oscillatory patterns were present (Extended Data Fig. 1d). Significant effects at the individual level are also shown (Fig. 1d), and these results were consistent with previous work[17] related to memory patterns in the anterior temporal lobe. In addition, the correlation between observed SME values revealed an expected relationship between low- and high-frequency SMEs (Extended Data Fig. 1e). We note that observed differences may be due to functional changes in narrowband oscillations or broadband power shifts (or a mix of the two). Extended Data Fig. 1f,g shows the results of an oscillation detection analysis, which indicated that narrowband oscillations were present in our data, and we comment on this issue in the Discussion. These 16 study participants then underwent a temporal lobectomy operation. This surgery was performed by a single surgeon (B.C.L.) using a technique that was standardized across these participants for obtaining tissue from BA38 (Fig. 1e). None of the individuals included in this study had gross or radiographic lesions, such as temporal sclerosis or cortical dysplasia. Participants with seizure onset in the temporal pole were not included in our data.

We generated whole-transcriptome RNA-sequencing (RNA-seq) data from the 16 BA38 samples. In addition to the 16 individuals with matched oscillatory signature measurements and gene expression data, we generated BA38 RNA-seq data from an additional 11 temporal lobectomies from individuals for whom we did not obtain oscillation measurements, and postmortem tissue from 12 healthy individuals and 8 patients with epilepsy to validate our predictions using permutations/bootstraps (Fig. 1f and Methods). Principal component analysis revealed that gene expression was uniform across samples, with no outliers (Extended Data Fig. 1h–m). Variance explained by technical, biological and sequencing covariates was analyzed and removed before further analyses (Extended Data Fig. 1n). These adjusted gene expression values were used to calculate gene–oscillatory signature correlations across individuals for each frequency band and co-expression networks.

Memory oscillatory signatures are correlated with gene expression. To determine the relationship between memory oscillatory signatures and gene expression, we performed a MVA followed by decomposition by brain oscillation using a Spearman's rank correlation that included the aforementioned permutations/bootstraps (Methods). Correlations between gene expression and brain oscillations were performed across participants, with each participant contributing a single gene-expression value and a single SME value per frequency band. The MVA detected a total of 753 genes with false-discovery rate (FDR)-corrected P values of <0.05 (Fig. 2a) for SME–gene expression correlations. The F-statistics for the significant genes we identified were robust and greater than for nonsignificant genes (Extended Data Fig. 2a). We next decomposed the MVA by a correlative analysis to identify genes whose expression correlated with memory-related oscillatory signatures in each of the six frequency bands ("SME genes"; multivariate, FDR < 0.05; Spearman's rank correlation ρ and permutations $P < 0.05$). Of the 753 genes detected by MVA, 300 genes were linked with memory effects in specific frequency bands, with a high proportion associated with 2–4 Hz delta band oscillations (Fig. 2a and Supplementary Table 2). The majority of the identified genes were specific to one frequency band, with primarily only a small number of genes shared by delta and one other frequency band (Fig. 2b). Spearman's ρ values were robust and greater than for random expectation (Extended Data Fig. 2b,c). These results further confirmed the significance of the identified genes.

Data from these 16 individuals also included a control behavioral paradigm in which individuals performed simple mathematical problems, which allowed us to observe oscillatory signatures linked to this separate cognitive domain (Methods). We performed the same analysis as above to test whether gene–oscillatory signature associations were specific for mnemonic processing. Our

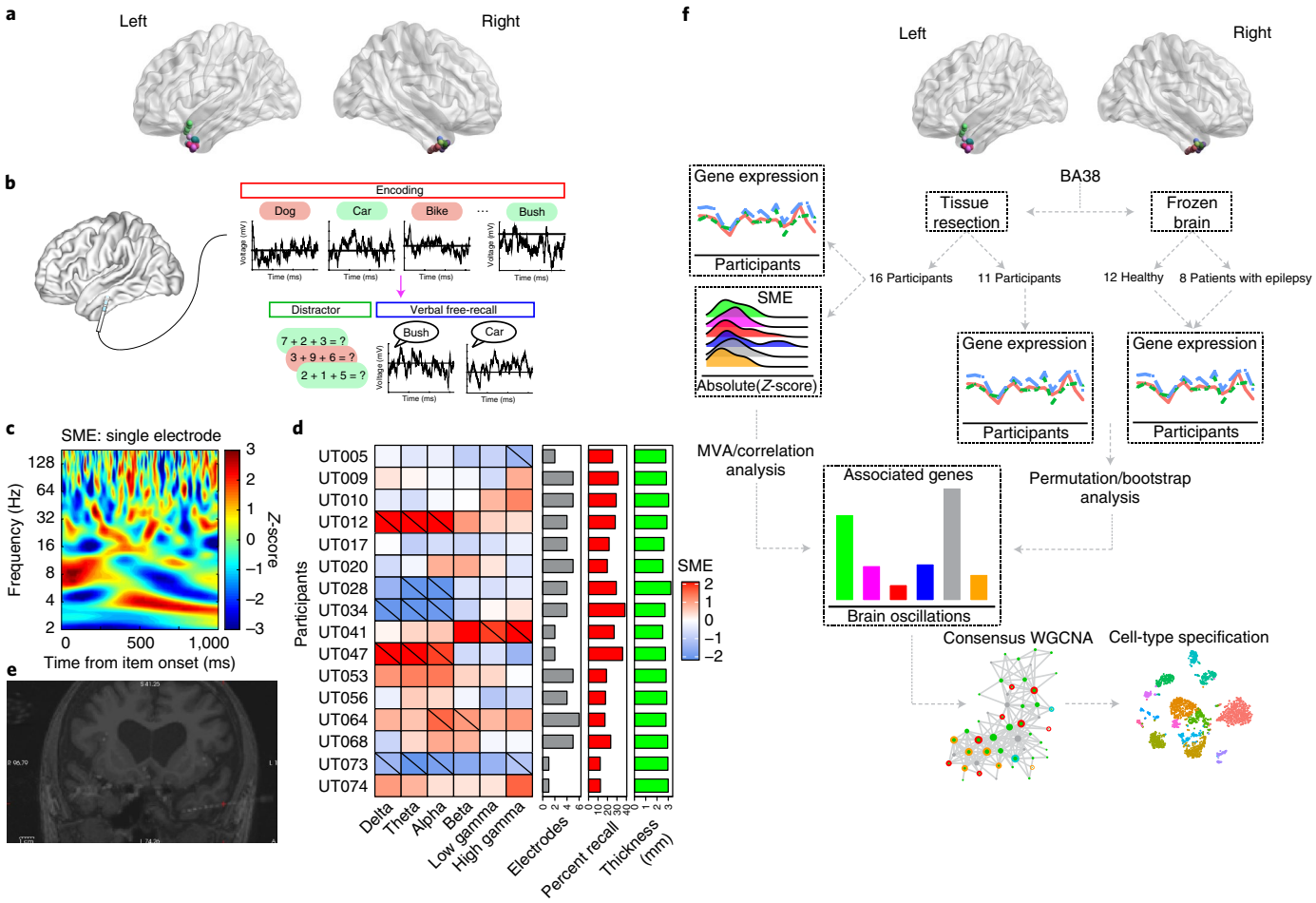


Fig. 1 | Within-individual study design and quality control. a, Representation of the position of each electrode in BA38 for the 16 participants. **b**, Schematic of iEEG memory testing. Intracranial electrodes were used to record oscillations as participants performed an episodic memory task. SMEs were calculated by contrasting brain activity recorded as individuals either remembered (green) or forgot (red) each item. **c**, Example SME recorded from BA38 (full time frequency representation, color axis represents the *z*-transformed *P* value for successful/unsuccessful contrast). **d**, Individual-level SME values in our data. A solidus indicates a significant SME ($P < 0.05$; two-sided Student's *t*-test with permutation procedure) at the individual level. Warm colors indicate power increases during successful encoding. The gray bar plot indicates the total number of electrodes localized to BA38 for each participant. The red bar plot indicates the recall fraction for each individual, and the green bar plot indicates the measured cortical thickness in mm as determined using the FreeSurfer volume-extraction routine. **e**, Postoperative MRI and computed tomography overlay image after implantation of intracranial electrodes; image was used for localization. **f**, Human BA38 RNA-seq data from resected tissue were integrated with brain oscillation data derived from SME analysis to identify protein-coding genes whose gene expression support SMEs (SME genes). Permutations/bootstraps with additional human BA38 samples from independent sources were performed. SME genes were prioritized using co-expression networks and specified at the cell-type level using snRNA-seq and snATAC-seq data from BA38.

dataset also included cortical thickness estimates for BA38 for each individual, which were extracted from our FreeSurfer processing routine, allowing us to perform an additional control analysis looking for genes correlated with this measurement. We did not observe an overlap with these alternative data, and all the genes highlighted below using co-expression network analysis were memory-specific (that is, gene–oscillation correlations were specific for memory-related oscillatory effects). Finally, we looked for gene correlations with memory performance (that is, behavioral data without regard to any oscillatory signature observations). Only one gene associated with oscillatory signatures overlapped with those identified in these control analyses, thereby reinforcing the unique memory-relevant information obtained by examining gene–oscillatory signature correlations (Fig. 2c).

Networks refine molecular pathways associated with memory. We sought to understand the functional properties of the genes identified as correlated with oscillatory signatures of successful memory encoding. We performed consensus weighted gene co-expression network analysis (WGCNA; Methods, Extended Data Fig. 3a and Supplementary Table 3) using gene expression from resected temporal lobe tissue together with the postmortem gene expression datasets. We placed the memory genes into a systems-level context to identify co-expression networks (for example, modules of highly correlated genes) linked with brain oscillations to further prioritize genes. We required that identified modules were robust across these multiple expression datasets (Methods), and this identified a total of 26 modules. Of these, six were significantly associated with oscillatory-signature-correlated genes (Fig. 3a and Extended Data Fig. 3b).

Two modules were significantly associated with delta oscillatory signatures, one module with both delta and low-gamma oscillations, and three modules were significantly associated with beta oscillatory signatures (Fig. 3a). Notably, we did not detect module

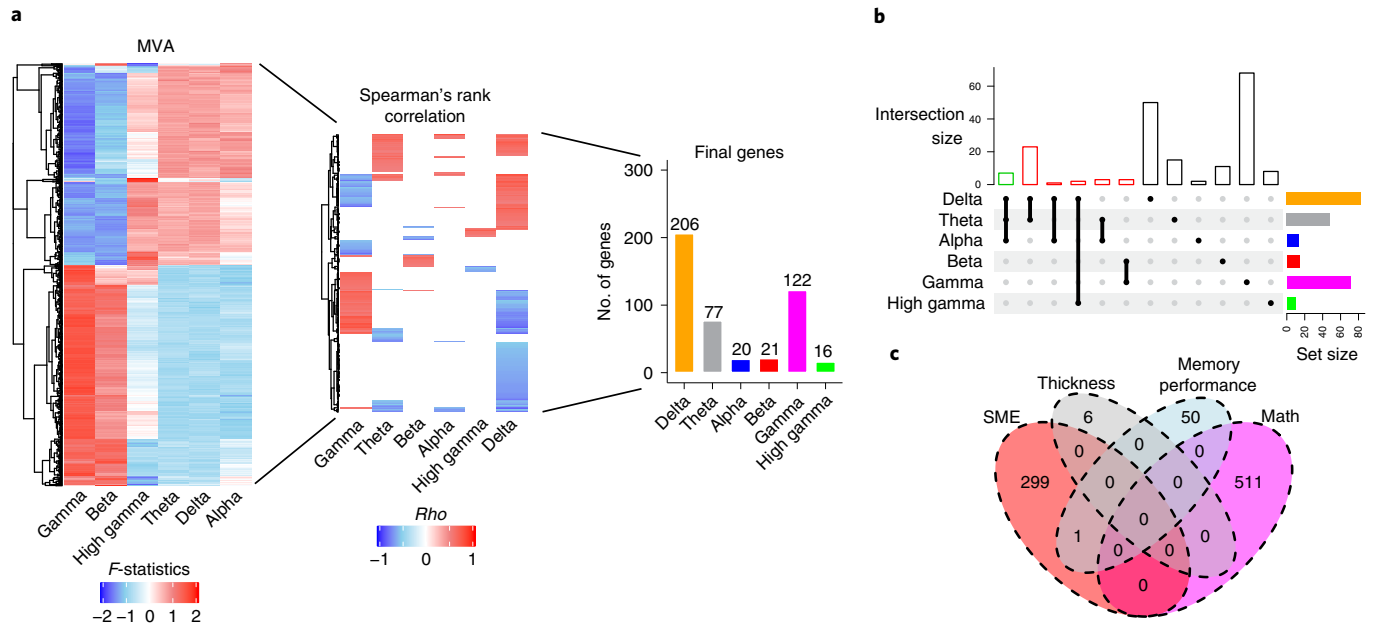


Fig. 2 | Genes associated with SMEs are distinct. a, Heatmap showing the *F*-statistics for the initial SME genes (753 genes; FDR < 0.05). We next used correlation analysis to decompose the genes by brain oscillations (Spearman's rank correlation, $\rho < 0.05$, permutations/bootstraps $P < 0.05$). We identified a total of 300 genes using this iterative filtering process that represents brain oscillation–gene expression correlations by wavelength. **b**, Upset plot indicating shared and specific SME genes between brain oscillations. Most genes are uniquely correlated to a specific wavelength. **c**, Venn diagram showing shared and specific genes between genes associated with memory effect (SME), genes associated with the math task (Math), genes associated with cortical thickness (Thickness) and genes associated with recalled words (Memory performance). Only one gene overlaps between SME and memory performance.

associations for genes correlated with cortical thickness or recall fraction (behavioral measurement), whereas genes correlated with oscillations during the mathematical tasks were associated with two independent modules, which provides further confirmation that genes associated with memory encoding and their networks are distinct (Fig. 3b). In addition, in three of these modules (WM4, WM12 and WM21), we identified significant enrichment for the SME genes (Extended Data Fig. 3c). Moreover, SME genes within these three modules showed higher connectivity compared with other genes, which suggests that oscillatory-signature-associated genes have a central role in the transcriptome of BA38 (Extended Data Fig. 3d). We also observed convergence of genes and modules associated with oscillatory-signature-correlated genes from our previous work[8], which examined gene–oscillatory signature associations across cortical regions (Fig. 3c). The convergence of these findings using different patient populations and methods gives confidence to our inferences regarding the link between these genes and mnemonic processes. The two modules positively associated with delta oscillatory signatures (WM4 and WM12) were enriched for genes implicated in ion channel activity (Fig. 3d). Notably, WM4 contained previously identified genes correlated with memory oscillatory signatures[8] (Fisher's exact test, FDR-corrected $P = 0.003$, odds ratio (OR) = 4.4), whereas WM12 is enriched for a previously identified[8] synaptic-related module correlated with memory oscillatory signatures (Fisher's exact test, FDR-corrected $P = 1.0 \times 10^{-07}$, OR = 4.1) (Fig. 3c).

Because 'hub genes', genes with high intramodular connectivity, have been previously shown to drive module and network structure[18], we examined the correlated modules for any hub genes that might have previously been linked to memory or cognition. *SHANK2*, one of the WM12 hub genes, encodes a synaptic scaffolding protein. Mutations in *SHANK2* have been linked with ASD, intellectual disability and schizophrenia[19,20]. Moreover, *SHANK2* has been associated with learning and memory deficits[21], thereby further confirming the pivotal role of this WM12 hub gene in memory encoding. Importantly, modules associated with different oscillatory frequency bands exhibited different functional properties. In contrast to the delta-associated modules, modules linked with beta oscillatory signatures (WM11 and WM22) were significantly associated with alternative splicing and chromatin remodeling (Fig. 3d). In accordance with previous results[8], we observed that both modules were enriched for genes in SME15, a module linked to beta oscillatory signatures containing genes implicated in splicing (Fisher's exact test, FDR-corrected $P = 2.2 \times 10^{-10}$, OR = 3.9 (WM11) and $P = 2.1 \times 10^{-05}$, OR = 3.8 (WM22)) (Fig. 3c). These data may support alternative splicing regulation as a mechanism for variation in oscillatory signatures observed across individuals.

Modules of memory oscillatory signatures are linked with neuropsychiatric disorders. We next investigated the association of SME modules with genomic data from brain disorders. Using comprehensive transcriptomic and genetic data from multiple disorders (Methods), we assessed enrichment for genes dysregulated in neuropsychiatric disorders and GWAS enrichment using linkage disequilibrium (LD) score regression. The WM4 module associated with delta oscillatory signatures was significantly enriched for downregulated genes in ASD (Fisher's exact test, FDR-corrected $P = 4.3 \times 10^{-4}$, OR = 2.95) and variants associated with ASD (Fisher's exact test, FDR-corrected $P = 0.001$) (Fig. 4a,b and Supplementary Table 4). WM12 showed enrichment for GWAS associated with attention-deficit/hyperactivity disorder (ADHD; FDR = 0.001), bipolar disorder (BD; FDR = 0.003), major depressive disorder (MDD; FDR = 0.006), schizophrenia (SCZ_2018; FDR = 5.8×10^{-6}) and variants associated with educational attainment (FDR = 0.03) and intelligence (FDR = 0.002) (Fig. 4b and Supplementary Table 4). Most importantly, we did not detect any significant enrichment for epilepsy-associated loci in the memory-related modules (Fig. 4b), and the enrichment for variants associated with non-brain-related

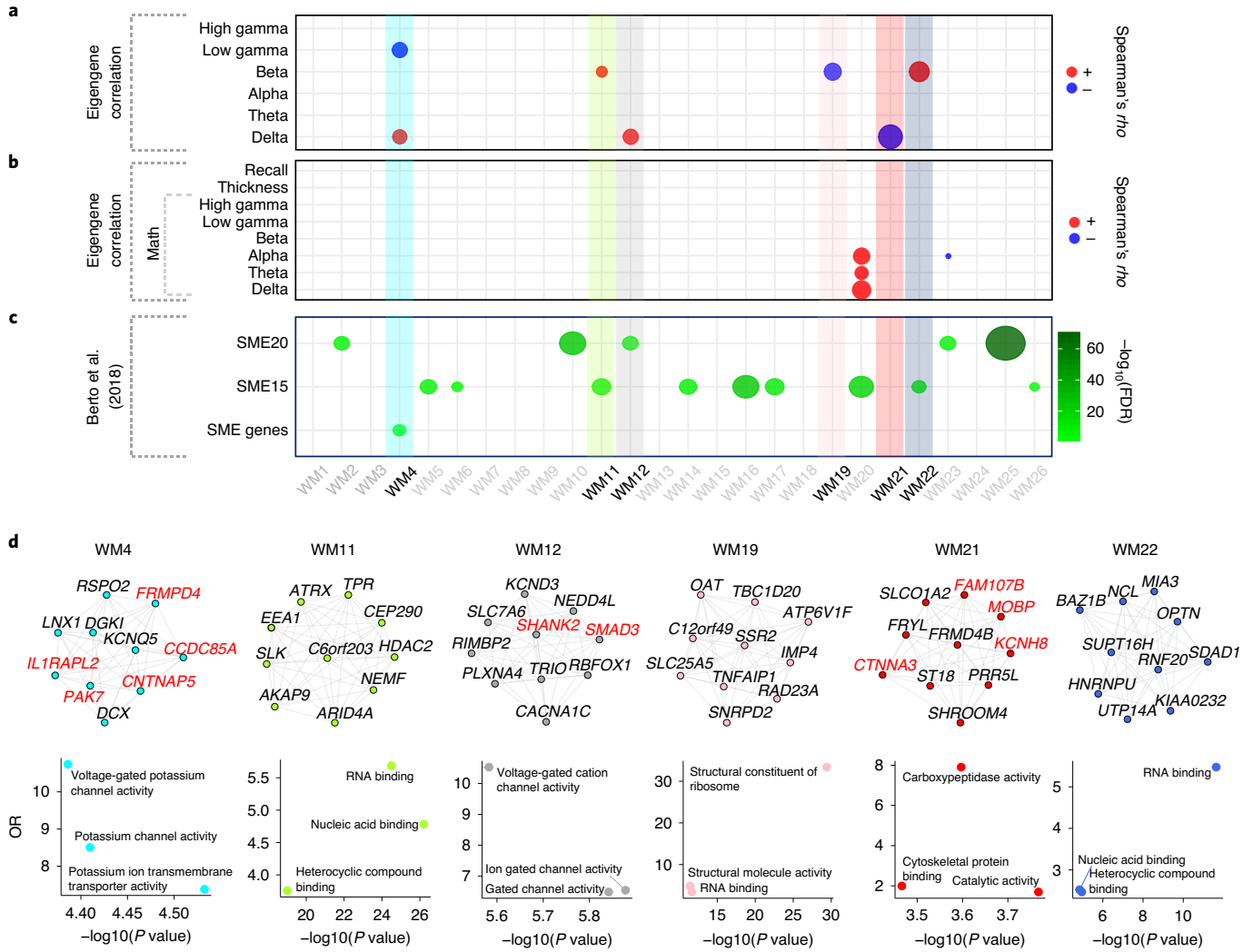


Fig. 3 | Gene co-expression networks highlight cellular processes implicated in memory encoding. a, Bubble chart showing the module eigengene association with brain oscillations (Spearman's $\rho > 0.5$, $P < 0.05$). Six modules have significant correlation. Positive correlations are indicated by red bubbles and negative correlations are indicated by blue bubbles. The size of the bubble corresponds to the strength of the correlation. **b**, Bubble chart showing the module eigengene association with math-associated brain oscillations, percentage of recall values and thickness values (Spearman's $\rho > 0.5$, $P < 0.05$). Only two modules are significantly associated with math, and no modules are significantly associated with memory performance or cortical thickness. **c**, Bubble chart showing the enrichment for previously identified SME genes from a population-based study[8]. The gradient color represents the $-\log_{10}$(FDR). Boldface type indicates the six modules significantly associated with brain oscillations as indicated in **a**. **d**, Representation of the top ten hub genes for the six modules significantly associated with SME signals. SME genes significantly associated with brain oscillations are highlighted in red. Edges represent co-expression ($\rho^2 > 0.25$). Scatterplots represent the top three molecular functions of each module as assessed by Gene Ontology analyses. The *y* axis represents the OR, while the *x* axis represents the $-\log_{10}$(FDR).

traits and disorders was minimal (Extended Data Fig. 4a). We also found an enrichment for WM4 (FDR = 8.9×10^{-07}, OR = 3.3) and WM12 (ASD: FDR = 3.9×10^{-06}, OR = 3.4; ASD (scored 1–3): FDR = 3.2×10^{-04}, OR = 4.5) in ASD-associated genes from the SFARI Gene database (Fig. 4c).

