UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE QUÍMICA

Programa de Pós-Graduação em Ciências Biológicas (Bioquímica)

DAVID ABRAHAM MORALES VICENTE

Caracterização da evolução dos RNAs longos não-codificadores no transcriptoma do cortex cerebral em desenvolvimento

Versão corrigida da Tese defendida

São Paulo

Data do Depósito na SPG:

07/10/2022

DAVID ABRAHAM MORALES VICENTE

Caracterização da evolução dos RNAs longos não-codificadores no transcriptoma do cortex cerebral em desenvolvimento

Tese apresentada ao Instituto de Química da Universidade de São Paulo para a obtenção do Título de Doutor em Ciências (Bioquímica)

Orientador: Prof. Dr. Sergio Verjovski-Almeida

São Paulo

2022

Ficha Catalográfica elaborada eletronicamente pelo autor, utilizando o programa desenvolvido pela Seção Técnica de Informática do ICMC/USP e adaptado para a Divisão de Biblioteca e Documentação do Conjunto das Químicas da USP

Bibliotecária responsável pela orientação de catalogação da publicação:
Marlene Aparecida Vieira - CRB - 8/5562

M828c

Morales-Vicente, David Abraham
Caracterização da evolução dos RNAs longos não-codificadores no transcriptoma do cortex cerebral em desenvolvimento / David Abraham Morales-Vicente. - São Paulo, 2022.
79 p.

Tese (doutorado) - Instituto de Química da Universidade de São Paulo. Departamento de Bioquímica.
Orientador: Verjovski-Almeida, Sergio

1. Cortex cerebral. 2. Elementos transponíveis. 3. Neurodesenvolvimento . 4. Neurônios glutamatérgicos. 5. RNAs longos não codificadores. I. T. II. Verjovski-Almeida, Sergio, orientador.

Universidade de São Paulo
Instituto de Química

"Caracterização da evolução dos RNAs longos não-codificadores no transcriptomado cortex cerebral em desenvolvimento"

DAVID ABRAHAM MORALES VICENTE

Tese de Doutorado submetida ao Instituto de Química da Universidade de São Paulo como parte dos requisitos necessários à obtenção do grau de Doutor em Ciências obtido no Programa Ciências Biológicas (Bioquímica) - Área de Concentração: Bioquímica.

_____

Prof. Dr. Sergio Verjovski de Almeida
(Orientador e Presidente)

APROVADO(A) POR:

_____

Profa. Dra. Bettina Malnic
IQ - USP

_____

Prof. Dr. Sergio Teixeira Ferreira
UFRJ

_____

Prof. Dr. Paulo de Paiva Rosa Amaral
INSPER

SÃO PAULO
16 de novembro de 2022

DEDICATION

This work is dedicated to my family.

In memoriam of my aunt Raquel.

Acknowledgment

I want to acknowledge Professor Sergio Verjovski-Almeida for his support and guidance during my doctoral years. Without his help and patience, I could not be able to carry on with this project. My stay in his lab is a life-changing experience.

I want to acknowledge Professor Irene Yan for her advice and gratitude for sharing her lab resources with me. Her help has been enormous.

To Ana Tahira, Ana Paula, Daisy, Lucas Maciel, Lucas Ferreira, Murillo, Gabriela, Gilbert, Adriana, Alexandre, João, Thalles, and all the other members of the lab for all their support throughout all these years, for being good companions in the long days of experiments, for the proliferative discussion about science and life, and for bringing joy to my life. You will be in my heart forever.

To my family for their love, patience, and support. All I am because of you.

To Martin, for being the best friend ever despite the distance and the time zones. To Ruth, Berna, Kim, Alexis, and Cristian, for being my family in Sao Paulo.

Finally, I would like to thank the CAPES agency for funding my scholarship. Without its funding, nothing could have been done.

Resumo

Morales-Vicente, D. A. **Caracterização da evolução dos RNAs longos não-codificadores no transcriptoma do cortex cerebral em desenvolvimento.** 2022. (80p). Tese (Doutorado) - Programa de Pós-Graduação em Bioquímica. Instituto de Química, Universidade de São Paulo, São Paulo.

As excelentes habilidades cognitivas humanas são computadas no córtex cerebral. Essa estrutura cerebral específica de mamíferos, dotada de notável plasticidade, tem sido o lugar de inovações biológicas massivas ao longo da evolução. Na última década, RNAs longos não codificadores (lncRNAs) surgiram como moléculas reguladoras. Eles apresentam maior especificidade tecidual e *turnover* evolutivo do que genes codificadores de proteínas e são altamente expressos em tecidos neurais, tornando-os candidatos interessantes para plasticidade, evolução e doença do cortex cerebral. Como as mudanças na expressão de lncRNAs ou a expressão *de novo* de lncRNAs impactaram o desenvolvimento do córtex cerebral? Continua sendo uma questão aberta. Para caracterizar as mudanças evolutivas de lncRNAs no córtex cerebral em desenvolvimento, usamos abordagens biológicas de sistema; primeiro, para anotar de forma abrangente o repertório de lncRNAs em humanos, macacos rhesus, camundongos e galinhas; segundo, para identificar a conservação sintênica do repertório de lncRNAs corticais na linhagem humana, classificando-os em grupos evolutivos em função da idade mínima prevista. Esses grupos de lncRNAs apresentaram diferenças nas características genéticas e na dinâmica de expressão, indicando uma diferença em sua funcionalidade. Ao combinar a análise de single-cell-RNA-seq e RNA-seq, o contexto celular da inovação do repertório de lncRNAs foi revelado; lncRNAs mais antigos mostraram expressão preferencial em estágios iniciais do neurodesenvolvimento e zonas germinativas; enquanto lncRNAs humanos específicos mostraram expressão preferencial em neurônios glutamatérgicos, foram enriquecidos em módulos coexpressos de genes específicos de humanos e desregulados em Transtorno do Espectro Autista. Os resultados destacam os lncRNAs como fontes genéticas da evolução do córtex cerebral e diversificação dos neurônios glutamatérgicos.

Palavras-chaves: Cortex cerebral. Elementos transponíveis. Evolução humana. Neurodesenvolvimento. Neurônios glutamatérgicos. RNAs longos não codificadores.

Abstract

Morales-Vicente, D. A. **Characterization of the evolution of long non-coding RNAs in the developing cerebral cortex transcriptome.** 2022. (80p). Tese (Doutorado) - Programa de Pós-Graduação em Bioquímica. Instituto de Química, Universidade de São Paulo, São Paulo.

The human outstanding cognitive abilities are computed in the cerebral cortex. This mammalian-specific brain structure has been the place of massive biological innovations throughout evolution. Over the past decade, long non-coding RNAs (lncRNAs) have emerged as gene regulatory elements with greater tissue-specificity and evolutionary turnover than mRNAs. lncRNAs are highly expressed in neural tissues, making them candidates for cerebral cortex plasticity, evolution, and disease. Whether changes in the expression or the *de novo* expression of lncRNAs have impacted the development and evolution of the human cerebral cortex remains an open question. To characterize the evolutionary changes of lncRNAs in the developing cerebral cortex, we used system biology approaches to comprehensively annotate the repertory of lncRNAs in humans, rhesus macaques, mice, and chickens; and to identify the syntenic conservation of the cortical lncRNA repertory in the human transcriptome, classifying human lncRNAs into evolutionary groups as a function of the predicted minimal age. Those groups of lncRNAs showed differences in genomic and regulatory features and expression dynamics, indicating differences in their functionality. By combining single-cell RNA-seq and weighted gene co-expression network analysis, the cellular context of the innovation of the lncRNAs repertory was unveiled; older lncRNAs showed preferential expression in early neurodevelopmental stages and germinative zones, while newer lncRNAs showed preferential expression in synaptogenic glutamatergic neurons and Human-specific gene co-expression modules. Additionally, newer lncRNAs were dysregulated in autism spectrum disorders, a Human-specific disease. These results highlight the *de novo* expression of lncRNAs as genetic sources of cerebral cortex evolution, especially for the diversification and dysfunction of glutamatergic neurons.

Keywords: Cerebral cortex. Glutamatergic neurons, Human evolution. Long non-coding RNAs. Neurodevelopment. Transposable elements.

List of abbreviations

aRGCs: Apical radial glial cells

bHLH: Basic helix-loop-helix

CDS: Coding sequence

CP: Cortical plate

CRs: Cajal-Retzius cells

DE: Differentially expressed

DEG: Differentially expressed genes

DL: Deep cortical layers

DVR: Dorsal ventricular ridge

ERV: Endogenous retrovirus

GlutN: Glutamatergic projection neurons

GO: Gene ontology

GZ: Germinative zone

HAR: Human accelerated region

IN: Interneurons

IPCs: Intermediate progenitor cells

kIN: intramodular connectivity

lncRNAs: Long non-coding RNAs

lincRNAs: Long intergenic non-coding RNAs

MYA: Million years ago

MZ: Mantle zone

mRNAs: messenger RNAs

NE: Neuroepithelial cell

OCRs: Open chromatin regions

OPC: Oligodendrocyte progenitor cells

ORF: Open reading frame

oRGCs: Outer radial glial cells

oSVZ: Outer subventricular zone

RBC: Red blood cells

RGCs: Radial glial cells

scRNA: single-cell RNA-sequencing

SVZ: Subventricular zone

TE: Transposable element

TF: Transcription factor

TIN: Transcript integrity number

UL: Upper cortical layers

UMAP: Uniform Manifold Approximation and Projection

UTR: Untranslated region

VZ: Ventricular zone

WGCNA: Weighted gene co-expression network analysis

Table of contents

1 Introduction

1.1 The cerebral cortex

1.1.1 The cerebral cortex development

The cerebral cortex is a primary information processing center of the central nervous system, key to the evolution of higher cognition, and affected in neurodevelopmental disorders. It comprises billions of excitatory projection neurons (glutamatergic) and inhibitory (GABAergic) interneurons assembled in local circuits intertwined with glial and vascular cells arranged in a six-layered architecture on the outer surface of the brain. This structure unique to mammals develops from the dorsal pallium in a precisely orchestrated process known as corticogenesis (Figure 1) (Libé-Philippot & Vanderhaeghen, 2021; Molnár et al., 2019).

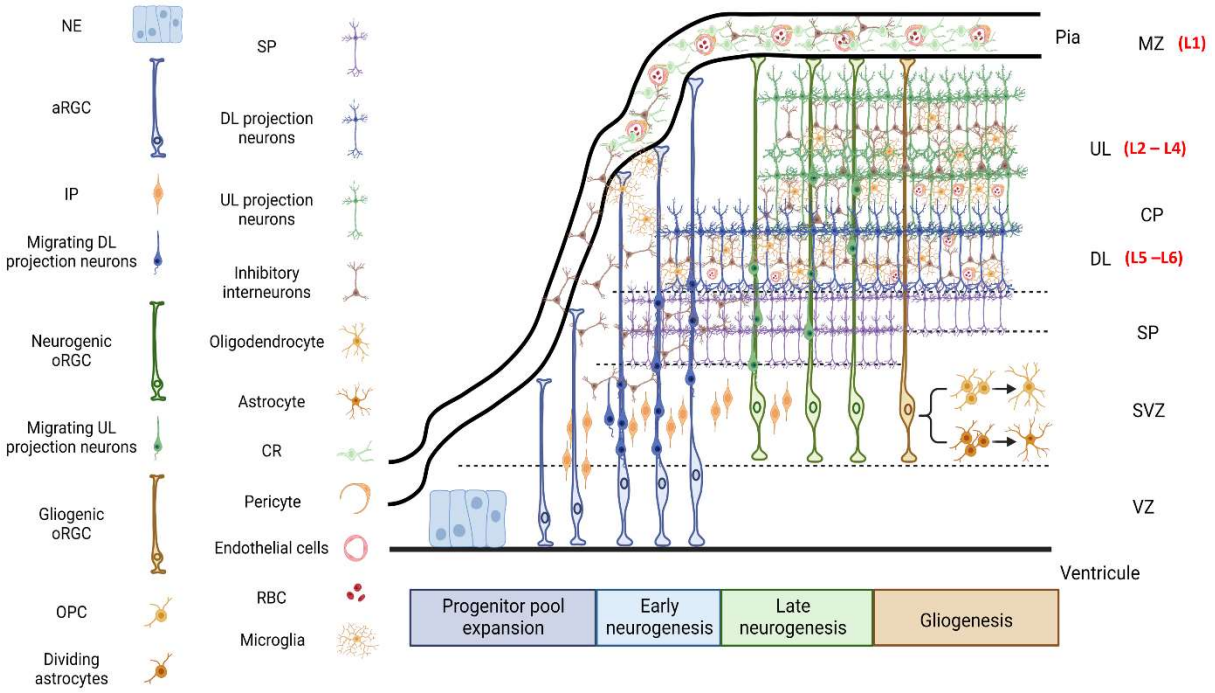

Figure 1. Schematic depiction of the distinct stages of cortical cell populations development. Apical progenitors on the dorsal pallium expand the cell progenitor pool by symmetric division at the beginning of corticogenesis. The neurogenic phase starts with the asymmetric division of aRGCs at the VZ, which directly produce immature projection neurons that will mainly populate the deep layers (DL) of the cortical plate, layers 5 and 6. Additionally, aRGCs give rise to IP cells that populate the SVZ and undergo a few symmetric divisions before generating immature projection neurons. At late neurogenic stages, aRGCs detach from the VZ and occupy the SVZ, becoming oRGCs that divide asymmetrically to produce upper layer (UL) cortical projection neurons, layers 2, 3, and 4. After the neurogenic period, oRGCs become gliogenic and start the generation of OPCs and dividing astrocytes. OPCs perdure in adulthood and continuously produce oligodendrocytes. Cajal-Retzius cells, inhibitory interneurons, microglia, and vascular cells that populate the cortical plate originate outside the dorsal pallium and migrate tangentially at different stages of corticogenesis to integrate into the developing cortex. **Abbreviations:** aRGCs, apical radial glial cells; CP, cortical plate; CR, Cajal-Retzius cells; DL, deep layer; IP, intermediate progenitors; MZ, mantle zone; NE, neuroepithelial cells; OPC, oligodendrocyte progenitor cells,

oRGCs; outer radial glial cells; RBC, red blood cells; SP, subplate neurons; SVZ, subventricular zone; UL, Upper layer; VZ, ventricular zone.

Corticogenesis begins with the progenitor pool expansion (Figure 1, dark blue box) by symmetric division of the neuroepithelial cells (NEs) in the ventricular zone (VZ) of the dorsal pallium. While dividing, NEs progressively transition into apical radial glial cells (aRGCs), which produce neurons directly or indirectly (Figure 1, light blue box). aRGCs can amplify themselves by symmetric division or divide asymmetrically to make immature glutamatergic neurons at the VZ. Apical progenitors located at the VZ are characterized by the expression of the paired-box transcription factor (TF) PAX6. Morphologically, aRGCs are bipolar cells that extend a short apical process to the ventral surface of the dorsal pallium and a long basal process to the pial surface, generating radial scaffolds that expand across the cortical thickness. Newly established immature neurons that express the basic helix-loop-helix (bHLH) TFs NEUROD2 and NEUROD6 use these scaffolds to migrate inside-out radially to their destination in the cortical plate (CP); neurons born at early neurogenesis will mainly produce transient subplate neurons (SP) at first, characterized by the expression of the TFs CRYM and NR4A2; and after that, projection neurons from deep cortical layers (DL) L6 and L5 that extend axons out of the telencephalon to the thalamus, brainstem, and spinal cord, and are molecularly characterized by the expression of the T-box brain TF TBR1 and the C2H2-type zinc finger TF BCL11B (Di Bella et al., 2021).

Additionally, the asymmetric division of aRGCs gives rise to intermediate progenitor cells (IPCs) that may undergo a few symmetric divisions before going into terminal divisions generating immature projection neurons. Unlike aRGCs, IPCs cells have multipolar morphology, express the T-box brain TF TBR2, and populate the subventricular zone (SVZ) that lies on top of the VZ. aRGCs progressively change their molecular identity, losing their apical process, detaching from the ventricle, and moving toward the SVZ, becoming basal RGCs or neurogenic outer RGCs (oRGCs) (Figure 1, green box). At a molecular level, oRGCs are like aRGCs but have a more robust expression of *MOXD1, HOPX, PTPRZ1, TNC,* and *FAM107A* genes (Pollen et al., 2015). oRGCs undergo symmetric divisions at the SVZ to increase the cortical progenitor pool and asymmetric divisions to produce late-born projection neurons that mainly populate the upper cortical layers (UL) L4, L3, and L2 that extend intratelencephalic axons to other cortical locations. At the molecular level, UL projection neurons express the homeobox TFs SATB2, CUX1, and CUX2. After their neurogenic period, oRGCs change their molecular identity to start the production of local glial progenitor cells

(Figure 1, brown box), which culminate in the production of oligodendrocyte progenitor cells (OPC), recognized by the expression of the bHLH TFs OLIG1 and OLIG2, which produce oligodendrocytes throughout the lifespan of the animal, and dividing astrocytes, that selectively express GFAP and the aquaporin AQP4 (Di Bella et al., 2021; Fan et al., 2018; Trevino et al., 2021; Zhong et al., 2018).

In addition to the locally originated excitatory projection neurons and glial cells, other cellular components of the cerebral cortex develop outside the dorsal pallium and migrate to be integrated into the developing cortical plate. Those cells include Cajal-Retzius cells (CRs), cortical interneurons, and microglia (Figure 1, left). CRs are developmental transient glutamatergic neurons that tangentially migrate from the pallial boundaries and populate the cortical mantle zone (MZ) or cortical layer L1 at the early stages of neurogenesis. They produce the extracellular matrix protein REELIN, which is crucial for proper cortical layering of glutamatergic neurons (Frotscher, 1998). GABAergic interneurons, the inhibitory neural component of the cortical circuitry, originate in the ventral pallium, mainly from the medial ganglionic eminence (MGE) and the caudal ganglionic eminence (CGE), from where they migrate tangentially to reach the developing cortical plate through different migratory streams. Interneurons typically express the homeobox TFs DLX1 and DLX2 and the glutamic acid decarboxylases GAD1 and GAD2. Meanwhile, microglia, the hematopoietic-derived cells that function as brain macrophages, originate from the yolk sac at the early embryonic stages, from where they migrate through the developing vasculature to infiltrate the cortical plate before the closure of the blood-brain barrier (Di Bella et al., 2021; Fan et al., 2018; Thion, Ginhoux, & Garel, 2018; Zhong et al., 2018).

1.1.2 Evolution of the cerebral cortex

Most of our understanding of corticogenesis comes from studies of model organisms, mainly from the developing cortex of mice. Although the cellular and molecular mechanisms of corticogenesis are conserved, clade-specific differences exist; understanding these differences at the molecular level is critical to unveiling the evolution of human higher cognition and having a deeper comprehension of how they are disrupted in disease (Silbereis, Pochareddy, Zhu, Li, & Sestan, 2016).

The six-layered cerebral cortex or neocortex evolved from the dorsal pallium of mammals after they diverged from sauropsids, which include reptiles and birds, around 300 million years ago (MYA). The dorsal pallium of sauropsids is arranged in different structures.

The anterior dorsal pallium is usually assembled in a three-layer cortex in reptiles. In birds, the pallium is organized as a series of nuclei, the largest of them being the dorsal ventricular ridge (DVR) located in the lateral and ventral pallium (Tosches, 2021), while the dorsal pallium is a reduced structure named hyperpallium. The remarkable plasticity of pallial architectures is not limited to extensive divergence periods; in mammals, the cerebral cortex is endowed with incredible plasticity, evident in the diverse neocortical sizes and shapes (Lui, Hansen, & Kriegstein, 2011; Silbereis et al., 2016). Those differences have resulted from molecular changes at the regulatory and genetic level in corticogenesis that tuned the conserved transcriptional landscape to fit the neural processes demand of the host species.

Primates present an expanded brain with an increased number of total neurons compared to most mammalian species. The human cerebral cortex has further expanded, differentiating us from our closest living relatives. These expansions are likely responsible for the augmented computational capacity of the human cerebral cortex and have been associated with the evolution of cognition in humans. At the cellular level, the human brain expansion is the result of an augmented proliferative capacity of RGCs, especially from the outer SVZ (oSVZ) (Lui et al., 2011). Using single-cell RNA-seq, it has been identified that human oRGCs express several protein-coding genes that self-sustain a niche that favors their self-renewal at the oSVZ (Pollen et al., 2015). Additionally, it has been found that modern segmental duplications lead to the evolution of Human-specific paralogs of the NOTCH signaling pathway, which are expressed in RGCs, enhancing their proliferative capacity (Fiddes et al., 2018).

The further selective expansion of the oRGCs in human corticogenesis leads to the expansion of the cerebral cortex, especially UL pyramidal glutamatergic neurons, mainly derived from late neurogenesis at the oSVZ. Human UL glutamatergic neurons present a high diversity measured at transcriptional and electrophysiological levels compared to the mouse (Berg et al., 2021). This expansion of the Human ULs has been associated with a more significant number of intratelencephalic neuron connections and the evolution of more efficient pyramidal neurons that leads to an increase in the computation capacity of the human brain.

This extensive work has been done in elucidating the molecular basis of the evolution of the cerebral cortex in the human lineage is mainly focused on changes in protein-coding genes; expanding this analysis to the human non-coding genome would improve our understanding of the gene-regulatory changes that have been taken place throughout corticogenesis and lead to the evolution of the human cerebral cortex.

1.2 Long non-coding RNAs

1.2.1 Genomic and molecular features of long non-coding RNAs

The central dogma of molecular biology state that genes are transcribed into messenger RNAs (mRNAs) to produce proteins, the functional blocks of life. Since the appreciation of the extensive transcription of the human genome outside protein-coding sequences, thousands of non-coding genes have been annotated. Long non-coding RNAs (lncRNAs) are non-coding genes transcribed into RNAs longer than 200 nucleotides that do not translate into functional proteins (Statello, Guo, Chen, & Huarte, 2021). lncRNAs represent a heterogenous group of RNAs and can be classified regarding their genomic position to the nearest protein-coding genes. lncRNAs transcribed in the opposite strand of protein-coding genes are known as antisense lncRNAs if they overlap the gene body, and bidirectional lncRNAs if they overlap the promoter region; similarly, lncRNAs transcribed from protein-coding introns are known as intronic lncRNAs, and those overlapping a protein-coding gene in the same direction are classified as overlapping lncRNAs. Finally, lncRNAs that do not overlap any protein-coding locus are long intergenic non-coding RNAs (lincRNAs).

This heterogeneous group of lncRNAs is transcribed by the RNA pol II and shares molecular features with mRNA, such as being 5' capped, spliced, and polyadenylated. Despite the molecular similarities with mRNAs, lncRNAs also present features that differentiate them, including increased tissue-specificity compared to mRNAs, distinct chromatin modifications at the promoter region, cell nucleus enrichment, inefficient splicing, and less stability than mRNAs (Ransohoff, Wei, & Khavari, 2018; Rinn & Chang, 2020). Although these features may point to lncRNAs as mere transcriptional noise, it has been shown in the past decade that at least a fraction of lncRNAs or the act of their transcription have gene-regulatory functions. An excellent example of functional lncRNAs is *XIST*, one of the most extensively studied lncRNAs, which through interactions with RNA binding proteins, directs chromosome X inactivation in mammals. By silencing the expression of most of the genes from the chromosomes where it is expressed, *XIST* has a gene-regulatory effect *in cis*. Nevertheless, other lncRNAs may have gene regulatory functions *in trans*, away from the loci where they are expressed (Rinn & Chang, 2020).

Interestingly, it has been shown that, after testes, neural tissues express the most significant number of lncRNAs in tetrapods (Anamaria Necsulea et al., 2014). Furthermore, several lncRNAs have been characterized as functional regulatory RNAs of different stages and

cell populations of mice corticogenesis, such as *Pnky, Sox2ot, Evf2,* and *Nr2f1-as1* (Ang et al., 2019; Cajigas et al., 2018; Knauss et al., 2018; Ramos et al., 2015).

1.2.2 Evolution of long non-coding RNAs

Unlike protein-coding genes that mainly have evolved by gene duplications, lncRNAs have preferentially evolved by *de novo* expression and exonization mediated by transposable elements (TEs). It has been shown that up to 75% of human lincRNAs transcripts harbor at least a fraction of a TE sequence in their transcript bodies. TEs contributes to every step of lncRNA biogenesis by providing sequence motif for splicing sites, polyadenylation signals, and regulatory sequences at promoter regions. Additionally, TE stretches in lncRNAs show more significant evolutionary constraints than non-TE sequences; even more, it has been proposed that TEs operate as functional modules of lncRNAs, by providing sequence motifs for interaction with RNA binding proteins. TEs also provide nuclear localization signals to lncRNAs, such as the primate-specific Alu sequences (Johnson & Guigó, 2014; Kapusta et al., 2013; Lubelsky & Ulitsky, 2018).

The *de novo* expression and the significant contributions of TEs to the evolution of lncRNAs explain the reduced constraint under which lncRNA evolved compared to protein-coding genes. This faster evolutionary turnover of lncRNAs compared to mRNAs and their gene-regulatory functions make lncRNAs good candidates for molecular drivers of biological innovations. In fact, the first identified highly evolving Human-specific region (HAR, human accelerated region) was the lncRNA *HAR1F*, expressed in Cajal-Retzius cells of the developing neocortex (Katherine S. Pollard et al., 2006). In addition, it has been shown in mammals that lncRNAs are a source of cellular plasticity due to the capacity to acquire new functional modalities (Guo et al., 2020). Overall, the unique features of lncRNAs and their increased tissue-specificity and enrichment in neural tissues indicate that lncRNAs are good candidates for genetic sources of evolution and diversification of the human cerebral cortex.

2 Aims and objectives

2.1 Aims

Characterize the molecular basis of the evolution of the lncRNA repertory expressed throughout human corticogenesis and investigate the potential role of lncRNAs in the evolution of the cerebral cortex.