We next compared the correlated modules with those found in a meta-analysis of transcriptomic data across neuropsychiatric disorders[22]. Both WM4 and WM12 are enriched for a module severely affected in ASD, with *RBFOX1* as a predominant hub (geneM1; FDR = 1.5×10^{-29}, OR = 7.9 (WM4) and FDR = 2.0×10^{-10}, OR = 3.92 (WM12)) (Extended Data Fig. 4b). Interestingly, *RBFOX1* is also a hub in WM12 (Fig. 3d), which provides further support for the role of this gene in neuropsychiatric disorders and memory. The beta module WM21 was enriched for schizophrenia variants (SCZ_2018; FDR = 0.03) (Fig. 4a,b and Supplementary Table 4), whereas the beta module WM22 was enriched for a splicing module affected in schizophrenia (geneM19; FDR = 2.6×10^{-09}, OR = 6.6) (Extended Data Fig. 4b). Overall, the association of delta and beta oscillatory-signature-correlated modules with neuropsychiatric disorders for which memory is impaired provide further support for the role of these genes and pathways in episodic memory.

Modules of memory oscillatory signatures are associated with specific cell types. To develop cell-type-specific associations for the identified correlated genes, we performed single-nucleus RNA-seq (snRNA-seq) analysis on tissue from six participants, four of whom contributed oscillatory data (Supplementary Table 1). We sequenced the transcriptomes of 17,632 nuclei (Extended Data Fig. 5a), detecting an overall median of 11,498 unique molecular identifiers (UMIs) and 4,069 genes (Extended Data Fig. 5b,c). We accounted

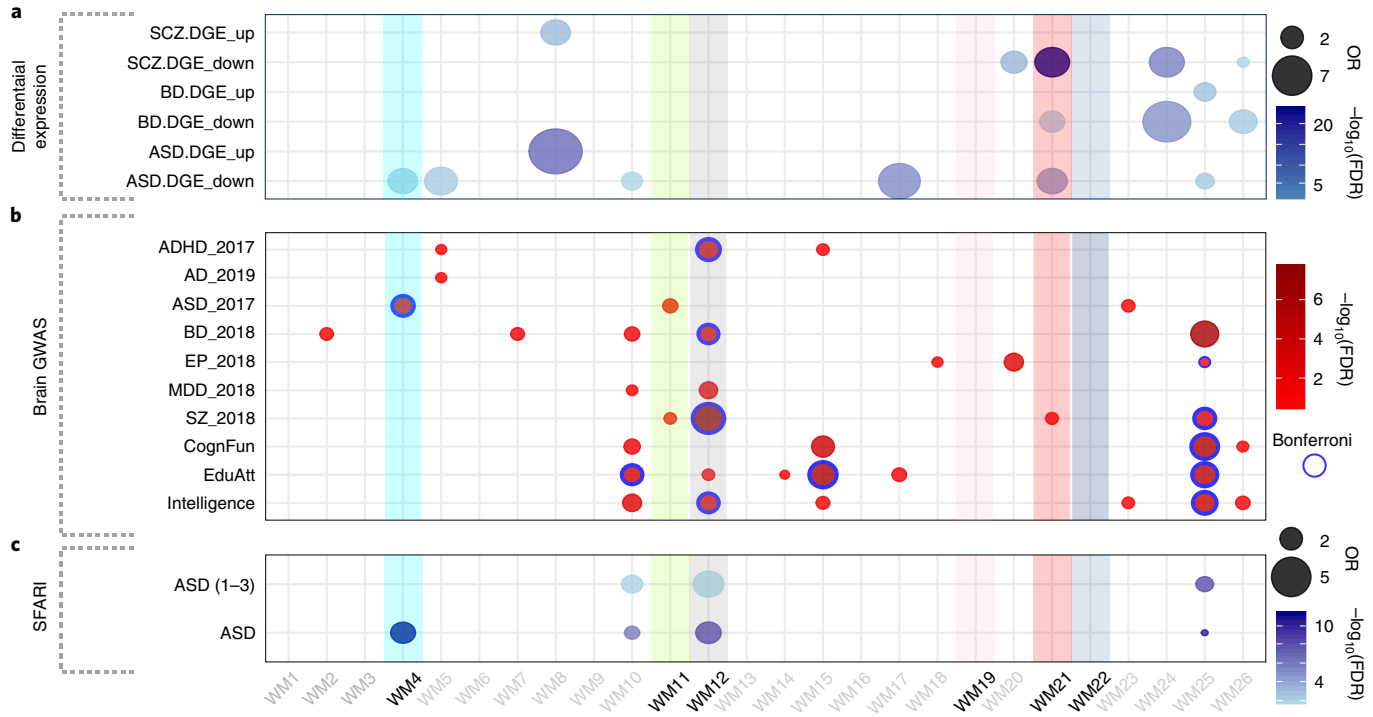


Fig. 4 | SME-specific modules capture genes dysregulated in neuropsychiatric disorders. a, Bubble chart showing the enrichment for genes dysregulated in multiple disorders. Up/down are genes that are upregulated or downregulated, respectively, in these disorders compared with healthy individuals. The gradient color represents the $-\log_{10}$(FDR) and the bubble size represents the OR from a Fisher's exact enrichment test of each module with disease-relevant gene lists. The *y* axis shows the acronyms for the disorders (ASD, BD and SCZ). The *x* axis indicates the modules from the present study. **b**, Bubble chart showing the enrichment for risk loci and loci associated with neuropsychiatric disorders and complex traits (see Methods for definitions of the acronyms). The gradient color represents the $-\log_{10}$(FDR) from LD gene set analysis performed using MAGMA. Blue border corresponds to the Bonferroni-correction threshold (Bonferroni $P < 0.05$). The *y* axis shows the acronyms for the GWAS data utilized for this analysis. The *x* axis shows the modules from the present study**. c**, Bubble chart showing the enrichment for genes associated with ASD in the SFARI Gene database. The gradient color represents the $-\log_{10}$(FDR) and the bubble size represents the OR from a Fisher's exact enrichment test. The *y* axis shows the acronyms for the complete SFARI database (ASD) and highly scored ASD genes (categories 1–3). The *x* axis shows the modules from the present study.

for technical and biological covariates before dimensionality reduction (Methods). We initially identified 24 clusters. We next used a publicly available snRNA-seq dataset from middle temporal gyri to further define our initial clusters by both cell-type and layer specificity (Methods and Supplementary Table 5). After the comparison based on marker enrichment (Methods), we focused on a robust set of 20 transcriptionally defined clusters (Fig. 5a). The proportion of cells were similarly distributed by participant in all clusters (Extended Data Fig. 5d,e). In total, we defined nine inhibitory neuron, eight excitatory neuron and three major non-neuronal clusters (Extended Data Fig. 5f,g). These clusters showed high expression of known major markers for their respective cell types (Fig. 5b and Supplementary Table 5).

We found that the delta-correlated modules WM4 and WM12 were strongly enriched for excitatory and inhibitory neurons (Fig. 5c). Specifically, WM4 and WM12 were highly enriched for combinations of *RORB*$^+$*THEMIS*$^+$*FEZF2*$^+$ deep-layer excitatory neurons. These deep-layer neurons have been associated with memory encoding circuitry receiving GABAergic inputs from the hippocampus[23]. In addition, delta rhythmicity might arise from deeper layer intrinsic bursting neurons[24] that project to other subcortical regions[25]. Therefore, these results further underscore the importance of these deep-layer excitatory neurons in episodic memory encoding. Moreover, both modules showed enrichments for combinations of *SST*$^+$*VIP*$^+$*PVALB*$^+$ inhibitory neurons. Interestingly, fast-spiking parvalbumin (PVALB)-containing basket cells decisively control excitatory output, and they are required for memory consolidation regulating neocortical–hippocampal circuitry[26]. Meanwhile, somatostatin (SST)-expressing neurons target distal dendrites of pyramidal cells[27], and they play a role in memory circuitry and cortical oscillatory synchronization[28]. While *SST*$^+$*PVALB*$^+$ interneurons specifically inhibit pyramidal neurons, *VIP*$^+$ neurons both inhibit and disinhibit pyramidal neurons[29,30] and might be implicated in working memory circuitry[31].

In addition, the module negatively associated with delta oscillatory signatures, WM21, was enriched for glia cells, with a predominance of oligodendrocyte-related genes (Fig. 5c), which provides support for a possible role for oligodendrocytes in memory circuits and neuronal synchrony as previously reported elsewhere[32]. Moreover, using snRNA-seq data from brain tissue of patients with ASD or Alzheimer disease (Methods), we found that WM4 is significantly enriched for genes dysregulated in layer 2–4 excitatory neurons and *SST*$^+$ inhibitory neurons in ASD, whereas WM21 is significantly enriched for oligodendrocyte markers affected in Alzheimer disease (Extended Data Fig. 5h,i). These results confirm the role of the modules associated with delta oscillatory signatures as linked to cognitive disorders at the cell-type level.

WM4 and WM12 are both enriched for delta-oscillatory-signature-correlated genes, cognitive-disease-related variants and multiple neuronal types. To validate our approach for the purpose of identifying targets for the future development of neuromodulation strategies specific to brain disorders and cell types, we selected one hub gene from one of the delta modules. *IL1RAPL2*, which encodes an interleukin-1 (IL-1) receptor accessory protein, is a

hub gene in the WM4 module. Intriguingly, along with its paralog IL1RAPL1, IL1RAPL2 promotes functional excitatory synapse and dendritic spine formation[33] and is associated with ASD[34]. Our snRNA-seq data showed that *IL1RAPL2* has the greatest expression in *RORB*$^+$ deep-layer excitatory neurons, but it is also expressed in *SST*$^+$*LAMP5*$^+$ upper layer inhibitory neurons (Fig. 5d). Fluorescence immunohistochemistry (IHC) analysis of independently obtained tissue resections showed that IL1RAPL2 has the greatest overlapping expression with a marker of excitatory neurons (CAMKII), some overlap with a marker of inhibitory neurons (GAD67) and no overlap with a marker of astrocytes (GFAP) or a marker of oligodendrocytes (OLIG2) (Fig. 5e,f). Along with its role in excitatory synapse formation, the snRNA-seq and memory oscillatory signature association indicated that *IL1RAPL2* might play an essential role in regulation of memory encoding in humans. Together, these results underscore the importance of further studies focused on the role of *IL1RAPL2* in memory and excitatory–inhibitory synaptic etiologies.

snATAC-seq reveals transcription factors as key regulators of memory-correlated modules. We next sought to understand what transcription factors (TFs) regulate modules of memory oscillatory signatures. We performed single-nucleus ATAC-seq (snATAC-seq) analysis on tissue obtained from three unique participants (Supplementary Table 1). We assessed the chromatin state of 22,177 nuclei (Extended Data Fig. 6a), with a median of 7,733 identified peaks (Extended Data Fig. 6b,c). We identified 17 clusters that were labeled by integrating the snATAC-seq data with the snRNA-seq data (Fig. 6a, Extended Data Fig. 6d and Methods). The proportion of nuclei from the three participants were similarly distributed among the clusters (Extended Data Fig. 6e). We noted differences between the resolution of the snRNA-seq and snATAC-seq datasets in terms of the cell types identified, with a high percentage of non-neuronal cells in the snATAC-seq dataset (Extended Data Fig. 6f,g). We speculate that this difference might be due to a bias in the snRNA-seq data caused by a larger amount of RNA and expressed genes in neuronal cell types[35]. Indeed, the glia-to-neuron ratio (GNR) in human gray matter varies between 1.13 and 1.64 (ref. [36]). The GNR resolved by snATAC-seq was in line with this assumption (~1), whereas snRNA-seq data underestimated the GNR (~0.15) (Extended Data Figs. 5g and 6f,g).

Overall, this multi-omics method allowed us to detect cell-type-specific regulatory loci whose accessibility profiles were consistent with the cell-type gene expression. Using motif analysis, we explored the enrichment of TFs in the cell-type-specific regulatory loci associated with the identified modules of memory oscillatory signatures. Among the modules with a cell-type association, motif enrichment was detected only in WM12 (Fig. 6b and Supplementary Table 6). Interestingly, we found that WM12 showed enrichment for *SMAD3* motifs, a WM12 hub gene (Fig. 6b and Supplementary Table 6). Remarkably, *SMAD3* motifs were observed in the promoter regions of other WM12 hub genes associated with neuropsychiatric disorders and memory such as *SHANK2* (ref. [20]) (Fig. 6c). In addition, WM12 contained genes associated with neuronal etiologies, and we found that *SMAD3* is primarily expressed in excitatory neurons (Fig. 6d). This result was further confirmed by fluorescence IHC analysis of independently obtained tissue resections (Fig. 6e,f). Overall, these results highlight the role of specific TFs in the regulation of the chromatin landscape necessary to express putative genes associated with memory oscillatory signatures and provide novel molecular entry points for understanding human memory.

Discussion

We set out to understand the genomic underpinnings of oscillatory patterns that support episodic memory encoding in humans, with the goal of identifying genes that are propitious targets for neuromodulation strategies to treat memory disorders. Using an unparalleled dataset from 16 human participants that included measurements of brain oscillations linked to successful episodic memory encoding and transcriptomic data from the temporal pole in the same individuals, we identified modules of genes that link specific cell types and cellular functions with memory-related oscillatory signatures.

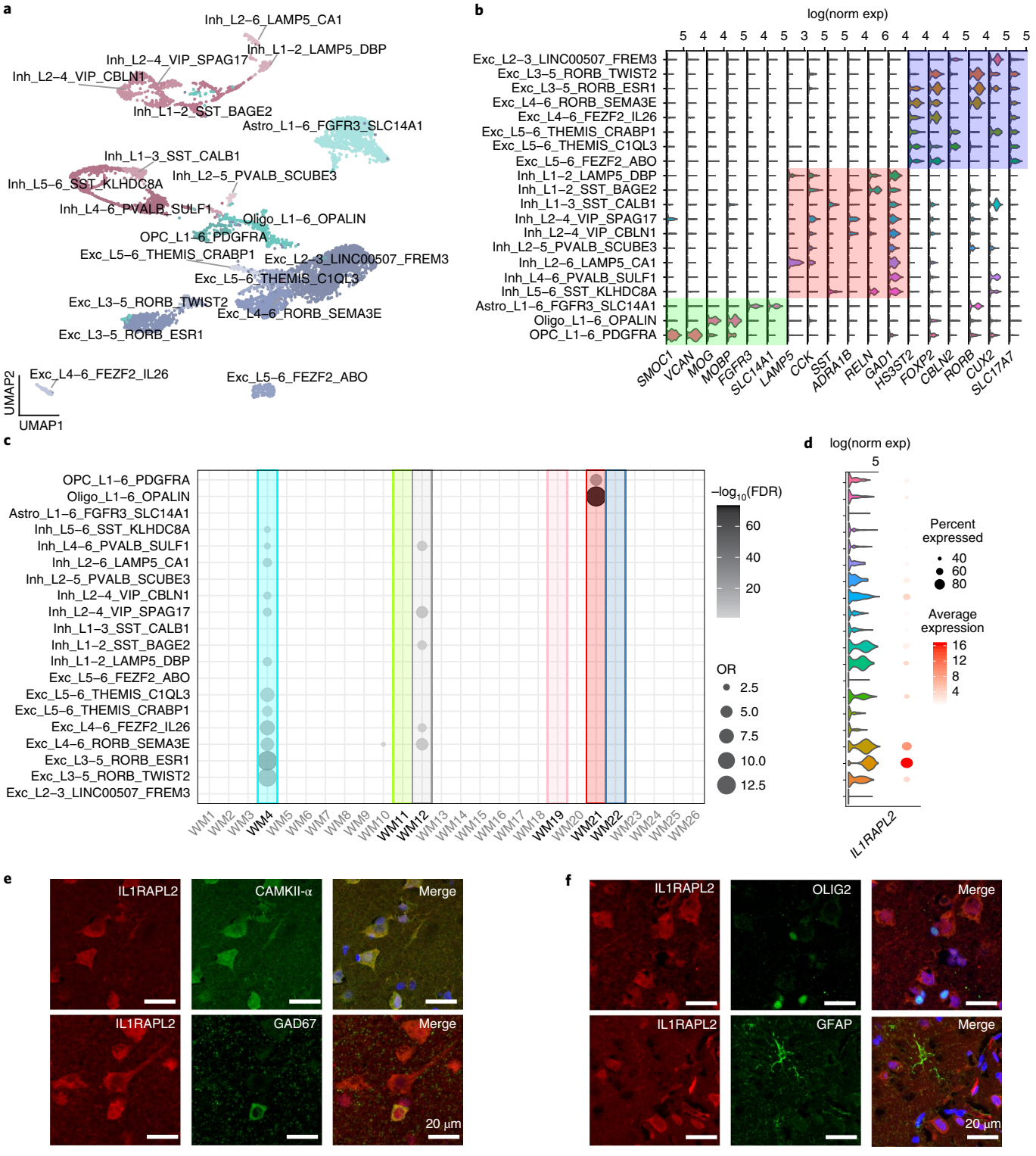
Our analysis is fundamentally different from previous attempts to correlate gene expression with behavioral measurements such as memory performance[37,38]. Oscillatory correlates of successful memory encoding represent an 'intermediate step' between gene regulation and memory behavior. Oscillations are localized to the brain region in which they are recorded using intracranial-depth electrodes and are dissociable into frequency bands with distinct properties. Linking neurophysiological measurements (such as these oscillatory signatures) with gene expression data will establish specific testable hypotheses in subsequent investigations for these identified genes. The hub genes described in Fig. 3d may represent the most propitious targets for subsequent testing using animal models or other approaches.

Our work sheds light on the molecular mechanisms that give rise to oscillatory correlates of successful memory encoding[39]. Our observation that delta oscillatory signatures are linked to ion channel genes and that these genes tend to be expressed in oligodendrocytes leads to the fascinating implication that the generation of low-frequency oscillatory patterns linked to mnemonic processing in humans is at least partially dependent on glial modulation of oscillations. This is based on our observations across all participants and on the single-nucleus expression analysis. This conclusion is supported by the role of oligodendrocytes in learning and memory acting on depolarization of membrane potential[32,40], which accelerates axonal conduction and ion channel activity as reflected by the delta-associated modules with positive association (WM4 and

Fig. 5 | SME-specific modules are enriched for excitatory and inhibitory neurons. a, UMAP representation of the 20 classes of cell types using the BA38 snRNA-seq data. Each dot represents a nucleus. Excitatory neurons (Exc) are highlighted in a dark blue gradient, the inhibitory neurons (Inh) in a red gradient and the non-neuronal cells (Astro, Oligo and OPC) in a light blue gradient. Cell types were annotated using a publicly available single-cell dataset. A Fisher's exact enrichment test between cell markers of the two datasets was performed. Major cell types tend to cluster near one another. **b**, Violin plots representing gene markers for the major cell types detected. The *y* axis represents the log-normalized expression (log(norm exp)) of each marker gene in each cluster. The markers for excitatory neurons (for example, *CUX* and *RORB*) are highlighted in blue. The markers for inhibitory neurons (for example, *GAD1* and *RELN*) are highlighted in red. The markers for non-neuronal cells (for example, *FGFR3, MOBP* and *VCAN*) are highlighted in green. **c**, Bubble chart showing the enrichment of the SME modules for cell-type markers defined using Seurat. The color gradient represents the $-\log_{10}$(FDR) and the bubble size represents the OR from a Fisher's exact enrichment test of genes in modules from this study with genes expressed in specific cell types defined by our snRNA-seq data. The *x* axis represents the SME-specific modules. The *y* axis represents the cell classes from the present study. Boldface type indicates the six modules significantly associated with memory-related brain oscillations. **d**, Violin plot representing the log-normalized expression level of *IL1RAPL2*. The adjacent dot plot represents the average expression (gradient) and percentage of cells (size) expressing *IL1RAPL2*. The order of cell types follows the labels of **c**. **e**,**f**, IHC of independent human temporal lobe specimens demonstrates the specific expression of IL1RAPL2 in excitatory (CAMKII-α^+) and inhibitory neurons (GAD67$^+$) in BA38 (**e**), but not in oligodendrocytes (OLIG2$^+$) or astrocytes (GFAP$^+$) (**f**).

WM12). Moreover, genes expressed in these positively associated modules were overrepresented in deep layers of excitatory neurons implicated in memory-encoding circuitry and delta-rhythmicity formation[41–43] and in *SST*$^{+}$*VIP*$^{+}$*PVALB*$^{+}$-expressing interneurons important for mediating cortical–hippocampal communication during memory encoding[44]. These results further support the role of the identified genes in memory encoding and specifically highlight cell types that might be implicated in episodic memory.

We observed interesting properties for genes correlated with delta oscillations, but not theta oscillations, which runs contrary to rodent data that universally implicate theta frequency activity in successful memory formation. However, in the human temporal lobe, oscillations outside the 4–9 Hz range routinely exhibit memory-relevant properties, including cross-frequency coupling; thus, our findings are in line with previous observations using oscillatory signatures of successful memory encoding in humans[45]. In humans, these low-frequency oscillations represent a consistent feature of oscillatory signatures of memory formation, including influence on the timing of single unit activity[45–47]. The significant representation of genes correlated with delta oscillatory signatures



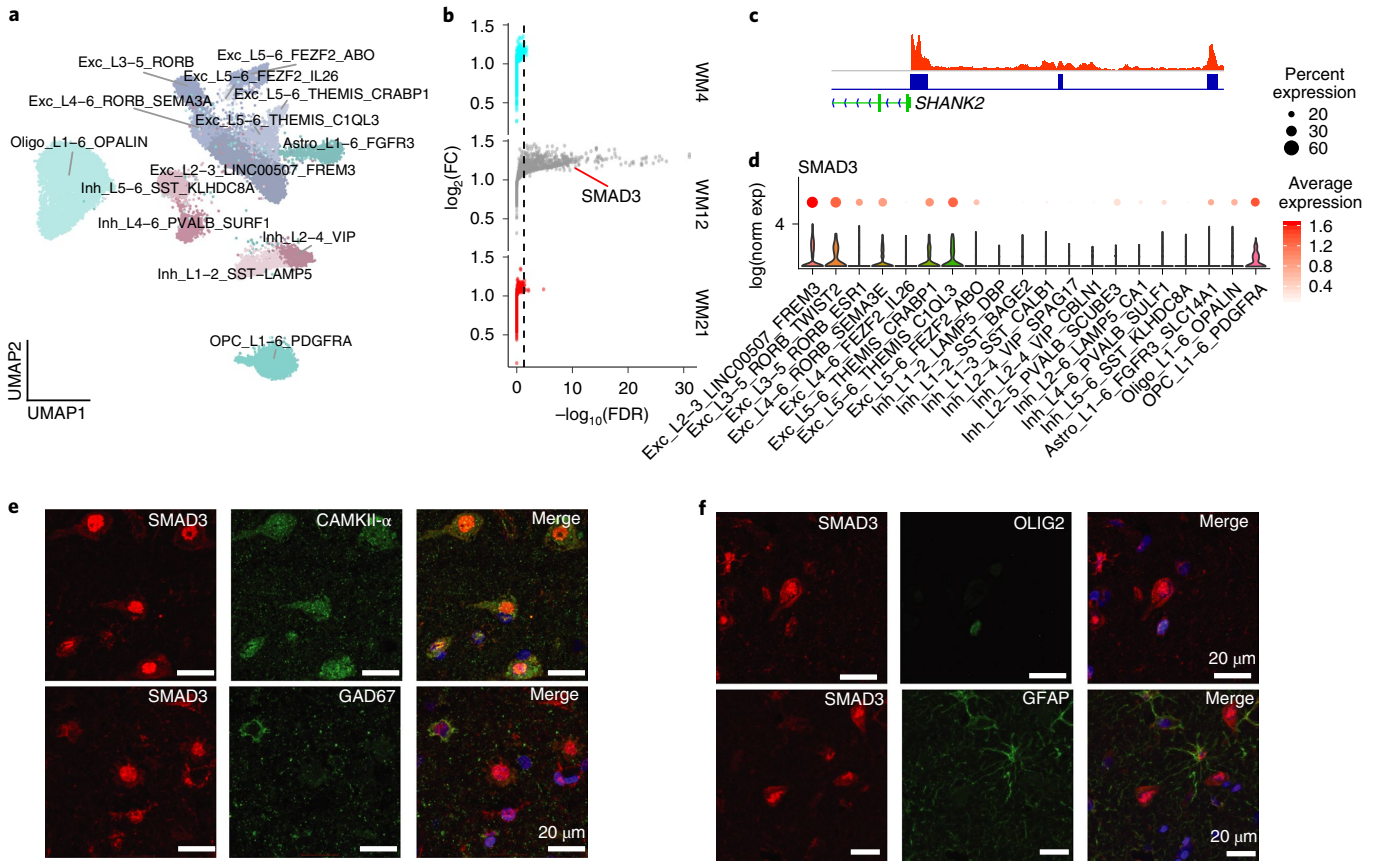


Fig. 6 | snATAC-seq highlights TFs regulating SME-correlated modules. a, Visualization of the 14 classes of cell types identified from BA38 snATAC-seq data. Nuclei are displayed based on UMAP. Each dot represents a nucleus. Cell classes were annotated using the BA38 snRNA-seq data generated in this study. Excitatory neurons are highlighted in a gradient of blue colors, the inhibitory neurons in a gradient of red and the non-neuronal cells in a gradient of turquoise. **b**, TF binding site enrichment for the three modules associated with cell types (WM4, WM12 and WM21). Only WM12 tends to have enrichment of TF binding motifs (dots to the right of the dashed line). SMAD3 is shown as a top TF whose motif is enriched in the WM12 module. The *y* axis represents the $-\log_2$(FC) of the motif enrichment reported by FindMotifs in Seurat. The *x* axis represents the $-\log_{10}$(FDR) of the motif enrichment reported by FindMotifs in Seurat. The dashed line corresponds to FDR = 0.05. **c**, Genome visualization tracks of snATAC-seq open chromatin regions representing SMAD3 binding sites in the promoter of the WM12 hub gene *SHANK2*. The red ridge plot represents the snATAC-seq data. The SMAD3 binding sites are indicated in blue. **d**, Violin plot representing the log-normalized expression level (*y* axis) of *SMAD3* for each cell type defined by snRNA-seq. The adjacent dot plot represents the average expression (gradient) and percentage of cells (size) expressing *SMAD3*. **e**,**f**, IHC of independent human temporal lobe specimens demonstrates the specific expression of SMAD3 in excitatory neurons (CAMKII-α$^+$; **e**), but not in inhibitory neurons (GAD67$^+$; **e**), oligodendrocytes (OLIG2$^+$; **f**) or astrocytes (GFAP$^+$; **f**) in BA38.

in our analysis may reflect the functional importance of these low-frequency components in humans.

A caveat in interpreting our data is that all the participants suffered from intractable epilepsy. Clearly, the use of such a specific population is necessary to generate these highly valuable data with both in vivo oscillations and gene expression data from the same individuals. However, several features of our analysis, such as under-enrichment for genetic variants associated with epilepsy and data integration with epileptic and healthy tissues, give us confidence that the insights we have uncovered represent more generalizable associations between gene expression and brain oscillations. Furthermore, numerous human studies have established that iEEG observations from participants have correlates using noninvasive studies of healthy participants and in animal models[48–50]. Moreover, we employed strict artifact-rejection criteria and eliminated electrodes located in the seizure-onset zone in our analysis, thereby reducing the impact of abnormal activity on observed oscillatory signatures[51]. We also integrated several control steps in our analysis, including incorporation of the duration of epilepsy to adjust gene expression values. Finally, several of the key genes we identified (for example, *IL1RAPL2* and *SMAD3*) have been independently shown to be linked to memory processing in data from non-epileptic individuals and genetically modified rodent models[34,52]. Even though these correlative analyses do not imply causality, these genes have been highlighted by stringent correlative statistics, by high connectivity in the modules associated with memory oscillations and by cell-type expression specificity. Using this analytical approach, we defined *IL1RAPL2* and *SMAD3* as genomic markers for episodic memory for further investigation at the molecular level in model systems.

An important issue one must consider when using participants who underwent neurosurgery to obtain both oscillation and gene expression data relates to timing. Specifically, the use of human participants simply does not allow collection of tissue specimens immediately after behavior-related oscillations are observed. Brain oscillations are dynamic, occurring during specific behavior, but gene expression snapshots are taken later in time, when the participants underwent temporal lobectomy. In practical terms, this means that genes we identify as being linked to oscillatory signatures of successful memory formation necessarily must persist in their

expression at least over a period of weeks, and that our study cannot identify genes whose expression is differentially induced (across participants) due to mnemonic stimuli over shorter time scales. We also note that while we use the term "oscillations" to describe power extracted in six predefined frequency bands, we acknowledge that the measurement of SMEs may reflect power differences that arise due to differences in both narrowband oscillations and broadband power shifts. We include examples of narrowband oscillations detected in our data using the multiple oscillation detection algorithm (MODAL; Methods). Future investigations may establish whether gene-expression correlation patterns are additionally correlated with such broadband power shifts during encoding[53], incorporating slope shifts or a quantification of episodes in which bursts of oscillations occur. Broadly stated, this area remains an active area of investigation in human electrophysiology[54].

Collectively, this translational work establishes an experimental and analytical approach for deconstructing human behavioral and cognitive traits, such as memory, using integrative physiological and multi-omics techniques. Integration of single-nucleus transcriptomic and epigenomic data allowed us to identify the cell-type specificity of the memory-related gene co-expression modules and potential regulators of these modules. This molecular characterization of human memory highlights key genes that can be further studied in model systems. We anticipate that this within-individual approach can be used in future studies to highlight molecular pathways of other human complex traits with the goal of identifying therapeutic targets and linking clinical and genomic data at the individual level. Importantly, investigations using animal and in vitro models will be necessary to definitively characterize the memory-related properties of the genes identified in our analysis.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-021-00803-x.

Received: 25 November 2019; Accepted: 19 January 2021;
Published online: 8 March 2021

References

1. Morgan, S. E. et al. Cortical patterning of abnormal morphometric similarity in psychosis is associated with brain expression of schizophrenia-related genes. *Proc. Natl Acad. Sci. USA* **116**, 9604–9609 (2019).
2. Lombardo, M. V. et al. Large-scale associations between the leukocyte transcriptome and BOLD responses to speech differ in autism early language outcome subtypes. *Nat. Neurosci.* **21**, 1680–1688 (2018).
3. Romero-Garcia, R., Warrier, V., Bullmore, E. T., Baron-Cohen, S. & Bethlehem, R. A. I. Synaptic and transcriptionally downregulated genes are associated with cortical thickness differences in autism. *Mol. Psychiatry* **24**, 1053–1064 (2019).
4. Wang, G. Z. et al. Correspondence between resting-state activity and brain gene expression. *Neuron* **88**, 659–666 (2015).
5. Patania, A. et al. Topological gene expression networks recapitulate brain anatomy and function. *Netw. Neurosci.* **3**, 744–762 (2019).
6. Le, B. D. & Stein, J. L. Mapping causal pathways from genetics to neuropsychiatric disorders using genome-wide imaging genetics: current status and future directions. *Psychiatry Clin. Neurosci.* **73**, 357–369 (2019).
7. Konopka, G. Cognitive genomics: linking genes to behavior in the human brain. *Netw. Neurosci.* **1**, 3–13 (2017).
8. Berto, S., Wang, G. Z., Germi, J., Lega, B. C. & Konopka, G. Human genomic signatures of brain oscillations during memory encoding. *Cereb. Cortex* **28**, 1733–1748 (2018).
9. Long, N. M., Burke, J. F. & Kahana, M. J. Subsequent memory effect in intracranial and scalp EEG. *NeuroImage* **84**, 488–494 (2014).
10. Mukamel, R. & Fried, I. Human intracranial recordings and cognitive neuroscience. *Annu. Rev. Psychol.* **63**, 511–537 (2012).
11. Sederberg, P. B. et al. Hippocampal and neocortical gamma oscillations predict memory formation in humans. *Cereb. Cortex* **17**, 1190–1196 (2007).
12. Nakamura, K. & Kubota, K. The primate temporal pole: its putative role in object recognition and memory. *Behav. Brain Res.* **77**, 53–77 (1996).
13. Hill, P. F., King, D. R., Lega, B. C. & Rugg, M. D. Comparison of fMRI correlates of successful episodic memory encoding in temporal lobe epilepsy patients and healthy controls. *NeuroImage* **207**, 116397 (2020).
14. Sederberg, P. B., Howard, M. W. & Kahana, M. J. A context-based theory of recency and contiguity in free recall. *Psychol. Rev.* **115**, 893–912 (2008).
15. Arora, A. et al. Comparison of logistic regression, support vector machines, and deep learning classifiers for predicting memory encoding success using human intracranial EEG recordings. *J. Neural Eng.* **15**, 066028 (2018).
16. Lin, J. J. et al. Theta band power increases in the posterior hippocampus predict successful episodic memory encoding in humans. *Hippocampus* **27**, 1040–1053 (2017).
17. Sederberg, P. B., Kahana, M. J., Howard, M. W., Donner, E. J. & Madsen, J. R. Theta and gamma oscillations during encoding predict subsequent recall. *J. Neurosci.* **23**, 10809–10814 (2003).
18. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
19. Berkel, S. et al. Mutations in the *SHANK2* synaptic scaffolding gene in autism spectrum disorder and mental retardation. *Nat. Genet.* **42**, 489–491 (2010).
20. Peykov, S. et al. Rare SHANK2 variants in schizophrenia. *Mol. Psychiatry* **20**, 1487–1488 (2015).
21. Won, H. et al. Autistic-like social behaviour in Shank2-mutant mice improved by restoring NMDA receptor function. *Nature* **486**, 261–265 (2012).
22. Gandal, M. J. et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362**, eaat8127 (2018).
23. Dembrow, N. C., Zemelman, B. V. & Johnston, D. Temporal dynamics of L5 dendrites in medial prefrontal cortex regulate integration versus coincidence detection of afferent inputs. *J. Neurosci.* **35**, 4501–4514 (2015).
24. Silva, L. R., Amitai, Y. & Connors, B. W. Intrinsic oscillations of neocortex generated by layer 5 pyramidal neurons. *Science* **251**, 432–435 (1991).
25. Kim, E. J., Juavinett, A. L., Kyubwa, E. M., Jacobs, M. W. & Callaway, E. M. Three types of cortical layer 5 neurons that differ in brain-wide connectivity and function. *Neuron* **88**, 1253–1267 (2015).
26. Xia, F. et al. Parvalbumin-positive interneurons mediate neocortical-hippocampal interactions that are necessary for memory consolidation. *eLife* https://doi.org/10.7554/eLife.27868 (2017).
27. Naka, A. et al. Complementary networks of cortical somatostatin interneurons enforce layer specific control. *eLife* https://doi.org/10.7554/eLife.43696 (2019).
28. Veit, J., Hakim, R., Jadi, M. P., Sejnowski, T. J. & Adesnik, H. Cortical gamma band synchronization through somatostatin interneurons. *Nat. Neurosci.* **20**, 951–959 (2017).
29. Pi, H. J. et al. Cortical interneurons that specialize in disinhibitory control. *Nature* **503**, 521–524 (2013).
30. Lee, S., Kruglikov, I., Huang, Z. J., Fishell, G. & Rudy, B. A disinhibitory circuit mediates motor integration in the somatosensory cortex. *Nat. Neurosci.* **16**, 1662–1670 (2013).
31. Kamigaki, T. & Dan, Y. Delay activity of specific prefrontal interneuron subtypes modulates memory-guided behavior. *Nat. Neurosci.* **20**, 854–863 (2017).
32. Pepper, R. E., Pitman, K. A., Cullen, C. L. & Young, K. M. How do cells of the oligodendrocyte lineage affect neuronal circuits to influence motor function, memory and mood? *Front. Cell. Neurosci.* **12**, 399 (2018).
33. Um, J. W. & Ko, J. LAR-RPTPs: synaptic adhesion molecules that shape synapse development. *Trends Cell Biol.* **23**, 465–475 (2013).
34. Kantojarvi, K. et al. Fine mapping of Xq11.1-q21.33 and mutation screening of *RPS6KA6*, *ZNF711*, *ACSL4*, *DLG3*, and *IL1RAPL2* for autism spectrum disorders (ASD). *Autism Res.* **4**, 228–233 (2011).
35. Tasic, B. et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19**, 335–346 (2016).
36. Azevedo, F. A. et al. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J. Comp. Neurol.* **513**, 532–541 (2009).
37. Zhang, Y. et al. Polymorphisms in human dopamine D2 receptor gene affect gene expression, splicing, and neuronal activity during working memory. *Proc. Natl Acad. Sci. USA* **104**, 20552–20557 (2007).
38. Papassotiropoulos, A. et al. Common *Kibra* alleles are associated with human memory performance. *Science* **314**, 475–478 (2006).
39. Jacobs, J. & Kahana, M. J. Direct brain recordings fuel advances in cognitive electrophysiology. *Trends Cogn. Sci.* **14**, 162–171 (2010).
40. Yamazaki, Y. et al. Oligodendrocytes: facilitating axonal conduction by more than myelination. *Neuroscientist* **16**, 11–18 (2010).
41. Baker, A. et al. Specialized subpopulations of deep-layer pyramidal neurons in the neocortex: bridging cellular properties to functional consequences. *J. Neurosci.* **38**, 5441–5455 (2018).
42. Carracedo, L. M. et al. A neocortical delta rhythm facilitates reciprocal interlaminar interactions via nested theta rhythms. *J. Neurosci.* **33**, 10750–10761 (2013).