2.2 Objectives

1. Development of a bioinformatics pipeline to comprehensively annotate the repertory of lncRNAs in the developing brain of humans, macaques, mice, and chickens.
2. Development of a bioinformatics methodology to identify the minimal evolutionary age of lncRNAs based on their syntenic conservation between species and cluster them into evolutionary groups.
3. Identify the set of lncRNAs expressed throughout human corticogenesis and characterize the genomic similarities and differences of the distinct evolutionary groups of cortical lncRNAs.
4. Examine the cellular and developmental context in which lncRNA innovations arise in human corticogenesis.
5. Assessment of the contribution of distinct evolutionary groups of lncRNAs to preserved and lineage-specific modules of co-expressed genes in human corticogenesis.
6. Assessment of the dysregulation of the distinct evolutionary groups of cortical lncRNAs in neurodevelopmental disorders.

3 Materials and Methods

3.1 Tissue collection, RNA extraction, and sequencing

Gallus gallus fertilized eggs were purchased from a local provider and incubated at 38°C and 50% humidity for seven and ten days. Embryos were collected and decapitated; brains were removed from the heads, and forebrains were further dissected. For developmental day 7 (E7), the whole pallium and subpallium were retrieved; for developmental day 10 (E10), the entire subpallium, the dorsolateral pallium, and the medial pallium were retrieved. Three brain sections from different embryos were pooled per working sample without considering biological sex. Brain sections were dissociated using pestles in 1 ml TRIzol and frozen at -80 °C until the day of RNA isolation.

For RNA isolation, 200 µL of chloroform was added per 1 ml TRIzol and centrifuged for 15 min and 16000 g at 4 °C; supernatants were transferred to new microtubes. One volume of ethanol 100% was added to each sample, then transferred to RNeasy Mini spin columns and the RNeasy Micro-kit protocol was followed.

RNA samples were quantified using Qubit2 Fluorometer (ThermoFisher), and their integrity was measured using Bioanalyzer 2100 (Agilent). RNA integrity number (RIN) of samples went from 7.5 to 8.5, which indicates the good quality of the samples. For each tissue-developmental window, four biological replicates were prepared and sent for sequencing to BGI Genomics (Shenzhen, China).

3.2 Bulk RNA-seq processing

Public libraries reads were retrieved from the SRA repository at GenBank (NCBI, USA) using *fasterq-dump* with the following parameter "--split-files"; the integrity of the data was checked using *vdb-validate*, and all files were identified as consistent. Mapped bam files for the rhesus macaque were retrieved from the synapsis repository (https://www.synapse.org/#!Synapse:syn17093056/tables/Rhesus mRNA-seq) using the repository API for UNIX, then transformed into fastq files using the *bedtools* (Quinlan & Hall, 2010) *bamtofastq* function. Raw fastq files generated in the present work and sequenced at BGI Genomics were retrieved from a dedicated AMAZON Web services account. Raw fastq files from all sources were then processed with *fastp* (Chen, Zhou, Chen, & Gu, 2018) to remove read adapters and to check read quality before and after trimming. Trimmed fastq files were mapped to the reference genome using *STAR* (Dobin et al., 2013) version 2.5.4b using the

following parameters "--outReadsUnmapped Fastx --chimSegmentMin 12 --chimJunctionOverhangMin 12 --alignSJDBoverhangMin 10 --alignMatesGapMax 100000 --alignIntronMax 100000 --chimSegmentReadGapMax 3 --alignSJstitchMismatchNmax 5 -1 5 5 --runThreadN 94 --outSAMstrandField intronMotif --outFilterMultimapNmax 20 --outFilterType BySJout --outFilterMismatchNoverReadLmax 0.04 --alignIntronMin 20 --outSAMtype BAM Unsorted". The latest reference genomes available: are hg38, rheMac10 (Warren et al., 2020), mm39 (Church et al., 2011), and galGal6 (Bellott et al., 2017; Steffen, Petti, Aach, D'Haeseleer, & Church, 2002). They were downloaded from the USCS Genome browser and used for humans, rhesus macaque, mice, and chickens, respectively. The resulting unsorted BAM files were sorted using *samtools* (H. Li et al., 2009).

3.3 Iso-seq long reads processing

For the rhesus macaque, raw unmapped bam files from the SRA project PRJNA476474 were downloaded directly from the SRA repository using GNU *wget*. Standard *Isoseq3* pipeline was used to obtain polished, high-quality fasta files for all processed samples. Additionally, for the chicken (*Gallus gallus*), fasta files from Iso-seq sequencing deposited at SRA were downloaded using *fasterq-dump*, as described above. Detailed information on all libraries used can be found in supplementary table 1.

Long reads fasta files were mapped to the reference genomes using *Minimap2* (H. Li, 2018) with the following parameters "-ax splice -uf --secondary=no -C5 -O6,24 -B4". Output sam files were converted to bam files, sorted, and indexed using *samtools*. All output sam files from the same species were collapsed into a gtf file using the function *collapse_isoforms_by_sam.py* from *Cupcake* with default parameters. Spurious transcripts were removed from the collapsed gtf file using the functions *sqanti3_qc.py and sqanti3_RulesFilter.py* from *Squnti3* (Tardaguila et al., 2018) with default parameters.

3.4 Bulk RNA-seq quantification and differential expression analysis

Gene expression was quantified by *FeatureCounts* from the *Rsubread* package (Liao, Smyth, & Shi, 2019) using the new assemblies for each assessed species as a reference with the following parameters "allowMultiOverlap = T, countMultiMappingReads = F, juncCounts = T, nthreads = 96". The raw expression matrix was batch corrected for humans using ComBat-seq (Zhang, Parmigiani, & Johnson, 2020) as it was done in the original PsychEncode publication (M. Li et al., 2018). TPM values were calculated from raw expression matrices, as previously shown by (Zhao et al., 2021):

$$TPMi = \frac{qi/li}{\sum_j (qj/lj)} * 10^6$$

Where *TPMi* is the TPM value of *gene i*, *qi* is the number of reads mapped in *gene i*, *li* is the length in kilobases of *gene i*, and $\sum_j (qj/lj)$ is the sum of counts/length ratios of all genes.

Only genes with a TPM value greater than 0.5 for all samples from developmental window/brain region pairs were kept for the subsequent analyses. Filtered TPM matrices were normalized using variance stabilization normalization (Huber, von Heydebreck, Sültmann, Poustka, & Vingron, 2002).

To identify differentially expressed genes (DEG), the R packages edgeR (McCarthy, Chen, & Smyth, 2012) was used. Briefly, lowly-expressed genes (less than 0.5 CPM in all samples of a variable) from raw count matrices were removed. Filtered matrices were used to identify DEG using the quasi-likelihood test; resulting P values were FDR corrected, and all genes with an FDR less than 0.05 were identified as DEGs.

3.5 Single-cell RNA-seq processing

Fastq files were retrieved from SRA using fasterq-dump as described above but with the following parameters “fasterq-dump -S -O /output/dir -e 94 --include-technical”. Fastq files were then mapped to the reference genome using *STARsolo* (Kaminow, Yunusov, & Dobin, 2021) version 2.7.9a with the following parameters “--soloType CB_UMI_Simple --soloCBwhitelist /barcodes/dir --soloBarcodeMate 0 --soloBarcodeReadLength 0 --soloCBstart 1 --soloCBlen 8 --soloUMIstart 9 --soloUMIlen 8 --readFilesCommand zcat --runThreadN 94 --soloStrand Forward --clipAdapterType CellRanger4 --readFilesIn READ1.fq READ2.fq --soloCellFilter None --soloFeatures Gene Velocyto GeneFull --soloMultiMappers PropUnique ”. Raw, sparse matrices from all samples were loaded into R (Team, 2018) and merged using *Seurat* (Hafemeister & Satija, 2019). Cells with less than ten thousand and more than one million UMIs, with less than a thousand detected genes, and with more than 5% of all counts mapped to mitochondrial genes were removed from further analysis. Raw, sparse expression matrix was normalized using SCT transformation while regressing by the percentage of expressed mitochondrial genes, cellular-cycle score, number of UMIs, and the number of identified genes using the following parameters “method = "glmGamPoi", vst.flavor = "v2", variable.features.n = 5000, vars.to.regress = c("percent.mt", "CC.Difference", "nCount_RNA", "nFeature_RNA")”.

Single-cell clusters were identified using the Seurat function FindNeighbors considering the first fifty dimensions and the function FindClusters with resolution *2.5*. Markers for all clusters were identified using the functions FindAllMarkers with the following parameters “assay = "SCT", test.use = "wilcox", only.pos = T, logfc.threshold = 0.25, min.pct = .25, return.thresh = 0.05, densify = T”. Known cell population gene markers in the literature, which are cited in the Introduction above, were used to annotate the identified clusters to different cortical cell-types.

3.6 GTF building

To generate consensus gene models from short-reads, sorted bam files from bulk RNA-seq libraries were processed using *scallop* (Shao & Kingsford, 2017) with the following parameters “--min_transcript_length_base 200 --min_mapping_quality 250 --min_splice_bundary_hits 1”. To choose the correct parameter for “--library_type” the type of library was assessed before running *scallop*; for the parameter “--min_num_hits_in_bundle” 10 was chosen if the library possesses less than 20 million uniquely mapped reads. Otherwise, 20 was used. After generating gtf files for every sample, the monoexonic transcripts were removed from unstranded libraries. Additionally, the monoexonic transcripts were removed from rRNA-depleted libraries using *gffread* (Pertea & Pertea, 2020) with the following parameter “gffread in_gtf_file -F -U -T -o /out_put/file”.

GTF files from all samples of the same developmental window/brain region were merged into a consensus transcriptome using *taco* (Niknafs, Pandian, Iyer, Chinnaiyan, & Iyer, 2017) before generating the final gtf files, to avoid overrepresentation of libraries of a tissue/brain region group, which would bias the construction of the final transcript model, with the following parameters “--gtf-expr-attr RPKM --filter-min-expr 0 --filter-min-length 200 --isoform-frac 0.01”. Consensus transcriptomes were merged into a final gtf file using *taco* with similar parameters. The output consensus file was filtered for readthrough, mapping errors, intron-retention, and run-on polymerase transcripts using *gffcompare* (Pertea & Pertea, 2020) with the species reference transcriptome as a model to generate the final consensus gtf file.

3.7 Coding potential identification

Coding potential was assessed for all transcripts in the final gtf files using *CPAT3* (Wang et al., 2013), *FEELnc* (Wucher et al., 2017), and *CPC2* (Kang et al., 2017). For CPAT and CPC2, gtf files were transformed into fasta files, first generating intermediate bed files with *gffread*, then using *getfasta* from *samtools* with the following parameters “-split -name -s”. The

fasta files were used as input to generate coding potential values for each transcript. For FEELnc, the reference gtf files for each evaluated species were used to train the random forest model, running the tool with the following parameters "-n 6000,6000 --learnorftype=3 --testorftype=3". Tables with output coding potential scores can be found in supplementary table 2.

3.8 Open reading frame identification and annotation

Transdecoder 5.5.0 (Douglas, 2018) was used to identify *bona fide* ORFs; additional Pfam and Uniprot matches were provided to improve the identification of ORFs. The *HMMER* 3.1b2 tool (Mistry, Finn, Eddy, Bateman, & Punta, 2013) was used to identify Pfam matches with the following parameters "hmmscan --domtblout pfam.domtblout --tblout file_name.tsv -E 1e-5". To identify Uniprot matches, *blastp* was used with the following parameters "-max_target_seqs 1 -outfmt 6 -evalue 1e-5". Final identification of ORFs was carried out using the function *TransDecoder.Predict* with the following parameters "--retain_pfam_hits pfam.domtblout --retain_blastp_hits blastp.outfmt6". Additionally, gff3 files were generated for each species containing the genomic coordinates of the ORFs, information that was added to the final consensus transcriptomes. Finally, eggNOG-mapper (Huerta-Cepas et al., 2017) matches were identified for all identified ORFs using the UNIX standalone tool *emapper* 2.0.1 with default parameters. Output annotations of identified ORFs are found in the supplementary table 2.

3.9 Transposable element content identification

Repetitive elements from each studied species were downloaded from the UCSC Genome Browser database (https://genome.ucsc.edu), keeping only the records from Transposable elements (TE) families SINE, LINE, LTR, DNA, Retroposon, and RC. TE tables were converted to bed files using custom *Rscripts* and sorted using UNIX *sort*.

For identifying ORFs with more than 50% of their gene body coming from a TE element, protein-coding sequence (CDS) bed files were intersected with TE bed files using *bedtools intersect* function with the following parameters "-s -wo" to ensure strand-specificity of the intersection. The total sum of TEs intersection was divided by the length of the CDS; CDS with more than 50% of their gene body coming from a TE element were tagged as "Transposable element".

To identify the TE class distribution in CDS, mRNA untranslated regions (UTRs), pseudogenes, and lncRNAs bed files from the new assemblies of each species were intersected with the TE bed file. TEs that intersected at least ten bp with a gene were kept for further analysis. The percentage of the gene body containing TEs was calculated as the ratio of the total length of all TEs intersecting with a gene to the gene length.

3.10 Identification of syntenic lncRNAs

To identify syntenic conserved lncRNAs between studied species, two approaches were used. Whole genome alignment and long transcript mapping. In both cases, all isoforms from a gene were merged into a metagene annotation generating new bed files of metagenes. Then, sequences matching with transposable elements were removed from the metagene coordinates using *bedtools subtract*. Fasta files for metagene annotations without TEs were generated using *bedtools getfasta,* as shown before. TEs were removed because their repetitive nature may complicate the mapping processes.

In the whole genome alignment approach (*liftover*), metagene bed files were lifted to the other species genome coordinates using the standalone *liftover* function from UCSC Genome Browser with the following parameters "-minBlocks=0.01 -minMatch=0.01".

In the transcript mapping approach, metagene fasta files were mapped to the genome of the other species using *Minimap2* with the following parameters "-ax splice -uf --secondary=no -C5 -O4,24 -A2 -B4 -G 100K". The output sam files were converted to bam sorted files using *samtools* and then to bed files using *bedtools bamtobed*.

Bed files containing mapped lncRNAs from one species to the other genome were joined, removing all transferred genes from the transcript mapping approach if they were transferred using the whole genome alignment approach. The final set of transferred genes coordinates was used to identify the syntenic lncRNAs, as described in the results sections and shown in Figures 7A, 7B, and 7C.

TEs enormously contribute to the gene body of lncRNAs (Kapusta et al., 2013); removing TEs from the gene body of lncRNAs might misidentify some syntenic conservation. So, the same approaches were undertaken but without removing the TE insertions from the lncRNA body. The syntenic information was added when it identified a strong syntenic homologous in a more distant species.

3.11 Identification of expression-matched genes

The R package "optmatch" (Hansen & Klopfer, 2006) was used to identify the set of expression-matched genes among the evaluated gene categories, using the mean variance stabilized expression of all samples from the same gene as input.

3.12 Identification of closest genes to lncRNAs

The *bedtools* function *closest* was used to identify the most proximal genes with the following parameters "-d", and for that, a metagene bed file for protein-coding genes and small RNAs was built for each species. Only genes located in the genome around 100 kb of lncRNAs were kept for the following analysis. Gene ontology (GO) analysis of the closest protein-coding genes was undertaken using the R package clusterProfiler (Wu et al., 2021; Yu, Wang, Han, & He, 2012), and using as input the list of closest protein-coding genes to a minimal age (MA) category and the list of all closest protein-coding genes as background.

3.13 ATAC-seq processing

To identify the open chromatin regions (OCR) of cortical tissues at mid-gestation stages, ATAC-seq libraries from the dbGaP project phs001438 were downloaded using *fasterq-dump*. Adapters were removed using *fastp* resulting in trimmed fastq files that were mapped to the human genome hg38 using *bwa mem* version 0.7.17-r1198-dirty (H. Li, 2013). Output sam files were further processed using *PicardTools* version 2.18.23-SNAPSHOT ("Picard toolkit," 2019). First, sam files were sorted using the function *SortSam*, then deduplicated using the function *MarkDuplicatesWithMateCigar*. Output deduplicated bam files were sorted using *samtools sort*. Peaks were called for germinative zone and cortical zone samples, separately, using *Genrich* version 0.6.1 (Gaspar, 2018) with the following parameters "-j -y -r -e chrM". Peaks identified overlapping the exclusion list from ENCODE (https://www.encodeproject.org/files/ENCFF356LFX/) were removed from further analysis.

3.14 Identification of promoter features

Chromatin modification peaks H3K27ac, H3K4me1, and H4K20me3 of cortical tissues at mid-gestation stages were downloaded from the GEO project GSE149268. For every gene, the promoter (2 kb downstream and 1 kb upstream the TSS of a transcript) from the longest isoform with the most significant number of exons was retrieved and used in the following analysis. Additionally, a similar number of non-redundant random regions that did not overlap with the identified promoters of the same length (3 kb) were produced using *bedtools shuffle* for enrichment analysis.

Promoters from expression-matched genes from all assessed categories were retrieved, and a set of an equal number of random genomic regions for further processing. OCRs, H3K27ac, H3K4me1, and H4K20me3 coordinates were intersected with the set of working promoters and random regions using *bedtools intersect*, and the number of chromatin features intersected for each gene category was assessed. Fisher hypergeometric test was performed between the set of random sequences and the promoters from a gene category. P values were FDR corrected, and all features with an FDR less than 0.05 were identified as enriched if the odds ratio is greater than one and depleted when the odds ratio is less than one.

Non-redundant remap 2022 data (Hammal, de Langen, Bergon, Lopez, & Ballester, 2022) were retrieved and intersected with the set of working promoters and random regions using *bedtools intersect*. The absolute number of different proteins bound to each gene category and random genomic regions was assessed and compared. Significant differences in the distribution of the gene categories were assessed using Wilcoxon test.

3.15 lncRNA expression dynamics modeling

The R package ggplot2 (Wickham, 2016) was used to plot the expression dynamics of the gene categories throughout human corticogenesis, using variance stabilized TPM as input values for the function *geom_smooth*.

3.16 Statistical Analysis

All statistical plots and tests were obtained using the statistical package R version 4.1.0 (Team, 2018).

4 Results

4.1 Development of bioinformatic pipelines to assemble new transcriptomes help to improve the annotation of lncRNAs

It is possible that lncRNAs from scarce cell types of the brain or lncRNAs expressed at a low level might not have been annotated yet, as it has been reported that neural tissues are among those that express the most tissue-specific lncRNAs (Hezroni et al., 2015; A. Necsulea et al., 2014; Sarropoulos, Marin, Cardoso-Moreira, & Kaessmann, 2019) . To avoid misidentifying the syntenic conservation of lncRNAs because of the differences in completeness of the transcriptome annotation among the studied species, it is necessary to use their most comprehensive set of annotated lncRNA genes. Several publications have annotated long non-coding RNAs in an extensive collection of species, based on short-reads RNA-seq libraries (Hezroni et al., 2015; A. Necsulea et al., 2014; Sarropoulos et al., 2019). These reconstructions of transcript models require mapping the short reads to a reference genome, the inference of the transcript model in the set of mapped reads, and finally, the generation of consensus transcriptomes based on combining different transcript models.

Following those guidelines, a bioinformatic pipeline was built for the present work using state-of-the-art bioinformatic tools (Figure 2A). First, for the adapter removal step, the fast *fastp* (Chen et al., 2018) tool was chosen; *fastp* additionally performs read quality assessment of raw and trimmed fastq files in parallel. Then, *STAR* (Dobin et al., 2013) was selected to map the trimmed fastq files to the reference species genomes using the two-pass mode with the ENCODE parameters to increase the number of identified splicing sites. *Scallop* was chosen to build transcript models for each mapped library because it shows a greater accuracy in reconstructing multiexon transcripts than similar tools (Shao & Kingsford, 2017). Finally, *TACO* was selected to generate a consensus transcriptome for each library set, as it outperforms similar tools in detecting the start and end of transcriptional sites (Niknafs et al., 2017). As the core tools of the pipeline are the *STAR* mapper, the transcriptome modeler *Scallop* and the consensus transcriptome builder *TACO*, the pipeline was named SST. As proof of the reliability of the SST pipeline, new sets of lncRNAs for the parasite *Schistosoma mansoni* and *Schistosoma japonicum* were built using the same set of bioinformatic tools (L. F. Maciel et al., 2019; Lucas F. Maciel, Morales-Vicente, & Verjovski-Almeida, 2020).

In recent years, third-generation high-throughput sequencers have been used to improve the quality of genomes and annotated genes of different species, including humans (Leung et

al., 2021; Nurk et al., 2022) , mice (Leung et al., 2021), macaques (Warren et al., 2020), and chickens (Kuo et al., 2017). The Iso-seq technology from PacBio, which generates long reads, has been used to identify new genes and improve transcriptome annotation in complex genomic regions that could not be solved in assemblies based on short-read technologies, especially for those species with poorly annotated transcriptomes. To help improve the quality of the reference transcriptomes of chicken and macaque that have poorly annotated reference transcriptomes, we used public Iso-seq libraries from these two species (Kuo et al., 2017; Warren et al., 2020) following a second pipeline (Figure 2B). For the macaque, unmapped bam files were downloaded, and the *ISOSEQ3* pipeline was used to obtain high-quality, full-length transcripts in fasta format; for the chicken, fasta files were downloaded from SRA. Fasta files were mapped to reference genomes using *Minimap2* (H. Li, 2018), concatenated with *Cupcake,* and spurious transcripts were filtered using *SQANTI3* (Tardaguila et al., 2018), thus generating full-length transcript models that were merged with the transcriptome from Ensembl to create new reference transcriptomes for both species.

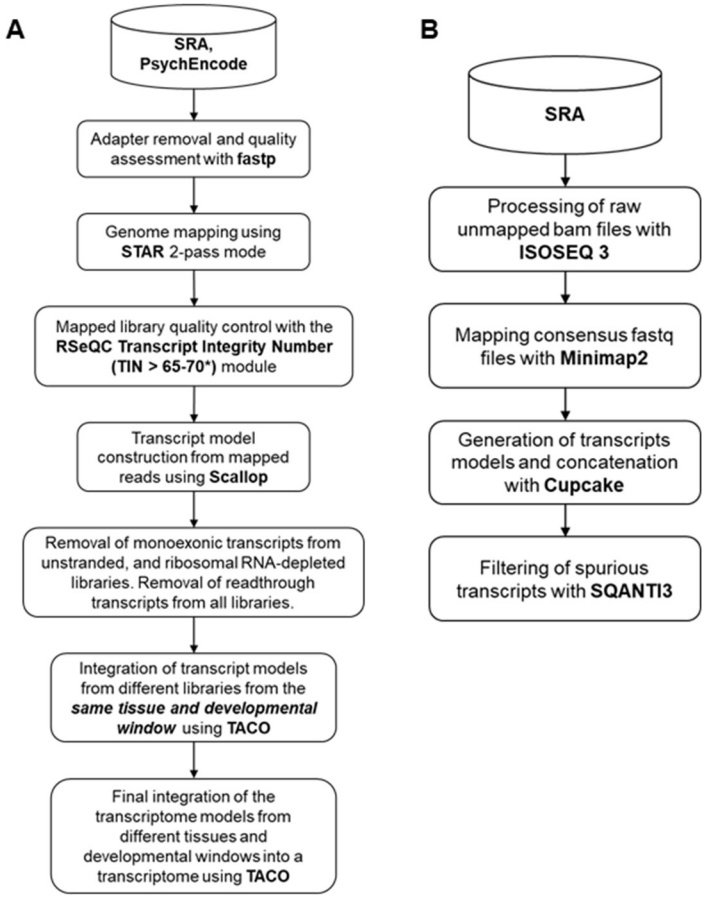

Figure 2. Depiction of computational pipelines used for assembling new transcriptomes. A. SST pipeline was developed to assemble new transcriptomes based on short reads. The pipeline uses *STAR* to map short reads,

Scallop to build transcriptional models for each library, and *TACO* to generate consensus transcriptomes based on a set of transcriptional models. **B.** Bioinformatic pipeline to build new transcriptomes based on Iso-seq long reads.

In addition to the core steps used in the SST pipeline, control-quality checks in the building of new transcriptomes from short reads were introduced to remove libraries with sequencing bias towards the 3' end, as they might not be able to reconstruct the entire length of new transcripts. To remove samples with the 3' bias, the gene body coverage and the TIN median score were calculated using *RSeQC* (Wang et al., 2016; Wang, Wang, & Li, 2012), and the relationship between the TIN median and the gene body coverages of libraries was inferred by visualizing the distribution of reads alongside the transcript body and their associated TIN median score (Figure 3 A-D).