43. Jinno, S. et al. Neuronal diversity in GABAergic long-range projections from the hippocampus. *J. Neurosci.* **27**, 8790–8804 (2007).
44. Kim, D. et al. Distinct roles of parvalbumin- and somatostatin-expressing interneurons in working memory. *Neuron* **92**, 902–915 (2016).
45. Jacobs, J. Hippocampal theta oscillations are slower in humans than in rodents: implications for models of spatial navigation and memory. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369**, 20130304 (2014).
46. Yaffe, R. B. et al. Reinstatement of distributed cortical oscillations occurs with precise spatiotemporal dynamics during successful memory retrieval. *Proc. Natl Acad. Sci. USA* **111**, 18727–18732 (2014).
47. Rutishauser, U., Ross, I. B., Mamelak, A. N. & Schuman, E. M. Human memory strength is predicted by theta-frequency phase-locking of single neurons. *Nature* **464**, 903–907 (2010).
48. Duzel, E., Penny, W. D. & Burgess, N. Brain oscillations and memory. *Curr. Opin. Neurobiol.* **20**, 143–149 (2010).
49. Kahana, M. J. The cognitive correlates of human brain oscillations. *J. Neurosci.* **26**, 1669–1672 (2006).
50. Vaz, A. P., Inati, S. K., Brunel, N. & Zaghloul, K. A. Coupled ripple oscillations between the medial temporal lobe and neocortex retrieve human memory. *Science* **363**, 975–978 (2019).
51. Lega, B., Burke, J., Jacobs, J. & Kahana, M. J. Slow-theta-to-gamma phase-amplitude coupling in human hippocampus supports the formation of new episodic memories. *Cereb. Cortex* **26**, 268–278 (2016).
52. Muñoz, M. D., Antolín-Vallespín, M., Tapia-González, S. & Sánchez-Capelo, A. Smad3 deficiency inhibits dentate gyrus LTP by enhancing GABA neurotransmission. *J. Neurochem.* **137**, 190–199 (2016).
53. Donoghue, T. et al. Parameterizing neural power spectra. *Nat. Neurosci.* **23**, 1655–1665 (2020).
54. Herweg, N. A., Solomon, E. A. & Kahana, M. J. Theta oscillations in human memory. *Trends Cogn. Sci.* **24**, 208–227 (2020).

Methods

Experimental model and participant details. *Participants and memory task.* The research protocol was approved by the Institutional Review Board at UT Southwestern, and informed consent was obtained from each participant. Participants contributing gene expression data were recruited from the UT Southwestern surgical epilepsy program during a preoperative visit before temporal lobectomy. These participants underwent iEEG to map seizure-onset location. Participants needed to complete a full session of the free-recall task with a minimum performance (recall fraction >10%) to be included. Participants performed a free-recall task consisting of multiple study–test cycles. During the study period, 12 words from a preselected pool of high-frequency, single-syllable common nouns were visually presented, one at a time, on a computer screen for a duration of 1.6 s followed by a blank screen of 4 s with 100 ms of random jitter. Participants were instructed to study each word as it appeared on the screen. The presentation of the last item in a list was followed by a 30-s period during which a math distractor task (A + B + C = ??) was performed to limit rehearsal. Participants were then instructed to verbally recall as many items as possible from the immediately prior list in no particular order. A full session consisted of 12 full study–test cycles and 1 practice study–test cycle that was excluded from analysis. One complete session yielded electrophysiological recordings from 144 word-encoding epochs (12 lists × 12 words) and a variable number of retrieval epochs. Participants performed between 1 and 9 sessions of the free-recall task over several days (median number of 2). Behavioral performance was measured by calculating the fraction of successfully recalled memory items. A list intrusion rate was measured for each participant (the proportion of recall attempts that were classified as list intrusions relative to veridical recall). List intrusions are items seen on previous lists incorrectly recalled on the list being tested.

iEEG processing. iEEG data were recorded using a Nihon Kohden EEG-1200 clinical system. Signals were sampled at 1,000 Hz and referenced to a common intracranial contact. Raw signals were subsequently re-referenced to an average reference montage, after excluding channels with frequent interictal activity or other subsequent noise. All analyses were conducted using Matlab with both built-in and custom-made scripts. We employed an automated artifact-rejection algorithm to exclude interictal activity and abnormal trials (kurtosis threshold >4) in line with previous publications using similar iEEG datasets[55].

We compared oscillatory power at 1,600 ms immediately following the study item presentation for subsequently recalled and non-recalled words. The iEEG signal from each encoding epoch along with a 1,500-ms flanking buffer was notch-filtered from 58 to 62 Hz to reduce possible line noise contamination (Butterworth, first order). The filtered signal was then subjected to spectral decomposition using the wavelet transform (width of 6) with log-spaced frequencies from 2 to 120 Hz. The decomposed spectral power values were then averaged across the entire 1,600-ms period. Oscillatory power values were divided into trials for which items were later remembered (recalled) and trials for which items were not remembered (non-recalled). Oscillatory power for the recalled trials was compared to non-recalled trials at each frequency using a two-sample *t*-test to determine the SME. We incorporated a permutation procedure, shuffling trial labels between the two classes 1,000 times to generate an unbiased estimate of the type 1 error rate[17]. We obtained an estimate of the magnitude of the SME by identifying the position of the true *t*-statistic from the distribution of 1,000 *t*-statistics resulting from the randomly shuffled recalled and non-recalled event labels to generate a *P* value. We then applied normal inverse transformation to the *P* value matrices of each electrode to convert them to SME *Z* values to combine across frequency bands. To limit the overall number of comparisons in our analysis, we averaged the SME *Z* values into six frequency bands (delta 2–4 Hz, theta 4–8 Hz, alpha 8–16 Hz, beta 16–30 Hz, low gamma 30–70 Hz and high gamma 70–120 Hz). Because our goal was to determine how variance in memory-related oscillatory patterns depend on differences in gene expression, there was no SME threshold applied to filter which electrodes were included in the gene correlation analysis (we included data from all the electrodes). We made the a priori decision to average all SME estimates (*Z* values) across the region of interest (BA38) within each participant before calculating gene correlations, which we believed was the most unbiased method for this analysis. Data distribution was assumed to be normal, but this was not formally tested.

We also measured oscillatory power differences for successful versus unsuccessful math trials for use in the control analysis (described below). This utilized the same shuffling procedure as described above and the same methods for extraction of signal across electrodes. To identify the presence of narrowband oscillations during successful events during memory encoding, we used an oscillation-detection algorithm. Artifact-free trials from BA38 region electrodes were used to identify peak frequencies using MODAL for frequency ranges from 2 to 50 Hz. This algorithm included a procedure to remove 1/f fit from the power spectrum and adaptively identify frequency bands[56,57].

Anesthetic. For all samples, we calculated the time under anesthetic before procurement of tissue using the time of initial anesthetic induction as documented in the "Anesthesia event" encounter in the software Epic patient care. All cases were the first of the day, and a standardized anesthetic induction procedure was utilized incorporating remifentanil, propofol and rocuronium, with desflurane as an inhalational agent during the procedure (0.5 MAC). All patients received dexamethasone before induction. The mean interval between induction and tissue processing from the temporal pole was 219 ± 21 min (95% confidence interval).

FreeSurfer segmentation. FreeSurfer extraction form T1 mprage volume acquisition was used to quantify the cortical thickness in the temporal pole[58]. Volume data for the temporal pole were identified from the aseg.stats files (in millimeters) for each participant (one value per participant).

Resected brain samples. All surgical samples included in this study were BA38 resections from patients with temporal lobe epilepsy. The brain specimen was dropped into ice-cold 1× PBS in a 50-ml conical tube immediately after removal from the patient. After four to five inversions, the tissue sample was transferred to a fresh tube with ice-cold 1× PBS for a second wash. The specimen was then moved to a Petri dish and grossly dissected by scalpel into ~12 subsamples and immediately frozen in individual Eppendorf tubes in liquid nitrogen as the tubes were filled. Care was taken to avoid major blood vessels. Gray matter was prioritized over tracts of white matter in an attempt to increase homogeneity and consistency of results across all samples. Time from removal of brain to flash freezing ranged from roughly 2 min for the first piece to about 7 min for the last subsample. Three to four of the subsamples were extracted for RNA, and the subsample with the highest RIN value was selected for RNA-seq. See Supplementary Table 1 for detailed demographic information.

Tissue preparation for sequencing. *Postmortem brain samples.* Twelve samples of BA38 were obtained from the Dallas Brain Collection. These tissue samples were donated from individuals without a history of neurological or psychiatric disorders, as previously published[59]. Eight samples of BA38 were obtained from the University of Maryland Brain and Tissue Bank. These samples were donated from individuals with epilepsy. See Supplementary Table 1 for detailed demographic information.

RNA-seq. Total RNA was purified using an miRNeasy kit (217004, Qiagen) following the manufacturer's recommendations. RNA-seq libraries from mRNA were prepared in-house as previously described[60]. Sequencing was performed on randomly pooled samples by the McDermott Sequencing Core at UT Southwestern on an Illumina NextSeq 500 sequencer. Single-end, 75-base-pair (bp) reads were generated. Data collection and analysis were not performed blind to the conditions of the experiments. No statistical methods were used to predetermine sample sizes because of the limitation of availability of human brain surgical tissues

Isolation of nuclei from resected brain tissues (snRNA-seq). Nuclei were isolated as previously described[61] (https://www.protocols.io/view/rapid-nuclei-isolation-from-human-brain-scpeavn). Surgically resected cortical tissue was homogenized using a glass Dounce homogenizer in 2 ml of ice-cold Nuclei EZ lysis buffer (EZ PREP NUC-101, Sigma) and was incubated on ice for 5 min. Nuclei were centrifuged at 500 × *g* for 5 min at 4 °C, washed with 4 ml ice-cold Nuclei EZ lysis buffer and incubated on ice for 5 min. Nuclei were centrifuged at 500 × *g* for 5 min at 4 °C. After centrifugation, the nuclei were resuspended in 1 ml of nuclei suspension buffer (NSB) consisting of 1× PBS, 1% BSA (AM2618, Thermo Fisher Scientific) and 0.2 U μl^{-1} RNAse inhibitor (AM2694, Thermo Fisher Scientific) and were filtered through a 40-µm Flowmi Cell Strainer (H13680-0040, Bel-Art). The concentration of nuclei was determined using 0.4% Trypan Blue (15250061, Thermo Fisher Scientific). The final concentration of 1,000 nuclei per µl was adjusted with NSB. Droplet-based snRNA-seq libraries for the first batch were prepared using Chromium Single Cell 3′ v.2 (120237, 10× Genomics) according to the manufacturer's protocol[62]. Libraries were sequenced using an Illumina NextSeq 500 at the McDermott Sequencing Core (UT Southwestern). Droplet-based snRNA-seq libraries for the second batch were prepared using Chromium Single Cell 3′ v.3 (1000075, 10× Genomics) according to the manufacturer's protocol. Libraries were sequenced using an Illumina NovaSeq 6000 at the North Texas Genome Center (UT Arlington).

Isolation of nuclei from resected brain tissue (snATAC-seq). For snATAC-seq, nuclei were isolated as described above. After lysis, the nuclei were washed once in 500 µl of nuclei wash buffer consisting of 10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM $MgCl_2$, 1% BSA and 0.1% Tween-20. Nuclei were resuspended in 500 µl of 1× Nuclei Buffer (10× Genomics). Debris was removed via density gradient centrifugation using Nuclei PURE 2M Sucrose Cushion Solution and Nuclei PURE Sucrose Cushion Buffer from a Nuclei PURE Prep Isolation kit (NUC201-1KT, Sigma Aldrich). Nuclei PURE 2M Sucrose Cushion Solution and Nuclei PURE Sucrose Cushion Buffer were first mixed in a 9:1 ratio. A total of 500 µl of the resulting sucrose buffer was added to a 2-ml Eppendorf tube. A total of 900 µl of the sucrose buffer was added to 500 µl of isolated nuclei in NSB. A total of 1,400 µl suspension of nuclei was layered to the top of sucrose buffer. This gradient was centrifuged at 13,000 × *g* for 45 min at 4 °C. Nuclei pellets were resuspended and washed once in nuclei wash buffer. The concentration and integrity of the nuclei were determined using ethidium homodimer-1 (E1169, Thermo Fisher Scientific).

Finally, nuclei were resuspended in 1× Nuclei Buffer at a concentration of 4,000 nuclei per µl for snATAC-seq. Droplet-based snATAC-seq libraries were prepared using Chromium Single Cell ATAC kit solution v.1.0 (10× Genomics) and following the Chromium Single Cell ATAC reagent kits user guide: CG000168 Rev B. The library was sequenced using an Illumina NextSeq 500 at the McDermott Sequencing Core at UT Southwestern.

Immunofluorescence staining of human tissue. Fresh surgically resected tissue was fixed in 4% paraformaldehyde in 1× PBS for 24–48 h at 4 °C and then cryoprotected in a 30% sucrose solution. The tissue was sectioned at 7 µm using a cryostat (Leica). Sections underwent heat-induced antigen retrieval in citrate buffer (pH 6.0) for 10 min at 95 °C. Sections were blocked with 2% fetal bovine serum (FBS) in 0.1 M Tris (pH 7.6) for 1 h at room temperature. After blocking, the sections were incubated with primary antibodies in 0.1 M Tris pH 7.6/2% FBS overnight at 4 °C and subsequently incubated with secondary antibodies in 0.1 M Tris pH 7.6/2% FBS for 1 h at room temperature. Sections were immersed in 0.25% Sudan Black solution to quench lipofuscin autofluorescence and counterstained with 4′-6-diamidino-2-phenylindole (DAPI). Sections were mounted and cover slipped using ProLong Diamond Antifade mountant (P36970, Thermo Fisher Scientific). The following antibodies and dilutions were used: goat anti-IL1RAPL2 (PA5-47039, Thermo Fisher Scientific 1:20); rat anti-SMAD3 (MAB4038, R&D Systems, 1:100); rabbit anti-CaMKII alpha (PA514315, Thermo Fisher Scientific, 1:50); chicken anti-GFAP (ab4674, Abcam, 1:400); mouse anti-GAD67 (MAB5406, Millipore, 1:200); mouse anti-OLIG2 (MABN50, Millipore, 1:200); species-specific secondary antibodies produced in donkey and conjugated to Alexa Fluor 488, Alexa Fluor 555 or Alexa Fluor 647 (Thermo Fisher Scientific, 1:800). Images were acquired using a ×63 oil objective on a Zeiss LSM 880 confocal microscope. Experiments using secondary antibody only were conducted for each antibody to ensure specificity. The anti-ILRAPL2 antibody was validated by Thermo Fisher Scientific using flow cytometry of human HepG2 cells. The anti-SMAD3 antibody was validated by R&D Systems using flow cytometry in human PC-3 cells and IHC in human pancreatic cancer tissue and human MDA-MB-231 cells[63]. The anti-CamKII antibody was validated by Thermo Fisher Scientific using IHC in human brain and western blotting in human 293 cells[64]. The anti-GFAP antibody was validated by Abcam across many species, including human. Protocol validations include IHC and immunofluorescence. A total of 211 references are provided for this antibody at https://www.abcam.com/gfap-antibody-ab4674.html. The anti-GAD67 antibody was validated by Millipore in human brain via IHC. Over 75 references are provided at http://www.emdmillipore.com/US/en/product/Anti-GAD67-Antibody-clone-1G10.2,MM_NF-MAB5406#anchor_BRO. The anti-OLIG2 antibody has been validated by Millipore in human via IHC, and 15 references are provided at http://www.emdmillipore.com/US/en/product/Anti-Olig2-Antibody-clone-211F1.1,MM_NF-MABN50#documentation. Immunofluorescence staining was performed in four different surgically resected tissues (data not shown), and a representative optimized image is shown in Figs. 5e,f, and 6e,f.

Computational methods. *Bulk RNA-seq mapping, quality control and expression quantification.* Quality control was performed using FastQC (v.0.11.9). Reads were aligned to the human hg38 reference genome using STAR (v.2.5.2b)[65]. For each sample, a BAM file including mapped and unmapped reads that spanned splice junctions was produced. Secondary alignment and multi-mapped reads were further removed using in-house scripts. Only uniquely mapped reads were retained for further analyses. Quality control metrics were performed using RSeQC (v.2.6.4)[66] with the hg38 gene model provided. These steps included the number of reads after multiple-step filtering, ribosomal RNA reads depletion and defining reads mapped to exons, untranslated regions and intronic regions. Picard tool was implemented to refine the quality control metrics (http://broadinstitute.github.io/picard/) and to calculate sequencing statistics. Gencode annotation for hg38 (v.24) was used for reference alignment annotation and downstream quantification. Gene level expression was calculated using HTseq (v.0.9.1)[67] using intersection-strict mode by gene. Counts were calculated based on protein-coding genes from the annotation file.

Covariate adjustment. Counts were normalized using counts per million reads (CPM) with edgeR (v.3.32.0) package in R[68]. Normalized data were $\log_2$-scaled with an offset of 1. Genes with no reads were removed. A total of 15,192 genes were used for the downstream analysis.

Normalized data were assessed for effects from known biological covariates (sex, age, race, ethnicity, hemisphere, epilepsy duration (EpDur)), technical variables related to sample processing (RNA integrity number (RIN) and batch). The postmortem interval (PMI) was not considered in the analysis because it was confounded with the brain resected data from living individuals.

Residualizations were calculated using the following model:

$$\text{Gene expression} \sim \text{age} + \text{sex} + \text{race} + \text{ethnicity} + \text{EpDur} + \text{RIN} + \text{hemisphere} + \text{batch}$$

Residuals were extracted and average gene expression added as follows:

$$\text{Adjusted gene expression} = \text{residuals} + \text{average gene expression}$$

We applied two residualizations: (1) resected tissues and (2) resected tissues plus frozen tissues.

The adjusted CPM from the 16 participants were used for SME correlation and quantile regression. The adjusted CPM from resected tissue and frozen tissue were used for the consensus WGCNA analysis and permutations/bootstraps analysis.

MVA. We performed a MVA based on the following model:

$$\text{Gene expression} \sim \text{SME} : \text{band} + \text{EpDur} + \text{RIN} + \text{batch} + (1/\text{participants})$$

Due to the limited sample size and because we did not want to over-parametrize the model, we utilized the three fixed covariates that explained the highest variance in the data: EpDur, RIN and batch. Contrasts were used to compare SME association between waves. Genes with FDR < 0.05 were considered to be differentially associated with SME. The analysis was performed using edgeR (v.3.32.0)[68]. These results were integrated with the correlative analysis to define the final 300 SME genes. The code used for this analysis is available at GitHub (https://github.com/konopkalab/Within_Subject).

Correlation analysis and permutation analysis. Spearman's rank correlation was performed between each of the six memory brain oscillations and gene expression. We also utilized this method for six math brain oscillations, thickness and behavioral performances.

For this analysis, we used within-individual bulk RNA-seq from BA38 resected tissue from the 16 participants with calculated SMEs (WrS).

We next performed permutations/bootstraps analysis using the following three datasets:

1. Additional participants: bulk RNA-seq from BA38 resected tissue without SMEs from an additional 11 participants (ArS).
2. Independent data healthy: bulk RNA-seq from BA38 frozen tissue from 12 participants (HfS).
3. Independent data epilepsy: bulk RNA-seq from BA38 frozen tissue from 8 participants with epilepsy (EfS).

Bootstrapping was applied by randomly subsampling 16 participants (as WrS) from the composite data and recalculating the correlation 100 times. We then calculated a Monte Carlo *P* value comparing the observed effect with the simulated effects for each gene as follows:

$$\text{sum}(\text{abs}(\text{simulated rho}) \geq \text{abs}(\text{observed rho}))/100$$

We calculated two Monte Carlo *P* values: (1) BootP, based on WrS + ArS (only resected tissues) and (2) BootP_All, based on WrS + ArS + HfS + EfS (resected tissues and frozen tissues).

We additionally applied a permutation approach, shuffling the gene expression of WrS and recalculating 100 times the correlation between oscillations and gene expression. We then calculated a Monte Carlo *P* value (Perm*P*). Nominal *P* value < 0.05, PermP < 0.05, Boot*P* < 0.05 and Boot*P*_All < 0.05 were used to filter for significant correlations, as reported in Supplementary Table 2.

Co-expression network analysis. To identify modules of co-expressed genes in the RNA-seq data, we carried out WGCNA (v.1.69)[18]. We applied a consensus analysis based on WrS + ArS + HfS + EfS data, defining modules highly preserved across multiple datasets. This method was applied to reduce the potential noise between different types of data. A soft-threshold power was automatically calculated to achieve approximate scale-free topology ($R^2 > 0.85$). Networks were constructed with the blockwiseConsensusModules function with biweight midcorrelation (bicor). The modules were then determined using the dynamic tree-cutting algorithm. To ensure the robustness of the observed network, we used a permutation approach, recalculating the networks 200 times and comparing the observed connectivity per gene with the randomized one. None of the randomized networks showed similar connectivity, thereby providing robustness to the network inference. Module sizes were chosen to detect small modules driven by potential noise on the adjusted data. A deep split of four was used to more aggressively split the data and create more specific modules. Spearman's rank correlation was used to compute module eigengene–memory oscillatory signature associations.

snRNA-seq analysis. snRNA-seq data from BA38 were processed using the mkfastq command from 10× Genomics CellRanger (v.3.0.1). Extracted paired-end fastq files (26/28-bp (v2, v3) long R1: cell barcode and UMI sequence information; 124-bp long R2: transcript sequence information) were checked for read quality using FastQC (v.0.11.9). Gene counts were obtained by aligning reads to the hg38 genome using an in-house pipeline. UMI tools (v.1.0.0)[69] was used to generate a whitelist of barcodes and to extract reads to match the detected barcodes. Reads were aligned to the human hg38 reference genome using STAR (2.5.2b)[65]. Gencode annotation for hg38 (v.24) was used as reference alignment annotation. Gene level expression was calculated using featureCounts (v.1.6.0)[70] by gene. UMIs per gene across all detected nuclei were further calculated using UMI tools. Two batches from three participants were processed for a total of six samples. Nuclei with >10,000 UMI and >5% of mitochondrial gene expressed were

removed. Downstream analysis was performed using Seurat (v.3.9.9)[71]. Briefly, we normalized the expression data and integrated the two different batches of sequencing by SCtransform, retaining 3,000 variable genes. We constructed a *k*-nearest neighbor graph based on the Euclidian distance in 30 principal component space and identified distinct clusters of cells using the Leiden algorithm (resolution of 0.8). Visualization of clusters was performed by applying the function RunUMAP() based on uniform manifold approximation and projection (UMAP)[72] in two dimensions. Cell-type markers were identified by Wilcoxon's rank-sum test (two-sided; Benjamini–Hochberg-adjusted; FDR < 0.05, $\log_2$(fold change) > 0.3, pct.1 > 0.5). Clusters were annotated based on marker enrichment, with markers defined by middle temporal gyrus data[73]. Briefly, data were downloaded from the Allen portal (https://portal.brain-map.org/atlases-and-data/rnaseq). Seurat was used to define the markers for each cluster by Wilcoxon's rank-sum test (two-sided; Benjamini–Hochberg-adjusted; FDR < 0.05, $\log_2$(fold change) > 0.3, pct.1 > 0.5). Statistics for the overlap between BA38 and middle temporal gyrus markers was performed by Fisher's exact test (one-sided with the alternative greater; Benjamini–Hochberg-adjusted). Labels for BA38 cell types were selected by using the highest significant enrichment defined by an OR with FDR < 0.05. These labels were used for all downstream analysis and snATAC-seq integration. The code used for this analysis is available at GitHub (https://github.com/konopkalab/Within_Subject).

snATAC-seq analysis. snATAC-seq data from BA38 of three participants were processed using the Cell Ranger ATAC pipeline. Seurat extension Signac (v.1.1.0)[71] was used for additional filtering, clustering and annotation. Cells with total fragments in peaks <1,500 or <15% of the total fragments were not considered for further analysis. Clustering and creating a gene activity matrix were done using the default parameters. Only the cells with >0.5 confidence in annotation were considered for downstream analysis. The gene activity matrix was produced by counting fragments in the gene body +2 kb upstream. Identified clusters were cross-referenced to the snRNA-seq data using Seurat integration workflow. Visualization of the clusters was performed by applying UMAP[72] in two dimensions. Motif enrichment testing was applied to the upstream regions of the genes in each module. Motif analysis was performed only for the modules with cell-type enrichment (WM4 and WM12: excitatory–inhibitory clusters; WM21: oligodendrocyte–oligodendrocyte progenitor cell clusters). Fragments for excitatory–inhibitory clusters and oligodendrocyte–oligodendrocyte progenitor cell (OPC) clusters were extracted separately from the Cell Ranger's fragments.tsv file. For each cut site, the fragments.tsv file was adjusted to contain 200 bp around the cut site, and peaks were called using MACS2 (v.2.1.1)[74]. The CIS-BP database for human was used for enrichment (http://cisbp.ccbr.utoronto.ca/index.php)[75]. Only TFs with directly determined motifs were kept. TFs were filtered for presence in >30% of cells in the cluster that was being tested for enrichment. A motif matrix (peaks in rows, motifs in columns) was created using CreateMotifMatrix from Signac. Using the FindMotifs function from Signac, the enrichment of each TF was tested for the upstream peaks of module genes versus upstream peaks of all genes using all the peaks as background. Peak visualization was done using IGV (v.2.8.13)[76]. The code used for this analysis is available at GitHub (https://github.com/konopkalab/Within_Subject).

Functional enrichment. Functional annotation of the genes within the modules was performed using GOstats (v.2.56.0)[77] and confirmed by ToppGene[78]. We used Gene Ontology and Kyoto Encyclopedia of Genes and Genomes databases. Expressed genes (15,192) were used as background. A one-sided hypergeometric test was performed to test overrepresentation of functional categories. A Benjamini–Hochberg-adjusted *P* value was applied as a multiple comparisons adjustment.

Neuropsychiatric genes. ASD-associated genes used for Fig. 4c were downloaded from SFARI Gene database[79]. ASD (1–3) are ASD genes with a score between 1 and 3. Modules and genes differentially expressed in ASD, SCZ and BD were downloaded from an independent source[22]. Differentially expressed cell-type markers from snRNA-seq of ASD and AD were downloaded from independent sources[80,81].

GWAS data and enrichment. We used genome-wide gene-based association analysis implementing MAGMA (v.1.07)[82]. We used the 19,346 protein-coding genes from human gencode v.19 as background for the gene-based association analysis. Single nucleotide polymorphisms (SNPs) were selected within exonic, intronic and untranslated regions as well as SNPs within 10-kb upstream/downstream of the protein-coding gene. SNP associations revealed 18,988 protein-coding genes with at least one SNP. Gene-based association tests were performed using LD between SNPs. Benjamini–Hochberg correction was applied, and significant enrichment is reported for FDR < 0.05. Summary statistics for GWAS of neuropsychiatric disorders and non-brain disorders were downloaded from the Psychiatric Genomics Consortium and the GIANT Consortium[83–97]. Supplementary Table 4 reports MAGMA statistics for each of the GWAS data analyzed. The following GWAS acronyms were used for the figures: AD, Alzheimer disease; ADHD, attention-deficit/hyperactivity disorder; ASD, autism spectrum disorder; BD, bipolar disorder; BMI, body mass index; CHD, coronary artery disease; CognFun, cognitive functions; DIAB, diabetes; EduAtt, educational attainment; EP, epilepsy; HGT, height; MDD, major depressive disorder; OSTEO, osteoporosis; SZ, schizophrenia.

Gene set enrichment. Gene set enrichment was applied to correlated genes and SME genes from our previous study as shown in Fig. 3c, SME genes from the current study as shown in Extended Data Fig. 3c, neuropsychiatric differentially expressed genes as shown in Fig. 4a and Extended Data Fig. 4b, ASD genes as shown in Fig. 4c, and cell-type markers as shown in Fig. 5c and Extended Data Fig. 5h,i. We used a Fisher's exact test in R with the following parameters: alternative = "greater", conf. level = 0.95. We reported OR and Benjamini–Hochberg-adjusted *P* values (FDR).

Statistical analysis and reproducibility. No statistical methods were used to predetermine sample sizes because of the limited availability of human brain surgical tissues. Participants were not randomly selected for inclusion in the study based on the availability of human tissue/oscillation data. However, tissue pieces were randomized for processing for RNA-seq. Data collection and analysis of human participants were not performed blind to the conditions of the experiments. However, separate individuals carried out wet-bench and dry-bench analyses. Thus, the researcher was blinded to patient/oscillation characterization while processing tissue for RNA and data analysis. The final analysis required including participant covariate information, so the researcher was unblinded to participant characteristics and oscillation data at that point. For SME values, bulk RNA-seq transcriptomic data, snRNA-seq transcriptomic data and scATAC-seq epigenomic data distributions were assumed to be normal, but this was not formally tested. Nonparametric tests were used to avoid uncertainty when possible. Data collection and analysis were not performed blind to the conditions of the experiments.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The RNA-seq dataset used for memory oscillatory signature analysis in this study are available at GEO with accession number GSE139914.

Code availability

Custom R codes for the quality control, MVA, correlative analysis, permutation/bootstraps, WGCNA, snRNA-seq analysis, snATAC-seq analysis, visualizations, functional enrichments, GWAS enrichment and gene set enrichments are available at https://github.com/konopkalab/Within_Subject.

References

55. Natu, V. S. et al. Stimulation of the posterior cingulate cortex impairs episodic memory encoding. *J. Neurosci.* **39**, 7173–7182 (2019).
56. Goyal, A. et al. Functionally distinct high and low theta oscillations in the human hippocampus. *Nat. Commun.* **11**, 2469 (2020).
57. Watrous, A. J., Miller, J., Qasim, S. E., Fried, I. & Jacobs, J. Phase-tuned neuronal firing encodes human contextual representations for navigational goals. *eLife* https://doi.org/10.7554/eLife.32554 (2018).
58. Fischl, B. et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* **33**, 341–355 (2002).
59. Ghose, S., Gleason, K. A., Potts, B. W., Lewis-Amezcua, K. & Tamminga, C. A. Differential expression of metabotropic glutamate receptors 2 and 3 in schizophrenia: a mechanism for antipsychotic drug action? *Am. J. Psychiatry* **166**, 812–820 (2009).
60. Takahashi, J. S. et al. ChIP-seq and RNA-seq methods to study circadian control of transcription in mammals. *Methods Enzymol.* **551**, 285–321 (2015).
61. Habib, N. et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017).
62. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
63. Yeh, Y. H. et al. Transforming growth factor-beta and oxidative stress mediate tachycardia-induced cellular remodelling in cultured atrial-derived myocytes. *Cardiovasc. Res.* **91**, 62–70 (2011).
64. Del Cid-Pellitero, E., Plavski, A., Mainville, L. & Jones, B. E. Homeostatic changes in GABA and glutamate receptors on excitatory cortical neurons during sleep deprivation and recovery. *Front. Syst. Neurosci.* **11**, 17 (2017).
65. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
66. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
67. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
68. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
69. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).

70. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
71. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
72. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2019).
73. Hodge, R. D. et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
74. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
75. Weirauch, M. T. et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
76. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
77. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257–258 (2007).
78. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37**, W305–W311 (2009).
79. Banerjee-Basu, S. & Packer, A. SFARI Gene: an evolving database for the autism research community. *Dis. Models Mech.* **3**, 133–135 (2010).
80. Velmeshev, D. et al. Single-cell genomics identifies cell type-specific molecular changes in autism. *Science* **364**, 685–689 (2019).
81. Mathys, H. et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).
82. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
83. Grove, J. et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
84. Jansen, I. E. et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).
85. International League Against Epilepsy Consortium on Complex Epilepsies. Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. *Nat. Commun.* **9**, 5269 (2018).
86. Lee, J. J. et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).
87. Savage, J. E. et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).
88. Bipolar Disorder and Schizophrenia Working Group of the Psychiatric Genomics Consortium.Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes. *Cell* **173**, 1705–1715.e16 (2018).
89. Davies, G. et al. Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nat. Commun.* **9**, 2098 (2018).
90. Wray, N. R. et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
91. Pardinas, A. F. et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).
92. Martin, J. et al. A genetic investigation of sex bias in the prevalence of attention-deficit/hyperactivity disorder. *Biol. Psychiatry* **83**, 1044–1053 (2018).
93. Sohail, M. et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife* https://doi.org/10.7554/eLife.39702 (2019).
94. Hoffmann, T. J. et al. A large multiethnic genome-wide association study of adult body mass index identifies novel loci. *Genetics* **210**, 499–515 (2018).
95. Morris, A. P. et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
96. Estrada, K. et al. Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat. Genet.* **44**, 491–501 (2012).
97. Schunkert, H. et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* **43**, 333–338 (2011).

Acknowledgements

We thank the patients for participating in the study and the donors and their families for the additional tissue samples. We also thank K. Gleason for assistance with postmortem samples. G.K. is supported by a Jon Heighten Scholarship in Autism Research at UT Southwestern. This work was supported by NIMH (F30MH105158) to M.R.F.; NIDA (5T32DA007290-25) and NHBLI (1T32HL139438-01A1) to F.A.; NINDS (NS106447), a UT BRAIN Initiative Seed Grant (366582), the Chilton Foundation, and the National Center for Advancing Translational Sciences of the NIH under the Center for Translational Medicine's award number UL1TR001105 to B.C.L. and G.K.; NINDS (NS107357) to B.C.L.; and NIMH (MH103517), The Chan Zuckerberg Initiative, an advised fund of Silicon Valley Community Foundation (HCA-A-1704-01747), and the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition—Scholar Award (220020467) to G.K. Postmortem human tissue samples were obtained from the NIH NeuroBioBank (The Harvard Brain Tissue Resource Center, funded through HHSN-271-2013-537 00030C; the Human Brain and Spinal Fluid Resource Center, VA West Los Angeles Healthcare Center; and the University of Miami Brain Endowment Bank) and the UT Neuropsychiatry Research Program (Dallas Brain Collection). We also thank the UT Southwestern Neuroscience Microscopy Facility for providing imaging resources.

Author contributions

S.B., B.C.L. and G.K. analyzed the data and wrote the paper. M.R.F. and C.D. collected surgical samples, processed RNA and generated bulk RNA-seq libraries. M.R.F. contributed to the design of the project. F.A. generated the snRNA-seq and snATAC-seq data and performed IHC. A.K. preprocessed the snRNA-seq data. E.C. preprocessed the snATAC-seq data. C.A.T. provided postmortem human brain tissue. S.S. analyzed the oscillation data. B.C.L. conducted all surgical procedures and memory testing. B.C.L. and G.K. designed and supervised the study and provided intellectual guidance. All authors discussed the results and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

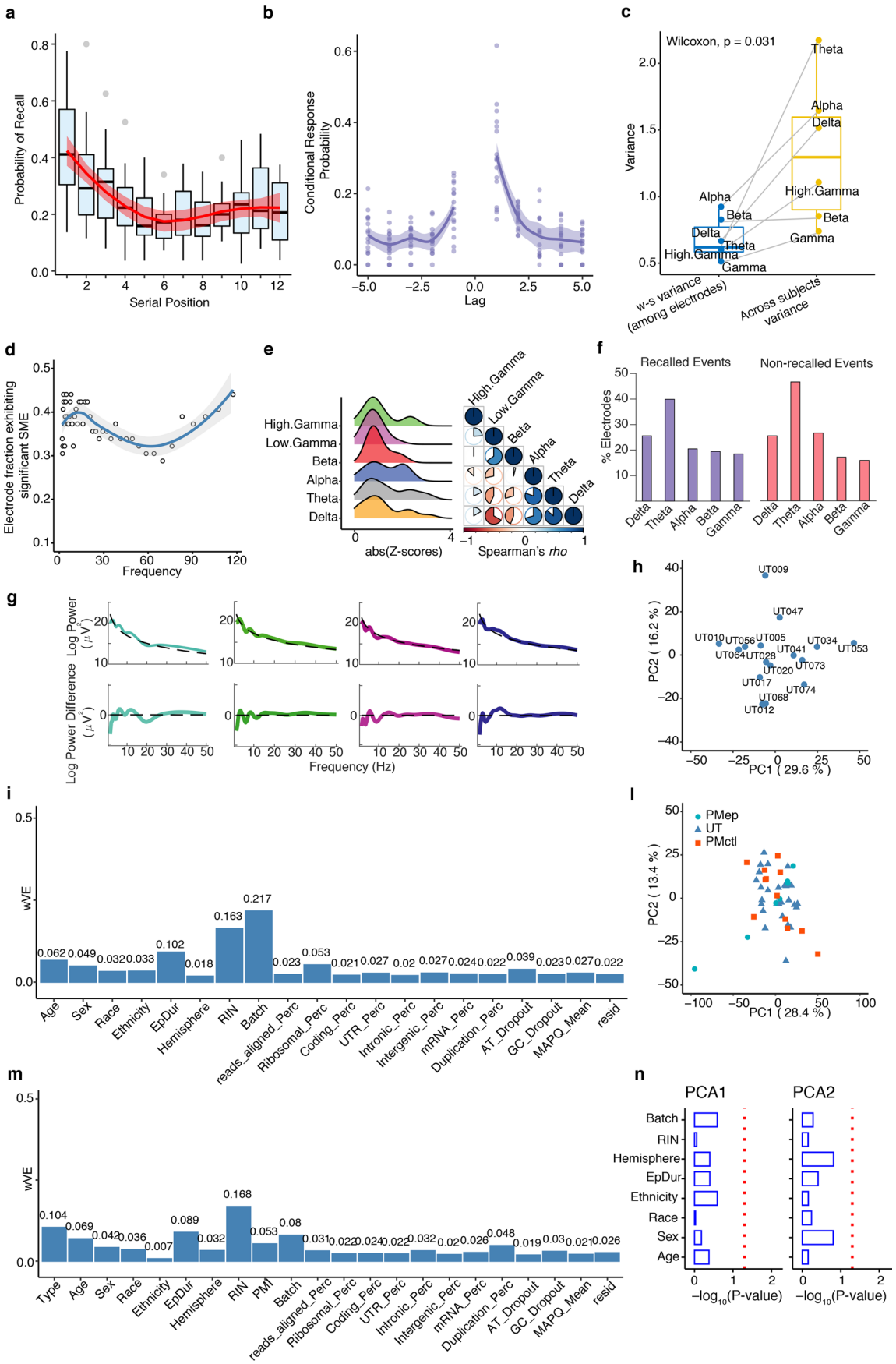
Extended data is available for this paper at https://doi.org/10.1038/s41593-021-00803-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-021-00803-x.

Correspondence and requests for materials should be addressed to B.C.L. or G.K.