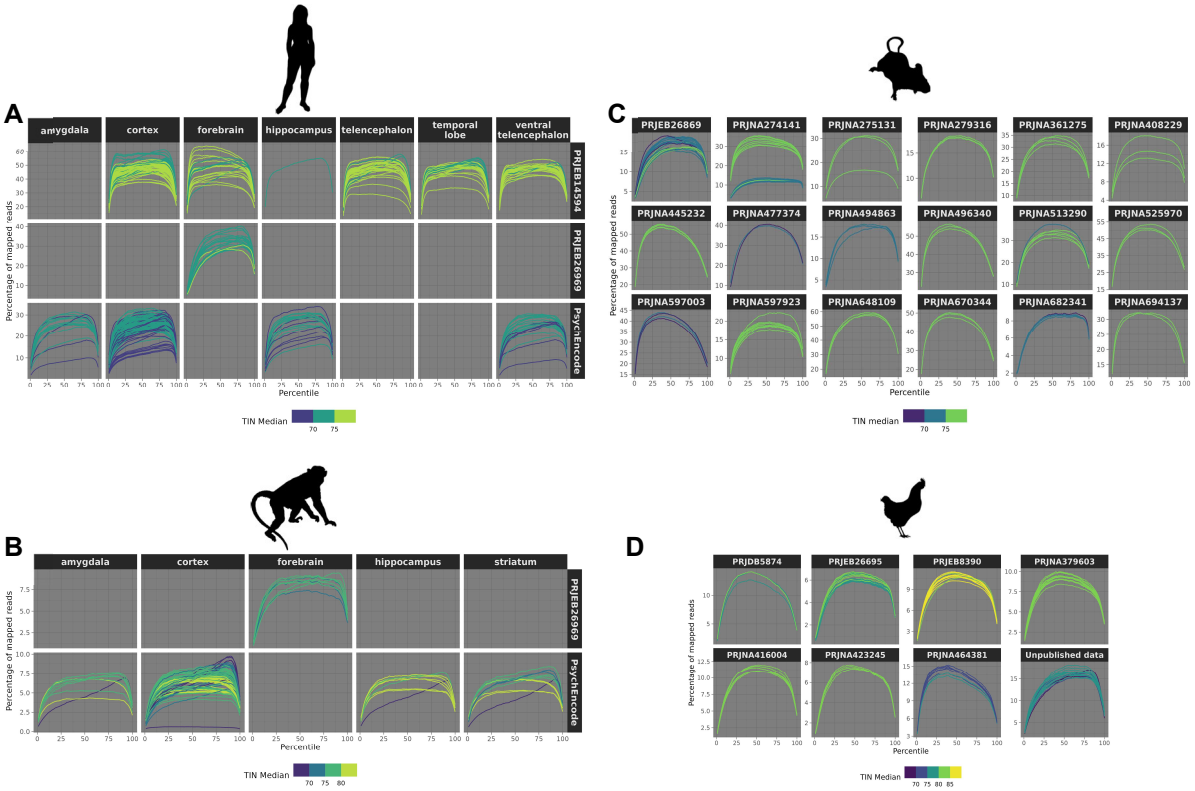

Figure 3. Mapping quality of samples used to generate new transcriptome assemblies. A. Averaged distribution of short read percentage along all transcript bodies in each library and colored depending on the TIN median score. Library samples were separated by the public project and the tissues of origin. **B.** Like A but in the rhesus macaque. **C.** Like A but in mice, and libraries were separated only by the origin of public data. **D.** Like A but in chicken, and libraries were separated only by the origin of public data.

Empirically, the threshold of the TIN median value was set at 65 for libraries from the PsychEncode project and at 70 for other libraries from humans and macaques, and a threshold TIN median value of 65 was set for all the libraries from mice and chickens. These thresholds made it possible to keep a high number of libraries with acceptable good quality (Figure 4 A-

D). Additionally, readthrough transcripts that expanded two different coding or mixed coding-non-coding loci were removed because they might complicate the large-scale annotation of lncRNAs at further steps, and monoexonic transcripts from unstranded libraries or rRNA-depleted libraries were removed because they might be spurious transcripts that arise from contamination of rRNA or degraded introns.

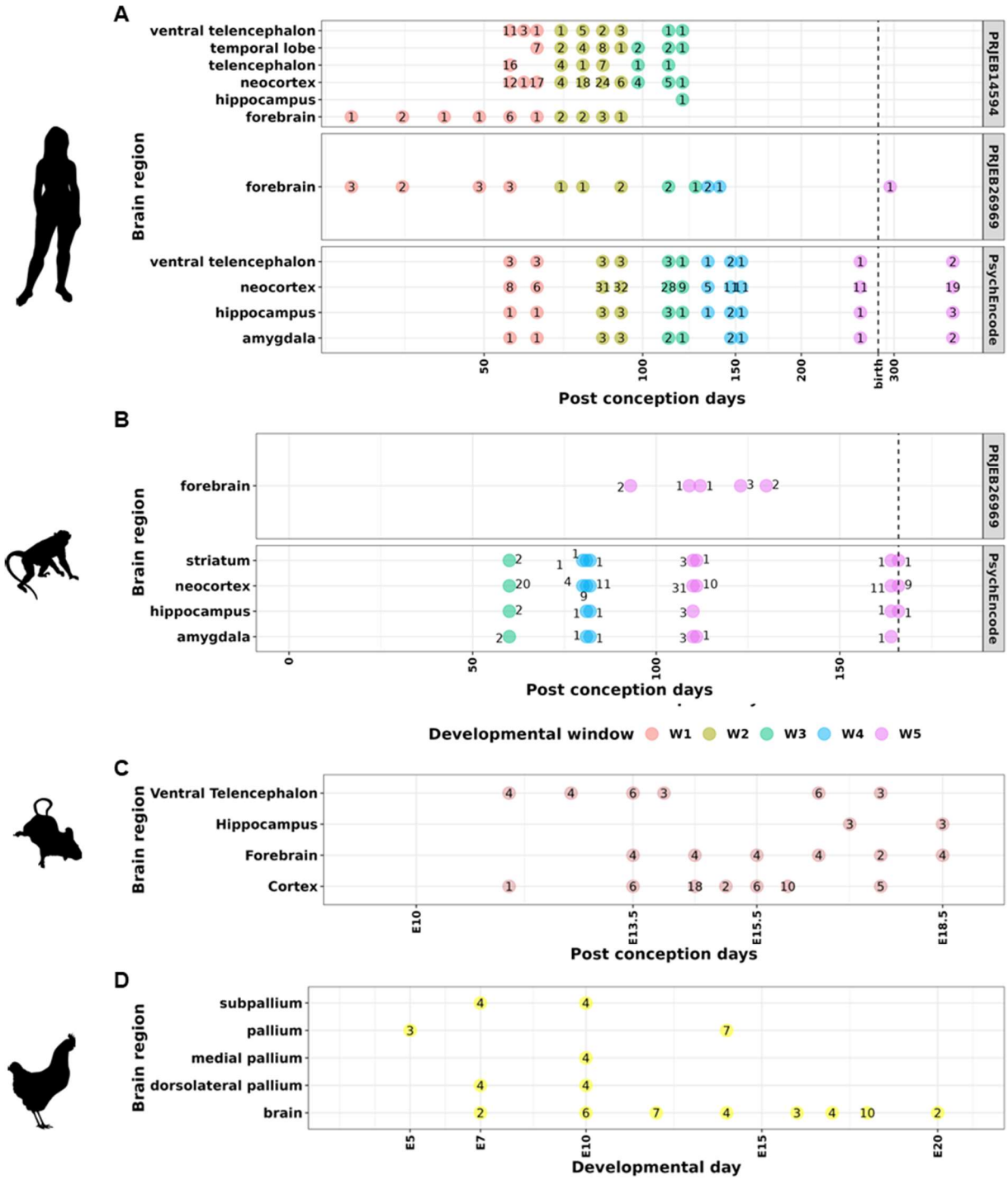

Figure 4. The number of samples used for annotating new transcriptome assemblies. A. Number of samples used for building the SST assembly of humans, grouped by the developmental time, tissue of origin, and the public data source; and colored by the PsychEncode developmental window. Developmental window W1, post-conception weeks 8 and 9; W2, post-conception weeks 12 and 13; W3, post-conception weeks 16 and 17; W4, post-conception weeks 19, 21 and 22; W5, post-conception week 37 and post-natal day 100. **B.** Like A, but for macaques. The PsychEncode developmental windows of macaque represent different developmental days but reflect similar developmental stages to humans. **C.** Like A, but in mice, where there is no identification of matched developmental stages to humans, and samples span the cortical proliferative, early, and late neurogenesis stages. **D.** Like A, but in chickens, where there is no identification of matched developmental stages to humans, and samples span the pallial proliferative, early and late neurogenesis, and gliogenic stages.

To produce the most comprehensive collection of cortical lncRNAs, it is essential to use a broad set of libraries from different regions of the developing brain. After filtering, a group of libraries that encompasses the neural progenitor pool expansion and neurogenic phases of cortical and pallial development was obtained for each species (Figure 4A-D). Additionally, samples from gliogenic stages of pallial development were obtained for all species except for the mouse (Figure 4A, 4B, and 4D), as gliogenesis starts at post-natal days in mice (Libé-Philippot & Vanderhaeghen, 2021). For humans and rhesus macaque, the PsychEncode (M. Li et al., 2018; Zhu et al., 2018) library sets contain information on developmental-matched windows (Figure 4A and 4B). Those developmental windows roughly correspond to the cortical pool expansion phase W1, early and late neurogenesis phase, W2-W3, W3-W4, respectively, and gliogenic period W5. For chickens, the developmental days E5, E7, and E10 correspond roughly to the progenitor pool expansion, the early neurogenic phase, and the late neurogenic stage, respectively. Meanwhile, the developmental age E14 of chicken contains gliogenic radial glia. The list of accession numbers of all libraries is shown in Supplementary Table 1.

New transcriptome assemblies were built for each species using the collected sets of libraries that expand the corticogenesis phases and the SST pipeline. Additional filters were applied to the raw assemblies to remove spurious transcripts that might arise from bioinformatic artifacts. First, the remnants of readthrough transcripts were identified and removed; then, the transcripts that likely originated from mapping errors, intron-retention, and polymerase run-on were identified using *GffCompare* (Pertea & Pertea, 2020) and removed from further analysis. After filtering, lncRNA genes were identified by combining two approaches: identifying annotated ORFs and calculating for each transcript the coding potential using three different tools. For the robust identification of ORFs, the *Transdecoder* (Douglas, 2018) tool was used. The ORFs were annotated using *eggNOG-mapper* (Huerta-Cepas et al., 2017) and *PFAM* databanks (El-Gebali et al., 2018). *CPAT* (Wang et al., 2013), *CPC2* (Kang et al., 2017), and *FEELnc* (Wucher et al., 2017) were used to calculate the coding potential of transcripts. Each

of them uses different machine learning approaches to predict the coding potential of transcripts based on the set of coding mRNAs and lncRNAs of reference transcriptomes.

Non-coding transcripts were identified as those that do not contain an annotated ORF, and at least two of the three coding potential calculators did not classify the transcripts as coding (Figure 5). When an annotated ORF had at least 50% of its gene body inside a TEs, the transcript was annotated as a "transposable element." When two out of the three coding potential calculators classified the transcript as coding or at least one classifies it as coding and has an annotated ORF, the transcript is annotated as coding. Otherwise, the transcript was classified as a transcript of unknown coding potential (TUCP). Genes with all their transcripts being a non-coding or transposable element in humans and mice or only non-coding in macaque and chicken were classified as lncRNA genes. Genes with at least one coding transcript were classified as coding genes. In chicken and macaque, genes with all transcripts annotated as transposable elements were classified as TE; meanwhile, in humans and mice, where the set of protein-coding genes is thoroughly annotated, TE genes were included in the group of lncRNAs, as it is known that TEs are a source of lncRNA evolution. Genes with a least one TUCP and no coding transcripts were classified as genes of unknown coding potential (GUCP). Additional manual curation for chicken and macaques was implemented to classify GUCP into coding, lncRNAs, or pseudogenes, when possible. Finally, to generate a comprehensive catalog of lncRNAs of each species, lncRNAs from public lncRNA databanks (Hezroni et al., 2015; Sarropoulos et al., 2019) that contain syntenic evolution information of lncRNAs for amniotes and that were not annotated in the reference and SST assemblies were incorporated into the final transcriptome (Figure 5).

Across all the species, lncRNAs represent the category with more genes (Figure 6 A-D), corroborating the widespread expression of lncRNAs in vertebrates (A. Necsulea et al., 2014). A large fraction of the annotated lncRNAs come from the SST pipeline in all species (Figure 6 E-H). These new transcriptomes also significantly improve the number of reads mapped to annotated features (Figure 6 I-M). In addition to contributing to the annotation of new lncRNAs, the implemented approach was able to identify new protein-coding genes, pseudogenes, and genes carrying TE-derived ORFs (grouped in figure 6 into the "other" gene types), showing the robustness of our approach to identifying coding and lncRNAs across different organisms, especially for species with poorly annotated transcriptomes, such as the chicken and rhesus macaque.

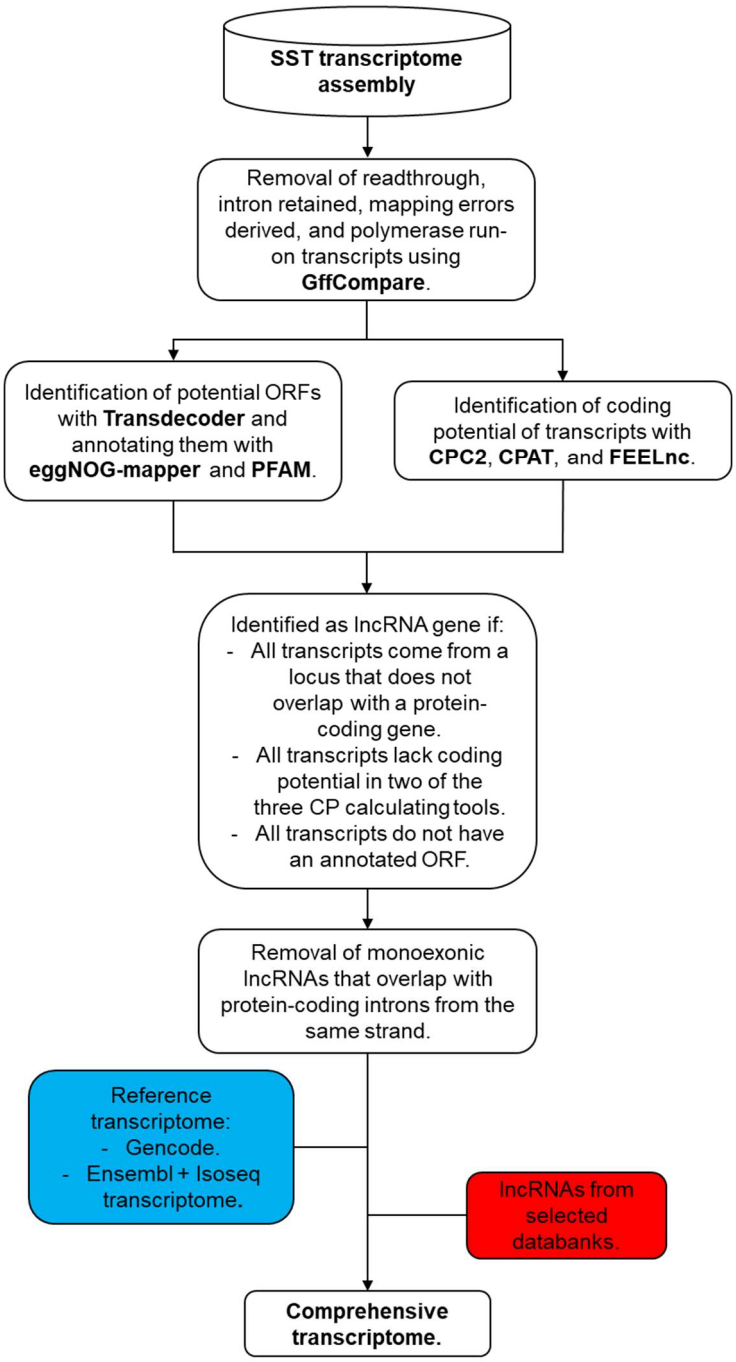

Figure 5. Scheme of the bioinformatic pipeline used for the annotation of lncRNAs. Raw assembled transcriptomes built using the SST pipeline underwent extensive filters, first to remove spurious transcripts; second to identify lncRNA genes and separate them from protein-coding isoforms. Additionally, transcripts from other public databases and the lncRNA set of reference transcriptomes were added to the final comprehensive transcriptome.

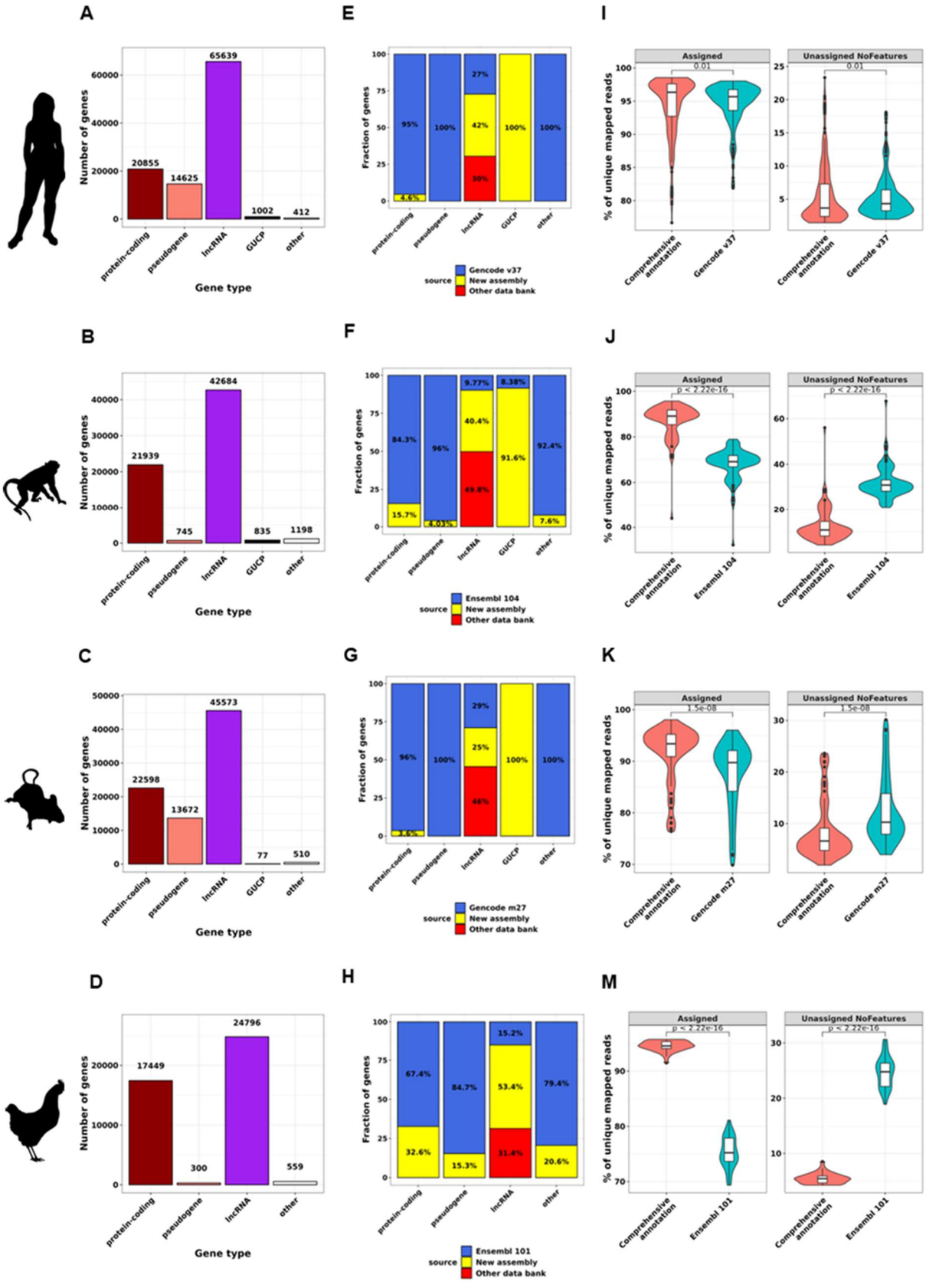

Figure 6. New comprehensive transcriptome assemblies improve the annotation of lncRNAs. A-D. Distribution of gene types in the new comprehensive transcriptomes annotated for humans, macaques, mice, and chickens, respectively. **E-H.** Percentage of genes from different sources across the different gene types for humans,

macaques, mice, and chickens, respectively. **I-M.** Percentage of uniquely mapped reads mapped to an annotated and unannotated region for humans, macaques, mice, and chickens, respectively. **Statistics:** All statistics are Wilcoxon-test.

4.2 Identification of *bona fide* minimal evolutionary age of human cortical lncRNAs based on syntenic conservation

To identify patterns of gene function and specialization of the lncRNAs throughout the evolution of the cerebral cortex, a methodology was developed to systematically classify lncRNAs into evolutionary ancestry groups based on syntenic conservation of lncRNAs among humans, rhesus macaque, mice, and chickens (Figure 7A and 7B).

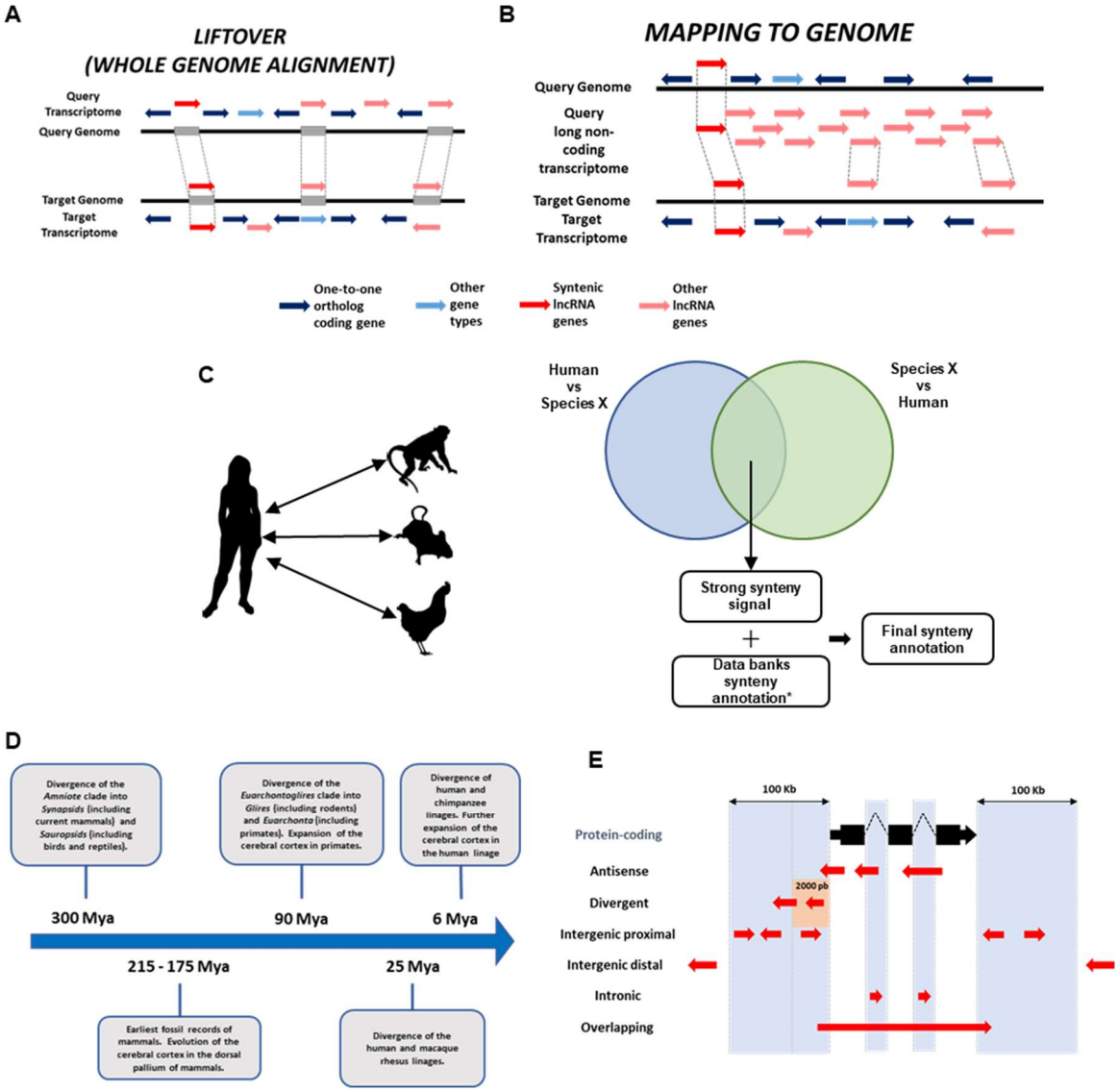

Figure 7. Identification of minimal evolutionary age and positional classification of lncRNAs. A. Depiction of the whole genome mapping approach used to identify syntenic lncRNAs between two species. **B.** Depiction of the genome mapping approach used for identifying syntenic lncRNAs between two species. **C.** Depiction of the final identification of synteny conservation of human lncRNAs. **D.** Landmarks of cerebral cortex evolution in the

human lineage. **E.** Depiction of the classification of lncRNAs based on their position regarding protein-coding genes.

The strategy takes advantage of two bioinformatic approaches; the first is based on genome-wide alignment between two species, followed by the transference of gene coordinates from one query species to the other reference species (liftover); the second approach makes use of long-read mappers to map the whole transcriptome set from one query species to the genome of the target species (genome mapping) (H. Li, 2018). After mapping the lncRNA gene set of the query species to the genome of the target species, the group of transferred lncRNA genes was intersected with the collection of lncRNAs of the target species considering the strand of genes. The nearest one-to-one orthologous of the intersected pairs of lncRNAs between species were assessed; lncRNA pairs surrounded by the same one-to-one orthologous genes and expressed from the same strand were identified as syntenic lncRNAs (Figures 7A and 7B). To reduce the chances of type I errors, it was required that the set of syntenic lncRNAs identified when using the human lncRNA transcriptome as query must be the same when using the other species as query (Figure 7C), thus allowing the annotation of strong syntenic conservation of lncRNAs between human and the other three species.

To improve the classification of human lncRNAs into minimal-evolutionary-age (MA) groups, the syntenic conservation data from lncRNAs of public databases were integrated into the identified syntenic conservation of this work. The human lncRNAs identified as syntenic to chicken, mice, and macaque lncRNAs were clustered into 300 million years ago (MYA), 90 MYA, and 25 MYA minimal-evolutionary-age groups, respectively. LncRNAs that did not share synteny with any species using our approach or the public databases were classified as Human-specific lncRNAs. LncRNAs that did not have strong syntenic signals and did not have syntenic conservation in a more distal species in public databases were grouped in the uncertain category. MA groups are clustered around landmarks of the cerebral cortex evolution (Figure 7D): before the evolution of the cerebral cortex from the dorsal pallium, before the expansion of the cerebral cortex in the primate linage, before the further expansion of the cerebral cortex in apes, and the current evolution of the cerebral cortex in humans, respectively. LncRNAs genes were also classified depending on their position concerning protein-coding genes. The different types of lncRNAs used in this work to typify them are depicted in Figure 7E.