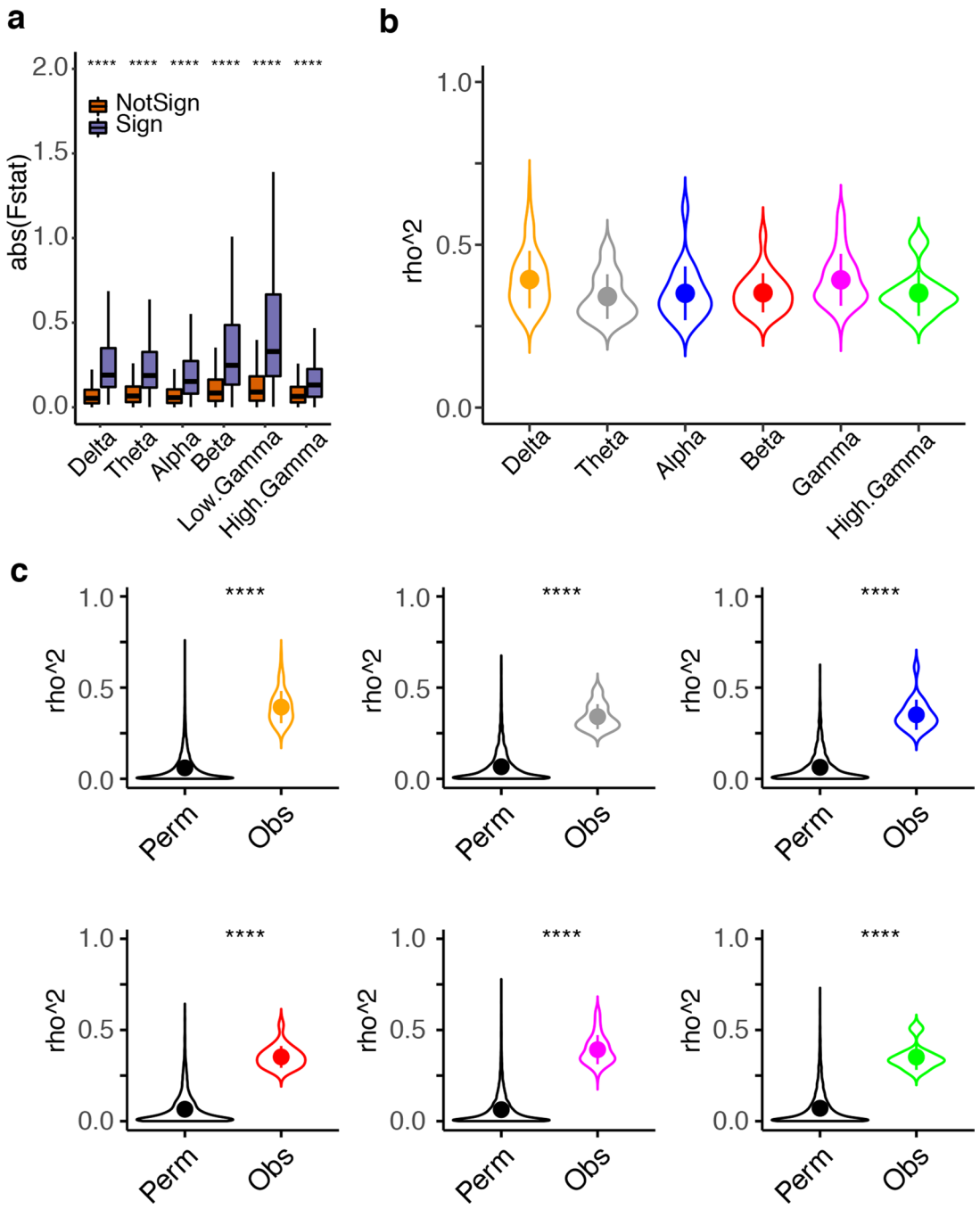
Peer review information *Nature Neuroscience* thanks Andrew Jaffe, Ueli Rutishauser, Ziv Williams, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

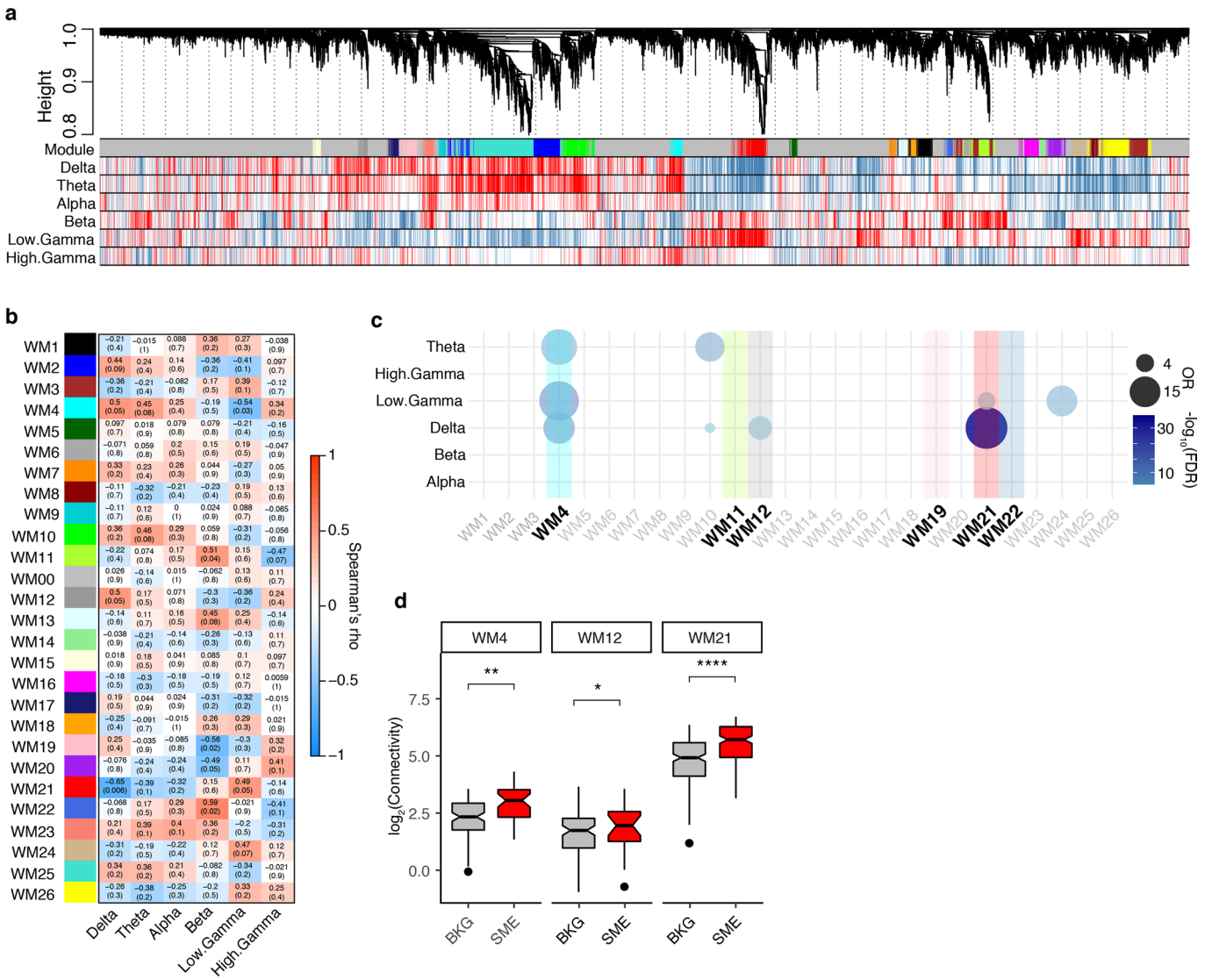


Extended Data Fig. 1 | See next page for caption.

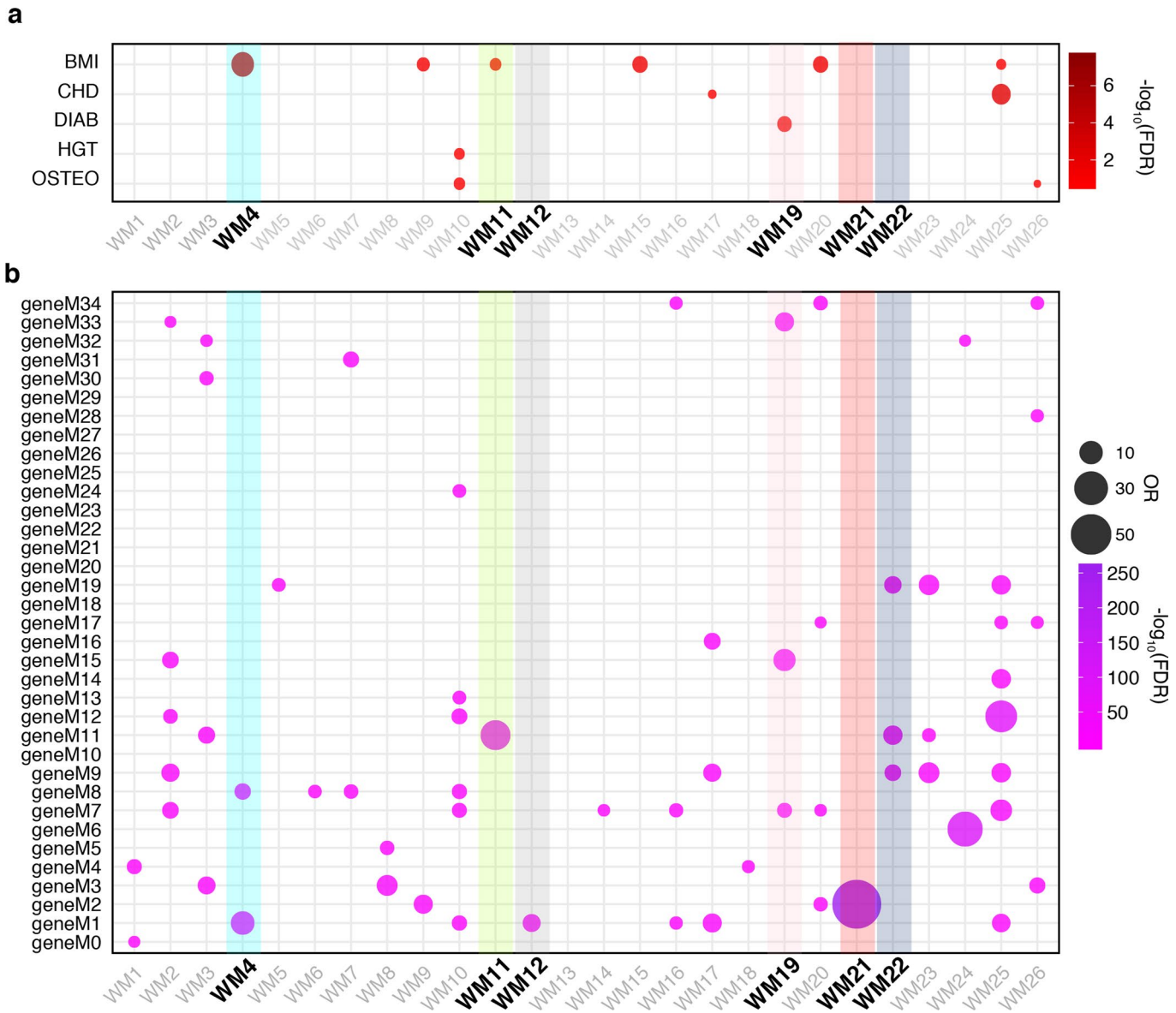
Extended Data Fig. 1 | Data quality control. a, Box plots depicting the probability of recall for items presented at each serial position. Primacy and recency effects are visible, consistent with expectations for performance in the free recall episodic memory paradigm. Whiskers on box plots represent maximum and minimum values. Boxes extend from the 25th to the 75th percentiles, the center lines represent the median. Loess regression with confidence intervals is superimposed to depict the overall distribution. Smooth curves are shown with 95% confidence bands **b**, Lag conditional response probability curves in our data (lag CRP), indicating expected temporal clustering behavior. Loess regression with confidence intervals depicts the overall distribution. Smooth curves are shown with 95% confidence bands. **c**, Boxplot showing the comparison of within-subject variance (across all measured electrodes at each band, blue box plot,) with the variance across subjects (at each band, yellow box plot). Across subjects variance is significantly greater than within-subject variance. Reported p-value from Wilcoxon rank sum test (one-sided with alternative greater). Boxplots extend from the 25th to the 75th percentiles, the center lines represent the median. **d**, Scatter plot showing the fraction of all BA38 electrodes exhibiting a significant subsequent memory effect at each frequency. We observed significant differences predicting recall success across the frequency spectrum, including the delta and gamma bands. Loess regression with confidence intervals depicts the overall distribution. Smooth curves are shown with 95% confidence bands. **e**, Distribution of SME values for each brain oscillation and cross-correlation based on Spearman's rank correlation. **f**, Barplots showing the fraction of electrodes at which oscillations were detected in each frequency band in the recalled and non-recalled conditions. 85% of electrodes exhibited an oscillation in at least one of the delta, theta, or alpha frequency bands. **g**, Scatter plot showing individual electrode examples of power curves used for oscillation detection via the MODAL algorithm, both before and after subtraction of the best fit line. **h**, Principal component analysis of the subjects used for the within-subject analysis. Variance explained by each principal component is highlighted in the axis. **i**, Barplot showing the variance explained by each covariate adjusted across 10 principal components (wVE) for the within-subject data. Technical, biological and sequencing covariates calculated by PICARD (see Methods) are included. **I**, Principal component analysis of all the subjects used in this study. PMep = post-mortem epileptic subjects, UT = within-subjects, PMctl = post-mortem healthy subjects. **m**, Variance explained by each covariate adjusted across 10 principal components (wVE). Type corresponds to the three different types of data included in the analysis (PMep, UT, PMctl). Technical, biological and sequencing covariates calculated by PICARD (see Methods) are included. **n**, Association between the first two components and covariates based on adjusted gene expression. X-axis corresponds to the -log10(P-value) from linear regression modeling between PCs and covariates.



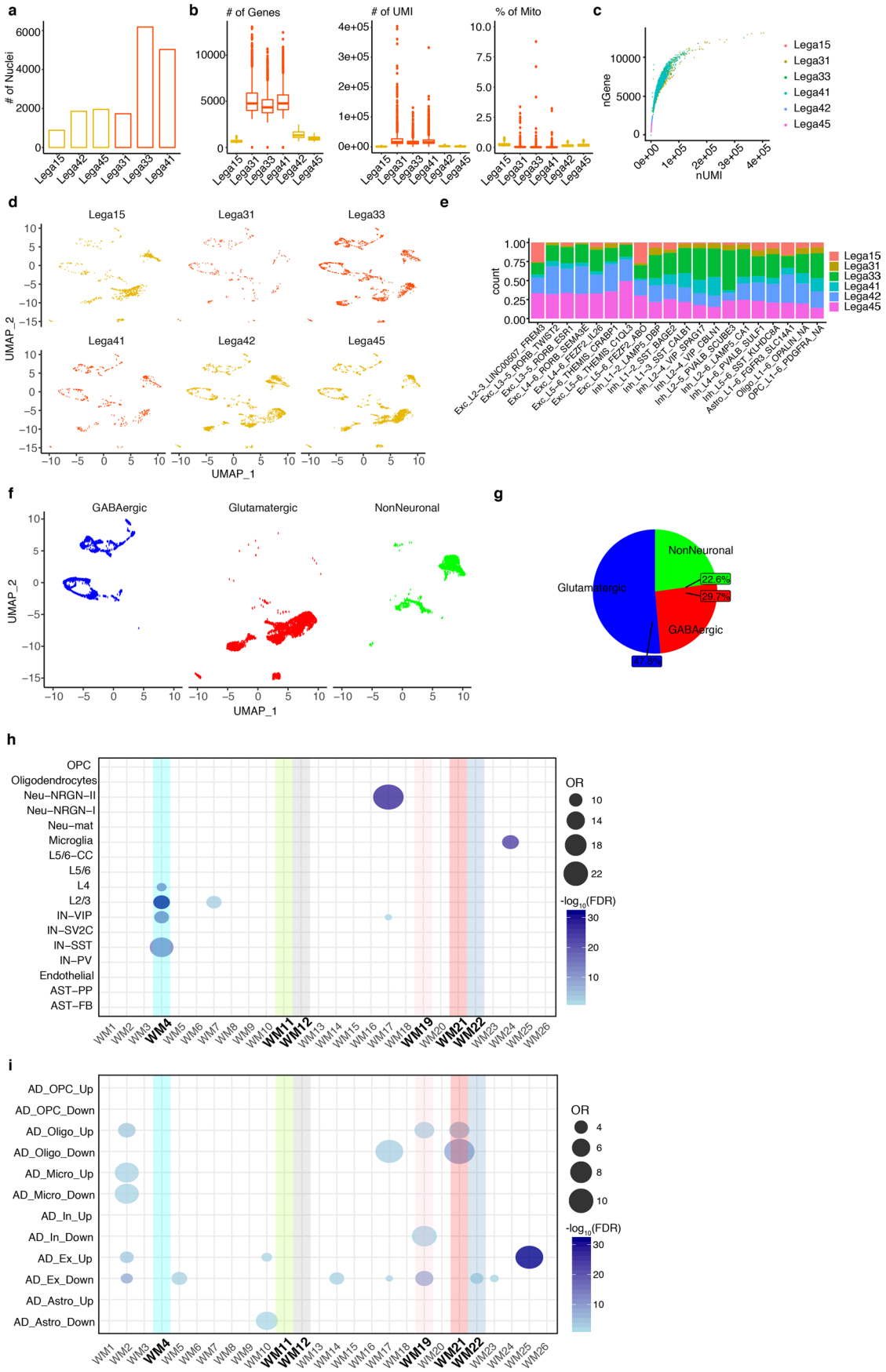
Extended Data Fig. 2 | SME gene robustness and overlap with other tasks. a, Boxplot showing the difference between F-statistics of the SME genes (Multivariate analysis) compared with the other genes. Stars correspond to the Wilcoxon's rank sum test (N, Sign = 753, NotSign = 14439; one-sided with alternative greater; $p < 0.0001$ = ****; Benjamini-Hochberg adjusted: Delta, $FDR = 2.3x10^{-249}$, Theta, $FDR = 3.2x10^{-205}$, Alpha, $FDR = 4.1x10^{-140}$, Beta, $FDR = 2.1x10^{-159}$, Low Gamma, $FDR = 7.2x10^{-207}$, High Gamma, $FDR = 1.3x10^{-63}$). Boxes extend from the 25th to the 75th percentiles and the center lines represent the median. **b**, Violin plots showing the $rho^{\hat{}2}$ of the genes significantly associated with each brain oscillation. Standard errors are calculated based on the $rho^{\hat{}2}$ distribution of the significantly correlated genes. Dots represent the median $rho^{\hat{}2}$ for the specific brain oscillation. **c**, Violin plots showing the $rho^{\hat{}2}$ of the genes significantly associated with each brain oscillation (Obs = observed) compared with $rho^{\hat{}2}$ derived from the permutation control analyses (Perm = Permutation). Standard errors are calculated based on the $rho^{\hat{}2}$ distribution of the significantly correlated genes. Dots represent the median $rho^{\hat{}2}$ for the specific brain oscillation. 100 random permutations were applied to calculate the Perm values (see Methods). Stars correspond to the Wilcoxon's rank sum test (unadjusted, one-sided with alternative greater; $p < 0.0001$ = ****).



Extended Data Fig. 3 | WGCNA highlights modules associated with memory oscillations. a, Representative network dendrogram for the consensus WGCNA. Heatmap shows the correlation between memory oscillatory signatures and genes. Red = positively correlated, Blue = negatively correlated. **b**, Heatmap showing the module association between memory oscillatory signatures and module eigengenes (Spearman's rank correlation). Warm colors represent positive correlations and cool colors represent negative correlations. P-values for each correlation together with exact correlation values are contained within each box. **c**, Bubble-chart showing the enrichment for 300 SME genes decomposed by brain oscillation. Gradient color represents the $-\log_{10}$(FDR) and bubble size represents the odds ratio (OR) from a Fisher's exact enrichment test of each module with disease-relevant gene lists. Y-axis shows the brain oscillations labels. X-axis indicates the modules of the present study. **d**, Boxplots showing the differential connectivity (for example number of edges) between SME genes and non-SME genes in the modules associated with memory oscillatory signatures with SME genes enriched. Stars correspond to the results of a Wilcoxon's rank sum test (one-side test with alternative greater; $p < 0.001$ = ****, $p < 0.01$ = **, $p < 0.05$ = *; Benjamini-Hochberg adjusted: WM4, *FDR* = 0.016, WM12, *FDR* = 0.048, WM21, *FDR* = 4.5×10^{-4}). Boxes extend from the 25th to the 75th percentiles and the center lines represent the median.

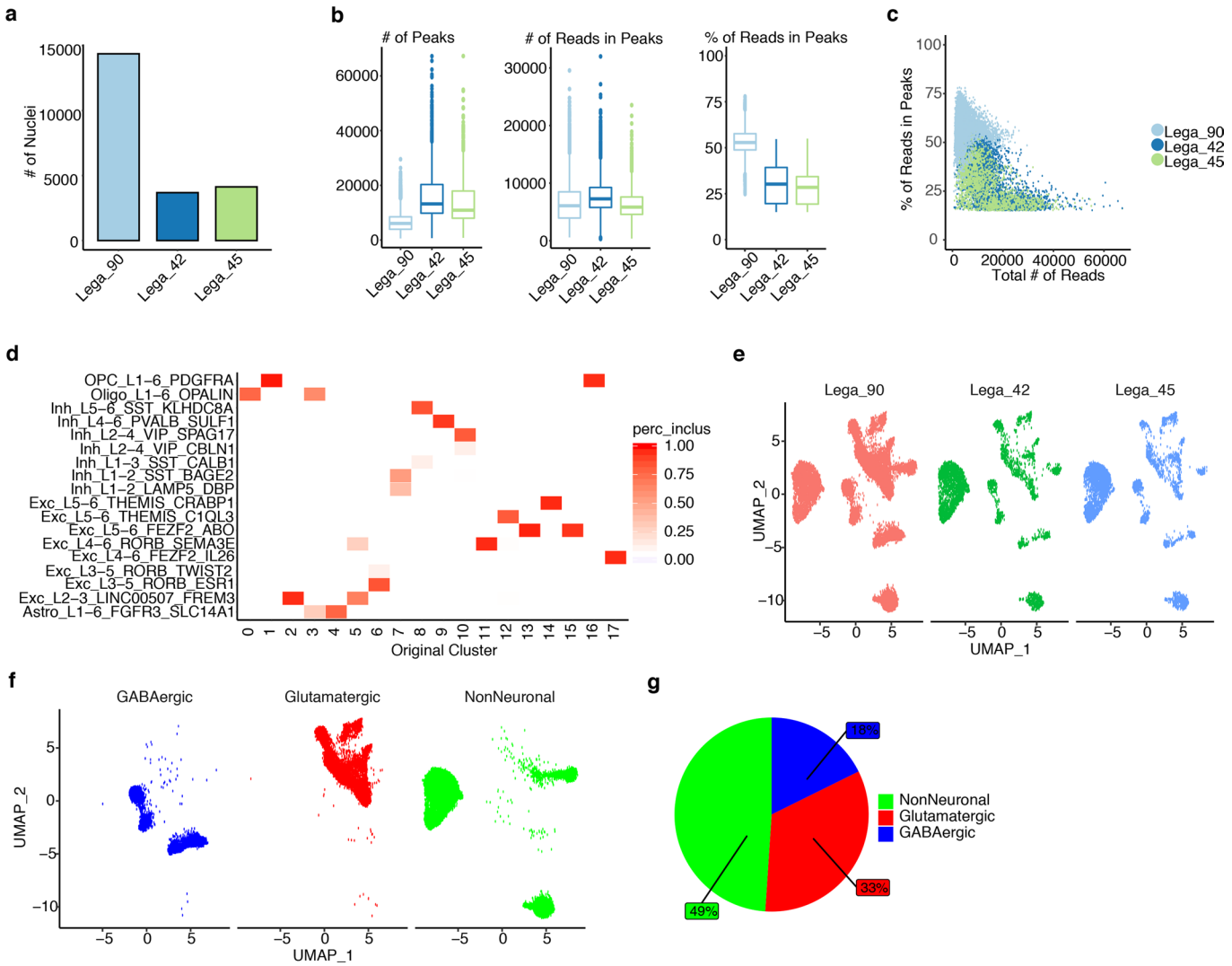


Extended Data Fig. 4 | Memory-related modules are enriched for gene co-expression modules associated with neuropsychiatric disorders. a, Bubble-chart showing the enrichment for loci associated with human traits used as negative controls. Gradient color represents the -$\log_{10}$(FDR) from linkage disequilibrium gene set analysis performed by MAGMA. Y-axis shows the acronyms for the GWAS data utilized for this analysis (see Methods). **b**, Bubble-chart showing the enrichment for modules of co-expressed genes dysregulated in ASD, SCZ or BD. Gradient color represents the -$\log_{10}$(FDR) and bubble size represents the odds ratio (OR) from a Fisher's exact enrichment test. Y-axis shows the acronyms for the modules associated with neuropsychiatric disorders utilized for this analysis (see Methods). X-axis shows the modules of the present study. Modules significantly correlated with memory-related oscillations are highlighted in bold text.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | snRNA-seq quality control metrices and module enrichment for cell-types dysregulated in cognitive disorders. a, Barplot showing the total number of nuclei identified per subject. Colors correspond to the two different batches. **b**, Quality control boxplots for snRNA-seq with number of genes detected, number of UMIs and percentage of mitochondrial genes. Colors correspond to the two different batches. Boxes extend from the 25th to the 75th percentiles and the center lines represent the median. Dots represent outliers. **c**, Scatter plot showing the relationship between number of UMIs (X-axis) and detected genes (Y-axis). Each sample is indicated in a different color. **d**, UMAP plots showing the distribution of nuclei in each subject. Colors correspond to the two different batches. **e**, Proportion of nuclei representing the identified clusters. Colors correspond to the six different subjects analyzed. **f**, UMAP plots showing the distribution of the three major cell-classes: GABAergic (blue), Glutamatergic (red), and non-neuronal (green). **g**, Pie chart showing the proportion of the three major cell-type classes (GABAergic, Glutamatergic, and non-neuronal cells). **h**, Bubble-chart showing the enrichment of the SME modules for cell-type markers dysregulated in ASD. Color gradient represents the $-\log_{10}$(FDR) and bubble size represent the odds ratio (OR) from a Fisher's exact enrichment test. Y-axis shows the acronyms for the cell-types defined in the ASD study. **i**, Bubble-chart showing the enrichment of the SME modules for cell-type markers dysregulated in Alzheimer disease (AD). Color gradient represents the $-\log_{10}$(FDR) and bubble size represent the odds ratio (OR) from a Fisher's exact enrichment test. Y-axis shows the acronyms for the cell-types defined in the AD study.



Extended Data Fig. 6 | snATAC-seq quality control metrices. a, Barplot showing the total number of nuclei identified per subject. **b**, Quality control boxplots for each snATAC-seq sample demonstrating the total number of peaks, the number of reads in the peaks and the percentage of reads in peaks. Boxes extend from the 25th to the 75th percentiles and the center lines represent the median. Dots represent outliers. **c**, Scatter plot showing the relationship between total number of reads (X-axis) and percentage of reads in the peaks (Y-axis). Each sample is indicated in a different color. **d**, Heatmap of the pairwise similarity between cluster identities. Y-axis shows the snRNA-seq clusters. X-axis shows the snATAC-seq clusters. Gradient corresponds to the percentage of cells for the corresponding prediction label. **e**, UMAP plots showing the distribution of nuclei in each subject. **f**, UMAP plots showing the distribution of the three major cell-classes: GABAergic (blue), Glutamatergic (red), and non-neuronal (green). **g**, Pie chart showing the proportion of the three major cell-type classes.

CHAPTER 6: Discussions and Future Directions

My thesis research focused on molecular understanding of human brain evolution using comparative genomics at cellular resolution. To this end, I compared epigenomes and transcriptomes of human, chimpanzee and rhesus macaque using single-nuclei sequencing. To ensure biologically relevant interpretation of my findings, I performed an in-depth analysis of other single-nuclei omics datasets as well as other data modalities (e.g., DNA sequence comparisons). My efforts revealed the pervasiveness of ambient RNA contamination in single-nuclei RNA-seq datasets, a previously underappreciated problem **(Chapter 3)**. Analyzing comparative datasets after accounting for this and other technical biases (e.g., doublets, barcode multiplets, inter-individual variation, post-mortem interval differences between species), my analyses revealed fundamental novelties in human brain evolution that were previously uncharacterized **(Chapter 2)**. In the next sections, I discuss the implications of my findings, gaps in the current knowledge and the potential future directions.

Compositional and functional evolution of oligodendrocyte lineage

We detected proportionally higher oligodendrocyte progenitor cells (OPCs) and lower mature oligodendrocytes in human brain compared to non-human primate brains. This observation is consistent across four cortical brain regions, different datasets and experiments (including single-molecular fluorescent in-situ hybridization (sm-FISH)) **(Chapter 2, Figure 1)**. Interestingly, we did not observe a human-specific increase of OPCs in two subcortical regions (dentate gyrus and caudate nucleus). Although our results follow a pattern of cortical and subcortical division, the spatial span of our

observation remains low to make a definitive conclusion. While most studies profile a single brain region or few brain regions, recent studies have expanded into many other brain regions. For example, one study profiled 106 regions across the human brain at cellular resolution[84]. Similarly, another study profiled 30 brain regions in rhesus macaque at cellular resolution[85]. While these studies provide useful references, they may not be directly utilized for rigorous comparative analyses which require brain regions to be anatomically matched across species to prevent region-specific differences from being interpreted as species-specific differences. It is therefore paramount to expand the region-matched cross-species comparisons to the entire brain for a comprehensive understanding of evolutionary changes in the cellular landscape of the human brain. There are also striking differences in oligodendrocyte maturation between gray and white matter; the rate of oligodendrogenesis reaches a plateau at ~5 years old in white matter but increases until ~40 years old in gray matter[64]. Since most comparative studies to date only focused on gray matter cortical regions, it is currently unknown whether oligodendrocytes also have human-specific proportional and/or different regulatory changes in the white matter and whether this would be heterogeneous across the white matter regions. Thus, future studies with increased spatiotemporal span and resolution can make significant contributions to our understanding of oligodendrocyte evolution in the human brain, provided that they are carefully designed to account for biological covariates.

Comparative studies also suffer from low temporal resolution. Most evolutionary comparisons, including ours, focus on ‘adult’ stage that spans from the end of

adolescence to death. This can be especially limiting for oligodendrocyte lineage; since i) oligodendrocytes mature later in the prefrontal cortex than in the posterior cortical regions[86] and ii) both myelin content and myelin related gene upregulation are prolonged beyond adolescence in human development compared to non-human primates[58,63]. Studies in ageing also show that gene expression is variable during adulthood and can dramatically change in old brains[87,88]. Changes in gene expression are not homogenous either; glial cell types undergo more gene expression changes than neuronal cell types[87,88]. Most comparative studies ignore the potential consequences of the age difference between species. For example, some studies show more human-specific changes in glia than other cell types[26,66]. But this might also be explained by the comparison of different age groups between species (i.e., they might be age-specific instead of species-specific) which is often not accounted for in the interpretation or analysis[26,66,75]. To circumvent this problem, we humanized the ages of non-human primates based on the life history traits[70,89]. Since life history traits are often developmental (e.g., weaning, sexual maturity), transcriptome or methylome based predictions could be more accurate measurements and these measurements have already been implemented to compare the developing tissues and organoids across species[58,60]. However, their accuracy and generalizability need to be tested in the adult brain. Another option is to generate data from all age groups and identify evolutionary changes based on whether and how they change with age across species. However, this would require a massive effort, especially considering the need for expanding the number of brain regions (as discussed above), species and sample size for a rigorous comparison.

As previously mentioned, myelination is prolonged past adolescence in human brain compared to non-human primate brains[58,63] and we show that adult individuals (>35 years old in human age or humanized age) have low proportions of mature oligodendrocytes. This indicates that humans could still have lower myelin content than other species during adulthood since studies in mice show that new myelination often correlates with new oligodendrocyte generation[90,91]. To test these predictions, myelin content can be compared between human and non-human primate brains across ages during adulthood. Since myelin content is the greatest in deep layers and least abundant in upper layers, it would be important to do this comparison across layers. Notably, we observed similar results for OPC and oligodendrocyte proportional differences between human and chimpanzee tissues when we compared across cortical layers **(Chapter 2, Extended Figure 4)** .

High OPC and low oligodendrocyte proportions in the human cortex implicate slower or less frequent maturation in human brains compared to non-human primate brains. iPSC-derived or primary culture systems can be used to understand whether there is an intrinsic difference of oligodendrocyte maturation clocks between the species[60]. If a phenotypic divergence is found, these in vitro systems can be faithful model systems to dissect the molecular mechanisms behind oligodendrocyte maturation. One could test some of the hypotheses of our study such as a potential causative link between lower cytoskeletal activity in human OPCs and oligodendrocyte maturation. If *in vitro* systems do not show phenotypic differences in oligodendrocyte maturation, transplantation of human and non-human primate OPCs into rodent brain might provide a better proxy for

in vivo conditions and also enable evolutionary comparisons of other features such as synapses between OPCs and axons in human OPCs and non-human primate OPCs[92,93].

Interestingly, both myelination and oligodendrogenesis are known to be induced by neuronal activity[93,94], indicating that lower oligodendrocyte and -potentially- myelin content in the human brain may result in increased potential for activity-induced myelination in humans. This could indicate increased potential of adult human brain to alter neural wiring by new myelination compared to non-human primates. It is also possible that activity-dependent oligodendrocyte generation and myelination might be intrinsically different in humans than chimpanzees. These hypotheses could be tested by experimentation on the iPS-derived co-cultures of oligodendroglia and neurons[95] across species.

Heterogeneity and activity dependent regulation in neurons

Single-nucleus sequencing has revealed the extreme heterogeneity of neurons and identified many neuronal subtypes with unknown function[74]. Not surprisingly, comparisons across brain regions show that some neuronal subtypes are region-specific or show changes in abundance across regions[75]. Although we did not identify significant changes in the neuronal subtype abundance between species (except for more abundant upper layer excitatory neurons in hominins compared to rhesus macaques, as previously reported[27,75]), widening the breadth of single nuclei sequencing across species can reveal significant species-specific differences in other brain regions.

Depolarization of neurons immediately induces gene expression changes in many transcription factors (TFs) – also a few non-TF genes – that are known as immediate early genes or early response genes[81]. These TFs subsequently regulate hundreds of genes

that regulate synaptic transmission known as late response genes[81]. Interestingly, recent studies show that late response genes are heterogeneous across different cell types and species (mouse or human)[96,97]. Since most early response genes are shared across cell types and species, this indicates that other trans (co-factors of early response genes) and / or cis (changes in the DNA sequence that alter TF binding) changes that determine the profile of late response genes. While single-nuclei transcriptomics can effectively capture global gene expression profiles, human and non-human primate brain tissues are almost exclusively from post-mortem brains, preventing identification of activity-dependent genes at cellular resolution. Indeed, we have also detected very low levels of immediate early genes in our study **(Chapter 2, Supplementary Table 3)**. However, immediate early genes are in open-chromatin state at basal levels, and while some late response genes require chromatin decondensation[98], many can be in open-chromatin state **(Supplementary Table 4)**. Therefore, single-nuclei ATAC-seq provides a unique opportunity to analyze whether there are differences in the chromatin accessibility of late response genes between species. An unbiased approach to this is to test whether enriched motifs among human-specific chromatin accessibility regions contain motifs of early response genes. Interestingly, we found significant enrichments for FOS and JUN motifs within human-specific chromatin accessibility gains, specifically among upper layer excitatory neurons **(Chapter 2, Figure 5A-B)**. This was also accompanied by a significant excess of human-specific DNA sequence changes within these regions, indicating a genetic causation **(Chapter 2, Figure 5C-F)**. These observations indicate a potential human-specific response to neuronal depolarization that can differentially regulate synaptic plasticity in humans. It is also possible that activity-regulated genes between

human and non-human primate largely overlap but differ in the degree of their upregulation due to differences in their chromatin accessibility at the basal state. Importantly, a study on human brain slices suggests that human upper layer excitatory neurons have more complex and graded action potentials than rodents[99]. These results indicate that more work is needed to uncover the evolutionarily divergent functional properties of human neurons and the underlying molecular mechanisms. Future studies can compare activity-dependent gene regulation and firing properties between human and monkey (macaque or marmoset) brain slices. iPSC-derived cultures (monolayer or organoid) can also be useful and more versatile models; however, they are more similar to prenatal neurons and offer limited cell type heterogeneity[100].

Insights from genotypic changes

Human-specific DNA sequence changes underlie the cognitive capabilities of the human brain among many other traits that evolved in the human body. Therefore, the challenge is to identify genetic changes that are causative to molecular mechanisms linked to the unique functions of the human brain. Since the publication of the chimpanzee genome, studies have identified stretches of genome that have an excess number of human-specific substitutions (human accelerated regions, HARs[101]). Notably, HARs were identified from the entire genome, and thus are not prioritized to be functional in the human brain. Our methodology can fill this gap since we utilized the single-nuclei ATAC-seq data from human, chimpanzee and rhesus macaque neocortex to identify chromatin accessible regions in that brain region. Testing for human-specific DNA sequence acceleration within these regions in comparison to the entire anthropoid lineage, we revealed ~3800 HARs that were i) significantly enriched in human-specific chromatin

accessibility changes, ii) significantly overlapped with the previously published HARs, and iii) revealed ~580 HARs that were within human-specific chromatin accessibility changes. Therefore, these findings provide a promising list of HARs that are associated with human-specific activity in cortical brain and can be further tested to dissect their mechanism.

To understand the functional impact of HARs in a high throughput manner, recent studies have compared human and chimpanzee sequences in massively parallel reporter assays (MPRA) in monolayer cell cultures[43,102]. These studies revealed that nearly 50% of the previously identified HARs are neurodevelopmental enhancers[43]. Since our HARs are defined within the neocortex, they provide a more relevant list for brain evolution that can be readily tested with MPRA approaches. Moreover, since our data also provide classification of these HARs based on their cell type activity, this list can be further refined to include HARs that are linked to human-specific chromatin activity in the cell type more similar to the cell type of the monolayer culture (e.g., glutamatergic monolayer culture, excitatory neurons). However, some cell types are challenging to transfect, hindering the interpretability of MPRAs. Viral injection into mouse brain followed by single-nuclei RNA-sequencing to capture cell type heterogeneity might be an alternative approach, but the scalability of this approach to hundreds of genomic regions with reliable signal from diverse cell types presents a technical challenge. Taken together, approaches driven by changes in the genomic sequence can provide unique opportunities to understand how genotypic changes bring evolutionary innovation in the human brain.

BIBLIOGRAPHY

1 Amster, G. & Sella, G. Life history effects on the molecular clock of autosomes and sex chromosomes. *Proc Natl Acad Sci U S A* **113**, 1588-1593, doi:10.1073/pnas.1515798113 (2016).

2 Premack, D. & Woodruff, G. Does the Chimpanzee Have a Theory of Mind. *Behav Brain Sci* **1**, 515-526, doi:10.1017/S0140525x00076512 (1978).