After classifying the human long non-coding transcriptome into lncRNA types and MA groups, the set of lncRNAs expressed throughout embryonic and fetal corticogenesis was

assessed. A collection of 189 libraries from the neocortex at developmental windows W1-W5 from the PsychEncode was used (Table 1).

Table 1. After filtering for low-quality sequencing, the number of bulk RNA-seq libraries from the PsychEncode data set per each developmental window (W1 to W5) and brain region were used for identifying the set of cortical lncRNAs. Developmental window W1, post-conception weeks 8 and 9; W2, post-conception weeks 12 and 13; W3, post-conception weeks 16 and 17; W4, post-conception weeks 19, 21 and 22; W5, post-conception week 37 and post-natal day 100.

Brain Region	W1	W2	W3	W4	W5
dorsolateral prefrontal cortex	2	6	4	4	3
medial prefrontal cortex	2	6	4	4	4
orbital prefrontal cortex	2	6	2	3	4
ventrolateral prefrontal cortex	0	6	4	4	4
primary motor-somatosensory cortex	2	0	3	0	0
primary motor (m1) cortex	0	6	1	3	4
primary somatosensory (s1) cortex	0	6	0	3	4
occipital neocortex	2	0	0	0	0
primary visual (v1) cortex	0	6	4	3	4
parietal cortex	1	0	0	0	0
posterior inferior parietal cortex	0	6	4	3	4
primary auditory (a1) cortex	0	6	4	3	4
inferior temporal cortex	0	6	2	2	4
superior temporal cortex	0	4	4	3	4

LncRNAs being expressed (> 0.5 TPM) in all the samples from the same developmental window/cortex region were classified as cortical lncRNAs. Because the lowly-expressed lncRNAs in bulk RNA-seq libraries may be highly expressed in sparse cell populations in the developing cerebral cortex, as was the case of the GABAergic-specific lncRNA DLX6-AS (Liu et al., 2016), it might be the case that those lncRNAs were removed from the set of expressed genes when the expression cut-off was applied. To accurately characterize the expressed genes in the developing cerebral cortex, including those lowly-expressed cell-type-specific genes, we re-analyzed public single-cell RNA-seq data sets of developing cerebral cortex cell populations (Gordon et al., 2015; Zhong et al., 2018) to identify coding and lncRNAs expressed in these cells, which were clustered according to their expression patterns (Figure 8A). We identified differentially expressed genes (DEGs) as markers of the cell clusters (Figure 8B) and added the newly identified genes to the final 40312 expressed genes. The collection of expressed lncRNAs in the developing cerebral cortex was named cortical lncRNAs. Using this set of expressed genes, the 189 samples were clustered. The samples from the developing cerebral cortex clustered mainly by the developmental window, with a clear transition from prenatal windows W1-W4 to the W5, where the prenatal-post natal transition occurs (Figure 8C), as is shown in

the PsychEncode leading publication (M. Li et al., 2018), indicating the excellent quality of the transcriptional data.

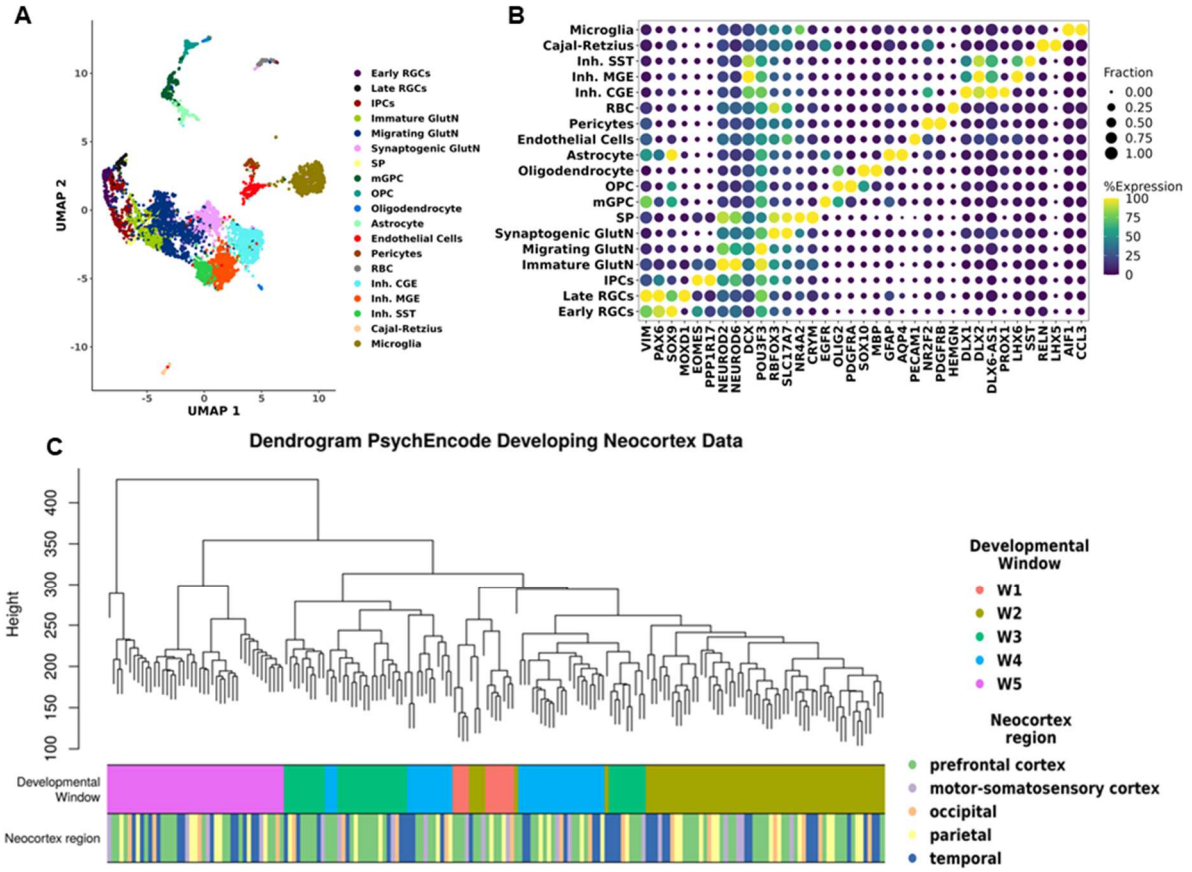

Figure 8. Bulk and single-cell RNA-seq datasets were employed to identify the human cortical lncRNAs. A. UMAP representation of the scRNA-seq data set of 5183 cells of the human cerebral cortex development from conception week 8 to conception week 26, which were clustered according to their gene expression patterns and labeled by cell type (colors). **B.** Dotplot shows the average expression of the cell population marker genes related to the cluster with the maximum expression and the fraction of cells from the cluster expressing the markers. **C.** Dendrogram of 187 bulk RNA-seq libraries grouped with unsupervised clustering according to their gene expression patterns; for information purposes, the developmental stage and the region of origin of each library is indicated at the bottom with color. **Abbreviations:** UMAP, Uniform Manifold Approximation, and Projection; early RGCs, early radial glial cells; late RGCs, late radial glial cells; IPCs, intermediate progenitor cells; Migrating GlutN, migrating pyramidal glutamatergic neurons; Maturing GlutN, maturing pyramidal glutamatergic neurons; Mature GlutN, synaptogenic prenatal pyramidal glutamatergic neurons; SP, subplate neurons; mGPC, multipotent glial progenitor cells; OPC, oligodendrocyte progenitor cells; RBC, red blood cells; Inh. CGE, inhibitory GABAergic interneurons derived from the caudal ganglionic eminences; Inh. MGE, inhibitory GABAergic interneurons derived from the medial ganglionic eminences; Inh. SST, inhibitory GABAergic interneurons expressing somatostatin.

Next, the phyloP score (K. S. Pollard, Hubisz, Rosenbloom, & Siepel, 2010) was used to test the conservation status among MA groups of cortical lncRNAs (Figure 9A). As expected, the conservation scores decreased throughout the evolution of cortical lncRNAs, indicating that

the implemented approach correctly classified lncRNAs into sequential evolutionary groups, with similar scores between the oldest MA group (300 Mya) and protein-coding UTRs.

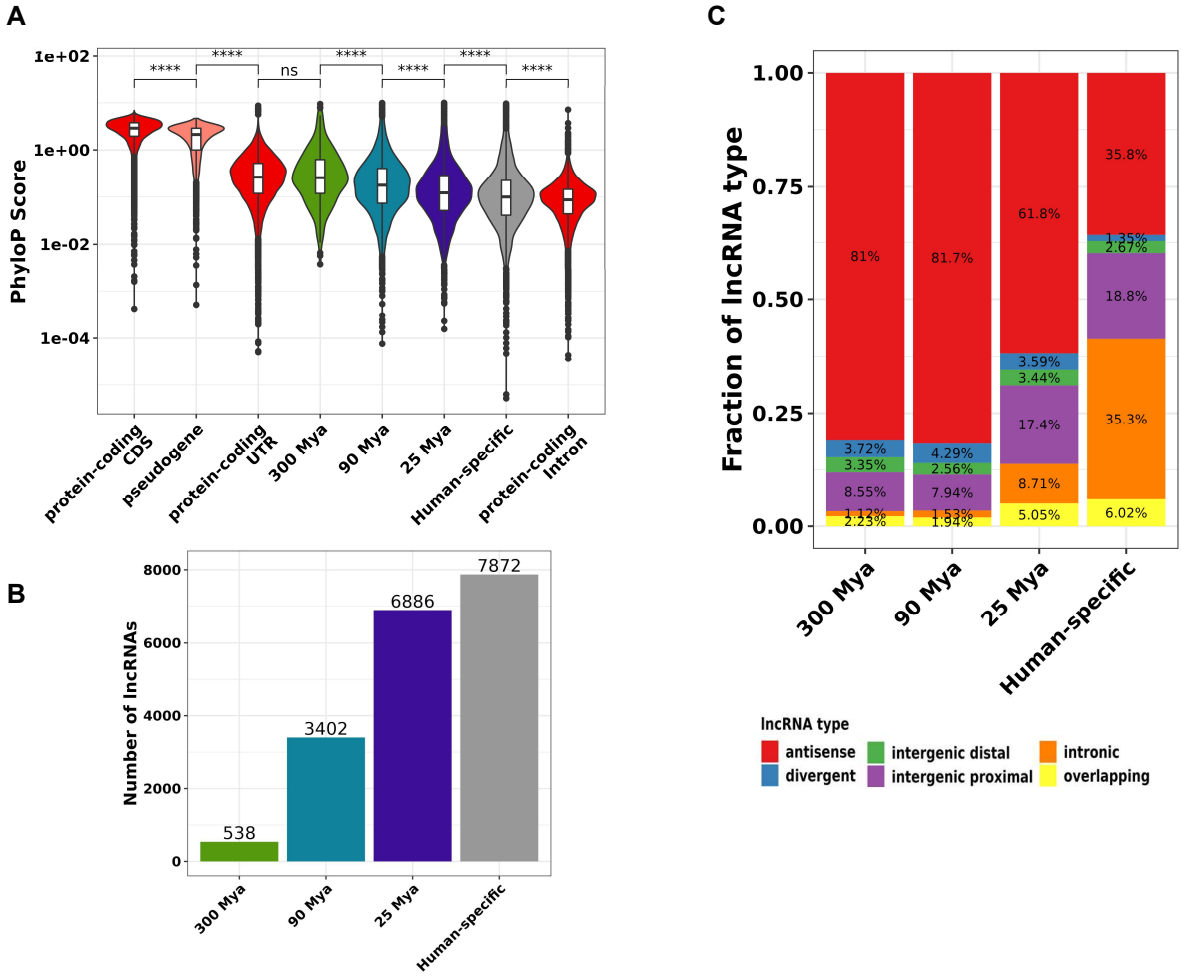

Figure 9. Syntenic classification of human cortical lncRNAs. A. Distribution of the lncRNA genes in each category of minimal-evolutionary-age (MA) groups. **B.** Distribution of lncRNA types among the MA groups. **C.** Mean PhyloP conservation scores of gene types and MA groups. **Statistics:** All statistics are one side (greater) Wilcoxon-test.

All MA groups generally showed higher conservation scores than protein-coding introns, indicating the positive selection of lncRNAs as a group is higher than expected by chance (Figure 9A). The number of cortical lncRNAs increases throughout the evolution of the human lineage, with only 2.88% of them identified as arising before the onset of the cerebral cortex and around 42% of them being specific to humans (Figure 9B). The distribution of lncRNA types also changes through evolution, with older lncRNAs being mostly antisense lncRNAs, while overlapping, intergenic proximal, and intronic lncRNA fractions increase in newer lncRNAs (Figure 9C). Significantly, the number of intronic lncRNAs explodes in the

human lineage, as 80% of the total intronic lncRNAs are human-specific and are depleted in older lncRNAs (Figure 9C).

4.3 Older lncRNAs have enhanced expression strength, splicing efficiency, and locus complexity

After proving the *bona fide* quality of the MA groups of cortical lncRNAs, differences in molecular features among them were assessed. Surprisingly, the MA groups form an expression gradient where older lncRNAs reach more robust expression levels than younger lncRNAs. Interestingly, the oldest group of cortical lncRNAs achieves similar expression levels to protein-coding genes (Figure 10A), which shows that the general lower expression of lncRNAs compared to protein-coding genes (Hezroni et al., 2015; A. Necsulea et al., 2014) is masked by the significant difference in gene expression levels between old and younger lncRNAs. These differences in expression levels were replicated when considering the lncRNA types, except for intronic lncRNAs that are depleted in the oldest MA group (Figures 9C and 10B), pointing to a negative selection in this type of lncRNAs in the oldest MA group.

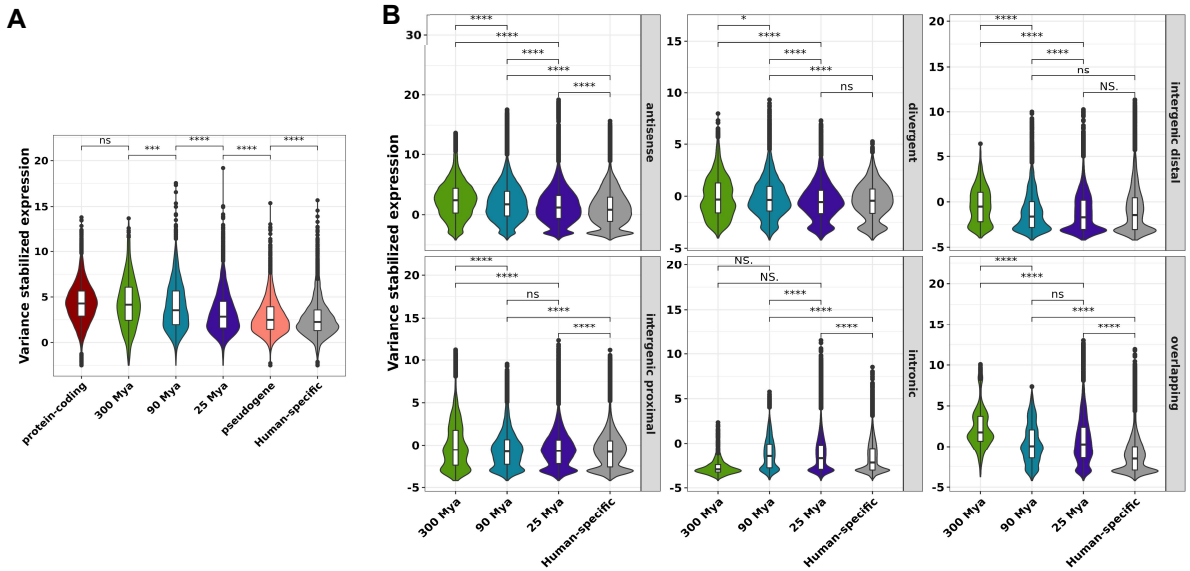

Figure 10. Expression level differences among human cortical lncRNA evolutionary groups. **A**. Expression level distribution of protein-coding, pseudogenes, and lncRNA MA groups. **B**. Expression level distribution of MA groups split by lncRNA type. **Statistics:** All statistics are one side (greater) Wilcoxon-test.

It has been shown that long intergenic non-coding RNAs (lincRNAs) are inefficiently transcribed by particular C-terminal-domain isoforms of Pol II, leading to reduced splicing and 3' UTR processing of lincRNAs (Schlackow et al., 2017); additionally, lncRNAs, in general, are shorter and have fewer exons than mRNAs (Hezroni et al., 2015). Previously documented differences between protein-coding and lncRNAs splicing efficiency might also be masked by the different MA groups, as is the case of gene expression levels. Therefore, differences in the splicing efficiency were assessed among the MA groups. The analyses were restricted to a set of expression and type-matched genes (Figures 11A and 11B), as differences in expression levels and lncRNA type may disguise the splicing differences among the MA groups.

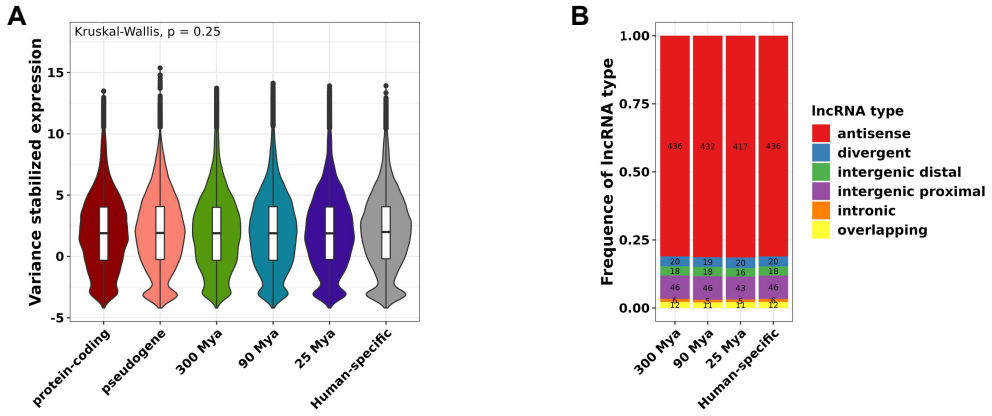

Figure 11. Set of expression and lncRNA type-matched cortical genes. A. Gene expression distribution of protein-coding genes, pseudogenes, and cortical lncRNA MA groups. **B.** Frequency of lncRNA types among the lncRNA MA groups.

The following splicing features were evaluated: number of exons, exon lengths, intron lengths, and splicing motif frequency. Like the different gene expression levels previously observed, MA groups form a gradient where older lncRNA populations have significantly more exons on average than the younger ones (Figure 12A). Still, differently to expression levels, protein-coding genes present a significantly higher number of exons than all lncRNA MA groups (Figure 12A). Simultaneously, older lncRNAs have shorter exon lengths and longer intron lengths, while protein-coding genes have, in general, shorter exon lengths than all lncRNA MA groups; interestingly, the oldest MA group has larger intron lengths than protein-coding genes (Figures 12B and 12C). Further, older lncRNAs present stronger splicing motifs than younger lncRNAs (Figure 12D). Together, they indicate that lncRNAs are less spliced than mRNAs, but older lncRNAs have gained splicing efficiency throughout evolution. This splicing efficiency enhancement of older lncRNAs might be associated with a gain in functionality. It

has been shown that longer transcripts have features of dynamics expression associated with lncRNA functionality (Sarropoulos et al., 2019). To further evaluate a possible increase in functionality of older lncRNAs, the length of the transcripts and the number of isoforms among the MA groups were assessed. Concomitant with the increase in exon number, older lncRNA populations have longer transcripts and more isoforms than younger lncRNAs (Figures 12E and 12F), indicating an increase in locus complexity for older lncRNAs; still, all lncRNAs have reduced locus complexity than protein-coding genes.

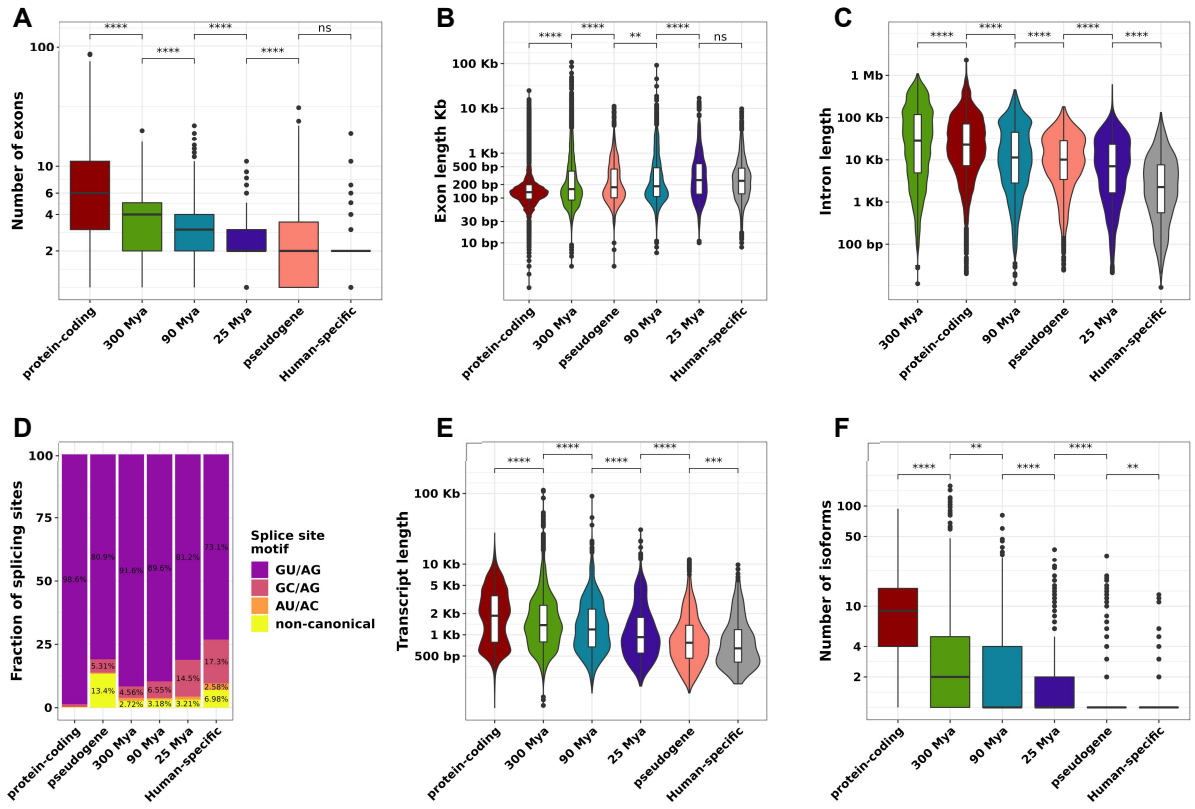

Figure 12. Splicing efficiency and locus complexity of human cortical lncRNA evolutionary groups. A. Number of exon distribution among protein-coding genes, pseudogenes, and lncRNA MA groups. **B-C.** Like B but showing the distribution of exon and intron length, respectively. **D.** Frequency of splicing motifs among protein-coding genes, pseudogenes, and lncRNA MA groups. **E-F.** Distribution of transcript length and the number of isoforms among protein-coding genes, pseudogenes, and lncRNA MA groups. **Statistics:** All statistics are one side (greater) Wilcoxon-test.

4.4 lncRNA evolutionary groups show distinct distributions of transposable element insertions but shared nuclear retention

Transposable elements are the main drivers of lncRNA diversification and evolution (Kapusta et al., 2013); therefore, they might be involved in the differences in transcript length and locus complexity among lncRNA MA groups, where different lncRNAs might contain distinct types of TE sequences that reflect their evolutionary ancestry and might affect their functionality. Several TE features were assessed among the lncRNA groups to test this hypothesis.

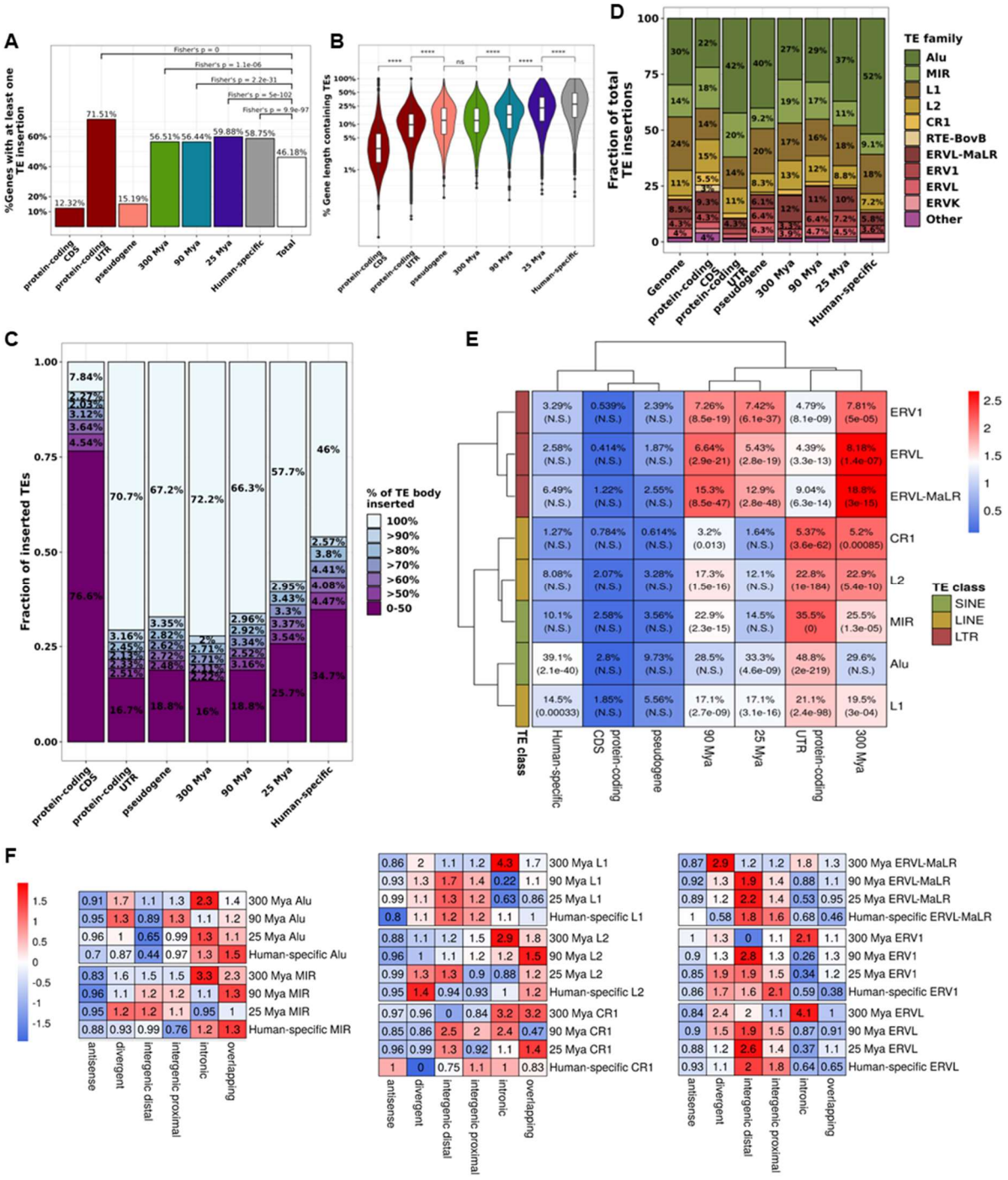

Figure 13. TE distribution in human cortical lncRNA evolutionary groups. A. Percentage of genes in a gene category carrying at least one TE insertion. **B.** Distribution of the percentage of the gene body made up of TEs in each gene category. **C.** Frequency of TE percentage inserted into the gene body of a gene category. **D.** Frequency of TEs families inserted into gene categories. **E.** Percentage of genes in a gene category carrying at least one TE family. **F.** Like E but separated by lncRNA type and Z-score normalized. The number shown represents the odds ratio concerning all lncRNA types together. **Statistics:** A and E mean Fisher hypergeometric tests, FDR corrected in E. B represents one side (greater) Wilcoxon-test.

First, the incidence of TE insertion was measured among the lncRNA groups; interestingly, all lncRNAs show a similar percentage of TE occurrence that differs from the depletion of TE insertion in pseudogenes and mRNA coding sequences (CDS); still, mRNA untranslated regions (UTRs) present increased incidence of TE insertions, even higher than lncRNAs (Figure 13A). Therefore, TE insertion is a shared feature among all non-coding sequences. Nevertheless, the extent of the gene body that contains a TE differs among lncRNA MA groups (Figure 13B). Older lncRNAs have a smaller percentage of their gene body made up of TEs than younger lncRNAs. Both untranslated and CDS mRNA regions have a lesser portion of their gene body made up of TEs, indicating that protein-coding genes are less tolerant to broad TE insertion than lncRNAs (Figure 13B). Although older lncRNAs have shorter TE patches related to their gene body, the size fraction of the TE that is inserted in older lncRNAs (72.2%) is larger than the fraction inserted in younger lncRNAs (57.7%) (Figure 13C). Together with the more extended transcript size of older lncRNAs, the data shows that older lncRNAs accept more extended TE insertions probably due to their larger gene body size compared to the shorter, younger lncRNAs.

It has been hypothesized that lncRNAs coopt TEs as functional domains (Johnson & Guigó, 2014). Thus, the differences in the type of TE insertion among lncRNA MA groups might point to differences in their functionality. The distribution of the TE families among the different gene categories (mRNAs CDS and UTRs, pseudogenes, and lncRNA MA groups) was assessed to evaluate this possibility. All gene categories showed deviations from the genomic distribution of TE families. The LINE family L1 was depleted from all considered gene types (Figure 13D), especially from protein-coding genes; interestingly, pseudogenes are genes with the highest frequency of L1 insertion (Figure 13D), indicating that L1 insertion is a mark of pseudogenization of coding genes.

Moreover, most of the lncRNA categories, except for the Human-specific lncRNAs, showed a marked increased frequency of endogenous retrovirus (ERVs); in particular, the ERVL-MaLR family, which is the most abundant ERV in the human genome (Figure 13D).

Furthermore, the occurrence of TE families among gene categories was also evaluated with similar results (Figure 13E); CDS and pseudogenes are depleted, while UTRs and all lncRNAs are rich in TE insertions. ERV families are enriched in lncRNA MA groups, except for Human-specific lncRNAs, and depleted from the other evaluated gene categories, indicating that ERVs insertions are a particular feature of lncRNAs (Figure 13E). Finally, L1 and Alu are the only TE families enriched in lncRNAs; in particular, Alu represents more than half of the TE insertion in the Human-specific group of lncRNAs (Figures 13D and 13E), which corroborates the novelty of these lncRNAs. Remarkably, Alu occurrence among lncRNA MA groups follows a gradient, where older lncRNAs are less tolerant, and younger lncRNAs are more pervasive to Alu insertions; inversely, ERVs are more prevalent in older than in younger lncRNAs.

To test whether the distinct patterns of distribution of lncRNA with respect to protein-coding genes observed in the different MA groups (Figure 9C) might affect the tolerance of TE insertion within lncRNAs throughout evolution, the occurrence of TE insertion among the lncRNA MA groups and lncRNA types was assessed (Figure 13F). The SINE families Alu and MIR are preferentially inserted in lncRNAs near protein-coding genes, as they are more common in intronic and overlapping lncRNAs (Figure 13F, left). However, older lncRNAs present an increased frequency of Alu in divergent lncRNA types (Figure 13F, left). The L1 family is more prevalent in lincRNAs (Figure 13F, middle), concomitant with the depletion of L1 insertions in protein-coding genes (Figure 13E). Likewise, ERVs are more commonplace in lincRNAs and divergent lncRNAs (Figure 13 F, right) and depleted from protein-coding genes (Figure 13E). In summary, these data show that protein-coding genes constrain the nature of TE insertion in cortical lncRNAs. The differences in TE family content and possible differences in functionality might partially be explained by the distinct distribution of lncRNAs around protein-coding genes among the different MA groups.

Several TE families have been recognized as signals of nuclear retention of lncRNAs (Carlevaro-Fita et al., 2019; Lubelsky & Ulitsky, 2018). As different lncRNA MA groups display distinct distributions of TEs, the nuclear retention feature of lncRNA might also differ among lncRNA MA groups. Thus, the distribution of lncRNAs among the nuclear and cytoplasmic compartments in fetal cortical tissues was assessed. Spuriously, the gene expression of all lncRNAs skewed toward the nuclear fraction (Figure 14A), and all lncRNA MA groups are proportionally enriched in the nucleus, despite differences in the percentage of genes enriched in that fraction (Figures 14 A and 14B).

It has been shown that the nuclear RNA exosome complex actively degrades lncRNAs, blocking their accumulation in the nucleoplasm (Schlackow et al., 2017). It is also partially responsible for the enrichment of lncRNAs in the chromatin fraction of human cell lines (Schlackow et al., 2017). The exosome complex might differentially recognize different lncRNA MA groups due to the differences in the molecular features among them. To test this, public libraries from HeLa cells where the nuclease component of the exosome complex *EXOSOME3* was knocked down were reanalyzed, testing only the set of cortical genes. Interestingly, the localization of all lncRNAs, both in the chromatin fraction and in the nucleoplasm, is affected after the knock-down of *EXOSOME3*, different from protein-coding genes, which are not dislocated under the same conditions (Figure 14C). In summary, independently of their evolutionary time, lncRNAs are enriched in the nucleus of fetal tissues and actively degraded by the exosome complex in cycling human cell lines.

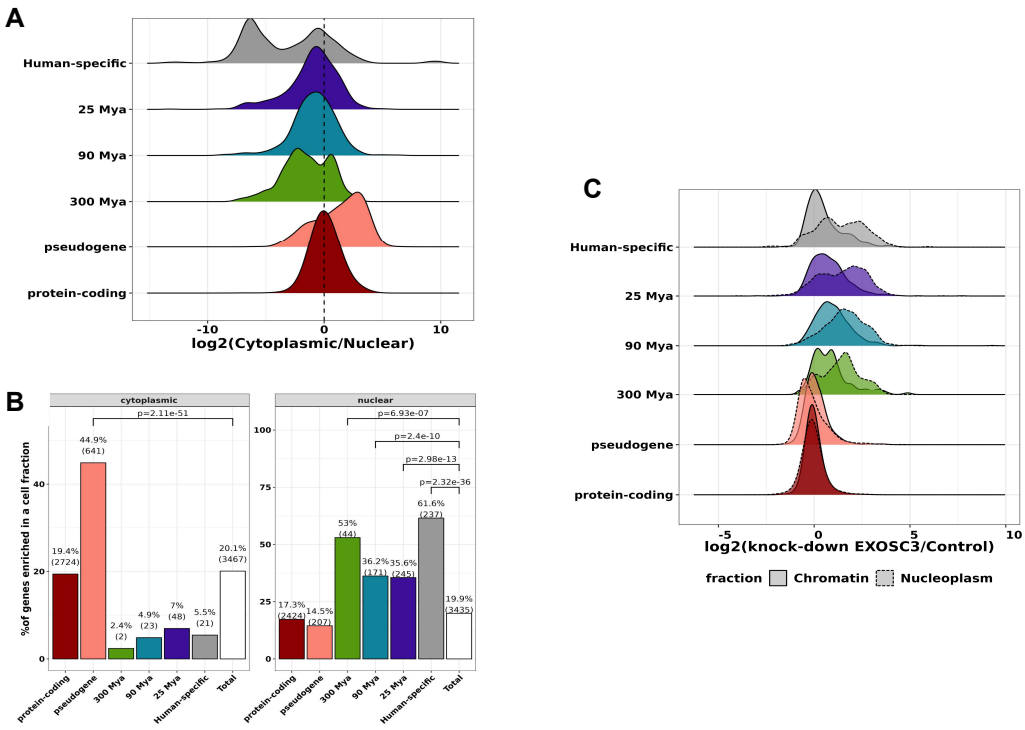

Figure 14. Nuclear enrichment of human cortical lncRNA evolutionary groups. A. Cytoplasmic and nuclear compartments fold change expression distribution among protein-coding genes, pseudogenes, and lncRNA MA groups. **B.** Enrichment of different gene types in a cellular compartment. **C.** Change in expression distribution among gene type categories after the knockdown of the exosome component EXOSC3. **Statistics:** Brackets in B mean Fisher hypergeometric tests P values.

4.5 lncRNA evolutionary groups show distinct genomic distributions that highlight a potential functional specialization

The lncRNA gene loci are intimately associated with their functionality, as shown for the topological anchor points RNAs (tapRNA), a group of syntenic conserved lncRNAs in mammals that co-expressed and regulated proximal developmental regulatory genes in a tissue-specific fashion (Amaral et al., 2018). Thus, systematically inspecting protein-coding genes proximal to lncRNAs of MA groups might help to shed light on the function of those cortical lncRNAs and how their roles have changed during evolution.

The nearest protein-coding genes were retrieved to a hundred kilobases surrounding the different lncRNA MA groups, then we assessed their enriched GO terms. Interestingly, it was found that MA groups evolved from loci near distinct types of developmental protein-coding genes that regulate various levels of neuron specification and maturation. Thus, ancient lncRNAs that appeared before the evolution of the cerebral cortex (300 MYA) are expressed from loci near to broad developmental genes, including TFs, which suggests a pleiotropic function of those lncRNAs (Figure 15A and supplementary table 3); meanwhile, lncRNAs that appeared before the expansion of primate cerebral cortex (90 MYA) are preferentially expressed from loci near to protein-coding genes associated to the development of axons; finally, Human-specific lncRNAs are preferentially expressed from loci proximal to genes associated to dendrite development, where synapsis are finely tuned (Figure 15A).

Furthermore, one of the enriched GO terms of protein-coding genes from the vicinity of antique cortical lncRNAs (300 MYA) was “DNA-binding transcription activator activity” (supplementary table 3); therefore, we tested whether older lncRNAs are enriched for TFs in their proximity. It was found that only the older MA group is slightly increased for proximal TFs (Figure 15B). Due to the considerable fraction of TFs as the closest coding gene (8.96 to 12.2%) and that TFs are the master regulators of biological functioning, we sought to identify the type of TFs proximal to lncRNAs of different MA groups. Surprisingly, lncRNAs of different MA groups are preferentially distributed around certain families of TFs. Older lncRNAs are preferentially expressed from loci close to the homeodomain-containing TFs (Figure 15C), master regulators of early development. Meanwhile, newer lncRNAs are preferentially expressed from loci near C2H2 zinger finger-containing (ZFs) TFs (Figure 15C) that present more specialized functions. It is essential to point out that hundreds of ZFs have evolved in the primate lineage in response to the expansion of retrovirus in primate genomes (Senft & Macfarlan, 2021), so it is plausible that a scenario where new ZFs protein-coding genes evolved in response to the expansion of TEs, led in turn to the evolution of new cortical lncRNAs around these new protein-coding genes in the primate lineage.

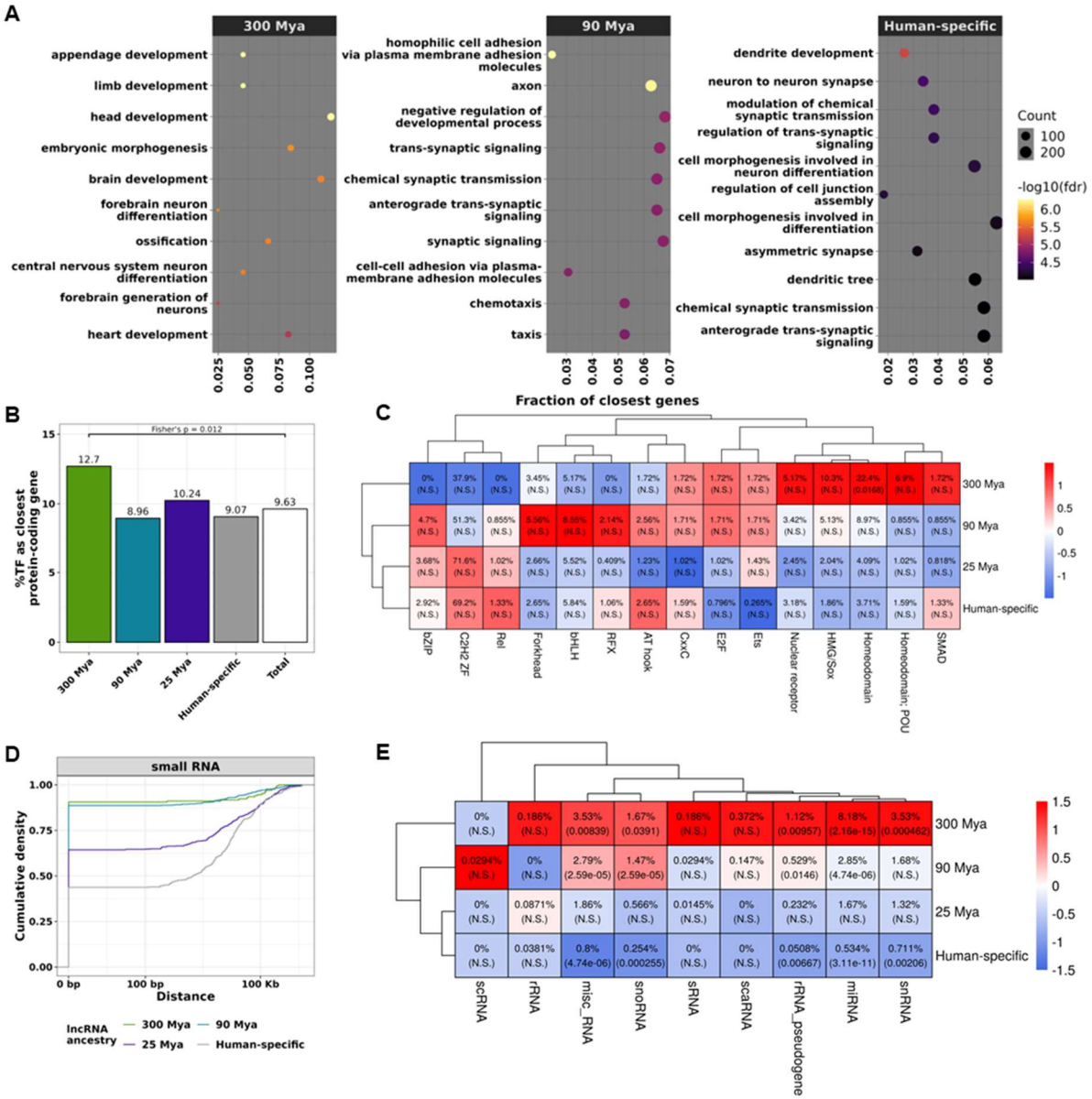

Figure 15. Genomic distribution of cortical lncRNA MA groups. A. Top10 gene ontology terms enriched in the closest protein-coding genes of different lncRNA MA groups. **B.** Percentage of transcription factors (TF) as the closest genes for each lncRNA MA group. **C.** Percentage of TF as the nearest gene separated by TF family; displayed number indicates the percentage. The FDR corrected Fisher hypergeometric P values are in parenthesis. **D.** Cumulative distribution of the distance to the nearest small RNA separated by lncRNA groups. **E.** Like C but for small RNAs as the closest gene (snRNA, small nuclear RNA; miRNA, microRNA; rRNA, ribosomal RNA; scaRNA, small Cajal body-specific RNA; sRNA, small RNA; snoRNA, small nucleolar RNA; scRNA, small cytoplasmic RNA).

Finally, it was observed that a large percentage of the oldest lncRNA MA groups (300 and 90 Mya) than of the younger lncRNAs (25 Mya and Human-specific) were proximal to small RNAs (Figure 15D); several lncRNAs have been identified as hosting small RNAs (Sun, Song, & Prasanth, 2021), therefore we tested whether old lncRNAs were enriched in these types of small RNA genes; remarkably, it was found that older lncRNAs preferentially host small

RNAs in their loci (Figure 15E), especially microRNAs, which points to a possible co-evolution of old cortical lncRNAs with microRNAs.

Collectively, these data indicate that cortical lncRNAs did not evolve randomly in the genome but have followed an evolutionary path that resembles the development of the cerebral cortex, which might be reflected in their different functionality.

4.6 Cortical lncRNAs shared chromatin features that differentiate them from other gene types

It has been shown that lincRNA promoters are depleted of most active chromatin marks compared to mRNAs (Mele et al., 2017). Still, they are particularly enriched in the chromatin repressive mark H3K9me3, which has been associated with the lower expression and tissue-specificity of lincRNAs in human cell lines; and they also display less TF diversity than mRNAs (Mele et al., 2017). To test whether these chromatin differences between lincRNAs and mRNAs stand in our set of cortical lncRNAs, are masked by evolutionary ancestry, and extend to other types of lncRNAs, we assessed several repressive and active chromatin modifications (Markenscoff-Papadimitriou et al., 2020), ATAC-seq data (de la Torre-Ubieta et al., 2018), and protein binding data (Hammal et al., 2022) of a group of promoters of expression and type-matched cortical genes at the mid developmental stages (Figures 16A and 16B).

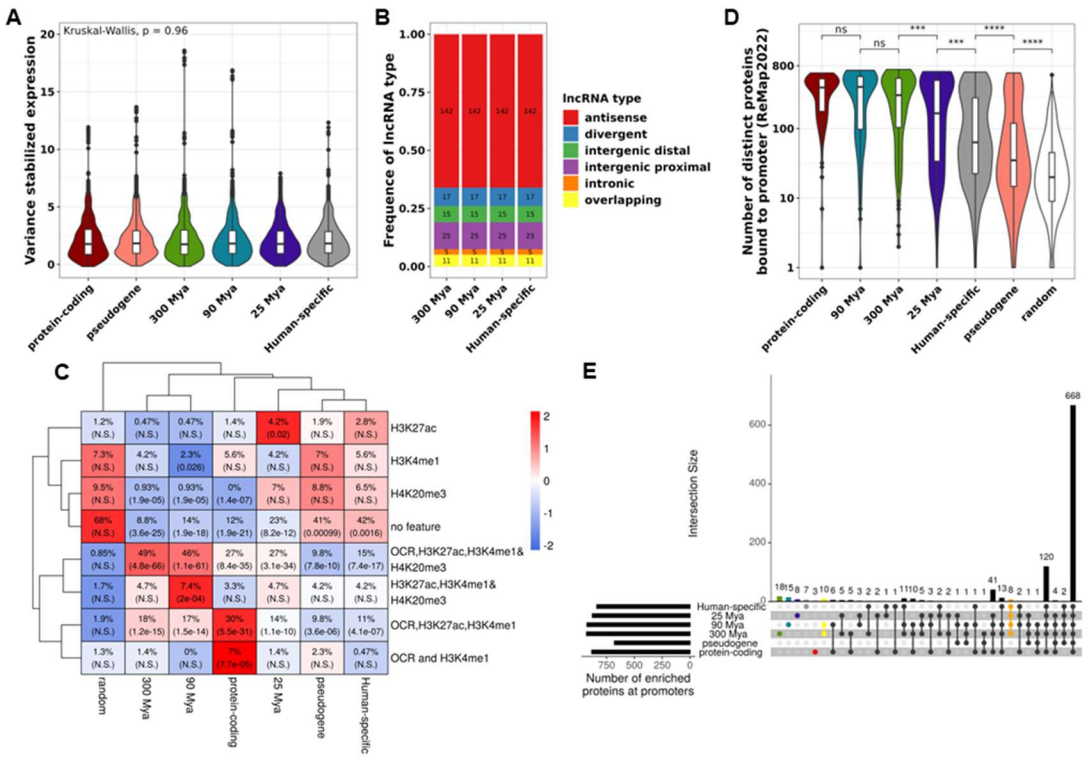

Figure 16. Chromatin features in human cortical lncRNAs MA groups. A. Set of expression-matched cortical genes in human cortical tissues at mid-gestation, week 15-17. **B**. Distribution of lncRNA types among lncRNA MA groups in the set of expression-matched genes. **C**. Proportion of promoters with respective chromatin features of expression-matched genes. In parenthesis, FDR corrected P values of the Fisher hypergeometric test between the gene category and a set of random sequences. **D**. Distribution of the number of distinct proteins bound to the promoter of expression-matched cortical genes. **E**. Upset plot of enriched proteins bound to promoters compared to a set of random sequences identified in each gene category. ORC, open chromatin regions.

In general, all gene type categories assessed showed an enrichment in chromatin modifications higher than expected by chance. Of note, this enrichment follows a gradient where older lncRNAs have fewer promoters with no chromatin modification features than younger lncRNAs (Figure 16C); older lncRNAs reached a similar proportion of promoters with no features (8.8 – 14%) as the mRNAs (12%), whereas Human-specific lncRNAs reached equal proportions as pseudogenes (41–42%) (Figure 16C), which indicated a gain in gene-regulation, and possible functionality of older cortical lncRNAs. mRNA promoters are depleted of the repressive mark H4K20me3 and most frequently contain active chromatin features (ORC, H3K27ac, and H3K4me1) (Figure 16C). Interestingly, the most abundant chromatin feature of all categories of lncRNAs is open chromatin regions (OCR) with chromatin bivalent marks (active: H3K27ac, H3K4me1, and repressive: H4K20me3) (Figure 16C), which differentiate lncRNAs from mRNAs, in agreement with the previous identification that lincRNA promoters are enriched in repressive chromatin marks (Mele et al., 2017).

Furthermore, the diversity of proteins bound at promoters of the cortical expression-matched genes was assessed. All lncRNA MA groups considered show an increased number of proteins bound to their promoter than a set of random genomic regions and pseudogenes (Figure 16D); indicating that the set of cortical lncRNAs presents more features of gene-regulation than expected by chance, reducing the possibility of most of them being bioinformatic artifacts. Remarkably, older lncRNAs have a similar diversity of proteins bound to their promoters than mRNAs, and a higher number than younger lncRNAs, showing an increased regulation for older lncRNAs, comparable to the degree of regulation of protein-coding genes (Figure 16D).

The enriched proteins bound at promoters of different gene categories were also assessed. Pseudogenes presented fewer enriched proteins, not identifying any protein specific to these genes (Figure 16E). Interestingly, lncRNA MA groups have group-specific proteins bound to their promoters, which form a gradient where older 300 Mya lncRNAs have eighteen group-specific proteins, 90 Mya lncRNAs have fifteen, and younger 25 Mya lncRNAs have seven group-specific proteins bound to their promoters (Figure 16E and Table 2); concomitant

with the increased regulatory marks of older lncRNAs. At the same time, expression-matched mRNAs showed a reduced number of these proteins, only three (Figure 16E), namely SIX4, NFIX, and ZNF408. Old lncRNAs (older than 90 Mya) share ten specific proteins bound to their promoters, including BDP1 (Table 2). BDP1 is a subunit of the Pol III transcription initiation factor III B that transcribes small RNAs in concordance to the increased number of older lncRNAs hosting small RNAs (Figures 15D and 15E).

Additionally, all lncRNAs (and not protein-coding genes or pseudogenes) have eight proteins in common (Figure 16E, orange dots, and Table 2). Remarkably, the DNA methyltransferase DNMT1 is among the proteins enriched in all cortical lncRNA MA groups (Table 2). This protein plays a significant role in DNA methylation maintenance and is found in regions of gene repression, in concordance to the elevated repressive chromatin marks found in all cortical lncRNA groups.

Table 2. Set of MA group-specific and shared proteins bound to the promoter of cortical lncRNAs. Summarized FDR aggregates the FDR from all collapsed gene categories using geometric mean.