3 Call, J. & Tomasello, M. Does the chimpanzee have a theory of mind? 30 years later. *Trends Cogn Sci* **12**, 187-192, doi:10.1016/j.tics.2008.02.010 (2008).

4 Povinelli, D. J. & Vonk, J. Chimpanzee minds: suspiciously human? *Trends Cogn Sci* **7**, 157-160, doi:10.1016/s1364-6613(03)00053-6 (2003).

5 Krupenye, C., Kano, F., Hirata, S., Call, J. & Tomasello, M. Great apes anticipate that other individuals will act according to false beliefs. *Science* **354**, 110-114, doi:10.1126/science.aaf8110 (2016).

6 Horschler, D. J., MacLean, E. L. & Santos, L. R. Do Non-Human Primates Really Represent Others' Beliefs? *Trends Cogn Sci* **24**, 594-605, doi:10.1016/j.tics.2020.05.009 (2020).

7 Ferrigno, S., Huang, Y. & Cantlon, J. F. Reasoning Through the Disjunctive Syllogism in Monkeys. *Psychol Sci* **32**, 292-300, doi:10.1177/0956797620971653 (2021).

8 Hill, A., Collier-Baker, E. & Suddendorf, T. Inferential Reasoning by Exclusion in Great Apes, Lesser Apes, and Spider Monkeys. *J Comp Psychol* **125**, 91-103, doi:10.1037/a0020867 (2011).

9 Laland, K. & Seed, A. Understanding Human Cognitive Uniqueness. *Annu Rev Psychol* **72**, 689-716, doi:10.1146/annurev-psych-062220-051256 (2021).

10 Goodall, J. Tool-Using and Aimed Throwing in a Community of Free-Living Chimpanzees. *Nature* **201**, 1264-1266, doi:10.1038/2011264a0 (1964).

11 Terrace, H. S., Petitto, L. A., Sanders, R. J. & Bever, T. G. Can an ape create a sentence? *Science* **206**, 891-902, doi:10.1126/science.504995 (1979).

12 Gardner, R. A. & Gardner, B. T. Teaching sign language to a chimpanzee. *Science* **165**, 664-672, doi:10.1126/science.165.3894.664 (1969).

13 Fouts, R. & Mills, S. T. *Next of kin : what chimpanzees have taught me about who we are*. 1st edn, (William Morrow, 1997).

14 Chomsky, N. *Syntactic structures*. (Mouton, 1965).

15 Sherwood, C. C., Bauernfeind, A. L., Bianchi, S., Raghanti, M. A. & Hof, P. R. Human brain evolution writ large and small. *Prog Brain Res* **195**, 237-254, doi:10.1016/B978-0-444-53860-4.00011-8 (2012).

16 Jerison, H. J. *Evolution of the brain and intelligence*. (Academic Press, 1973).
17 Herculano-Houzel, S. The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proc Natl Acad Sci U S A* **109 Suppl 1**, 10661-10668, doi:10.1073/pnas.1201895109 (2012).
18 Deaner, R. O., Isler, K., Burkart, J. & van Schaik, C. Overall brain size, and not encephalization quotient, best predicts cognitive ability across non-human primates. *Brain Behav Evol* **70**, 115-124, doi:10.1159/000102973 (2007).
19 Herculano-Houzel, S. The human brain in numbers: a linearly scaled-up primate brain. *Front Hum Neurosci* **3**, 31, doi:10.3389/neuro.09.031.2009 (2009).
20 Elston, G. N., Benavides-Piccione, R. & DeFelipe, J. The pyramidal cell in cognition: a comparative study in human and monkey. *J Neurosci* **21**, RC163, doi:10.1523/JNEUROSCI.21-17-j0002.2001 (2001).
21 Bianchi, S. *et al.* Synaptogenesis and development of pyramidal neuron dendritic morphology in the chimpanzee neocortex resembles humans. *Proc Natl Acad Sci U S A* **110 Suppl 2**, 10395-10401, doi:10.1073/pnas.1301224110 (2013).
22 Sherwood, C. C. *et al.* Invariant Synapse Density and Neuronal Connectivity Scaling in Primate Neocortical Evolution. *Cereb Cortex* **30**, 5604-5615, doi:10.1093/cercor/bhaa149 (2020).
23 Bianchi, S. *et al.* Dendritic morphology of pyramidal neurons in the chimpanzee neocortex: regional specializations and comparison to humans. *Cereb Cortex* **23**, 2429-2436, doi:10.1093/cercor/bhs239 (2013).
24 Fang, R. *et al.* Conservation and divergence of cortical cell organization in human and mouse revealed by MERFISH. *Science* **377**, 56-62, doi:10.1126/science.abm1741 (2022).
25 Sherwood, C. C. *et al.* Evolution of increased glia-neuron ratios in the human frontal cortex. *Proc Natl Acad Sci U S A* **103**, 13606-13611, doi:10.1073/pnas.0605843103 (2006).
26 Khrameeva, E. *et al.* Single-cell-resolution transcriptome map of human, chimpanzee, bonobo, and macaque brains. *Genome Res* **30**, 776-789, doi:10.1101/gr.256958.119 (2020).
27 Ma, S. *et al.* Molecular and cellular evolution of the primate dorsolateral prefrontal cortex. *Science*, eabo7257, doi:10.1126/science.abo7257 (2022).
28 Raghanti, M. A. *et al.* Differences in cortical serotonergic innervation among humans, chimpanzees, and macaque monkeys: a comparative study. *Cereb Cortex* **18**, 584-597, doi:10.1093/cercor/bhm089 (2008).
29 Raghanti, M. A. *et al.* Cortical dopaminergic innervation among humans, chimpanzees, and macaque monkeys: a comparative study. *Neuroscience* **155**, 203-220, doi:10.1016/j.neuroscience.2008.05.008 (2008).

30 Raghanti, M. A. *et al.* Cholinergic innervation of the frontal cortex: differences among humans, chimpanzees, and macaque monkeys. *J Comp Neurol* **506**, 409-424, doi:10.1002/cne.21546 (2008).
31 Sierpowska, J. *et al.* Comparing human and chimpanzee temporal lobe neuroanatomy reveals modifications to human language hubs beyond the frontotemporal arcuate fasciculus. *Proc Natl Acad Sci U S A* **119**, e2118295119, doi:10.1073/pnas.2118295119 (2022).
32 Geschwind, D. H. & Rakic, P. Cortical evolution: judge the brain by its cover. *Neuron* **80**, 633-647, doi:10.1016/j.neuron.2013.10.045 (2013).
33 Preuss, T. M. Human brain evolution: from gene discovery to phenotype discovery. *Proc Natl Acad Sci U S A* **109 Suppl 1**, 10709-10716, doi:10.1073/pnas.1201894109 (2012).
34 Dumas, G., Malesys, S. & Bourgeron, T. Systematic detection of brain protein-coding genes under positive selection during primate evolution and their roles in cognition. *Genome Res* **31**, 484-496, doi:10.1101/gr.262113.120 (2021).
35 Pinson, A. & Huttner, W. B. Neocortex expansion in development and evolution-from genes to progenitor cell biology. *Curr Opin Cell Biol* **73**, 9-18, doi:10.1016/j.ceb.2021.04.008 (2021).
36 Consortium, C. S. A. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87, doi:10.1038/nature04072 (2005).
37 Franchini, L. F. & Pollard, K. S. Human evolution: the non-coding revolution. *BMC Biol* **15**, 89, doi:10.1186/s12915-017-0428-9 (2017).
38 Prabhakar, S., Noonan, J. P., Paabo, S. & Rubin, E. M. Accelerated evolution of conserved noncoding sequences in humans. *Science* **314**, 786, doi:10.1126/science.1130738 (2006).
39 Pollard, K. S. *et al.* An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**, 167-172, doi:10.1038/nature05113 (2006).
40 Capra, J. A., Erwin, G. D., McKinsey, G., Rubenstein, J. L. & Pollard, K. S. Many human accelerated regions are developmental enhancers. *Philos Trans R Soc Lond B Biol Sci* **368**, 20130025, doi:10.1098/rstb.2013.0025 (2013).
41 Gittelman, R. M. *et al.* Comprehensive identification and analysis of human accelerated regulatory DNA. *Genome Res* **25**, 1245-1255, doi:10.1101/gr.192591.115 (2015).
42 Whalen, S. *et al.* Machine-learning dissection of Human Accelerated Regions in primate neurodevelopment. *bioRxiv*, 256313, doi:10.1101/256313 (2022).
43 Girskis, K. M. *et al.* Rewiring of human neurodevelopmental gene regulatory programs by human accelerated regions. *Neuron* **109**, 3239-3251 e3237, doi:10.1016/j.neuron.2021.08.005 (2021).

44 Uebbing, S. *et al.* Massively parallel discovery of human-specific substitutions that alter enhancer activity. *Proc Natl Acad Sci U S A* **118**, doi:10.1073/pnas.2007049118 (2021).

45 Klein, J. C. *et al.* A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat Methods* **17**, 1083-1091, doi:10.1038/s41592-020-0965-y (2020).

46 Charrier, C. *et al.* Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* **149**, 923-935, doi:10.1016/j.cell.2012.03.034 (2012).

47 Prufer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43-49, doi:10.1038/nature12886 (2014).

48 Weiss, C. V. *et al.* The cis-regulatory effects of modern human-specific variants. *Elife* **10**, doi:10.7554/eLife.63713 (2021).

49 Lai, C. S., Fisher, S. E., Hurst, J. A., Vargha-Khadem, F. & Monaco, A. P. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* **413**, 519-523, doi:10.1038/35097076 (2001).

50 Enard, W. *et al.* Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* **418**, 869-872, doi:10.1038/nature01025 (2002).

51 Enard, W. *et al.* A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. *Cell* **137**, 961-971, doi:10.1016/j.cell.2009.03.041 (2009).

52 Konopka, G. *et al.* Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature* **462**, 213-217, doi:10.1038/nature08549 (2009).

53 Trujillo, C. A. *et al.* Reintroduction of the archaic variant of NOVA1 in cortical organoids alters neurodevelopment. *Science* **371**, doi:10.1126/science.aax2537 (2021).

54 Pinson, A. *et al.* Human TKTL1 implies greater neurogenesis in frontal neocortex of modern humans than Neanderthals. *Science* **377**, eabl6422, doi:10.1126/science.abl6422 (2022).

55 Somel, M. *et al.* Transcriptional neoteny in the human brain. *Proc Natl Acad Sci U S A* **106**, 5743-5748, doi:10.1073/pnas.0900544106 (2009).

56 Liu, X. *et al.* Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques. *Genome Res* **22**, 611-622, doi:10.1101/gr.127324.111 (2012).

57 Bianchi, S. *et al.* Synaptogenesis and development of pyramidal neuron dendritic morphology in the chimpanzee neocortex resembles humans. *Proc Natl Acad Sci U S A* **110 Suppl 2**, 10395-10401, doi:10.1073/pnas.1301224110 (2013).

58 Zhu, Y. *et al.* Spatiotemporal transcriptomic divergence across human and macaque brain development. *Science* **362**, doi:10.1126/science.aat8077 (2018).

59 Marchetto, M. C. *et al.* Species-specific maturation profiles of human, chimpanzee and bonobo neural cells. *Elife* **8**, doi:10.7554/eLife.37527 (2019).
60 Kanton, S. *et al.* Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574**, 418-422, doi:10.1038/s41586-019-1654-9 (2019).
61 Schornig, M. *et al.* Comparison of induced neurons reveals slower structural and functional maturation in humans than in apes. *Elife* **10**, doi:10.7554/eLife.59323 (2021).
62 Gordon, A. *et al.* Long-term maturation of human cortical organoids matches key early postnatal transitions. *Nat Neurosci* **24**, 331-342, doi:10.1038/s41593-021-00802-y (2021).
63 Miller, D. J. *et al.* Prolonged myelination in human neocortical evolution. *Proc Natl Acad Sci U S A* **109**, 16480-16485, doi:10.1073/pnas.1117943109 (2012).
64 Yeung, M. S. *et al.* Dynamics of oligodendrocyte generation and myelination in the human brain. *Cell* **159**, 766-774, doi:10.1016/j.cell.2014.10.011 (2014).
65 Caglayan, E., Ayhan, F., Liu, Y., Vollmer, R., Oh, E., Sherwood, C. C., Preuss, T. M., Yi, S., Konopka, G. . *Molecular features driving cellular complexity of human brain evolution* (2023).
66 Jorstad, N. L. *et al.* Comparative transcriptomics reveals human-specific cortical features. *bioRxiv*, 2022.2009.2019.508480, doi:10.1101/2022.09.19.508480 (2022).
67 Castelijns, B. *et al.* Hominin-specific regulatory elements selectively emerged in oligodendrocytes and are disrupted in autism patients. *Nat Commun* **11**, 301, doi:10.1038/s41467-019-14269-w (2020).
68 Konopka, G. *et al.* Human-specific transcriptional networks in the brain. *Neuron* **75**, 601-617, doi:10.1016/j.neuron.2012.05.034 (2012).
69 Sousa, A. M. M. *et al.* Molecular and cellular reorganization of neural circuits in the human lineage. *Science* **358**, 1027-1032, doi:10.1126/science.aan3456 (2017).
70 Berto, S. *et al.* Accelerated evolution of oligodendrocytes in the human brain. *Proc Natl Acad Sci U S A* **116**, 24334-24342, doi:10.1073/pnas.1907982116 (2019).
71 Kozlenkov, A. *et al.* Evolution of regulatory signatures in primate cortical neurons at cell-type resolution. *Proc Natl Acad Sci U S A* **117**, 28422-28432, doi:10.1073/pnas.2011884117 (2020).
72 Paolicelli, R. C. *et al.* Synaptic pruning by microglia is necessary for normal brain development. *Science* **333**, 1456-1458, doi:10.1126/science.1202529 (2011).

73 Buchanan, J. *et al.* Oligodendrocyte precursor cells ingest axons in the mouse neocortex. *Proc Natl Acad Sci U S A* **119**, e2202580119, doi:10.1073/pnas.2202580119 (2022).

74 Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61-68, doi:10.1038/s41586-019-1506-7 (2019).

75 Bakken, T. E. *et al.* Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111-119, doi:10.1038/s41586-021-03465-8 (2021).

76 Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol Cell* **49**, 825-837, doi:10.1016/j.molcel.2013.01.038 (2013).

77 Vermunt, M. W. *et al.* Epigenomic annotation of gene regulatory alterations during evolution of the primate brain. *Nat Neurosci* **19**, 494-503, doi:10.1038/nn.4229 (2016).

78 Reilly, S. K. *et al.* Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**, 1155-1159, doi:10.1126/science.1260943 (2015).

79 Jeong, H. *et al.* Evolution of DNA methylation in the human brain. *Nat Commun* **12**, 2021, doi:10.1038/s41467-021-21917-7 (2021).

80 He, Y. & Ecker, J. R. Non-CG Methylation in the Human Genome. *Annu Rev Genomics Hum Genet* **16**, 55-77, doi:10.1146/annurev-genom-090413-025437 (2015).

81 Yap, E. L. & Greenberg, M. E. Activity-Regulated Transcription: Bridging the Gap between Neural Activity and Behavior. *Neuron* **100**, 330-348, doi:10.1016/j.neuron.2018.10.013 (2018).

82 Toma, K., Kumamoto, T. & Hanashima, C. The timing of upper-layer neurogenesis is conferred by sequential derepression and negative feedback from deep-layer neurons. *J Neurosci* **34**, 13259-13276, doi:10.1523/JNEUROSCI.2334-14.2014 (2014).

83 Golson, M. L. & Kaestner, K. H. Fox transcription factors: from development to disease. *Development* **143**, 4558-4570, doi:10.1242/dev.112672 (2016).

84 Siletti, K. *et al.* Transcriptomic diversity of cell types across the adult human brain. *bioRxiv*, 2022.2010.2012.511898, doi:10.1101/2022.10.12.511898 (2022).

85 Chiou, K. L. *et al.* A single-cell multi-omic atlas spanning the adult rhesus macaque brain. *bioRxiv*, 2022.2009.2030.510346, doi:10.1101/2022.09.30.510346 (2022).

86 van Tilborg, E. *et al.* Origin and dynamics of oligodendrocytes in the developing brain: Implications for perinatal white matter injury. *Glia* **66**, 221-238, doi:10.1002/glia.23256 (2018).

87 Soreq, L. *et al.* Major Shifts in Glial Regional Identity Are a Transcriptional Hallmark of Human Brain Aging. *Cell Rep* **18**, 557-570, doi:10.1016/j.celrep.2016.12.011 (2017).
88 Allen, W. E., Blosser, T. R., Sullivan, Z. A., Dulac, C. & Zhuang, X. Molecular and spatial signatures of mouse brain aging at single-cell resolution. *Cell* **186**, 194-208 e118, doi:10.1016/j.cell.2022.12.010 (2023).
89 de Magalhaes, J. P. & Costa, J. A database of vertebrate longevity records and their relation to other life-history traits. *J Evol Biol* **22**, 1770-1774, doi:10.1111/j.1420-9101.2009.01783.x (2009).
90 Hill, R. A., Li, A. M. & Grutzendler, J. Lifelong cortical myelin plasticity and age-related degeneration in the live mammalian brain. *Nat Neurosci* **21**, 683-695, doi:10.1038/s41593-018-0120-6 (2018).
91 Hughes, E. G., Orthmann-Murphy, J. L., Langseth, A. J. & Bergles, D. E. Myelin remodeling through experience-dependent oligodendrogenesis in the adult somatosensory cortex. *Nat Neurosci* **21**, 696-706, doi:10.1038/s41593-018-0121-5 (2018).
92 Revah, O. *et al.* Maturation and circuit integration of transplanted human cortical organoids. *Nature* **610**, 319-326, doi:10.1038/s41586-022-05277-w (2022).
93 Knowles, J. K., Batra, A., Xu, H. & Monje, M. Adaptive and maladaptive myelination in health and disease. *Nat Rev Neurol* **18**, 735-746, doi:10.1038/s41582-022-00737-3 (2022).
94 Kougioumtzidou, E. *et al.* Signalling through AMPA receptors on oligodendrocyte precursors promotes myelination by enhancing oligodendrocyte survival. *Elife* **6**, doi:10.7554/eLife.28080 (2017).
95 Blanchard, J. W. *et al.* APOE4 impairs myelination via cholesterol dysregulation in oligodendrocytes. *Nature* **611**, 769-779, doi:10.1038/s41586-022-05439-w (2022).
96 Qiu, J. *et al.* Evidence for evolutionary divergence of activity-dependent gene expression in developing neurons. *Elife* **5**, doi:10.7554/eLife.20337 (2016).
97 Pruunsild, P., Bengtson, C. P. & Bading, H. Networks of Cultured iPSC-Derived Neurons Reveal the Human Synaptic Activity-Regulated Adaptive Gene Program. *Cell Rep* **18**, 122-135, doi:10.1016/j.celrep.2016.12.018 (2017).
98 Su, Y. *et al.* Neuronal activity modifies the chromatin accessibility landscape in the adult brain. *Nat Neurosci* **20**, 476-483, doi:10.1038/nn.4494 (2017).
99 Gidon, A. *et al.* Dendritic action potentials and computation in human layer 2/3 cortical neurons. *Science* **367**, 83-87, doi:10.1126/science.aax6239 (2020).

100 Mertens, J. *et al.* Age-dependent instability of mature neuronal fate in induced neurons from Alzheimer's patients. *Cell Stem Cell* **28**, 1533-1548 e1536, doi:10.1016/j.stem.2021.04.004 (2021).
101 Hubisz, M. J. & Pollard, K. S. Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. *Curr Opin Genet Dev* **29**, 15-21, doi:10.1016/j.gde.2014.07.005 (2014).
102 Whalen, S. *et al.* Machine learning dissection of human accelerated regions in primate neurodevelopment. *Neuron* **111**, 857-873 e858, doi:10.1016/j.neuron.2022.12.026 (2023).