Protein	Gene category	Summarized FDR	Protein	Gene category	Summarized FDR
CASZ1	300 Mya	1.58E-04	ZNF248	Human-specific	0.0037
ZNF197	300 Mya	4.36E-04	LHX2	Human-specific	0.00692
ZNF681	300 Mya	5.64E-04	ZFP41	Human-specific	0.0242
PTTG1	300 Mya	6.73E-04	ZNF132	Human-specific	0.0242
ZKSCAN8	300 Mya	0.00176	ZNF430	Human-specific	0.0242
ZNF485	300 Mya	0.00176	ZNF747	Human-specific	0.0242
ETS2	300 Mya	0.00241	TOP2A	Human-specific	0.0276
ZNF304	300 Mya	0.003	ZNF624	All lncRNAs	1.36E-06
SOX3	300 Mya	0.00622	FOSB	All lncRNAs	3.44E-06
GLI2	300 Mya	0.0127	DNMT1	All lncRNAs	8.23E-06
ZNF17	300 Mya	0.0128	ZNF548	All lncRNAs	8.59E-05
SMC4	300 Mya	0.0242	ZNF488	All lncRNAs	6.35E-04
ZFP90	300 Mya	0.0242	MBD4	All lncRNAs	9.42E-04
ZNF155	300 Mya	0.0242	ZNF266	All lncRNAs	0.00338
ZNF432	300 Mya	0.0242	ZNF565	All lncRNAs	0.0187
ZNF484	300 Mya	0.0242	KAT8	old lncRNAs	5.13E-06
ZNF582	300 Mya	0.0242	PRMT5	old lncRNAs	1.33E-05
ZNF7	300 Mya	0.0242	ZNF510	old lncRNAs	3.35E-04
MCM2	90 Mya	0.0037	ZNF85	old lncRNAs	0.00103
TRIP13	90 Mya	0.0037	BDP1	old lncRNAs	0.00555
ZNF267	90 Mya	0.0037	ZNF776	old lncRNAs	0.00622
ZNF426	90 Mya	0.00575	ZIK1	old lncRNAs	0.00945
TP73	90 Mya	0.0122	ZNF77	old lncRNAs	0.00945
ZNF808	90 Mya	0.0128	NUFIP1	old lncRNAs	0.0242
ZNF585B	90 Mya	0.0162	PPARA	old lncRNAs	0.0248

FOXF2	90 Mya	0.0221	GLYR1	young lncRNAs	3.10E-05
HEY2	90 Mya	0.0242			
PPARGC1A	90 Mya	0.0242			
SALL1	90 Mya	0.0242			
ZNF483	90 Mya	0.0242			
ZNF493	90 Mya	0.0242			
ZNF658	90 Mya	0.0242			
MCM5	90 Mya	0.0487			
ZNF669	25 Mya	0.0037			
ZNF136	25 Mya	0.00622			
ZNF26	25 Mya	0.0235			
TFCP2	25 Mya	0.0242			
ZNF138	25 Mya	0.0242			
ZNF250	25 Mya	0.0242			
ZNF280C	25 Mya	0.0242			
ZNF491	25 Mya	0.0242			

4.7 lncRNA ancestry has a strong effect on the lncRNA expression dynamics

Here, it was shown that lncRNAs from different MA groups possess diverse splicing efficiency, locus complexity, TE content, diversity of proteins bound to their promoters, and genome distribution that might be reflected in distinct functionalities. These possible functional differences should be mirrored in their expression dynamics during the development of the cerebral cortex. To further explore it, the expression data from cortical tissues of PsychEncode was used (Table 1). The trend of expression of lncRNAs throughout the prenatal development of the human cerebral cortex along the post-conception days was plotted, grouped by type and MA group (Figure 17A). Of note, the most significant visual differences in the profiles are among the lncRNA types; there are differences among MA groups in their expression tendency, even among lncRNAs of the same type, which indicates that lncRNAs from similar MA groups may share similar expression dynamics that differ from other MA groups.

To further corroborate the visual differences in the expression profiles, we tested whether the evolutionary ancestry (MA groups) and the lncRNA types can model the expression dynamics of the genes summarized by the Principal Component 1 (PC1), which accounts for 83.79% of the variance of expression of the genes (Figure 17B). After visualizing the PC1 distribution among MA groups, it was decided to use a GLM gamma model to account for the skewed distribution of the PC1 seen in Figure 17B; additionally, different models considering the lncRNA type and MA group were explored, being the best model the full interaction model, which includes the interaction between type and evolutionary ancestry (Figure 17C). After

modeling the PC1 based on the lncRNA type and MA group, the ANOVA test was used to see whether the ancestry has a significant effect on the model of expression dynamics, which was the case (Figure 17D). Accordingly, a cortical lncRNA's evolutionary ancestry impacts its expression dynamics throughout prenatal cortical development.

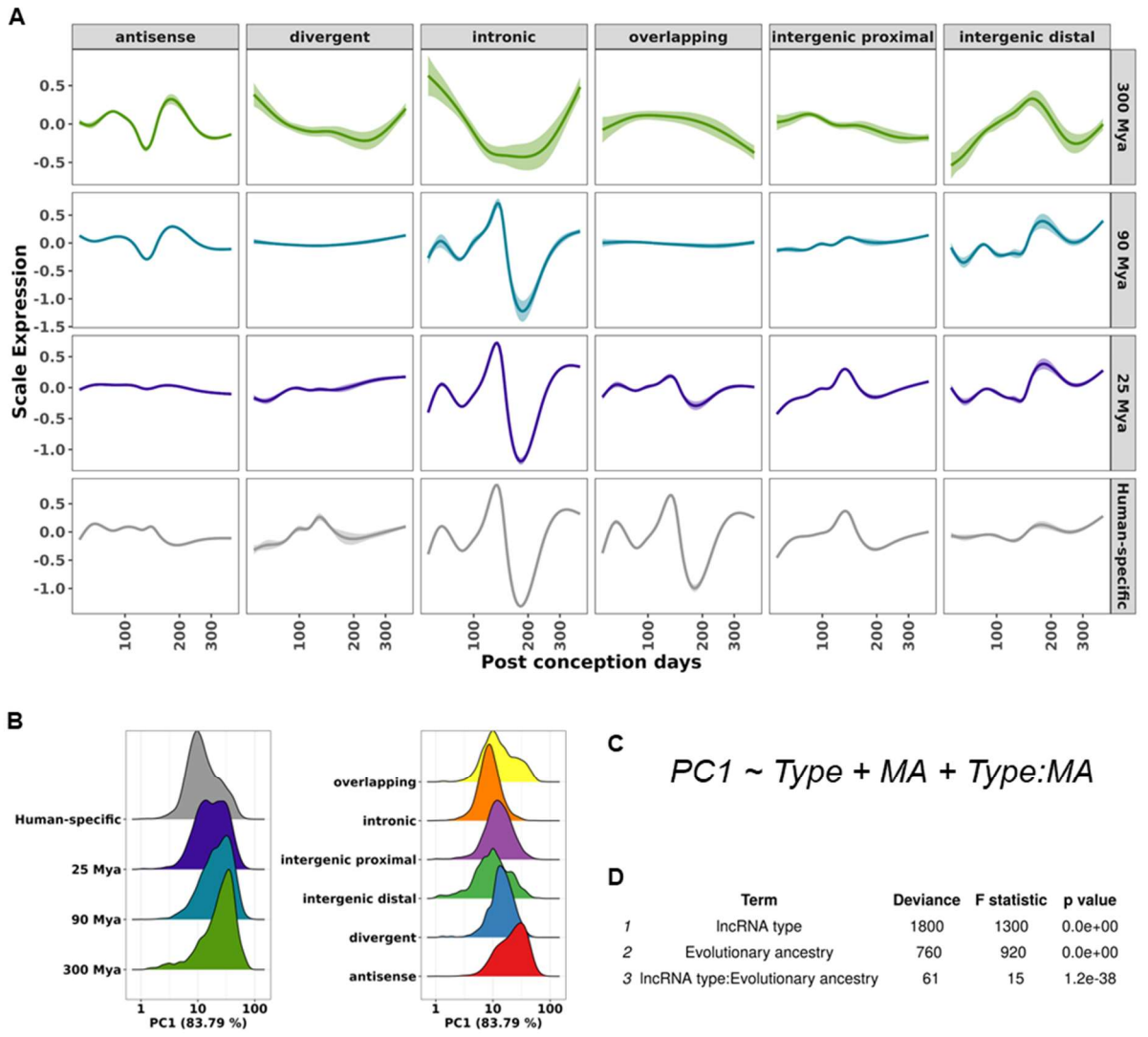

Figure 17. lncRNAs of different evolutionary ancestry display distinct gene expression dynamics throughout the cerebral cortex development. A. Expression trend of lncRNA genes from 189 bulk RNA-seq libraries throughout the prenatal and early postnatal development of the human cerebral cortex along the post-conception days, clustered by minimal age (MA) (lines) and lncRNA type (columns). Transparent lines around solid lines represent the 95% confidence intervals. **B.** PC1 distribution (that accounts for 83.79% of the variation of the bulk RNA-seq data set) of lncRNAs clustered by MA (left panel) and lncRNA type (middle panel). **C.** GLM gamma model of the gene expression dynamics represented by the PC1 as a dependent factor of the lncRNA type (Type), of the evolutionary ancestry (MA, minimal age), and of the interaction of lncRNA Type:MA. **D.** ANOVA test of the full model to identify significant terms in the model.

4.8 cortical lncRNAs are conspicuously expressed in glutamatergic neurons

Cortical lncRNAs display commonalities that differentiate them from protein-coding and pseudogenes and show differences among MA groups that point to functional specialization throughout evolution. However, the extent to which those features have impacted the evolution of the cerebral cortex has not been answered yet. For that, it is essential to identify which cells are expressing the different types of lncRNAs, and how those lncRNAs, in turn, modify ancestral gene modules to tune the molecular identity that impacts the diversification of cell types that give rise to the human cerebral cortex.

Single-cell RNA-sequencing (scRNA-seq) has been used to elucidate the molecular landscape of the developing cerebral cortex at a cell-type resolution (Fan et al., 2018; Zhong et al., 2018). Thus, public scRNA-seq data from the developing human cerebral cortex was used to map the cortical lncRNAs to the cellular populations of the developing cerebral cortex. All identified cell populations (Figures 8A and 8B) express at least one member of each of the MA groups, indicating widespread expression of all cortical lncRNAs. However, several cell types preferentially expressed different cortical lncRNA MA groups (Figure 18A). The older MA group is enriched in all interneuron cell populations (Inh. CGE, Inh. SST, Inh. MGE), late (outer) RGCs, and migrating glutamatergic neurons; instead, younger lncRNAs (90 Mya, 25 Mya, Human-specific) and pseudogenes are preferentially enriched in the synaptogenic glutamatergic neurons (Figure 18A). Additionally, lncRNAs from the 25 Mya and 90 Mya MA groups are enriched in Cajal-Retzius cells, while younger lncRNAs are enriched in vascular cell types (Pericytes, RBC, Endothelial cells), especially Human-specific cortical lncRNAs (Figure 18A). Interestingly, cortical lncRNAs are depleted from cycling cell types (early and late RGCs, mGPC, IPCs, Microglia, Oligodendrocytes, Astrocytes, OPC); at the same time, protein-coding genes are enriched in those cell populations (Figure 18A).

It has been shown that lowly expressed lncRNAs are specifically active in one cell population in the developing cerebral cortex (Liu et al., 2016). The specificity of cortical lncRNA expression was assessed in all gene categories to test whether this is a feature particular to a lncRNA MA group or shared among them. Of all gene categories assessed, protein-coding genes are the most broadly expressed in cell clusters, as only 20% of genes from this category are specific to a single cell type (Figure 18B). Among lncRNA MA groups, the oldest lncRNAs showed to be more broadly expressed among many cell clusters (only 31% of lncRNAs from this MA group are specific to a single cell cluster). Younger lncRNAs share a similar percentage of cell-type specificity (53% – 55%) (Figure 18B). The cell-type enrichment of the cluster-

specific lncRNAs was assessed, finding that Cajal-Retzius cells and synaptogenic glutamatergic neurons conspicuously express higher fractions of cell type-specific lncRNAs compared with other cells (Figure 18C). In the case of Cajal-Retzius cells, the 90 Mya and 25 Mya are the lncRNA MA groups that are enriched, meanwhile for synaptogenic glutamatergic neurons, the 25 Mya and, especially, the Human-specific lncRNAs are the enriched MA groups (Figure 18C), indicating that glutamatergic neurons are particularly enriched in the expression of cell-type specific cortical lncRNAs.

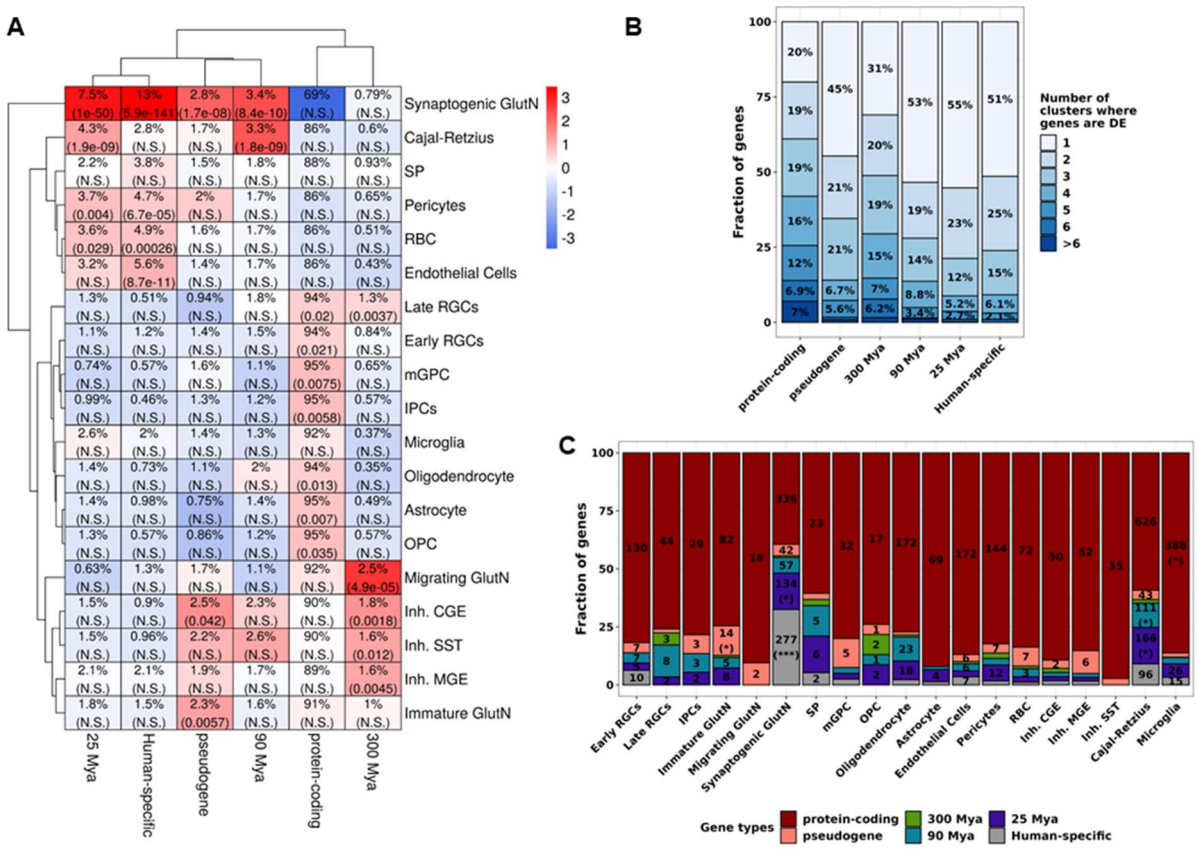

Figure 18. Glutamatergic neurons conspicuously express cortical lncRNAs. A. Percentage of lncRNA groups differentially expressed (DE) in a given cell type (indicated on the right); in parenthesis are shown FDR corrected P values of Fisher hypergeometric tests. **B.** Frequency of the cell type-specificity of different gene types; for each gene type (column), the fraction of those genes that are differentially expressed (DE) in 1 to 6 or more cell clusters (as indicated by the color) is given inside the boxes; these DE genes are specific markers of those cell clusters. **C.** Frequency of cluster-specific genes colored by gene type. *, FDR is less than 0.05; **, FDR is less than 10^{-5}; ***, FDR is less than 10^{-10}. (GlutN, glutamatergic neurons; SP, subplate cells, RBC, red blood cells; RGCs, radial glial cells; mGPC, multipotent glial progenitor cells; IPCs, intermediate progenitor cells; OPC, oligodendrocyte progenitor cells; Inh. CGE, inhibitory GABAergic neurons derived from caudal ganglionic eminences; Inh. SST, inhibitory GABAergic neurons expressing somatostatin; Inh. MGE, inhibitory GABAergic neurons derived from medial ganglionic eminences).

4.9 Developing glutamatergic neurons pseudotime analysis identifies putative mechanistic players of differential cellular distribution of cortical lncRNA MA groups

The Remap data were inspected to understand better why cortical lncRNAs are highly expressed in glutamatergic neurons compared to other clusters. First, the total number of lncRNA MA group-specific TFs enriched at the promoters of cortical lncRNAs at mid-gestational cerebral cortices (Table 2) were intersected with the scRNA-seq DE data (supplementary table 5), showing that TFs specific to 300 Mya and 90 Mya groups are preferentially expressed in cycling cortical progenitors (Figure 19A). In contrast, young cortical lncRNAs display similar numbers of specific TFs at promoters of cycling progenitors and postmitotic cells (Figure 19A). Consequently, the oldest MA group is enriched in germinal zones compared to the cortical plate at the mid-gestational stages of corticogenesis (Figure 19B). Besides, cortical lncRNAs form a gradient where older lncRNAs are proportionally more abundant in the germinative zone than in the cortical plate (Figure 19B).

Although younger lncRNAs show a similar number of specific TFs in RGCs as in synaptogenic glutamatergic neurons at mid-gestation (Figure 19A), it is in the later cell population that younger lncRNAs are highly expressed. Additionally, cortical lncRNAs are depleted in cycling progenitors. A possible explanation for this is the presence of a mechanism that represses the expression of cortical lncRNAs in mitotic progenitors. It was previously shown that the DNA methyl transferase DNMT1 is enriched in the promoters of all cortical lncRNAs (Table 2); it was also identified that *DNMT1* is differentially expressed in cycling cell populations: early RGCs, IPCs, mGPC, and OPC. This is different from the DNA methyl transferase *DNMT3A*, which is preferentially expressed in postsynaptic cell populations: synaptogenic glutamatergic neurons, CRs, oligodendrocytes, pericytes, and MGE interneurons (supplementary table 4). Moreover, it has been shown that DNMT1 reads the chromatin mark H4K20me3, which is enriched in promoters of all cortical MA groups at mid-gestation (Figure 16C), to reinforce repressive chromatin marks (Ren et al., 2021). Then, DNMT1 might be repressing the expression of cortical lncRNAs at the early stages of corticogenesis.

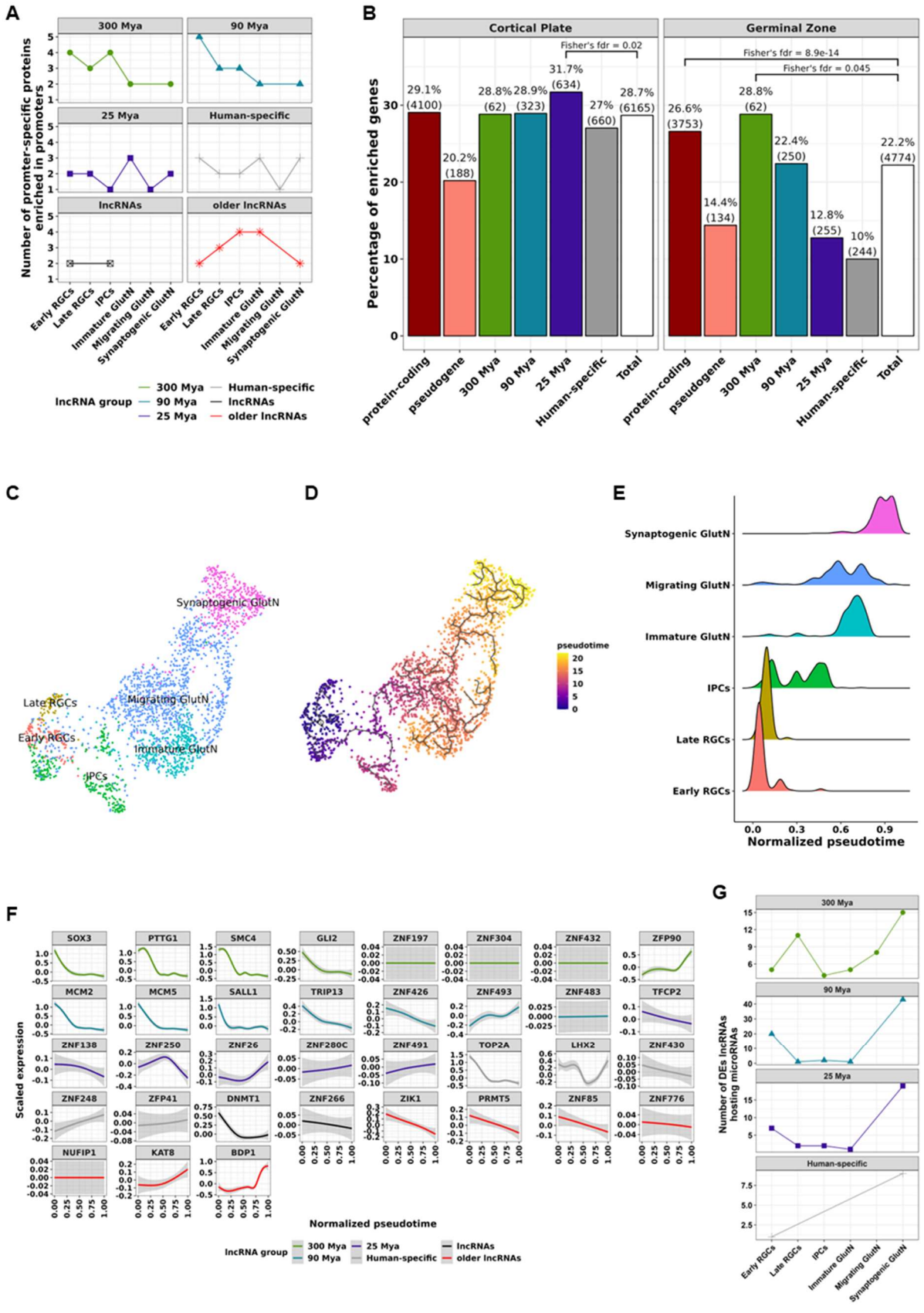

Figure 19. Identifying regulatory proteins involved in the differential expression pattern of lncRNA MA groups in the cortical glutamatergic lineage. A. DE genes were identified as lncRNA MA-specific TFs in the

cell population of the developing glutamatergic neurons. **B.** Percentage of DE genes by gene category in germinal or cortical plate zones of mid-gestation cerebral cortices. P values on the enriched gene categories come from Fisher's hypergeometric test. **C.** UMAP reduction of the cell population of the glutamatergic neuron lineage. **D.** Pseudotime path interpolated using Monocle3. **E.** Normalized pseudotime values distribution on cells from different cell clusters on C. **F.** Expression dynamics across the normalized pseudotime space. **G.** Number of DE cortical lncRNAs hosting microRNAs in glutamatergic cell populations.

To corroborate this, a pseudotime analysis using Monocle3 (Trapnell et al., 2014) was conducted in the glutamatergic neuron lineage (Figures 19D, 19E) to visualize the expression dynamics of *DNMT1* throughout the specification of glutamatergic neurons. Cell populations are distributed along the predicted normalized pseudotime scores according to the expected biological distribution. Mitotic cortical progenitors present low pseudotime scores, and post-mitotic neurons spread sequentially: immature, migrating, and synaptogenic glutamatergic neurons (Figure 19E). *DNMT1* expression is reduced alongside the pseudotime course (Figure 19F), in harmony with the specific expression of *DNMT1* in cortical progenitors. These results highlight DNMT1 as a potential active repressor of lncRNAs in cycling cell populations of the cerebral cortex.

We further examined the expression dynamics of all lncRNA-specific TFs identified in at least 5% of the cells in the pseudotime space, finding that older specific TFs preferentially follow a descending expression path. In contrast, Human-specific lncRNAs preferentially follow an ascending path (Figure 19F). This is an indication that post-mitotic activation of positive regulators might be a mechanism of high expression of Human-specific lncRNAs in glutamatergic neurons. Remarkably, *BDP1*, a small RNA activator, is highly expressed at later stages of glutamatergic neurons specification (Figure 19F), as well as most of the cortical lncRNAs hosting microRNAs (Figure 19G), suggesting that *BDP1* is a potential activator of antique microRNA-hosting cortical lncRNAs in synaptogenic glutamatergic neurons.

4.10 Cortical lncRNAs are sources of molecular innovation in the developing cerebral cortex.

To understand how lncRNAs have impacted the evolution of human corticogenesis gene expression networks, weighted gene co-expression network analysis (WGCNA) (Langfelder & Horvath, 2008) was performed. A hundred eighty-seven bulk RNA-seq samples from the cerebral cortex, expanding the prenatal and early post-conception days, were assessed after filtering for libraries outliers; forty-four co-expressed modules were identified.

First, the question of how central the different lncRNA MA groups are for the cortical modules was inspected by comparing the intramodular connectivity (kIN) distribution. Protein-coding genes are significantly more central than all lncRNA MA groups (Figure 20A). Cortical lncRNAs follow a gradient where older lncRNAs are more central than the younger lncRNAs, with Human-specific lncRNAs following similar distribution to pseudogenes (Figure 20A). These results indicate that protein-coding genes are pivotal for maintaining the networks, while lncRNA MA groups usually build up around central protein-coding genes following an ontological pattern, tunning the expression modules and potentially playing a role in plasticity.

The role of lncRNAs in the cerebral cortex gene network plasticity was assessed by preservation network analyses, comparing seven network properties (Ritchie et al., 2016) of cortical gene modules with the co-expression networks of developing tissues of the human, macaque, mouse, and chicken forebrains. We found that modules M24, M29, M38 were not preserved in any other of the assessed species (Figure 20B, red dots in chicken pallium, mouse forebrain and cortex, and rhesus cortex); as well, modules M1, M17, and M34 were only weakly preserved with the rhesus macaque (Figure 20B, yellow dots only in rhesus cortex); therefore we identified modules M24, M29, and M38 as Human-specific, meanwhile the modules M1, M17, and M34 as primates-specific (Figure 20B). Interestingly, gene modules M1, M17, and M24 are enriched in younger lncRNAs, while M29 is enriched in pseudogenes (Figure 20C). Furthermore, when intersecting with the scRNA-seq DE data, it was possible to identify the cortical cell types impacted by these more divergent modules; in particular, synaptogenic glutamatergic neurons are highly associated with the Human-specific module M24, and the primate-specific modules M1, and M17 (Figures 20B and 20C), all of them enriched in Human-specific lncRNAs, and module M1 and M17 also being enriched in primate-specific (25 Mya) lncRNAs (Figure 20C); indicating that lncRNAs evolution have preferentially impacted the molecular diversification of glutamatergic neurons.

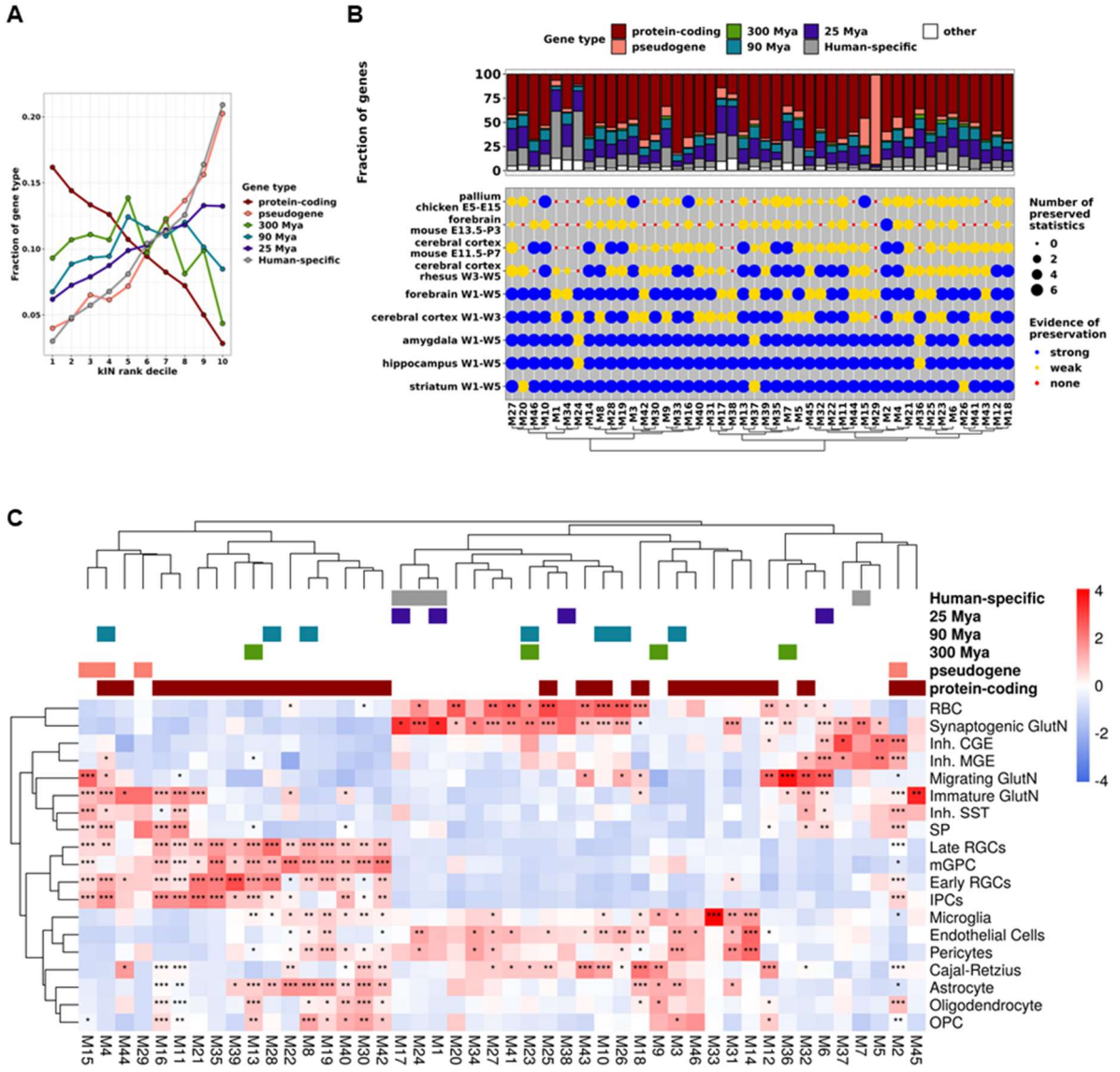

Figure 20. Cortical lncRNA are molecular sources of diversification. A. Intramodular connectivity distribution of protein-coding genes, pseudogenes, and lncRNA MA groups. **B.** Preservation analysis of 7 module statistics of the cortical modules in co-expression networks from forebrain tissues of humans, macaque, mice, and chicken. Strong preserved modules: all 7 module statistics were found preserved (Bonferroni transformed P value < 0.05); weak: between 6 and 1 module statistics were identified as preserved; none: no module feature was found preserved. **C.** Intersection of cortical modules and scRNA-seq DE data, where heatmap displays odds ratio scaled by row. FDR corrected Fisher hypergeometric P value: *, < 0.05; **, < 10^{-5}; ***, < 10^{-10}.

Cortical glutamatergic neurons are broadly clustered into deep layer (DL) and upper layer (UL) pyramidal neurons. Deep layers send axons out of the telencephalon, and upper layers send axons to intra-telencephalic structures. In primates, particularly humans, UL glutamatergic neurons are widely diversified, adopting distinct cellular shapes with different electrophysiological properties (Libé-Philippot & Vanderhaeghen, 2021). Here it was shown

that glutamatergic neurons are the cortical cell type where lncRNAs have played a pivotal role in their diversification. To further investigate whether lncRNAs have distinctly impacted the evolution of DL and UL neurons, the three synaptogenic glutamatergic subpopulations identified in the scRNA-seq data set were examined. DE analysis identified that the selective markers of UL neurons (*CUX1*, *CUX2*, and *RORB*) (Di Bella et al., 2021) were DE in the Synaptogenic GlutN subpopulation and the selective markers of DL neurons (*BCL11B* and *TLE4*) (Di Bella et al., 2021) were DE in the Synaptogenic GlutN2 subpopulation (Figure 21A, supplementary table 5). Afterwards, enrichment in lncRNAs MA groups was assessed, identifying that the Synaptogenic GlutN subpopulation, enriched in UL markers, is enriched in Human-specific lncRNAs (Figure 21B). GO analysis was further assessed among those populations, finding that the Synaptogenic GlutN subpopulation enriched in UL markers and Human-specific lncRNAs also expresses genes involved in dendritic development (Figure 21C). Together, these results suggest that human-specific lncRNAs have played a role in the diversification of UL glutamatergic neurons in the Human lineage, potentially by tuning the development of dendrites.

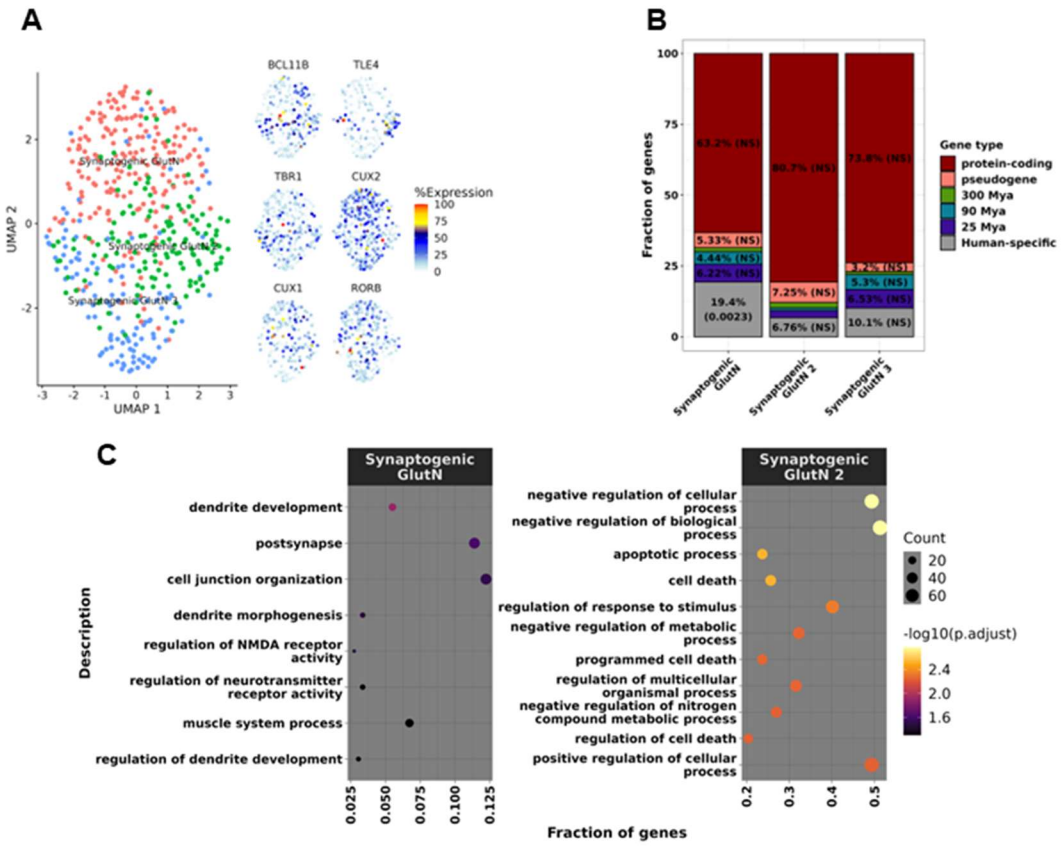

Figure 21. Upper layer glutamatergic neurons are enriched in Human-specific lncRNAs. A. UMAP reduction of three subpopulations of glutamatergic neurons. Synaptogenic GlutN (red), Synaptogenic GlutN2 (green), Synaptogenic GlutN3 (blue). Next to the UMAP plot, is the gene expression of upper (CUX1, CUX2, RORB) and deep layer markers (BCL11B, TLE4, TBR1) in those cells. **B.** Distribution of different gene categories of DE genes in each glutamatergic subpopulation. **C.** Gene ontology analysis of subpopulations enriched in UL and DL markers, respectively.

4.11 Human-specific lncRNAs are molecular readouts of autism spectrum disorder (ASD)

Several studies have shown that neuropsychiatric disorders disrupt the homeostasis of post-mitotic developing glutamatergic neurons (M. Li et al., 2018; Parikshak et al., 2016; Ziffra et al., 2021). As lncRNAs are conspicuously expressed in this cell population and involved in its molecular diversification, we further investigated whether cortical lncRNAs are dysregulated in neuropsychiatric disorders. Public bulk RNA-seq data from prefrontal cortical tissues of normal and affected specimens from three studies (Figure 22A, three different SRA projects) were the subject of DE analysis. Remarkably, it was identified that Human-specific lncRNAs are highly dysregulated in ASD (Figure 22B). These results indicate that different upstream regulators of Human-specific cortical lncRNAs might be dysregulated in ASD and converged into the widespread upregulation of human-specific lncRNAs.

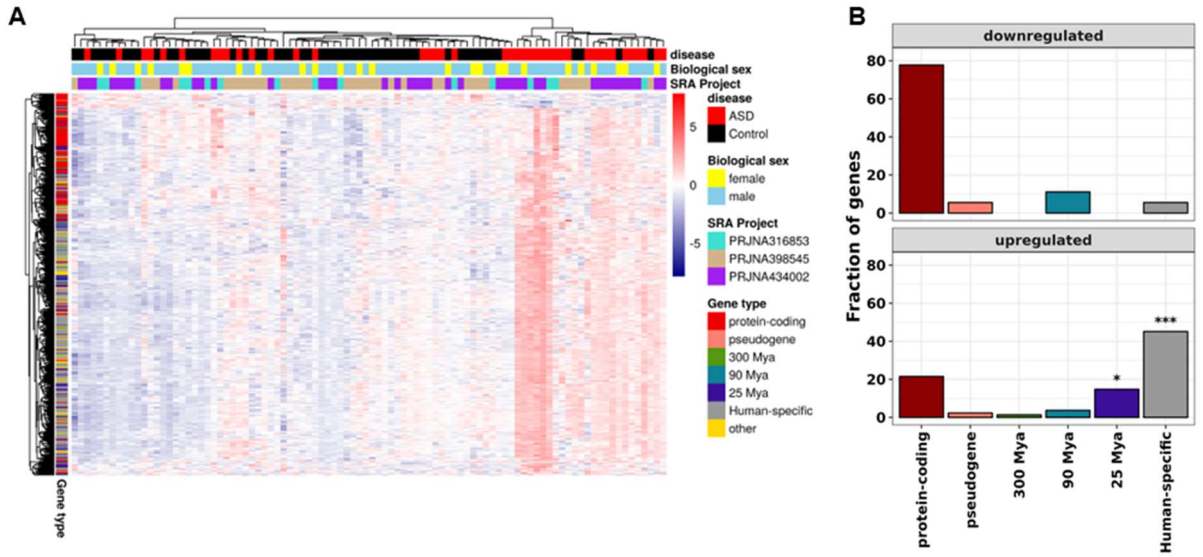

Figure 22. Human-specific lncRNAs are highly expressed in autism spectrum disorder (ASD). A. Heatmap (unsupervised clustering) of DE genes in autism spectrum disorders **B.** Frequency of DE genes belonging to protein-coding genes, pseudogenes, or one of lncRNA MA groups. FDR corrected Fisher hypergeometric P value: *, < 0.05; **, < 10^{-5}; ***, < 10^{-10}

5 Discussion

The low expression and tissue-specificity features of lncRNAs make them challenging to annotate. Additionally, corticogenesis is a highly dynamic process that integrates many cells from different developmental regions; in particular, human corticogenesis is lengthy and comprises more diverse cells than other mammals (Libé-Philippot & Vanderhaeghen, 2021; Molnár et al., 2019). Consequently, many lowly expressed lncRNAs or lncRNAs restricted to rare embryonic cell types might not be annotated in public references such as Gencode, which is detrimental to studying the evolution of lncRNAs. We used the vast PsychEncode RNA-seq database that spans a significant period of pre-natal windows and brain regions (Figure 4A) to annotate a comprehensive set of cortical lncRNAs in humans and the rhesus macaque. Furthermore, we collected additional RNA-seq public data from forebrain structures throughout the corticogenesis in humans, rhesus macaque, and mice; in addition, we combined RNA-seq public data from pallial development of chicken with data obtained in the present work for two critical developmental stages that were missing in the public repository, namely E7 and E10 to generate similar comprehensive transcriptome annotations for all species (Figure 4). Those new transcriptome assemblies notably increased the number of annotated lncRNAs (Figures 6 A-H). In addition, a new set of protein-coding genes and pseudogenes were annotated for all species (Figures 6 E-H). In summary, our new assemblies improved the annotation both quantitatively and qualitatively. Especially, they significantly helped to improve the annotation of macaque rhesus and chicken transcriptomes, which notoriously increased the number of reads mapped to an annotated feature (Figures 6M and 6J).

Thoroughly annotated transcriptomes were used to identify strong syntenic homologous genomic regions between humans and the other species, and jointly with the syntenic information from public databases, helped us to correctly classify lncRNAs into minimal evolutionary age (MA) groups, thus identifying that old lncRNAs show greater phyloP conservation scores than young lncRNAs (Figure 9A). lncRNA MA groups did not equally distribute; the 300 Mya group accounts for only 2.9% of the total cortical lncRNAs. In contrast, 90 Mya, 25 Mya, and Human-specific lncRNAs account for 18.2%, 36.8%, and 42.1%, respectively (Figures 9A, 9B, and 9C). A similar gradient of conservation and distribution of lncRNAs into different MA groups was identified by Necsulea *et al*. 2014, who used a different approach to identify evolutionary groups among lncRNAs; this shows that the methods built here and used for syntenic lncRNA identification yielded accurate results and that cortical lncRNAs follow similar evolutionary rules as those of lncRNAs expressed on other tissues.

Interestingly, the oldest group of lncRNA, those that appeared before the divergence of amniotes 300 Mya, showed similar conservation scores to UTRs, and pseudogenes showed similar scores to CDS (Figure 9A), indicating that coding and non-coding sequences have similar evolutionary turnover, independently of the gene type. The large fraction of primate and Human-specific lncRNAs suggests a rapid expansion of *de novo* expression of lncRNAs, especially those expressed near protein-coding genes, as seen by the increased gradient of intronic and overlapping lncRNAs in young lncRNAs (Figure 9C).

Identified lncRNA MA groups were used to evaluate shared and divergent molecular features of cortical lncRNAs, aiming to identify specialization signals of lncRNAs through evolution. First, we determined that lncRNAs form a gene expression gradient, where older lncRNAs have higher expression than younger lncRNAs, and this gradient is maintained when comparing lncRNAs of the same type (Figure 10). Additionally, we identified in a population of expression- and type-matched genes that older lncRNAs contain a greater diversity of TFs bound to their promoters than younger lncRNAs (Figure 16D); concomitant, older lncRNAs have an increased fraction of promoters containing activating chromatin features than younger lncRNAs (Figure 16C). Furthermore, an increasing number of exons and longer mature transcript sizes in conserved lncRNAs have been previously identified as signals of functional gain of lncRNAs (Hezroni et al., 2015; Sarropoulos et al., 2019). We identified similar differences when we assessed expression and type-matched lncRNAs. Cortical lncRNAs exhibit a gradient, where older lncRNAs have shorter exon lengths, higher exon numbers, increased intron sizes, larger mature transcript sizes, and a significantly higher number of isoforms (Figure 12). Those increases in locus complexity could be explained by the rise in the frequency of strong splicing signals in older lncRNAs (Figure 12D). Altogether, the results show an increase in gene regulation and locus complexity throughout lncRNAs evolution that might be associated with a gain of function of older lncRNAs.

TEs are significant drivers of lncRNA evolution, as they extensively integrate into lncRNA loci, providing new regulatory sequences, splicing sites, polyadenylation signals, and RNA-binding sequences (Kapusta et al., 2013). Differences in loci complexity among lncRNAs throughout evolution could be explained by different distributions of TE insertion. Although we did not identify enrichment in the fraction of genes carrying at least one TE insertion among lncRNA MA groups (Figure 13E), we found differences in the distribution of TE families throughout evolution (Figure 13F). Remarkably, we found that ERVs are enriched specifically in lncRNAs, except Human-specific lncRNAs, and the percentage of lncRNAs from a MA

group carrying ERVs form a gradient where older lncRNAs are more tolerant to their insertion. At the same time, primate-specific Alu sequences are enriched in young lncRNAs, especially Human-specific lncRNAs, following an inverse gradient to ERVs (Figure 13E). The constrain exerted by protein-coding genes partially explained those differences in the distribution of TEs as is patent in UTRs that are depleted of ERVs and enriched in Alu sequences, indicating tolerance of mRNA loci to Alu insertions but not for ERVs. Intronic and overlapping lncRNAs, expressed near protein-coding genes (Figure 7E), are reduced in ERVs sequence insertions and augmented in Alu insertions (Figure 13F). Human-specific lncRNAs are enriched in intronic and overlapping types; therefore, Alu insertions into a Human-specific lncRNAs or primate-specific lncRNAs (25 Mya) have been more tolerated. Intronic ERVs are more prevalent in the 300 Mya groups (Figure 13F), but the expression levels of those intronic lncRNAs are highly suppressed (Figure 10B), arguing in favor of the active repression of ERVs at protein-coding loci. It has been shown that ERVs inserted into lncRNAs expressed in embryonic stem cells and fetal development (Bakoulis, Krautz, Alcaraz, Salvatore, & Andersson, 2022; Wilson et al., 2020). At the same time, Alu sequences were identified as dysregulated in autism spectrum disorder and Alzheimer's disease, disorders that affect the homeostasis of neurons (Cheng et al., 2021; Saeliw et al., 2018). Those results indicate that different distributions of TE insertions across the evolution might lead to functional differentiation of lncRNAs.

Despite differences in the distribution of TE families in the gene body of lncRNAs, all lncRNAs shared a similar percentage of genes with at least one TE insertion, around 56.51%-59.88% (Figure 13A). The insertion of Alu, MIR, and L2 TE families into lncRNAs has been associated with nuclear enrichment of lncRNAs (Carlevaro-Fita et al., 2019; Lubelsky & Ulitsky, 2018). Notwithstanding the differential distribution of TE families throughout the evolution of cortical lncRNAs, we did not find differences when we examined the fraction of genes of lncRNA MA groups DE in the cytoplasm. However, the abundance of Human-specific lncRNAs in the nucleus is markedly higher than of other lncRNAs (Figure 14). Additionally, we found that all cortical lncRNAs are actively degraded by the exosome complex in the nucleus and enriched in the chromatin of HeLa cells, indicating that independently of the evolutionary age, all cortical lncRNAs share nuclear retention features and are actively degraded in human cell lines.

lncRNAs might regulate the expression of neighbor genes *in cis* and be sources of different types of small RNAs (Statello et al., 2021); therefore, we decided to inspect the loci surrounding them, identifying that different lncRNA MA groups preferentially evolved near

distinct types of genes. The oldest group of lncRNAs are expressed near genes involved in embryonic regulation, including homeobox TFs and microRNAs (Figures 15C and 15E). Together with the high insertion of ERVs, typically expressed in embryonic and fetal development, and increased gene regulation of their promoters, these findings suggest that antique lncRNAs have evolved to function in development, possibly as TF regulatory tapRNAs, sources of microRNAs, or uncharacterized functionalities. Recently, a study on lncRNA evolution has reached a similar conclusion about conserved lncRNAs but proposed that they function primarily in an RNA-independent manner (Darbellay & Necsulea, 2020). We disagree with that because, along with the increase in promoter regulation, antique lncRNAs have also gained loci complexity and strong splicing signals, indicating that RNAs have gained functional domains, although under a more relaxed evolutionary constraint than CDS (Figure 9A). Further functional analysis of such lncRNAs might help to elucidate whether conserved cortical lncRNAs have functionality and whether the relaxed pressure under which they have evolved plays a role in the plasticity of the brain.

In contrast, lncRNAs that evolved in mammals (at least 90 MYA onwards) show preferential expression near genes involved in neuron maturation and synapsis formation. The older 90 Mya group is preferentially expressed near genes involved in axon development, and younger Human-specific lncRNAs are expressed near genes involved in dendrite development (Figure 15 A), concomitant with the enrichment of those lncRNA groups in synaptogenic glutamatergic neurons and Cajal-Retzius cells (Figure 18A), which indicates that functional lncRNAs of these categories might be co-expressed with proximal coding genes and regulate their expression *in cis*. We further evaluated the expression dynamics of lncRNAs throughout development and found that lncRNA evolutionary ancestry is highly influential in the expression dynamics of cortical lncRNAs (Figure 17). Overall, these differences in locus complexity, TE distribution, genomic distribution, and expression dynamics point to the specialization of the function of lncRNAs throughout evolution.

It has been shown that lncRNAs are enriched in the repressive chromatin mark H3K9me3 in human cell lines (Mele et al., 2017); here, we extend those observations to the heterochromatin repressive mark H4K20me3 (Figure 16C). We also identify the enrichment of DNA methyltransferase DNMT1 in the promoter regions of cortical lncRNAs (Table 2). DNMT1 and H4K20me3 crosstalk might reinforce the repressive marks at lncRNAs in cortical cycling progenitors, as we showed that DNMT1 is preferentially expressed in the dividing cell types of the cerebral cortex (supplementary table 4), and using pseudotime analysis, we found

a marked reduction in the gene expression of *DNMT1* in post-mitotic glutamatergic cells (Figure 19F). The nature of the active repression of lncRNAs in mitotic cells is not well understood. However, it has been associated with increased tissue-specificity of lncRNAs; therefore, there might exist a mechanism by which the expression of lncRNAs is finely regulated (Mele et al., 2017). Alternatively, it could be a collateral result of the extensive insertions of TEs in lncRNAs, as it has been shown that DNMT1 and H4K20me3 help to reinforce the repression of L1 elements in embryonic stem cells (Bulstrode et al., 2017; Ren et al., 2021). Another possibility would be that cortical lncRNAs might be repressed in cycling cells to avoid the formation of R-loops between lncRNAs and the replisome, which generates DNA damage, and is potentially toxic to the cell (Statello et al., 2021). Further mechanistic studies, namely DNMT1 knock-out experiments, are needed to understand why lncRNA promoters are actively repressed and how at a molecular level, this is achieved.

We further examined the functional specialization of lncRNAs throughout evolution at single-cell resolution. Interestingly, antique lncRNAs are the only lncRNAs evolutionary group enriched in a cycling population, outer (late) RGCs (Figure 18A), which correlate with the enrichment of this group in the germinative zone of the developing cerebral cortex at mid-gestation stages (Figure 19B), and with the enrichment of homeobox TF binding sites at promoters of conserved lncRNAs (A. Necsulea et al., 2014). Besides that, we identified that lncRNAs as a group are depleted from the progenitor cell populations and preferentially expressed in mature neurons (Figure 18A). Remarkably, the oldest group of lncRNAs is enriched in inhibitory GABAergic neurons, which are conserved cellular populations in the pallium of amniote (Colquitt, Merullo, Konopka, Roberts, & Brainard, 2021; Tosches et al., 2018), indicating that antique lncRNAs expressed in interneurons are part of a profoundly conserved gene regulatory program that leads to the development and identification of GABAergic neurons of the cerebral cortex.

On the other hand, the younger groups of lncRNAs are highly expressed and specific to glutamatergic neurons (Figure s18A and 18C), which are among the most divergent cell types of the cerebral cortex. This high plasticity of glutamatergic neurons is not only present at a significant evolutionary scale (Tosches et al., 2018), but it is evident in the cerebral cortex evolution of the human lineage, with the UL glutamatergic neurons showing a greater diversification than other hominids (Berg et al., 2021). The enrichment of 90 Mya, 25 Mya, and Human-specific lncRNA MA groups in the glutamatergic neurons raised the possibility that lncRNAs evolution after the divergence of amniotes has impacted the plasticity of this cell

population. When we combined preservation analysis and scRNA-seq expression data, we could identify that gene module M24 was not preserved in other of the studied species, was highly enriched in synaptogenic glutamatergic neurons, and was enriched in Human-specific lncRNAs (Figures 20B and 20C). At the same time, modules M1 and M17 were weakly preserved in the rhesus macaque, were enriched in synaptogenic glutamatergic neurons, and were enriched in younger lncRNAs (Figures 20B and 20C). UL cells were the glutamatergic cell subcluster with the most significant set of Human-specific lncRNA (Figure 21). Altogether indicating that lncRNAs *de novo* expression in primates and humans has contributed to the molecular innovation of transcriptional landscape in corticogenesis and might significantly impact the diversification of glutamatergic neurons. Further functional studies of the impact of Human-specific lncRNAs in the specification of UL cortical glutamatergic neurons are required to probe the role of lncRNAs in the rapid evolution of glutamatergic neurons.

Finally, we not only identified younger lncRNAs as sources of glutamatergic neuron diversification, but we found that these cortical lncRNAs are upregulated in ASD, which raises the possibility of using younger lncRNAs, and especially Human-specific lncRNAs, as molecular readouts to diagnose the disorder.

6 Conclusions

The cerebral cortex is endowed with remarkable plasticity as it is patent in the myriad diversity of neocortical structures across mammals. It has been proposed that this plasticity is responsible for the evolution of human cognitive abilities. lncRNAs have faster evolutionary turnover than coding genes, are more tissue-specific, and, in tetrapods, are enriched in neural tissues, which make lncRNAs good candidates for genomic sources of cortical plasticity, evolution, and disease. With that in mind, in this thesis, we assessed the evolution of the lncRNA repertory of the human developing cerebral cortex. The assessed lncRNA evolutionary groups were mapped to the cellular and molecular dynamics of corticogenesis to identify the regulatory mechanism that drives lncRNA expression and the impact of lncRNAs in the evolution of the human cerebral cortex.

lncRNAs have gained different genetic features throughout evolution, namely: enhanced splicing efficiency, increased gene regulation and functional chromatin features, differential distribution of TEs, and expression dynamics that points to different functional specialization.

Antique lncRNAs that appeared before the evolution of the cerebral cortex in mammals showed preferential expression in germinative zones and early stages of development. They are regulated by genes preferentially expressed in mitotic cortical progenitors. They are proximal to developmental regulatory genes such as homeodomain TFs and microRNAs and enriched in GABAergic neurons, a conserved neural type of the cerebral cortex. These features indicate that lncRNAs are part of conserved genetic programs regulating embryonic and fetal development, particularly the development of cortical GABAergic neurons.

On the other hand, lncRNAs that evolved in parallel with the rise of the cerebral cortex in the dorsal pallium of mammals are enriched in transient (Cajal-Retzius) and mature glutamatergic neurons while depleted from mitotic cortical progenitors. They are expressed near genes involved in the specification of glutamatergic neurons, whereas the oldest (90 Mya) are expressed near genes involved in axon development. Meanwhile, Human-specific lncRNAs are expressed near genes involved in dendrite development. Additionally, young lncRNAs substantially contribute to primate and Human-specific gene modules of the developing cerebral cortex, which are enriched in synaptogenic glutamatergic neurons and, to a lesser extent, in vascular cell types. Human-specific lncRNAs are DE in the UL-like glutamatergic cells, the most divergent cell type of the cerebral cortex, and upregulated in ASD, a Human-

specific neurodevelopmental disease. These indicate that the recent evolution of lncRNAs has impacted the diversification and disease of cortical glutamatergic neurons, although further mechanistic research is needed to understand how this is achieved.

Finally, lncRNAs shared chromatin features, including enrichment in H3K9me3, H4K20me3 repressive histone marks, and concomitant DNA methylation, inferred from the collective enrichment of DNMT1 at their promoters. This crosstalk between chromatin modification and DNA methylation represents a new exiting mechanism of gene regulation of lncRNAs that requires further research to understand their implication on homeostasis and disease.

7. References

Amaral, P. P., Leonardi, T., Han, N., Viré, E., Gascoigne, D. K., Arias-Carrasco, R., . . . Kouzarides, T. (2018). Genomic positional conservation identifies topological anchor point RNAs linked to developmental loci. *Genome Biology, 19*(1), 32. doi:10.1186/s13059-018-1405-5

Ang, C. E., Ma, Q., Wapinski, O. L., Fan, S., Flynn, R. A., Lee, Q. Y., . . . Chang, H. Y. (2019). The novel lncRNA lnc-NR2F1 is pro-neurogenic and mutated in human neurodevelopmental disorders. *Elife, 8*. doi:10.7554/eLife.41770

Bakoulis, S., Krautz, R., Alcaraz, N., Salvatore, M., & Andersson, R. (2022). Endogenous retroviruses co-opted as divergently transcribed regulatory elements shape the regulatory landscape of embryonic stem cells. *Nucleic Acids Research, 50*(4), 2111-2127. doi:10.1093/nar/gkac088

Bellott, D. W., Skaletsky, H., Cho, T. J., Brown, L., Locke, D., Chen, N., . . . Page, D. C. (2017). Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. *Nat Genet, 49*(3), 387-394. doi:10.1038/ng.3778

Berg, J., Sorensen, S. A., Ting, J. T., Miller, J. A., Chartrand, T., Buchin, A., . . . Lein, E. S. (2021). Human neocortical expansion involves glutamatergic neuron diversification. *Nature, 598*(7879), 151-158. doi:10.1038/s41586-021-03813-8

Bulstrode, H., Johnstone, E., Marques-Torrejon, M. A., Ferguson, K. M., Bressan, R. B., Blin, C., . . . Pollard, S. M. (2017). Elevated FOXG1 and SOX2 in glioblastoma enforces neural stem cell identity through transcriptional control of cell cycle and epigenetic regulators. *Genes Dev, 31*(8), 757-773. doi:10.1101/gad.293027.116

Cajigas, I., Chakraborty, A., Swyter, K. R., Luo, H., Bastidas, M., Nigro, M., . . . Kohtz, J. D. (2018). The Evf2 Ultraconserved Enhancer lncRNA Functionally and Spatially Organizes Megabase Distant Genes in the Developing Forebrain. *Mol Cell, 71*(6), 956-972.e959. doi:10.1016/j.molcel.2018.07.024

Carlevaro-Fita, J., Polidori, T., Das, M., Navarro, C., Zoller, T. I., & Johnson, R. (2019). Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs. *Genome Res, 29*(2), 208-222. doi:10.1101/gr.229922.117

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics, 34*(17), i884-i890. doi:10.1093/bioinformatics/bty560

Cheng, Y., Saville, L., Gollen, B., Veronesi, A. A., Mohajerani, M., Joseph, J. T., & Zovoilis, A. (2021). Increased Alu RNA processing in Alzheimer brains is linked to gene expression changes. *EMBO Rep, 22*(5), e52255. doi:10.15252/embr.202052255

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., . . . Hubbard, T. (2011). Modernizing reference genome assemblies. *PLoS Biol, 9*(7), e1001091. doi:10.1371/journal.pbio.1001091

Colquitt, B. M., Merullo, D. P., Konopka, G., Roberts, T. F., & Brainard, M. S. (2021). Cellular transcriptomics reveals evolutionary identities of songbird vocal circuits. *Science, 371*(6530), eabd9704. doi:10.1126/science.abd9704

Darbellay, F., & Necsulea, A. (2020). Comparative Transcriptomics Analyses across Species, Organs, and Developmental Stages Reveal Functionally Constrained lncRNAs. *Molecular Biology and Evolution, 37*(1), 240-259. doi:10.1093/molbev/msz212

de la Torre-Ubieta, L., Stein, J. L., Won, H., Opland, C. K., Liang, D., Lu, D., & Geschwind, D. H. (2018). The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. *Cell, 172*(1-2), 289-304.e218. doi:10.1016/j.cell.2017.12.014

Di Bella, D. J., Habibi, E., Stickels, R. R., Scalia, G., Brown, J., Yadollahpour, P., . . . Arlotta, P. (2021). Molecular logic of cellular diversification in the mouse cerebral cortex. *Nature, 595*(7868), 554-559. doi:10.1038/s41586-021-03670-5

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15-21. doi:10.1093/bioinformatics/bts635

Douglas, P. (2018). TransDecoder https://github.com/TransDecoder/TransDecoder/wiki. In.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., . . . Finn, R. D. (2018). The Pfam protein families database in 2019. *Nucleic Acids Research, 47*(D1), D427-D432. doi:10.1093/nar/gky995

Fan, X., Dong, J., Zhong, S., Wei, Y., Wu, Q., Yan, L., . . . Tang, F. (2018). Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell RNA-seq analysis. *Cell Research, 28*(7), 730-745. doi:10.1038/s41422-018-0053-3

Fiddes, I. T., Lodewijk, G. A., Mooring, M., Bosworth, C. M., Ewing, A. D., Mantalas, G. L., . . . Haussler, D. (2018). Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. *Cell, 173*(6), 1356-1369.e1322. doi:10.1016/j.cell.2018.03.051

Frotscher, M. (1998). Cajal-Retzius cells, Reelin, and the formation of layers. *Curr Opin Neurobiol, 8*(5), 570-575. doi:10.1016/s0959-4388(98)80082-2

Gaspar, J. M. (2018). Genrich: detecting sites of genomic enrichment. Retrieved from https://github.com/jsh58/Genrich

Gordon, S. P., Tseng, E., Salamov, A., Zhang, J., Meng, X., Zhao, Z., . . . Wang, Z. (2015). Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing. *PLOS ONE, 10*(7), e0132628. doi:10.1371/journal.pone.0132628

Guo, C. J., Ma, X. K., Xing, Y. H., Zheng, C. C., Xu, Y. F., Shan, L., . . . Chen, L. L. (2020). Distinct Processing of lncRNAs Contributes to Non-conserved Functions in Stem Cells. *Cell, 181*(3), 621-636.e622. doi:10.1016/j.cell.2020.03.006

Hafemeister, C., & Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology, 20*(1), 296. doi:10.1186/s13059-019-1874-1

Hammal, F., de Langen, P., Bergon, A., Lopez, F., & Ballester, B. (2022). ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Research, 50*(D1), D316-D325. doi:10.1093/nar/gkab996

Hansen, B. B., & Klopfer, S. O. (2006). Optimal Full Matching and Related Designs via Network Flows. *Journal of Computational and Graphical Statistics, 15*(3), 609-627. doi:10.1198/106186006X137047

Hezroni, H., Koppstein, D., Schwartz, M. G., Avrutin, A., Bartel, D. P., & Ulitsky, I. (2015). Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell reports, 11*(7), 1110-1122. doi:10.1016/j.celrep.2015.04.023

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., & Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics, 18*(suppl_1), S96-S104. doi:10.1093/bioinformatics/18.suppl_1.S96

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution, 34*(8), 2115-2122. doi:10.1093/molbev/msx148

Johnson, R., & Guigó, R. (2014). The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *Rna, 20*(7), 959-976. doi:10.1261/rna.044560.114

Kaminow, B., Yunusov, D., & Dobin, A. (2021). STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *bioRxiv*, 2021.2005.2005.442755. doi:10.1101/2021.05.05.442755

Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., & Gao, G. (2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Research, 45*(W1), W12-W16. doi:10.1093/nar/gkx428

Kapusta, A., Kronenberg, Z., Lynch, V. J., Zhuo, X., Ramsay, L., Bourque, G., . . . Feschotte, C. (2013). Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. *PLOS Genetics, 9*(4), e1003470. doi:10.1371/journal.pgen.1003470

Knauss, J. L., Miao, N., Kim, S.-N., Nie, Y., Shi, Y., Wu, T., . . . Sun, T. (2018). Long noncoding RNA Sox2ot and transcription factor YY1 co-regulate the differentiation of cortical neural progenitors by repressing Sox2. *Cell Death & Disease, 9*(8), 799. doi:10.1038/s41419-018-0840-2

Kuo, R. I., Tseng, E., Eory, L., Paton, I. R., Archibald, A. L., & Burt, D. W. (2017). Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics, 18*(1), 323. doi:10.1186/s12864-017-3691-9

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics, 9*(1), 559. doi:10.1186/1471-2105-9-559

Leung, S. K., Jeffries, A. R., Castanho, I., Jordan, B. T., Moore, K., Davies, J. P., . . . Mill, J. (2021). Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell reports, 37*(7), 110022. doi:https://doi.org/10.1016/j.celrep.2021.110022

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. In. arXiv: Genomics.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics, 34*(18), 3094-3100. doi:10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Li, M., Santpere, G., Imamura Kawasawa, Y., Evgrafov, O. V., Gulden, F. O., Pochareddy, S., . . . Sestan, N. (2018). Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science, 362*(6420), eaat7615. doi:10.1126/science.aat7615

Liao, Y., Smyth, G. K., & Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research, 47*(8), e47-e47. doi:10.1093/nar/gkz114

Libé-Philippot, B., & Vanderhaeghen, P. (2021). Cellular and Molecular Mechanisms Linking Human Cortical Development and Evolution. *Annu Rev Genet, 55*, 555-581. doi:10.1146/annurev-genet-071719-020705

Liu, S. J., Nowakowski, T. J., Pollen, A. A., Lui, J. H., Horlbeck, M. A., Attenello, F. J., . . . Lim, D. A. (2016). Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biology, 17*(1), 67. doi:10.1186/s13059-016-0932-1

Lubelsky, Y., & Ulitsky, I. (2018). Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature, 555*(7694), 107-111. doi:10.1038/nature25757

Lui, J. H., Hansen, D. V., & Kriegstein, A. R. (2011). Development and evolution of the human neocortex. *Cell, 146*(1), 18-36. doi:10.1016/j.cell.2011.06.030

Maciel, L. F., Morales-Vicente, D. A., Silveira, G. O., Ribeiro, R. O., Olberg, G. G. O., Pires, D. S., . . . Verjovski-Almeida, S. (2019). Weighted Gene Co-Expression Analyses Point

to Long Non-Coding RNA Hub Genes at Different Schistosoma mansoni Life-Cycle Stages. *Front Genet, 10*, 823. doi:10.3389/fgene.2019.00823

Maciel, L. F., Morales-Vicente, D. A., & Verjovski-Almeida, S. (2020). Dynamic Expression of Long Non-Coding RNAs Throughout Parasite Sexual and Neural Maturation in Schistosoma Japonicum. *Non-Coding RNA, 6*(2), 15. doi:10.3390/ncrna6020015

Markenscoff-Papadimitriou, E., Whalen, S., Przytycki, P., Thomas, R., Binyameen, F., Nowakowski, T. J., . . . Rubenstein, J. L. (2020). A Chromatin Accessibility Atlas of the Developing Human Telencephalon. *Cell, 182*(3), 754-769 e718. doi:10.1016/j.cell.2020.06.002

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research, 40*(10), 4288-4297. doi:10.1093/nar/gks042

Mele, M., Mattioli, K., Mallard, W., Shechner, D. M., Gerhardinger, C., & Rinn, J. L. (2017). Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Res, 27*(1), 27-37. doi:10.1101/gr.214205.116

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., & Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research, 41*(12), e121-e121. doi:10.1093/nar/gkt263

Molnár, Z., Clowry, G. J., Šestan, N., Alzu'bi, A., Bakken, T., Hevner, R. F., . . . Kriegstein, A. (2019). New insights into the development of the human cerebral cortex. *J Anat, 235*(3), 432-451. doi:10.1111/joa.13055

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., . . . Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature, 505*(7485), 635-640. doi:10.1038/nature12943

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., . . . Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature, 505*(7485), 635-640. doi:10.1038/nature12943

http://www.nature.com/nature/journal/v505/n7485/abs/nature12943.html#supplementary-information

Niknafs, Y. S., Pandian, B., Iyer, H. K., Chinnaiyan, A. M., & Iyer, M. K. (2017). TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat Methods, 14*(1), 68-70. doi:10.1038/nmeth.4078

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., . . . Phillippy, A. M. (2022). The complete sequence of a human genome. *Science, 376*(6588), 44-53. doi:10.1126/science.abj6987

Parikshak, N. N., Swarup, V., Belgard, T. G., Irimia, M., Ramaswami, G., Gandal, M. J., . . . Geschwind, D. H. (2016). Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature, 540*(7633), 423-427. doi:10.1038/nature20612

Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. *F1000Research, 9*(304). doi:10.12688/f1000research.23297.1

. Picard toolkit. (2019). Broad Institute, GitHub repository: Broad Institute. Retrieved from https://broadinstitute.github.io/picard/

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res, 20*(1), 110-121. doi:10.1101/gr.097857.109

Pollard, K. S., Salama, S. R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J. S., . . . Haussler, D. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. *Nature, 443*(7108), 167-172. doi:10.1038/nature05113

Pollen, A. A., Nowakowski, T. J., Chen, J., Retallack, H., Sandoval-Espinosa, C., Nicholas, C. R., . . . Kriegstein, A. R. (2015). Molecular identity of human outer radial glia during cortical development. *Cell, 163*(1), 55-67. doi:10.1016/j.cell.2015.09.004

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*. doi:10.1093/bioinformatics/btq033

Ramos, A. D., Andersen, R. E., Liu, S. J., Nowakowski, T. J., Hong, S. J., Gertz, C., . . . Lim, D. A. (2015). The long noncoding RNA Pnky regulates neuronal differentiation of embryonic and postnatal neural stem cells. *Cell Stem Cell, 16*(4), 439-447. doi:10.1016/j.stem.2015.02.007

Ransohoff, J. D., Wei, Y., & Khavari, P. A. (2018). The functions and unique features of long intergenic non-coding RNA. *Nat Rev Mol Cell Biol, 19*(3), 143-157. doi:10.1038/nrm.2017.104

Ren, W., Fan, H., Grimm, S. A., Kim, J. J., Li, L., Guo, Y., . . . Song, J. (2021). DNMT1 reads heterochromatic H4K20me3 to reinforce LINE-1 DNA methylation. *Nat Commun, 12*(1), 2490. doi:10.1038/s41467-021-22665-4

Rinn, J. L., & Chang, H. Y. (2020). Long Noncoding RNAs: Molecular Modalities to Organismal Functions. *Annu Rev Biochem, 89*, 283-308. doi:10.1146/annurev-biochem-062917-012708

Ritchie, S. C., Watts, S., Fearnley, L. G., Holt, K. E., Abraham, G., & Inouye, M. (2016). A Scalable Permutation Approach Reveals Replication and Preservation Patterns of Network Modules in Large Datasets. *Cell Syst, 3*(1), 71-82. doi:10.1016/j.cels.2016.06.012

Saeliw, T., Tangsuwansri, C., Thongkorn, S., Chonchaiya, W., Suphapeetiporn, K., Mutirangura, A., . . . Sarachana, T. (2018). Integrated genome-wide Alu methylation and transcriptome profiling analyses reveal novel epigenetic regulatory networks associated with autism spectrum disorder. *Molecular Autism, 9*(1), 27. doi:10.1186/s13229-018-0213-9

Sarropoulos, I., Marin, R., Cardoso-Moreira, M., & Kaessmann, H. (2019). Developmental dynamics of lncRNAs across mammalian organs and species. *Nature, 571*(7766), 510-514. doi:10.1038/s41586-019-1341-x

Schlackow, M., Nojima, T., Gomes, T., Dhir, A., Carmo-Fonseca, M., & Proudfoot, N. J. (2017). Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs. *Mol Cell, 65*(1), 25-38. doi:10.1016/j.molcel.2016.11.029

Senft, A. D., & Macfarlan, T. S. (2021). Transposable elements shape the evolution of mammalian development. *Nature Reviews Genetics, 22*(11), 691-711. doi:10.1038/s41576-021-00385-1

Shao, M., & Kingsford, C. (2017). Accurate assembly of transcripts through phase-preserving graph decomposition. *Nature Biotechnology, 35*, 1167. doi:10.1038/nbt.4020

Silbereis, John C., Pochareddy, S., Zhu, Y., Li, M., & Sestan, N. (2016). The Cellular and Molecular Landscapes of the Developing Human Central Nervous System. *Neuron, 89*(2), 248-268. doi:10.1016/j.neuron.2015.12.008

Statello, L., Guo, C.-J., Chen, L.-L., & Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology, 22*(2), 96-118. doi:10.1038/s41580-020-00315-9

Steffen, M., Petti, A., Aach, J., D'Haeseleer, P., & Church, G. (2002). Automated modelling of signal transduction networks. *BMC Bioinformatics, 3*. doi:10.1186/1471-2105-3-34

Sun, Q., Song, Y. J., & Prasanth, K. V. (2021). One locus with two roles: microRNA-independent functions of microRNA-host-gene locus-encoded long noncoding RNAs. *Wiley Interdiscip Rev RNA, 12*(3), e1625. doi:10.1002/wrna.1625

Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F. J., Del Risco, H., . . . Conesa, A. (2018). SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res, 28*(3), 396-411. doi:10.1101/gr.222976.117

Team, R. C. (2018). R: A Language and Environment for Statistical Computing. Retrieved from https://www.R-project.org/

Thion, M. S., Ginhoux, F., & Garel, S. (2018). Microglia and early brain development: An intimate journey. *Science, 362*(6411), 185-189. doi:10.1126/science.aat0474

Tosches, M. A. (2021). From Cell Types to an Integrated Understanding of Brain Evolution: The Case of the Cerebral Cortex. *Annual Review of Cell and Developmental Biology, 37*(1), 495-517. doi:10.1146/annurev-cellbio-120319-112654

Tosches, M. A., Yamawaki, T. M., Naumann, R. K., Jacobi, A. A., Tushev, G., & Laurent, G. (2018). Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science, 360*(6391), 881. doi:10.1126/science.aar4237

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., . . . Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology, 32*(4), 381-386. doi:10.1038/nbt.2859

Trevino, A. E., Müller, F., Andersen, J., Sundaram, L., Kathiria, A., Shcherbina, A., . . . Greenleaf, W. J. (2021). Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell, 184*(19), 5053-5069.e5023. doi:https://doi.org/10.1016/j.cell.2021.07.039

Wang, L., Nie, J., Sicotte, H., Li, Y., Eckel-Passow, J. E., Dasari, S., . . . Kocher, J.-P. A. (2016). Measure transcript integrity using RNA-seq data. *BMC Bioinformatics, 17*(1), 58. doi:10.1186/s12859-016-0922-z

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J.-P., & Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research, 41*(6), e74-e74. doi:10.1093/nar/gkt006

Wang, L., Wang, S., & Li, W. (2012). RSeQC: quality control of RNA-seq experiments. *Bioinformatics, 28*(16), 2184-2185. doi:10.1093/bioinformatics/bts356

Warren, W. C., Harris, R. A., Haukness, M., Fiddes, I. T., Murali, S. C., Fernandes, J., . . . Eichler, E. E. (2020). Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science, 370*(6523), eabc6617. doi:10.1126/science.abc6617

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis: Springer-Verlag New York. Retrieved from http://ggplot2.org

Wilson, K. D., Ameen, M., Guo, H., Abilez, O. J., Tian, L., Mumbach, M. R., . . . Wu, J. C. (2020). Endogenous Retrovirus-Derived lncRNA BANCR Promotes Cardiomyocyte Migration in Humans and Non-human Primates. *Developmental Cell, 54*(6), 694-709.e699. doi:10.1016/j.devcel.2020.07.006

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., . . . Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation, 2*(3). doi:10.1016/j.xinn.2021.100141

Wucher, V., Legeai, F., Hédan, B., Rizk, G., Lagoutte, L., Leeb, T., . . . Derrien, T. (2017). FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Research, 45*(8), e57-e57. doi:10.1093/nar/gkw1306

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology, 16*(5), 284-287. doi:10.1089/omi.2011.0118

Zhang, Y., Parmigiani, G., & Johnson, W. E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics, 2*(3), lqaa078. doi:10.1093/nargab/lqaa078

Zhao, Y., Li, M.-C., Konaté, M. M., Chen, L., Das, B., Karlovich, C., . . . McShane, L. M. (2021). TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. *Journal of Translational Medicine, 19*(1), 269. doi:10.1186/s12967-021-02936-w

Zhong, S., Zhang, S., Fan, X., Wu, Q., Yan, L., Dong, J., . . . Wang, X. (2018). A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature, 555*(7697), 524-528. doi:10.1038/nature25980

Zhu, Y., Sousa, A. M. M., Gao, T., Skarica, M., Li, M., Santpere, G., . . . Sestan, N. (2018). Spatiotemporal transcriptomic divergence across human and macaque brain development. *Science, 362*(6420), eaat8077. doi:10.1126/science.aat8077

Ziffra, R. S., Kim, C. N., Ross, J. M., Wilfert, A., Turner, T. N., Haeussler, M., . . . Nowakowski, T. J. (2021). Single-cell epigenomics reveals mechanisms of human cortical development. *Nature, 598*(7879), 205-213. doi:10.1038/s41586-021-03209-8

8. Supplementary information

Anexo 1: Curriculum vitae

Supplementary table 1: List of all public libraries used for all analysis in this thesis.

Supplementary table 2: Transcriptome annotation summary tables containing ORF, coding potential, and gene type data.

Supplementary table 3: Closest coding gene GO analysis results.

Supplementary table 4: List of the DEGs in single-cell clusters.

Supplementary table 5: List of the DEGs in synaptic glutamatergic neurons subclusters.